COMPUTER SIMULATION OF BIOMOLECULAR SYSTEMS

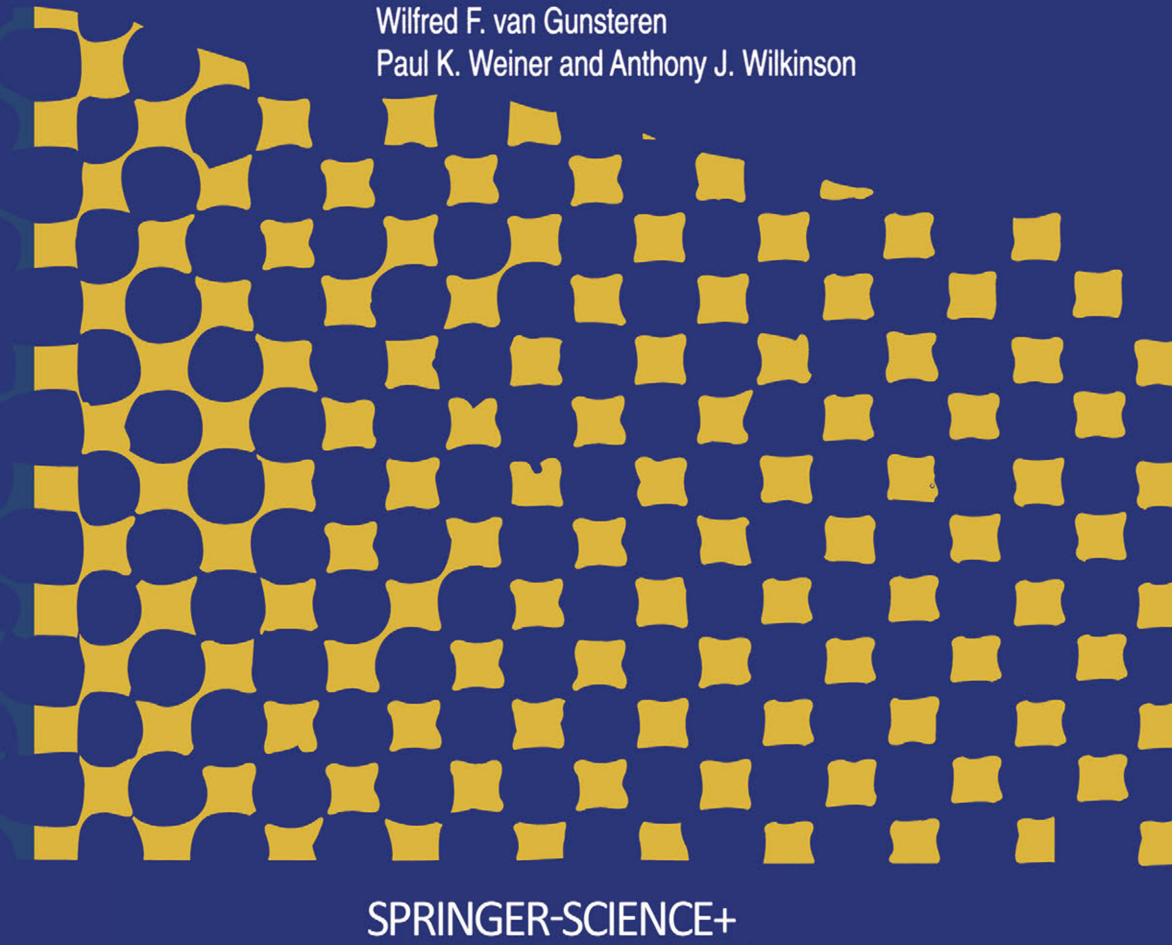
THEORETICAL AND EXPERIMENTAL APPLICATIONS

Vol. 3

Editors

Wilfred F. van Gunsteren

Paul K. Weiner and Anthony J. Wilkinson



SPRINGER-SCIENCE+
BUSINESS MEDIA, B.V.

COMPUTER SIMULATION OF BIOMOLECULAR SYSTEMS

THEORETICAL AND EXPERIMENTAL APPLICATIONS

Volume 3

COMPUTER SIMULATION OF BIOMOLECULAR SYSTEMS

THEORETICAL AND EXPERIMENTAL APPLICATIONS

Volume 3

Edited by

Wilfred F. van Gunsteren
Laboratory of Physical Chemistry
Swiss Federal Institute of Technology
ETH Zentrum, CH-8092 Zürich, Switzerland

Paul K. Weiner
Amdyn Systems Inc.
28 Tower Street
Somerville, Massachusetts 02143, U.S.A.

and

Anthony J. Wilkinson
Zeneca Pharmaceuticals
Alderley Park, Macclesfied
Cheshire SK10 4TG, U.K.

Springer-Science+Business Media, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-90-481-8528-3 ISBN 978-94-017-1120-3 (eBook)
DOI 10.1007/978-94-017-1120-3

Printed on acid-free paper

Preface

The third volume in the series on Computer Simulation of Biomolecular Systems continues with the format introduced in the first volume [1] and elaborated in the second volume [2]. The primary emphasis is on the methodological aspects of simulations, although there are some chapters that present the results obtained for specific systems of biological interest. The focus of this volume has changed somewhat since there are several chapters devoted to structure-based ligand design, which had only a single chapter in the second volume.

It seems useful to set the stage for this volume by quoting from my preface to Volume 2 [2].

"The long-range goal of molecular approaches to biology is to describe living systems in terms of chemistry and physics. Over the last fifty years great progress has been made in applying the equations representing the underlying physical laws to chemical problems involving the structures and reactions of small molecules. Corresponding studies of mesoscopic systems have been undertaken much more recently. Molecular dynamics simulations, which are the primary focus of this volume, represent the most important theoretical approach to macromolecules of biological interest."...

"Two attributes of molecular dynamics simulations have played an essential part in their explosive development and wide range of applications. Simulations provide individual particle motions as a function of time, so they can be used to answer detailed questions about the properties of a system, often more easily than experiments. For many aspects of biomolecule function, it is these details which are of interest (e.g., by what pathways does oxygen enter into and exit from the heme pocket in myoglobin). Of course, experiments play an essential role in validating the simulation methods; that is, comparisons with experimental data can serve to test the accuracy of the calculated results and to provide criteria for improving the methodology. This is particularly important because theoretical estimates of the systematic errors inherent in the simulations have not been possible; i.e., the errors introduced by the use of empirical potentials are difficult to quantify. Another important aspect of simulations is that, although the potentials employed in simulations are approximate, they are completely under the user's control, so that by removing or altering specific contributions their role in determining a given property can be examined. This is most graphically demonstrated in 'computer alchemy' – transmuting the potential from that representing one system to another during a simulation – in calculating free energy differences.

There are three types of applications of simulation methods in the macromolecular area, and in other areas as well. The first uses the simulation simply as a means of sampling configuration space. This is involved in the utilization of molecular dynamics, often with simulated annealing protocols, to determine or refine structures with data obtained from experiments, such as X-ray diffraction. The second uses simulations to determine equilibrium averages, including structural and motional properties (e.g. atomic mean-square fluctuation amplitudes) and the thermodynamics of the system. For such applications, it is necessary that the simulations adequately sample configuration space, as in the first application, with the additional condition that each point be weighted by the appropriate Boltzmann factor. The third area employs simulations to examine the actual dynamics. Here not only is adequate sampling of configuration space with appropriate Boltzmann weighting required, but it must be done so as to properly represent the time development of the system. For the first two areas, Monte Carlo simulations can be utilized, as well as molecular dynamics. By contrast, in the third area where the motions and their time development are of interest, only molecular dynamics can provide the necessary information. Because of their requirements, three sets of applications make increasing demands on the simulation method in terms of the accuracy that is required.

Now that molecular dynamics of macromolecules is a flourishing field and many people are working in that area, serious questions have to be asked concerning what more can be done with the methodology. What is the present and future role of molecular dynamics in the development of our knowledge of macromolecules of biological interest? How does the methodology need to be improved to make it applicable to important problems? The present volume is concerned with providing some answers."

Improvements in methodology are needed for present day applications primarily in two areas. The first is concerned with the potential energy function for the systems of interest and the second with the length of simulations and the sampling of the configuration space that is required to obtain meaningful results concerning the problems under investigation.

There are two reviews of empirical energy functions. The one by Hünenberger and van Gunsteren (Chapter 1) is a detailed overview of the form of potential functions, and the other by Kollman et al. (Chapter 2) stresses recent developments in one particular program. Considerable discussion is given on the validation of such functions, but how best to do this is still an unresolved problem. This means that most applications of empirical energy functions have an unknown systematic error, in addition to the statistical errors which can be estimated by relatively standard procedures.

The empirical potential functions that are used in molecular dynamics simulations cannot be used for bond making or bond breaking without introducing special terms. An alternative is to represent part of the system (that which undergoes a reaction, for example) by quantum mechanics and the remainder of the system by molecular mechanics. Such QM/MM methods are 'coming of age' and are being used increasingly, particularly for the determination of polarization effects on solvation and for the study of enzymatic reactions. The present status of the QM/MM methodology and its applications is reviewed by Cunningham and Bash (Chapter 6).

Because explicit treatment of the solvent often requires most of the computer time in actual simulations (e.g., for a protein consisting of a 1000 atoms, adequate solvation may require 10000 atoms of solvent), approaches to simplified solvation models are important. Two of these, by Gilson (Chapter 7) and by Elcock, Potter, and McCammon (Chapter 9), describe the use of continuum models based on the Poisson–Boltzmann equation. The former is concerned primarily with the use of a known structure to calculate pK_a values, a rapidly developing area, while the latter emphasizes the use of forces obtained from Poisson–Boltzmann calculations in molecular dynamics simulations. One of the known shortcomings of continuum models is that they usually do not treat explicitly water molecules that play a direct role as part of the 'structure' of the macromolecule. An interesting analysis of methods to include only the essential explicit solvent and represent the rest in terms of boundary effects is presented by Brower and Kimura (Chapter 8). Unfortunately, no results are included to permit evaluation of this type of approach.

Methods for extending the time scales of simulations are needed, even with the continuing increase in computer power. A review of approaches to increasing the basic time step in simulations that make use of a full atomic representation is given by Barth, Mandziuk, and Schlick (Chapter 3), with particular emphasis on recent implementations of implicit methods. Turner et al. (Chapter 4) describe the present stage of development of methods that include all atoms of the system but introduce reduced representations (e.g., treating helices as 'bodies') to speed up the calculations. It is shown by Case (Chapter 12) how a related approach, normal mode dynamics, can be extended to calculate not only the modes with respect to a single minimum, but to usefully analyze results from molecular dynamics simulations of multiminimum potential surfaces.

Simulations of protein folding require the treatment of large-scale motions that cannot yet be achieved with standard molecular dynamics methods. One way of guiding the simulation to the native structure is by the introduction of experimental information. This type of approach has become a fundamental part of X-ray structure refinement and NMR structure determination. An overview of recent advances in crystallographic applications is given by Schiffer (Chapter 10) with emphasis on fitting the data with an ensemble of structures. Although most refinements have been made without explicit solvent (i.e., it is customary to use somewhat special potential energy functions, often without any electrostatic terms, corresponding to an infinite dielectric constant), the introduction of simplified solvent models for this purpose is described by Braun (Chapter 11). Simplified or effective solvation models, sometimes combined with simplified representations of the protein, are also being used for studying protein folding without including data concerning the target structure. Lattice models are described by Skolnick and Kolinski (Chapter 15) and more detailed off-lattice models by Abagyan (Chapter 14); the latter chapter presents an overview of the protein folding problem and suggests possible solutions, the utility of most of which remains to be determined.

Only two chapters focus specifically on comparisons with and analyses of experiments. This is a little surprising since such applications are probably the primary objective of the ever-increasing number of papers that are being published on computer simulations of biomolecules. Smith (Chapter 13) provides an in-depth description of the use of molecular dynamics for the study of X-ray, neutron and infrared experiments, an area that is particularly important because detailed comparisons between simulations and experiment are possible. Another more specialized illustration of molecular dynamics and free energy perturbation calculations is presented by Kuramochi and Singh (Chapter 20), who describe studies of modifications of a specific position of a B-DNA duplex.

As already mentioned, an innovation in the present volume is the increase in the number of chapters devoted to various computational approaches to ligand design. I use the wording 'ligand design' because little is said about the relation between a good ligand and a real drug. Rational ligand design is an important and growing field, although the documented successes have so far not lived up to the expectations of many of the pharmaceutical companies. Robson (Chapter 19) discusses different computer languages and features that may aid in ligand design. Although improvements in computer languages certainly can play a role, the physical and chemical problems that still need to be resolved to achieve viable ligand and drug design are of even greater importance. One aspect considered by Grant and Pickup (Chapter 5) is the evaluation of molecular similarity; this plays an important role in QSAR and in pharmacophore generation and can also aid in the docking of related ligands. Two chapters survey developments in ligand design for known receptors. Rejto et al. (Chapter 17) are concerned primarily with ligand docking and make some interesting, although not yet verified, suggestions as to possible improvements in methodology, while Green and Johnson (Chapter 16) review the many methods that have been proposed for ligand design. One aspect of many ligand design methods [3] is that they have an intrinsic combinatorial character based on the fact that they employ a stepwise design procedure. The use of methods like the combination of MCSS [4] and HOOK [5] made possible combinatorial chemistry on the computer before it was invented in the laboratory. Moreover, the computational methods can survey 10^{15} or so ligands with ease, while laboratory methods are limited to 10^6 or less in a given experiment. Of course, computationally designed molecules often are difficult to synthesize and accurate predictions of their actual binding constants are not yet possible. Thus, a combination of combinatorial chemistry on the computer and in the laboratory may be the best way to approach the design problem. The question of evaluating binding constants is considered in detail by Timms and Wilkinson (Chapter 18). It is clear that it is relatively easy to use fitting procedures to find methods that correctly rank the binding constants of closely related ligands for a given receptor, but approaches that are accurate for general systems and fast enough to be applied to many ligands still remain to be developed.

Overall, the volume presents current material that can be useful to both the novice and the expert reader. Particularly for the former, Volume 2 in this series would serve as a useful introduction to the present volume.

3 June 1997

M. Karplus
Laboratoire de Chimie Biophysique
ISIS, Institut le Bel
Université Louis Pasteur
F-67000 Strasbourg, France
and
Department of Chemistry
Harvard University
Cambridge, MA 02138
U.S.A.

References

1. Van Gunsteren, W.F. and Weiner, P.K. (Eds.) Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, ESCOM, Leiden, 1989.
2. Van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993.
3. Miranker, A. and Karplus, M., Proteins Struct. Funct. Genet., 23(1995)472.
4. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins Struct. Funct. Genet., 19(1994)199.

Contents in brief

Part III Structure refinement

Part IV Simulation of large systems

Part V Protein folding

Part VI Structure-based design

Indexes

Contents

M. Karplus
Laboratoire de Chimie Biophysique, ISIS, Institut le Bel,
Université Louis Pasteur, F-67000 Strasbourg, France and
Department of Chemistry, Harvard University,
Cambridge, MA 02138, U.S.A.

Part I Methodology

P.H. Hünenberger and W.F. van Gunsteren
Laboratory of Physical Chemistry, ETH-Zentrum, CH-8092 Zürich,
Switzerland

Chapter 3. A separating framework for increasing the timestep in molecular dynamics

E. Barth, M. Mandziuk and T. Schlick
Department of Chemistry and Courant Institute of Mathematical Sciences, The Howard Hughes Medical Institute and New York University, 251 Mercer Street, New York, NY 10012, U.S.A.

Chapter 4. Reduced variable molecular dynamics

J. Turner[a], P.K. Weiner[a], B. Robson[b], R. Venugopal[c], H. Schubele III[c] and R. Singh[c]
[a]Amdyn Systems Inc., 28 Tower Street, Somerville, MA 02143, U.S.A.
[b]Proteus Molecular Design Ltd., Proteus House, Lyme Green Business Park, Macclesfield, Cheshire SK11 0JL, U.K.
[c]Dynacs Engineering Company Inc., 28870 U.S. Highway 19 North, Suite 405, Clearwater, FL 34621, U.S.A.

Part II Electrostatics and solvation

Chapter 7. Modeling protonation equilibria in biomolecules

M.K. Gilson
Center for Advanced Research in Biotechnology, National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850-3479, U.S.A.

Chapter 8. Semi-explicit bag model for protein solvation

R.C. Brower[a,b] *and S.R. Kimura*[a,c]
[a]*Center for Computational Science,*
[b]*Electrical and Computer Engineering Department, and*
[c]*Biomedical Engineering Department, Boston University, Boston, MA 02215, U.S.A.*

A.H. Elcock, M.J. Potter and J.A. McCammon
Department of Chemistry and Biochemistry, Department of Pharmacology, University of California at San Diego, La Jolla, CA 92093-0365, U.S.A.

Part III Structure refinement

C.A. Schiffer
Genentech Inc., 460 Point San Bruno Boulevard,
South San Francisco, CA 94080, U.S.A.

Part IV Simulation of large systems

Chapter 13. Dynamics of biomolecules: Simulation versus X-ray, neutron and infrared experiment

J.C. Smith
Molecular Simulation Group, SBPM/DBCM, Commissariat à l'Energie Atomique, CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France

Part V Protein folding

Chapter 14. Protein structure prediction by global energy optimization

R.A. Abagyan
The Skirball Institute of Biomolecular Medicine, Biochemistry Department, NYU Medical Center, New York University, 540 First Avenue, New York, NY 10016, U.S.A.

Chapter 15. Monte Carlo lattice dynamics and the prediction of protein folds

J. Skolnick[a] and A. Kolinski[a,b]
[a]Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.
[b]Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Part VI Structure-based design

S.M. Green and A.P. Johnson
Institute for Computer Applications in Molecular Sciences,
School of Chemistry, University of Leeds, Leeds LS2 9JT, U.K.

Chapter 17. New trends in computational structure prediction of ligand–protein complexes for receptor-based drug design

P.A. Rejto, G.M. Verkhivker, D.K. Gehlhaar and S.T. Freer
Agouron Pharmaceuticals Inc., 3565 General Atomics Court,
San Diego, CA 92121-1121, U.S.A.

Chapter 18. Estimation of binding affinity in structure-based design

D. Timms and A.J. Wilkinson
Zeneca Pharmaceuticals, Alderley Park, Macclesfield,
Cheshire SK10 4TG, U.K.

Chapter 19. Computer languages in pharmaceutical design

B. Robson
The Dirac Foundation, The Royal Veterinary College, University of London, Royal College Street, London NW1 0TU, U.K.
Current address: Principal Scientist, MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, U.S.A.

Indexes

Part I
Methodology

Empirical classical interaction functions for molecular simulation

P.H. Hünenberger and W.F. van Gunsteren
Laboratory of Physical Chemistry, ETH-Zentrum, CH-8092 Zürich, Switzerland

1. Introduction

With the continuing increase of the power of computers, the past decades have seen a rapid increase in the number, performance and accuracy of theoretical computational methods in chemistry [1,2]. One can distinguish three major classes of methods for the theoretical study of molecular properties, listed in order of decreasing computational expenses: (i) *ab initio* molecular orbital methods [3]; (ii) semiempirical molecular orbital methods [4,5]; and (iii) empirical classical force-field methods. The computational expenses of *ab initio* methods are of order $O(N_f^4)$ (Hartree–Fock level) or higher (configuration interaction, many-body perturbation theory), N_f being the number of basis functions used. Density functional approaches and semiempirical methods scale as $O(N_f^3)$ or lower. The costs of empirical methods scale as $O(N_a^2)$ down to nearly $O(N_a)$, where N_a stands for the number of elementary particles (atoms or groups of atoms). Independently of the scaling with the system size, the evaluation of an empirical interaction function remains usually much cheaper than any other method (size of the prefactor to the scaling) and currently allows for the simulation of systems typically up to 10^5–10^6 atoms.

Since the available computing resources are most often the true limiting factor to numerical calculations, it has become clear that there is no universal method able to solve all possible problems, but that one should rather select the method that is the most suitable to a problem of interest. As is schematically represented in Fig. 1, the properties of the observable(s) and system under consideration that will, together with the available computing power, largely determine which type of method can be used are [6]:

A. *the required system size*;

B. *the required volume of conformational space* that has to be searched or sampled (in terms of dynamics, the required timescale);

C. *the required resolution in terms of particles* (determined by the smallest entity, subatomic particle, atom or group of atoms, treated explicitly in the model);

D. *the required energetical accuracy* of the interaction function.

These requirements may be incompatible, in which case the observable cannot be computed adequately with the currently available computer resources [7]. When requirements A and B, together mostly determining the computational effort, are in conflict with requirements C and D, this conflict may be resolved by the design of

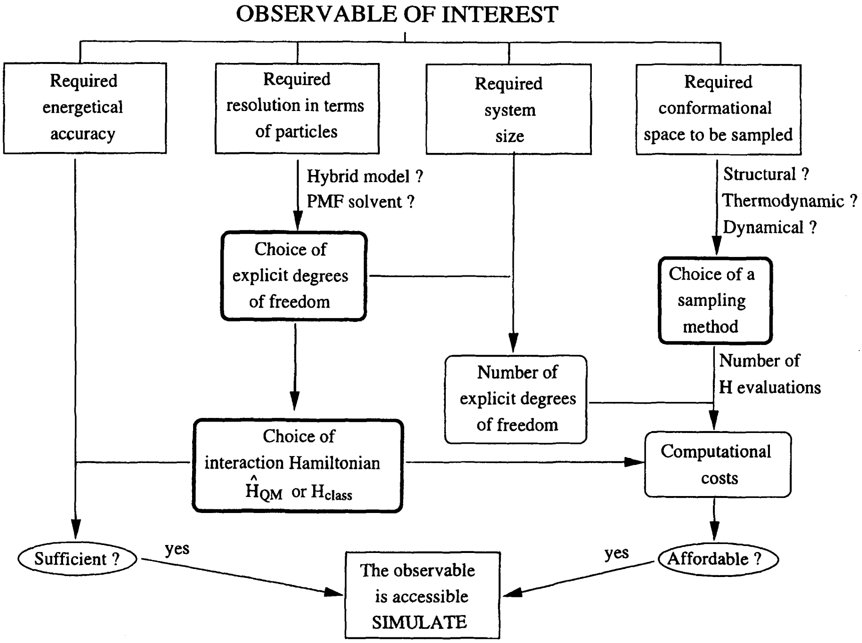

OBSERVABLE OF INTEREST
Required energetical accuracy
Required resolution in terms of particles
Required system size
Required conformational space to be sampled
Hybrid model ?
PMF solvent ?
Structural ?
Thermodynamic ?
Dynamical ?
Choice of explicit degrees of freedom
Choice of a sampling method
Number of explicit degrees of freedom
Number of H evaluations
Choice of interaction Hamiltonian
$\hat{H}_{QM}$ or H_{class}
Computational costs
Sufficient ?
yes
The observable is accessible SIMULATE
yes
Affordable ?

Fig. 1. Schematic representation of the basic choices made while building a model of the molecular system in order to simulate an observable of interest. The thick-line boxes represent the three essential choices and the global scheme of the present text.

hierarchical or hybrid models, where only the most relevant degrees of freedom are treated with a more expensive, higher resolution method. This is often done, for example, in the study of acid- or base-catalysed, organic, or enzymatic reactions in the bulk phase [8–11]. Another example is the use of a potential of mean force representation for the solvent, which includes its average effect without including its degrees of freedom explicitly [12]. Mean fluctuations in the solvent may also be included through a modification of the equations of motion as in stochastic dynamics [6,13].

Molecular orbital methods are well suited for the study of small molecules or small clusters of molecules (supermolecule) in vacuum [14], or within an averaged solvent environment [15–19], and give access to properties such as equilibrium geometries, vibrational frequencies, heats of formation, relative energies of conformers and isomerization barriers. These problems are also addressed with increasing accuracy by empirical methods [20–23]. Due to the size of the problem and the volume of accessible conformational space, the simulation of organic molecules or macromolecules in the condensed phase is the domain of atom-based empirical classical force fields [6]. Long-timescale (or long relaxation time) problems involving large systems, such as protein folding or *de novo* protein design, can currently be addressed only by residue-based force fields [24–28]. Finding an accurate description of the interaction at this low particle resolution (i.e. a sufficient energetical resolution) is, however, a major difficulty. Current areas of development with respect to the treatment of degrees of freedom are briefly discussed in Sec. 2.

Choosing the explicitly handled degrees of freedom is the first step in an empirical force-field calculation (Fig. 1). The second is the choice of a method to search or sample the conformational space [29–35]. This choice will also depend on the information required to compute the observable(s) of interest, namely:

A. *Structural information (searching) (Sec. 3.1):* The purpose of these methods is to search conformational space for one or a number of relevant low-energy conformations. In the latter case, the conformations obtained are not related by any well-defined probabilistic or dynamical relationship, and the method of choice is the one that searches the largest extent of conformational space, returning the highest number of low-energy structures.

B. *Structural and thermodynamic information (sampling) (Sec. 3.2):* The purpose of these methods is to sample conformational space or part of it in order to get a collection of conformations which build a correct statistical ensemble, that is, an ensemble in which the conformations appear with a Boltzmann probability. The sequence of the conformations is not relevant and the method of choice is the one which achieves the highest sampling efficiency.

C. *Structural, thermodynamic and dynamical information (simulating) (Sec. 3.3):* The purpose of these methods is to simulate the motion in conformational space or part of it, in order to get a sequence of conformations which build a correct statistical ensemble, but are also consecutive in time (dynamics). In this case, equations of motion which explicitly contain time are required, such as the Dirac, Schrödinger, Newton, Lagrange, Hamilton, Langevin or Liouville equations of motion.

The third choice to be made in an empirical force-field calculation is the one of an interaction function (or, together with the kinetic energy, a Hamiltonian) corresponding to the selected explicit degrees of freedom (Fig. 1). In principle, empirical force fields are constructed using experimental information (possibly complemented with theoretical results) and their only justification is their ability to reproduce or predict a large amount of experimental observables. It is, however, instructive to try to relate the empirical description to the underlying quantum mechanical reality. Empirical classical force fields are formally based on a generalization of the Born–Oppenheimer approximation, that is, on an averaging of the quantum mechanical Hamiltonian over implicit degrees of freedom (electronic and possibly also of individual atoms) to obtain an analytical interaction function depending solely on the explicit degrees of freedom of the model. Due to this averaging process, the interaction will be called a *potential of mean force* or *effective interaction function.* Averaging occurs at three levels:

A. averaging of the quantum mechanical interaction over the implicit degrees of freedom of the model (Sec. 4.1);

B. averaging of a force-field term over the different chemical/topological environments present in different molecules (Sec. 4.2);

C. averaging of a force-field term corresponding to an internal coordinate over the other force-field terms depending on the same coordinate, that is, over different geometrical environments (Sec. 4.3).

Table 1 *Hierarchy of explicit degrees of freedom included in the model*

Elementary unit	Phase	Type of interaction (operator/function)	Degrees of freedom averaged out	Reference
Electrons and nuclei	Gas phase	*Ab initio*, density functional: First-principles quantum mechanical Hamiltonian, Born–Oppenheimer surface	None	[3]
		Semiempirical: Approximated Hamiltonian	None	[4,5]
	Explicit solvent	*Idem*, supermolecule methods	None	[14]
	Implicit solvent	*Idem*, additional reaction field contribution	Solvent	[15–19]
United atoms	Gas phase			
All atoms		Classical empirical interaction function	Electronic	[64, 77]
United atom (aliphatic groups only)		*Idem*	Aliphatic H	[64, 77]
United atom (all CH_n groups)		*Idem*	All H bound to C	[64, 77]
United atoms (all)		*Idem*	All H	[64, 77]
Idem	Explicit solvent	*Idem*, including explicit solvent terms	*Idem*	[12]
Idem	Implicit solvent	*Idem*, possible corrections in the functional form, parameters, by additional terms or in the equations of motion	Solvent	[12]
Atom groups as 'bead(s)' E.g. amino acids in proteins represented by one or a few beads	Implicit solvent (or crystal)	Statistics-based interaction function	Side chain	[25]
Molecules				
Represented by a sphere, a rod or a disk	Liquid phase (or crystal)	Average intermolecular interaction function	Intramolecular	[257]

The present text will discuss possible options in the three basic choices outlined in Fig. 1 (thick-line boxes) and will mainly concentrate on the functional representation of the interaction function in atom- and united-atom-based force fields (Secs. 5 and 6). The list of methods is by far not exhaustive and the description is somewhat biased towards condensed-phase simulations and the simulation of large molecules (biomolecules). Finally, the problem of force-field parametrization will be briefly discussed (Sec. 7).

2. Choice of the explicit degrees of freedom of the model

The choice of an elementary unit (i.e. the particle that will have no explicit internal degrees of freedom) is the first step in the design of an empirical classical force field. Possible alternatives for the elementary unit and explicitly treated degrees of freedom, together with the corresponding type of interaction function, are summarized in Table 1. This choice will determine or strongly influence [6,7,36] the following:

A. The number of degrees of freedom that will have to be handled explicitly for describing a specific molecular system, and thus the computational effort.

B. The extent of conformational space that can be searched (or in terms of molecular dynamics, the reachable timescale). Because available computing power is most often a limiting factor, for a system of a given size, the number of possible evaluations of the potential energy function will rapidly decrease with the number of explicit degrees of freedom.

C. The maximum resolution, in terms of particles (e.g. subatomic particles, atoms, group of atoms, or molecules) and processes (e.g. conformational changes, chemical reactions) that can be achieved by the force field.

D. The type of functions that are likely to describe the interaction between elementary units in an adequate manner, that is, with a reasonable energetical accuracy.

E. The type of observables the force field may be able to describe correctly, and those which will necessarily stay inaccessible. Accessible observables will be those for which the extent of searchable conformational space (B), the force-field resolution in terms of particles (C) and the force-field accuracy (D) are sufficient.

Current developments in empirical classical force fields mainly follow five basic lines in terms of degrees of freedom [10,12,20,21,25,37], which will be described in Secs. 2.1–2.5. Note that in Secs. 2.3–2.5, the number of explicit degrees of freedom is reduced essentially by decreasing the force-field resolution in terms of particles. An alternative way to reduce the size of the conformational space to be searched is to limit the dimensionality or to discretize the coordinates (lattice methods, see e.g. Ref. 38). These methods will not be discussed here.

2.1. Gas-phase force fields

The primary purpose of gas-phase force fields is the accurate description of molecules in vacuum [20–23,39–41]. These force fields may be used to either complete or replace more expensive *ab initio* molecular orbital calculations [22], or to predict experimental gas-phase properties such as equilibrium geometries, vibrational frequencies, heats of formation, relative energies of conformers and energy barriers for isomerization [41]. Rapid progress in the design of such force fields is made possible by (i) the absence of intermolecular forces, (ii) the increasing amount and reliability of data from *ab initio* molecular orbital calculations, and (iii) the use of systematic and relatively inexpensive procedures for parameter calibration using both theoretical and experimental data (Sec. 7.5). These force fields, sometimes called class II force fields [22,23], are usually characterized by a detailed description of covalent degrees of freedom, involving anharmonic (nonquadratic) potential energy terms and terms that couple the internal coordinates (nondiagonal energy terms). Typical examples are the force fields CFF [42–44] and a recently modified version [45,46], CVFF [47–50], EFF93 [51,52], MM2 [20,53], MM3 [20,54–56] and QMFF/CFF93 [22,23,41].

The term gas-phase force field does not mean that such force fields cannot be extended for applications in condensed-phase simulations. Experimental information on crystal structures is sometimes used in the parametrization procedure [43,46,52]. For applications in liquid-phase problems, however, these force fields will suffer from the same difficulties in parametrization as condensed-phase force fields (Sec. 2.2), and whether the significantly improved accuracy gained in the gas phase by inclusion of anharmonic and off-diagonal terms will result in a significant increase of accuracy in the simulated condensed-phase properties is still a matter of discussion.

2.2. Condensed-phase force fields

The primary purpose of condensed-phase force fields is the accurate description of liquids, solutions of organic compounds or macromolecules and crystals [6,57–59]. Progress in the development of such force fields is slow, since: (i) the dominant forces in the condensed phase are intermolecular forces which are not easily described and parametrized adequately; (ii) the relevance of data from *ab initio* molecular orbital calculations in vacuum (even when reaction-field corrections are applied) is limited, and the parametrization has to rely mostly on a small amount of experimental data concerning the condensed phase; and (iii) the design of systematic optimization procedures is in general not possible (see, however, Sec. 7.4). One major reason for this impossibility is that the estimation of observables to be compared to experimental results generally requires a large number of evaluations of the potential energy function, and is therefore computationally expensive. In these force fields, the main effort is aimed at the description of nonbonded forces and torsional potential energy terms. Potential energy terms involving other covalent internal coordinates are often either quadratic-diagonal (so-called class I force fields) or simply zeroed by the use of constraints. Typical examples are the force fields AMBER [60–63], CHARMM [64–67], CHARMm/QUANTA [68], DREIDING [69], ECEPP/3 [70], ENCAD [71–73], EREF [74], GROMOS [75,76], MAB [77], MacroModel [78], OPLS [79], Tripos [80], UFF [81] and YETI [82].

2.3. Mean-solvent force fields

The purpose of a mean-solvent force field is the description of molecules in solution, but without an explicit treatment of the solvent degrees of freedom [12]. Although an accurate description of the structure, mobility, dynamics and energetics of molecules in solution generally requires an explicit treatment of the solvent, the omission of all or almost all solvent degrees of freedom dramatically reduces the computational expenses, e.g. by a factor of 10–50 for biomolecules in solution. The explicit influence of the solvent is approximated here by its mean effect, and possibly also the effect of its mean fluctuations, as in stochastic dynamics [13,83]. The main implicit influences of solvent, i.e. hydrophobic or structural effect, dielectric screening, random fluctuations and viscous drag, are mimicked by a modification of the interaction function (different functional form, additional terms, see e.g. Refs. 84 and 85) and of the equations of motion (the Langevin equation).

2.4. Low-resolution force fields

The purpose of low-resolution force fields is the study of large systems, while addressing long-timescale phenomena, such as fold recognition in proteins, protein folding, *de novo* protein design and protein–protein association. With the currently available computing power, these problems are difficult to address, using force fields at atomic resolution [7,86,87]. Force fields at the amino acid residue level are being developed for peptides and proteins [24–28]. The main difficulty is to find an adequate expression for the interaction between residues that provides a sufficient energetical resolution to discriminate correct from incorrect structures. Once a functional form is selected, the interaction function parameters are usually calibrated via a statistical analysis of native (and nonnative) protein structures. The effects of solvent are normally treated by a mean force term (Sec. 2.3). A correct description of the dynamics is not to be expected from such models.

2.5. Hybrid force fields

A whole variety of models include the combination of a treatment of a few degrees of freedom at a high particle resolution and a treatment of the others at a lower resolution. For instance, the first or first few hydration shells of a macromolecule may be included explicitly in a simulation, the bulk solvent being modelled through a mean force (Sec. 2.3). Another typical example is the simulation of chemical, or acid- or base-catalysed reactions, in solution or in enzymes [8–11,88,89]. Clearly, a quantum mechanical description of the electrons or the protons is required. However, due to the computational costs, such a treatment cannot be applied to the full system under study, and only a few relevant degrees of freedom can be treated in this way. Finding the proper interface between the different degrees of particle resolution in such hybrid models is the main difficulty here.

3. Choice of a method to sample conformational space

In addition to the present brief description, good reviews on this topic are available in the literature [29–35,90].

3.1. Methods that provide structural information

The purpose of these *search* methods is to obtain one or a number of relevant (low-energy) conformations for a given molecular system. Among the major issues of these approaches, one can cite: (i) the conformation analysis of open-chain and cyclic molecules [30]; (ii) the study of docking problems and application in drug design [91,92a,93]; (iii) the prediction of oligopeptide and protein tertiary structures and the study of the folding problem [32,33]; and (iv) structure refinement based on experimental (NMR spectroscopy, crystal X-ray diffraction) data [90,94]. In addition to the methods described below, the ones listed in Secs. 3.2 and 3.3 of course also provide structural information. However, if a correct thermodynamic and dynamical description of the system is not required, they may not be the most efficient techniques. One of the ultimate aims of these structural search techniques (except Sec. 3.1.1) is to try to locate the global energy minimum in conformational space. In the general case, this unique structure need not be functional, that is, it may have a very low statistical weight at nonzero temperature, due to entropic effects. When the energy hypersurface is highly frustrated (as is typically the case for molecules of medium and large size, like macromolecules and polymers), the global energy minimum, if narrow, may be of little relevance for the description of the macroscopic properties of the system, and other (higher but broader) minima are likely to be significantly populated. For large systems, it is even unclear whether the correct statistical ensemble corresponding to the free energy minimum will be centred on the global minimum of the energy hypersurface. In such cases, at least a collection of the lowest energy conformations (lowest local minima) should be considered [95–97] or, ultimately, a complete statistical ensemble (Secs. 3.2 and 3.3). Due to the steep increase in the density of states when raising the energy above the global minimum, the populations of higher energy conformers will increase with the temperature, which is, for example, responsible for the reversible thermal denaturation of proteins. In addition to these entropic effects, the environment of the molecule will have an influence on the preferred conformations. The lowest local minima will generally not be the same for an isolated molecule, a molecule in solution, in a crystalline environment or bound to a receptor. Unfortunately, most of the methods described below (especially those in Secs. 3.1.2–3.1.4) are not well suited for applications in the presence of explicit solvent. The question of the relevance of isolated molecule conformations for a solvated or a receptor-bound molecule may be a concern, especially for large molecules (hydrophobic effects), or molecules with polar or charged groups (charge solvation, hydrogen bonds). To some extent, mean-solvent approaches may improve the results (e.g. Refs. 32, 33 and 98).

The key problem of searching conformational space is that of dimensionality. If the size of the conformational space is estimated as the number of points in a grid defined by n discrete values in each degree of freedom (sampling density), the scaling of the conformational analysis problem is of order $O(n^N)$, where N is the number of degrees of freedom of the system. This exponential increase is sometimes referred to as the *combinatorial explosion* problem. Thus, for all but the smallest systems, searching such a grid entirely is an extremely expensive task. One can distinguish two types of approaches to this searching problem. In *nonheuristic* approaches, a set of trial conformations is generated either systematically or stochastically, and then refined by energy minimization to the closest local minimum. Since this refinement step is the expensive part of the calculation, one may try to use specific *filters* (sometimes referred to as constraints) in order to screen out, prior to energy minimization, any trial conformation which is unlikely to lead to a relevant minimum, or likely to lead to a minimum already encountered. In *heuristic* approaches, the way configurations are generated already follows more or less arbitrary *rules* (*heuristic rules*) that prevent the appearance of nonrelevant (high energy or in contradiction with experimental results) conformations or the reappearance of already known conformations. In both *heuristic* and *nonheuristic* approaches, the rules (or filters, respectively) are usually derived from physical or experimental information. One can distinguish the following cases.

A. *Structural and energetical rules or filters:* Structures containing van der Waals contact violations (highly unfavourable nonbonded contacts), structures not satisfying ring closure or having disfavourable transannular contacts [99], structures with inverted chiral centres [95,100] or incorrect bridgehead isomerism [101], protein loops with wrong terminal atom position [102] or containing fragments in a low-probability conformation (e.g. peptide units in an unallowed region of the Ramachandran map) and, in general, all high-energy structures (e.g. *heuristically* in MC and MD) should be avoided.

B. *Nondegeneracy (memory) rules or filters:* Structures which are close to an already discovered structure should be avoided. Some methods, like random searches, MC or MD, tend to generate similar structures many times, which leads to inefficiency. Typical examples of *heuristic* methods avoiding the generation of duplicate structures are the local elevation method in MD [103] or the combined use of normal and retrace pulses in the RIPS algorithm [104].

C. *Learning-based rules or filters:* Knowledge may be accumulated from previous searches for molecules that share a common structural element, and used to avoid structures of a new molecule presenting this element in an unfavourable conformation. A typical example is the distance between functionally important groups in a set of pharmacophores binding to the same receptor, which should be conserved in the active conformations of all molecules in the set [91,105]. The learning may also be performed by a so-called *Expert System*, as in WIZARD [106].

D. *Use of information from NMR experiments as rules or filters:* Typical examples are the requirement that nuclear Overhauser enhancement (NOE)-derived distances or J-coupling constants from nuclear magnetic resonance (NMR) are satisfied [94,107–109]. Methods of choice when such information is available are distance-geometry (DG) calculation and (possibly time-averaged or subsystem-averaged) distance-restrained MD refinement.

E. *Use of information from crystal X-ray diffraction experiments as rules or filters:* The results of X-ray crystallographic measurements on a series of compounds can, after a statistical analysis, be used to build conformation libraries for molecular fragments (e.g. a protein side-chain conformation library, see e.g. Ref. 110). Alternatively, known protein structures can serve as tertiary templates for other amino acid sequences, a process called *threading* (e.g. Ref. 28). In both cases, the X-ray-derived information is used as a molecule-adapted way of discretizing conformational space. It should, however, be kept in mind [30,106] that (i) crystal packing forces come into play in these solid state experiments, (ii) databases may not contain a representative sample of compounds (e.g. all must be crystallizable), and (iii) not all representative conformations of a given structural element may be present. In a different approach, the electron density map derived from the X-ray measurement is directly used during the searching phase by inclusion of a penalty term into the interaction function depending on the time-averaged simulated electron density [111–113].

When experimental data is used (D and E), one may face two problems. First, experimental errors may be present (e.g. erroneous assignments of NOE peaks to atom pairs, or of electron density peaks from X-ray crystallography to groups of atoms). Second, experimental measurements correspond in general to properties averaged over a large number of molecules and a long period of time, and the requirements that the derived constraints be satisfied in a single structure may be unrealistic. Time [90,94,114] or subsystem [115] averaging of NOE distances may be a way to somewhat relax this difficulty in MD. The use of weak coupling is also possible, although not recommended due to heating effects [116]. Time averaging has also been applied to X-ray crystallographic refinement [111–113]. When time averaging is applied, the correct dynamics of the system may to some extent be preserved.

As a further distinction between search algorithms, we shall call *consecutive* (walk) methods those which generate one molecular conformation from the previous one, thereby generating a path in conformational space, and *nonconsecutive* those which do not meet this criterion. In general, consecutive methods (typically MC, MD or SD) will have difficulties to cross energy barriers on the potential energy hypersurface. These difficulties may be partially relaxed by using various tricks (Sec. 3.1.5 and Ref. 31). Finally, structure search methods are also characterized by the use of different coordinate systems.

A. *External (Cartesian) coordinates:* The Cartesian coordinate system ($3N_{at} - 6$ coordinates, where N_{at} is the number of atoms) can be used for random searches (Sec 3.1.3) and is the standard coordinate system for MD-related methods (Sec. 3.1.5). Algorithms working with Cartesian coordinates are often easier to implement and the inclusion of constraints (e.g. bond length or ring closure) for *consecutive* methods (Sec. 3.1.5) is easy, e.g. by using the iterative algorithm SHAKE [117].

B. *Internal (torsional) coordinates:* The torsional coordinate system ($\leq N_{at} - 3$ coordinates) is often used together with fixed values for the other valence internal coordinates. This reduces considerably the number of degrees of freedom to be handled but limits the ability of a molecule to relax nonbonded strain [118]. Ring closure constraints are not easily handled in torsional space [30,95] except when algorithms such as corner flapping [119] or torsional flexing [120] are used. Since values for torsional dihedral angles are bound to $[-\pi; \pi]$, this coordinate system is well suited for systematic search (Sec. 3.1.2). Torsional coordinates may also be used in random search (Sec. 3.1.3).

C. *Interatomic distances:* Distance-geometry (DG) coordinates (matrices of interatomic distances) are particularly well suited when experimental NOE information can be included in the search [121,122].

A few tentative appraisals of the various search algorithms listed below do exist in the literature, using, for example, cyclotetradecane and 11-hydroxydecanoic lactone [95], cycloheptadecane [97], alanine dipeptide [123] and cyclosporin A [103,124,125] as (vacuum) test systems, application to the determination of side-chain conformations in proteins [110,126], or application of diverse distance-geometry methods to polyalanine chains [127]. Finally, a set of benchmark molecules for performing evaluations has been proposed [128]. The next subsections list a selection of common methods used in searching conformational space.

3.1.1. Downhill energy search methods

The aim of downhill energy search methods is to find the nearest low-energy conformation starting from a trial conformation. A wealth of energy minimization (EM) algorithms are available [129], which will find the closest local minimum in the potential energy surface, using information from the potential energy function itself and possibly its first (force vector) or second (Hessian matrix) derivative with respect to the coordinates. The second-derivative information at the minimum can be used to perform harmonic analysis in order to characterize the nearest surroundings of the minimum and to get crude estimates for the thermodynamic properties in the harmonic approximation. Other techniques can be used to find conformations (e.g. transition states) along optimal pathways connecting minima on the potential energy surface.

3.1.2. Systematic or exhaustive search methods

The aim of these methods is to exhaustively enumerate conformations in all or a significant fraction of conformational space [91,97,102,107,118,128,130,131]. The coordinates have to be discretized (grid search methods). The local minimum closest to a given grid point is then located by energy minimization. Due to the *combinatorial explosion* problem, systematic search is only tractable for systems of small and medium size [30]. A number of tree-searching algorithms have been proposed to bypass the combinatorial explosion by systematically discarding (*pruning*) through a filtering algorithm whole groups of conformations (*branches of the tree*) not satisfying a set of given (physical or experimental) constraints, prior to energy minimization [30,118]. Systematic search is usually performed in torsional space and may suffer from the inefficiency of ring closure constraints. An algorithm based on a Fourier representation of the atomic coordinates has, however, been proposed to generate systematically Cartesian coordinates for ring systems [99]. Although the idea of an exhaustive enumeration is appealing, a trade-off has to be found between grid resolution and computational efficiency, which may lead to the missing of some low-energy minima if the grid resolution is too low to provide a starting geometry in the vicinity of each local minimum.

3.1.3. Random or stochastic search methods

These methods are also sometimes referred to as Monte Carlo procedures, but we prefer to keep this term for the Metropolis algorithm (Sec. 3.2.1). They are based on the following common scheme [95,100,101,104,132–136]. Starting from a given current structure, a new structure is generated through a random change (*kick* or *pulse*) in the coordinates. The distorted structure is then energy minimized and added to the pool of generated structures. Then, a new current structure is taken from the pool, and a new iteration of the procedure is started. The working hypothesis is that low-energy conformers are generally more closely related to each other than to higher energy conformers. This assumption, although reasonable, also introduces a bias against finding low-energy conformers which are very different in geometry from the previously discovered ones. The different algorithms proposed in the literature vary with respect to the following points [30,97]:

A. *Coordinate system:* The random change may be performed in a Cartesian [100,101,104,133–135] or internal (torsional) coordinate system [95,132,136]. Changes in Cartesian coordinates tend to generate higher energy structures, and concerted torsional motions which in real dynamics would interconvert the conformers are unlikely [95]. In contrast, torsional angle changes should facilitate the sampling of low-energy regions, but relaxation of strain is limited if bond lengths and angles are frozen.

B. *Selection of the current structure:* The new current structure selected from the pool can be a random structure, the last generated structure [95,100,136], one of the lowest energy structures [95] or the currently least used structure [95]. This choice will affect the performance of the algorithm, either for finding the global minimum or for searching a large amount of conformational space.

C. *The maximum kick size:* The maximum size of the random change will influence the probability of transition to a new minimum [133] and the probability of rejection of the structure (e.g. due to van der Waals contact violations). If torsional coordinates are used, the number of torsional dihedral angles that are changed at each step may either be fixed or chosen at random [95].

D. *The filtering rules:* Various filtering rules may be applied to structures prior to the (time-consuming) energy minimization in order to discard unreasonable or already generated structures. In some variants, generated conformations are accepted or rejected according to a Metropolis scheme, thereby introducing an additional parameter, the temperature, and annealing schemes may be designed [136].

E. *The termination criterion:* The number of iterations to be performed for a complete search is difficult to estimate. The yield of new structures will decrease as the search progresses, but there is no straightforward convergence test to apply as a termination criterion [30,101]. Empirical ways to estimate convergence have been proposed based on the number of unsuccessful moves [100,104], the number of times each local minimum has appeared during the search [95], the sets of conformations generated in runs of different lengths [95] or using different annealing schemes [136].

3.1.4. Nonconsecutive heuristic search methods

Here, we consider *nonconsecutive* methods, that is, methods that generate conformations by fragments or by embedding (reduction of the dimensionality), but use no path in conformational space.

A. *Use of molecular models:* Hand-held (Dreiding, CPK) or interactive (computer graphics) molecular models may be used. In the latter case, assembly of fragments taken from a database (CSB or PDB) may be realized using various graphics interfaces. Although some insight may be gained by an examination of such models, the method is not systematic and becomes impractical for multiconformation problems.

B. *Use of artificial intelligence (AI):* In these methods (e.g. WIZARD [30,106, 137,138]), a molecule is first analysed in terms of a set of *conformational units*, which are fragments whose conformational behaviour the program has knowledge of. Each known conformation of a conformational unit is attributed a *symbol*, and corresponds to a set of coordinates or a *template*. The conformational space is then searched systematically by successively joining conformational units in all possible conformations. At any step of this buildup procedure, each trial conformation is *criticized*, first at a symbolic level and then at the coordinate level, by a so-called *Expert System*. If problems occur at this stage, an attempt can be made to resolve the problem (e.g. relax strain by a given adaptation of internal coordinates). Finally, only guesses that the Expert System has approved are minimized. Criticism may be based on (i) chemist's supplied rules (i.e. historically known unfavourable assemblies), (ii) self-learned rules (i.e. knowledge based on past experience of the program in previous searches on related compounds), or (iii) physical rules (van der Waals contact violations, effect of other intramolecular forces). The screening by an Expert System largely improves the performance of the search, but also increases the likelihood that minima are missed. The main difficulties encountered in these types of methods are (i) the design of the Expert System algorithm, (ii) the choice of the representative templates, and (iii) the generation of guesses, which may become time-consuming with respect to the minimization step.

C. *Stepwise buildup procedure:* This recursive procedure, used essentially for proteins, is based on the assumption that short-range interactions play a dominant role in determining the final conformation of a polypeptide chain [32,33,98,139,140]. The final conformation is built up stepwise, starting from known conformations of the *conformational units* (for proteins, residues). At each step, two *fragments* (peptides) containing N_1 and N_2 conformational units, respectively, are assembled to generate an enlarged fragment of length M ($\leq N_1 + N_2$). Possible conformations for the enlarged fragment are generated by systematically combining possible low-energy conformations of the two smaller fragments. The combinations are energy minimized and only those conformations are stored for which the energy is below an *energy cutoff*. To reduce the high memory requirements of the method, one may choose to store only one representative side-chain conformation for a given peptide backbone conformation. The choice of the energy cutoff and of the side-chain conformation in the single representative structure of a fragment is not straightforward since high-energy conformations at the fragment stage can still correspond to low-energy conformations in the assembled molecule. Finally, the choice of the fragment assembly procedure (i.e. where should the procedure be started along the polypeptide chain and how should it proceed) and the number of overlap residues ($M - N_1 - N_2$) may influence the result of the search.

D. *Distance geometry (metric matrix method):* Distance geometry (DG) is at the origin a purely geometric method that does not require a force field or a starting geometry [94,121,122,127,141]. If a system of N atoms is specified solely by $\frac{1}{2}N(N - 1)$ pairwise interatomic distances (distance matrix), and all the specified distances satisfy triangular, tetrangular, etc. inequalities, a single solution (set of atomic coordinates) exists in $N - 1$ dimensions. In this high-dimension space, the metric matrix (matrix of the dot products of the atomic coordinates) is easily generated from the distance matrix. This metric matrix is then gradually *embedded* in the lower dimensional spaces, that is, lower dimensional Cartesian coordinates are generated so that higher dimensional coordinates are closest to zero. In practice, upper and lower bounds are specified to each distance, based on the closest allowed van der Waals contacts, covalent coordinates (standard geometries), distance information from NOE, and possible other problem-specific constraints. The embedding is then performed for various random (or chosen according to certain criteria) distance values within the bounds (possibly after smoothing to satisfy the triangle inequalities). Since bounds to distances are often unevenly spread over the structure and may contain errors, and since not all random combinations give rise to a reasonable solution in three dimensions, generated structures have to be refined. This may be done by minimizing an *error function* that describes the quality of the structure, either by EM or by other techniques (MD). The error function includes distance bound violations, chirality violations and possibly an empirical interaction term. The second term is required since chirality cannot unambiguously be defined for dimensions higher than three and will often be incorrect in embedded structures. The method requires storage of $O(N^2)$ and is intrinsically slow for generating new structures with respect to random or systematic methods, but becomes competitive when distance constraints (NOE) are available for nonbonded atom pairs. The probability distribution of structures in the final ensemble may be a concern.

E. *Genetic algorithms:* Genetic algorithms (GA) [92,126,142–144] are optimization procedures inspired from the natural genetic evolution. A *population* of *individuals* (conformers), described by a *symbolic* encoding (*string* or *chromosome*), is maintained and evolves from *generation* to generation, keeping its size constant. The symbols may represent, for example, discrete values of a given dihedral angle. In addition, each individual is characterized by its *fitness* (e.g. negative energy or Boltzmann factor). Evolution of the population from one generation to the next occurs through four types of processes (*genetic operators*), each regulated by acceptance criteria (*selection*) based on the fitness function and possibly also on a stochastic element: (i) *replication*, i.e. a direct copy of a high-fitness individual to the next generation; (ii) *elimination*, i.e. removal of a low-fitness individual from the population; (iii) *mutation*, i.e. random changes (occurring with a low probability) in some symbols of an individual; and (iv) *crossovers*, i.e. the interchange of regions between a pair of high-fitness individuals (*parents*) generating mixed individuals (*children*). The crossover mechanism allows for the combination into a unique structure of favourable substructures that evolved separately, so that larger and larger good substructures will tend to stay in the population. On the other hand, the mutations preserve the diversity in the population and prevent premature convergence. In this sense, GA is an implicit buildup scheme, with a few additional advantages: (i) crossover may imply changes in dihedrals which are not close along the chain and thus include nonlocal conformational preferences; (ii) every individual in the population is a complete structure, and the energy has a more unambiguous meaning than for a fragment; and (iii) the selection of good substructures is fully automatic and involves no choices and human intervention. In principle, the structure should be energy minimized for the estimation of the fitness. The method may then, however, become expensive. Another inconvenience is the difficulty to deal with cyclic molecules. The efficiency of the algorithm will depend largely on the size of the population, the number of generations, the number of symbols per fragment (discretization of conformations) and the mutation rate.

3.1.5. Consecutive heuristic search methods

These methods, which generate one molecular conformation from a previous one, are generally based on MC, MD or SD schemes (Secs. 3.2 and 3.3). The major problem of these three techniques is crossing high-energy barriers ($\gtrsim k_BT$). Since for systems of medium and large size, the potential energy surface is generally complex with many barriers, all but the smallest ones are surmounted very infrequently and the above methods have a small radius of convergence. The aim of the algorithms described below is thus to combine them with modifications that spoil their ability to give a correct thermodynamic or dynamical description of the system, but substantially enhance their search power by lowering barriers or allowing them to be circumvented [31]. In general, the corresponding paths in conformational space are non-Newtonian and energy will not be conserved. Some kind of temperature regulation has thus to be applied (e.g. Ref. 125).

A. *Smoothing of the potential energy function in order to reduce barriers*

1. *Reverse collapse methods:* These types of methods [33,145,146] have been applied successfully to Lennard-Jones clusters, and are, in principle, generalizable to any interaction function which is a sum of pairwise-distance-dependent terms. The

potential energy hypersurface is deformed by modifying the atom–atom Lennard-Jones interaction through a *deformation parameter* γ, such that the real surface is recovered at $\gamma = 0$, and the surface collapses into a single basin around the origin when $\gamma = 1$. This is done either by deformation of the Lennard-Jones interaction by the diffusion equation method [147] or by shifting the pairwise distance entering in the Lennard-Jones function by a γ-dependent offset [33,145]. Since the global minimum of the collapsed surface is known, the reverse procedure has a meaningful starting point. γ is progressively decreased from 1 and the structure is energy minimized for every new γ decrement. Apart from rotation and translation and when a single permutation of identical atoms is considered, the method is deterministic (independent of any starting configuration). Although there are indications that no bifurcation occurs when the deformation parameter is decreased (i.e. that the final structure is the global minimum), no formal proof of this exists.

2. *Potential energy scaling:* In this approach, the magnitude of specific terms of the interaction function is scaled down by a *scaling parameter* β [148,149]. If all the terms are scaled, the search power is increased in a similar way as upon increasing the temperature. Since the temperature is a property of the overall system, the former method is advantageous for explicit solvent simulations, since the intramolecular terms may be scaled selectively. The original interaction is recovered by letting β go to 1. The method may also be used as an annealing scheme.

3. *Use of a soft-core potential:* The steepest barriers on the potential energy surface are due to van der Waals repulsion between atoms, which increases steeply when atoms start overlapping each other, and gives rise to a singularity when atoms occupy the same location. In the soft-core method [150], the functional form of the nonbonded interaction (the Lennard-Jones and electrostatic term) is changed through a *soft-core parameter* α. For any $\alpha > 0$, the interaction becomes finite at zero interatomic distance, and its magnitude decreases on increasing α. When α is large enough, atoms may pass through each other, which significantly increases the search power. At any time, the interaction can be relaxed to its original form by letting α decrease to 0.

4. *Extension of the dimensionality:* Extension of the dimensionality is a way to reduce the number of local minima, and provide energetically tractable pathways to pass barriers which could not be crossed in three dimensions. The 4D-MD refinement method [125,151] takes advantage of this. The interaction function is modified through a *4D coupling parameter* μ, so that when $\mu = 1$ the atoms interact according to energy terms based on their four-dimensional coordinates, while when $\mu = 0$ only their three-dimensional coordinates are used, the fourth coordinate being uncoupled. The interaction can, at any time of the search, be relaxed to the original interaction by letting μ decrease to 0 (dimensionality annealing). In order to limit the increase in the accessible volume of conformational space in the higher dimension, the dimensionality is not increased above 4 and a harmonic restraint prevents atoms from moving too far away from the three-dimensional hyperplane. Four-dimensional refinement is also used in DG calculations [122].

B. *Inclusion of a time-dependent memory term into the potential energy function*

In the local elevation (LE) method [103] the problem of repeatedly visiting the same low-energy regions of conformational space is addressed. To persuade the system to visit new areas, a penalty term is included into the interaction function in the form of a time-dependent memory function. The relevant degrees of freedom (torsional) are discretized and the penalty term is defined as a sum of truncated Gaussian functions, centred at each grid point, whose magnitude is proportional to the number of times the neighbourhood of the grid point was visited. Molecular dynamics is then used to integrate the equations of motion and the trajectory progressively maps out low-energy regions of conformational space. The method is memory intensive and requires a fast storage/comparison routine. It is therefore only applicable to a limited number of degrees of freedom, which may be only a subset of the degrees of freedom of the real system. Finally, a trade-off has to be found here between grid spacing, search power and memory requirement.

C. *Scaling of system parameters*

1. *Simulated temperature annealing:* The term annealing describes the process of slowly cooling a system [152–155]. High-temperature dynamics improves the search power of MC or MD, but also favours the selection of high-energy, high-entropy conformations. This behaviour is improved in simulated annealing (SA). In SA, one starts from a high temperature, where transitions out of local minima are facilitated, and then progressively decreases the temperature to almost zero. The probability that the system ends up in a very low energy conformation is high if the cooling is carried out slowly enough.

2. *Scaling of the atomic masses:* In this method [156], the atomic masses and the temperature of the system are scaled by a common factor. The increased kinetic energy and inertia lead to a larger amplitude in vibrational motions and thus increase the probability of torsional dihedral angle transitions. On the other hand, by equipartition, the atomic mean-square velocities, $< v_i^2 > = 3k_BT/m_i$, are unchanged and there is no need to use a shorter timestep for integrating the equations of motion, which is not the case if only the temperature is raised (SA).

3. *Potential energy annealing:* In PEACS [124], new dynamical laws (equations of motion) are defined, which meet the requirements for a good search method. The potential energy of the system is weakly coupled to an external bath. The reference energy level of the bath is slowly decreased (annealed) during the simulation, which should result in low potential energy structures at the end. A velocity correction (along the force vector) is made to relax the potential energy V to a value V_0 using a weak-coupling-type equation (first-order relaxation). The annealing may be performed automatically by slowly decreasing the reference level to the minimal potential energy value encountered during the simulation. The RUSH algorithm [123] is based on a similar principle.

D. *Direct search methods*

A direct search method has been used to attack the multiple minima problem [157]. The method is essentially a simplex method, but the size of the simplex is adapted to the shape of the potential energy surface, so that barriers can be crossed.

3.2. Methods that provide structural and thermodynamic information

The purpose of these methods is to *sample* conformational space, or part of it, in order to obtain a collection of conformations that represent a statistical mechanical ensemble from which thermodynamic quantities can be derived. Classical statistical mechanics uses the concept of an (infinite) ensemble of N_{sys} systems, $E \equiv \{(\mathbf{q}_i, \mathbf{p}_i), i = 1, \ldots, N_{sys}\}$ with $N_{sys} \rightarrow \infty$, where $\mathbf{q}_i \equiv \{q_{i\alpha}, \alpha = 1, \ldots, 3N_{at}\}$ is the generalized coordinate vector and $\mathbf{p}_i \equiv \{p_{i\alpha}, \alpha = 1, \ldots, 3N_{at}\}$ is the generalized momentum vector of system i. When the composition, volume and temperature are held constant (NVT or canonical ensemble), the probability distribution of systems in the ensemble, $\rho(\mathbf{q}, \mathbf{p})$, obeys the distribution

$$\rho(\mathbf{q}, \mathbf{p}) = \frac{e^{-\mathscr{H}(\mathbf{q},\mathbf{p})/k_BT}}{\int \cdots \int d\mathbf{q}\, d\mathbf{p}\, e^{-\mathscr{H}(\mathbf{q},\mathbf{p})/k_BT}} \tag{3.2.1}$$

where $\mathscr{H}(\mathbf{q}, \mathbf{p})$ is the Hamiltonian (total energy) of a system $(\mathbf{q}, \mathbf{p})$ (see Sec. 3.3.4), k_B is the Boltzmann constant and T is the absolute temperature. Under the assumptions that the kinetic energy term of the Hamiltonian contains the only dependence of the Hamiltonian on $\mathbf{p}$ (this is not true when constraints are applied), the kinetic energy contribution can be integrated and the description limited to an ensemble of N_{conf} conformations, $C \equiv \{\mathbf{q}_i, i = 1, \ldots, N_{conf}\}$ with $N_{conf} \rightarrow \infty$. The probability distribution of conformations in C, $p(\mathbf{q})$, obeys

$$p(\mathbf{q}) = \frac{e^{-V(\mathbf{q})/k_BT}}{\int \cdots \int d\mathbf{q}\, e^{-V(\mathbf{q})/k_BT}} \tag{3.2.2}$$

where $V(\mathbf{q})$ is the total potential energy of the system in conformation $\mathbf{q}$. Finally, ensembles obtained from finite simulations are of finite size. Under the assumption that a representative fraction of conformational space has been sampled, $p(\mathbf{q})$ may be written in a discrete form

$$p(\mathbf{q}_i) = \frac{e^{-V(\mathbf{q}_i)/k_BT}}{\sum_j e^{-V(\mathbf{q}_j)/k_BT}} \tag{3.2.3}$$

An ensemble of conformations that satisfy this equation will be called a Boltzmann ensemble.

Since the sequence of the conformations is not relevant here, the methods of choice are the ones which satisfy Eq. 3.2.3 but achieve the highest sampling efficiency. In addition to the methods described below, the ones listed in Sec. 3.3 also provide structural and thermodynamic information. However, if a correct dynamical description of the system is not required, they may not be the most efficient techniques.

3.2.1. (Metropolis) Monte Carlo sampling methods

In the Metropolis Monte Carlo algorithm [38,158], random steps are taken and accepted with a probability

$$p(\Delta\mathbf{q}) = \min\{1, e^{-\Delta V(\Delta\mathbf{q})/k_BT}\} \tag{3.2.1.1}$$

It can be shown that the method generates a Boltzmann ensemble. Steps may cross barriers higher than k_BT in the potential energy surface, provided that these are narrow. Internal or Cartesian coordinates can be used, but the method may be inefficient in Cartesian coordinate space for systems with many covalent bonds, due to the poor trial configuration acceptance probability if steep contributions to the potential energy (e.g. from bonds) are present.

3.2.2. Methods that generate a biased statistical ensemble

These methods do not generate a correct thermodynamic ensemble, but a biased ensemble that can easily be converted to a Boltzmann ensemble. The prototype is umbrella sampling, in which a well-defined potential energy term is added to the physical potential energy function in order to restrict the accessible conformational space.

3.2.3. Methods that provide differences in thermodynamic observables

The so-called *coupling parameter approach* can be used for calculating differences in thermodynamic observables between two states A and B (i.e. two Hamiltonians) of a given system when individual values of the observable at A and B are inaccessible. In practice, the method is used to calculate differences in free energies. In the general case, if the Hamiltonian along a given pathway (often unphysical) between states A and B can be cast in the form

$$\mathscr{H}(\mathbf{q},\mathbf{p};\lambda) \quad \text{with } \mathscr{H}(\mathbf{q},\mathbf{p};0) = \mathscr{H}_A(\mathbf{q},\mathbf{p}) \text{ and } \mathscr{H}(\mathbf{q},\mathbf{p};1) = \mathscr{H}_B(\mathbf{q},\mathbf{p}) \qquad (3.2.3.1)$$

and if the proper ensemble averages can be sampled along this pathway, differences between observables which are state functions may be computed. To achieve a higher sampling efficiency, any of the smoothing parameters defined in Sec. 3.1.5A may be introduced (soft-core, potential energy scaling, extension of the dimensionality) with a specified λ-dependence so that, at $\lambda = 0$ or 1, the original interaction function is recovered [148,150,151]. The *thermodynamic cycle* approach essentially follows the same principle: a direct pathway which cannot be sampled accurately is replaced by an indirect pathway for which sampling is easier.

3.3. Methods that provide structural, thermodynamic and dynamical information

In this case, one wants to *simulate* the motion, and equations of motion which explicitly contain time are required. Possible techniques are classical [83] molecular dynamics (MD), quantum [9] molecular dynamics (QMD) or stochastic [83] dynamics (SD) simulations. These methods generally have a small radius of convergence for potential energy surfaces with many barriers higher than k_BT.

3.3.1. Time-dependent Schrödinger equation

The time-dependent Schrödinger equation describes the nonrelativistic evolution of a quantum system in terms of its time-dependent wave function $\psi(\{\mathbf{r}_i\}, t)$:

$$\hat{\mathscr{H}}(\{\mathbf{r}_i\})\psi(\{\mathbf{r}_i\},t) = -i\hbar\frac{\partial}{\partial t}\psi(\{\mathbf{r}_i\},t) \tag{3.3.1.1}$$

where $\{\mathbf{r}_i\}$ denotes the coordinate vectors of all particles in the system and

$$\hat{\mathscr{H}}(\{\mathbf{r}_i\}) = \hat{K}(\{\mathbf{r}_i\}) + V(\{\mathbf{r}_i\}) = \sum_i^N \frac{\hbar^2}{2m_i}\frac{\partial^2}{\partial \mathbf{r}_i^2} + V(\{\mathbf{r}_i\}) \tag{3.3.1.2}$$

is the quantum mechanical Hamiltonian operator of the system, assumed here to be independent of time (isolated system). The second equality is only valid in a Cartesian coordinate system, but can be easily generalized. Integration of this equation of motion, where ψ is expanded as a linear combination of basis functions, is called quantum molecular dynamics (QMD). Possible discretization schemes have been reviewed elsewhere [159]. Due to the difficulty of the procedure, QMD is only directly applicable to very small systems or small parts of larger systems, the other degrees of freedom being treated classically (hybrid methods). Finding the proper coupling between the quantum and classical subsystems is then an important area of research.

3.3.2. Newton's equations of motion

Newton's (classical) equations of motion are valid only in Cartesian coordinates $\mathbf{r}$. If all forces in the system are conservative and derive from a potential $V(\mathbf{r})$, which is normally a good approximation at the atomic level, Newton's equations of motion can be expressed as

$$m_i\ddot{\mathbf{r}}_i = \mathbf{F}_i(\{\mathbf{r}_i\}) = -\frac{\partial V(\{\mathbf{r}_i\})}{\partial \mathbf{r}_i} \tag{3.3.2.1}$$

where $\mathbf{r}_i$ is the Cartesian coordinate vector of particle i, m_i is its mass, $\mathbf{F}_i$ is the force on atom i and a double dot on a quantity denotes its second derivative with respect to time. An equivalent expression for the overall system is the following second-order differential equation:

$$\mathbf{M}\ddot{\mathbf{r}} = \mathbf{F}(\mathbf{r}) = -\frac{dV(\mathbf{r})}{d\mathbf{r}} \tag{3.3.2.2}$$

where $\mathbf{r} \equiv \{\mathbf{r}_i, i = 1, \ldots, N\}$ is the 3N-dimensional vector describing the Cartesian coordinates of all N particles and $\mathbf{M}$ is a diagonal $3N \times 3N$ matrix containing the masses. If the velocity vector is introduced, an alternative formulation involves two first-order differential equations

$$\mathbf{M}\dot{\mathbf{v}} = -\frac{dV(\mathbf{r})}{d\mathbf{r}} \quad \text{and} \quad \dot{\mathbf{r}} = \mathbf{v} \tag{3.3.2.3}$$

Numerical integration of Newton's equations of motion is usually performed using this last form. Possible discretization schemes are described elsewhere [160,161]. Integration of Eq. 3.3.2.2 with respect to coordinates from $\mathbf{r}^0(t^0)$ to $\mathbf{r}(t)$, followed by derivation with respect to time, leads to the result of energy conservation:

$$\frac{d}{dt}[\tfrac{1}{2}{}^{T}\dot{\mathbf{r}}\mathbf{M}\dot{\mathbf{r}} + V(\mathbf{r})] = \frac{d}{dt}[K(\dot{\mathbf{r}}) + V(\mathbf{r})] = 0 \tag{3.3.2.4}$$

where $K(\dot{\mathbf{r}})$ is the kinetic energy.

3.3.3. Lagrange's equations of motion

Lagrange's equations of motion are a generalization of Newton's equations of motion applicable to any coordinate system $\mathbf{q}$ (generalized coordinates). They involve a function called the Lagrangian $L(\mathbf{q}, \dot{\mathbf{q}})$ of the system:

$$L(\mathbf{q}, \dot{\mathbf{q}}) = K(\mathbf{q}, \dot{\mathbf{q}}) - V(\mathbf{q}) \tag{3.3.3.1}$$

where $\mathbf{q} \equiv \{q_\alpha, \alpha = 1, \dots, 3N\}$ is the generalized coordinate vector describing the system, $\dot{\mathbf{q}} \equiv \{\dot{q}_\alpha, \alpha = 1, \dots, 3N\}$ is the corresponding generalized velocity vector, and $K(\mathbf{q}, \dot{\mathbf{q}})$ is the kinetic energy, which now also depends on generalized coordinates. In terms of this function, the equation of motion is a second-order differential equation

$$\frac{d}{dt}\left(\frac{\partial L(\mathbf{q}, \dot{\mathbf{q}})}{\partial \dot{\mathbf{q}}}\right) = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}})}{\partial \mathbf{q}} \tag{3.3.3.2}$$

This equation is easily converted to Newton's equation of motion in the special case where Cartesian coordinates are used, but has a wider range of applicability. This formalism is well suited for the inclusion of holonomic, i.e. time-independent, constraints (freezing of a specified generalized coordinate) and for the inclusion of additional artificial degrees of freedom to the system (extended Lagrangian). Formulae derived in generalized coordinates are most often converted and applied in Cartesian coordinates.

3.3.4. Hamilton's equations of motion

Hamilton's equations of motion form a symmetrization of Lagrange's equation of motion. The conjugate momentum vector $\mathbf{p}$ associated to the generalized coordinate vector $\mathbf{q}$ replaces here the conjugate velocity vector $\dot{\mathbf{q}}$. The vector $\mathbf{p}$ is defined by

$$\mathbf{p} = \frac{\partial L(\mathbf{q}, \dot{\mathbf{q}})}{\partial \mathbf{q}} \tag{3.3.4.1}$$

The Hamiltonian $\mathscr{H}(\mathbf{q}, \mathbf{p})$ of the system is then defined by

$$\mathscr{H}(\mathbf{q}, \mathbf{p}) = K(\mathbf{q}, \mathbf{p}) + V(\mathbf{q}) \tag{3.3.4.2}$$

where we recognize the total energy of the system, see Eq. 3.3.2.4. In terms of this function, the equations of motion are two first-order differential equations

$$\frac{\partial \mathscr{H}(\mathbf{q}, \mathbf{p})}{\partial \mathbf{p}} = \dot{\mathbf{q}} \quad \text{and} \quad \frac{\partial \mathscr{H}(\mathbf{q}, \mathbf{p})}{\partial \mathbf{q}} = -\dot{\mathbf{p}} \tag{3.3.4.3}$$

This equation is easily converted to Newton's equation of motion in the special case where Cartesian coordinates are used, but has a wider range of applicability. **p** is then the usual linear momentum vector.

3.3.5. Langevin's equations of motion

Langevin's equations of motion provide a better description of the dynamics when some degrees of freedom are not treated explicitly in a model, but one would like to include the effect of their mean fluctuations. A typical example of application are implicit solvent models. Here, nonconservative forces are included, and the energy will not be conserved. The following second-order equation is used:

$$m_i\ddot{\mathbf{r}}_i = \mathbf{F}_i(\{\mathbf{r}_i\}) = -\frac{\partial V_{\mathrm{mean}}(\{\mathbf{r}_i\})}{\partial \mathbf{r}_i} + \mathbf{R}_i - m_i\gamma_i\dot{\mathbf{r}}_i \tag{3.3.5.1}$$

where V_{mean} is a potential of mean force including the average effect of the omitted degrees of freedom, $\mathbf{R}_i$ is a stochastic force accounting for the effect of random collisions, and the last term accounts for dissipative effects and is proportional to a friction coefficient γ_i. Integrating this equation of motion is called stochastic dynamics (SD, see Ref. 83). To a first approximation, $\mathbf{R}_i$ may be assumed to obey a simple Gaussian distributed probability. In this case, the width of the Gaussian is related to the temperature so that the energy introduced by the random force balances the energy removed from the system by the stochastic force, i.e.

$$< \mathbf{R}_i^2 > = 6m_i\gamma_i k_B T \tag{3.3.5.2}$$

where k_B is the Boltzmann constant and T is the temperature. This explicit relation to the temperature introduces an explicit coupling to a heat bath. More elaborate treatments may introduce spatial and/or time correlation in the stochastic force (the generalized Langevin equation). When the inertial term, the left-hand side in Eq. 3.3.5.1, is small compared to other forces, the equation reduces to the following first-order differential equation:

$$m_i\gamma_i\dot{\mathbf{r}}_i = \mathbf{F}_i(\{\mathbf{r}_i\}) = -\frac{\partial V_{\mathrm{mean}}(\{\mathbf{r}_i\})}{\partial \mathbf{r}_i} + \mathbf{R}_i \tag{3.3.5.3}$$

Integration of this equation of motion is called Brownian dynamics (BD) or diffusive stochastic dynamics.

3.3.6. Liouville's equations of motion

In this representation, interest is focused on the density of states $\rho(\mathbf{q}, \mathbf{p}, t)$, i.e. the proportion of systems in an infinite statistical ensemble belonging to the hypercube of edge $(d\mathbf{q}, d\mathbf{p})$ around the point $(\mathbf{q}, \mathbf{p})$ in phase space. The conservation of the total number of systems in the ensemble at any time can be expressed by the following equation:

$$\frac{\partial \rho(\mathbf{q},\mathbf{p},t)}{\partial t} = -\sum_{i}^{3N}\left(\frac{\partial \rho}{\partial q_i}\dot{q}_i + \frac{\partial \rho}{\partial p_i}\dot{p}_i\right) = -\sum_{i}^{3N}\left(\frac{\partial \rho}{\partial q_i}\frac{\partial \mathscr{H}}{\partial p_i} - \frac{\partial \rho}{\partial p_i}\frac{\partial \mathscr{H}}{\partial q_i}\right) \tag{3.3.6.1}$$

where the second equality arises from the Hamiltonian equations of motion, Eq. 3.3.4.3. Equation 3.3.6.1 forms the basis for the simulation of nonequilibrium processes [162,163]. It is easily seen that Eq. 3.2.1 is a solution of Eq. 3.3.6.1 in the equilibrium case, which indicates that the classical equations of motion generate a Boltzmann ensemble.

4. Assumptions underlying empirical classical interaction functions

The only justification of empirical classical atomic interaction functions resides in their ability to reproduce and predict a vast amount of experimental results. Usually, most of the information used in their design and calibration comes from experiment and not from quantum mechanical calculations. Thus, no theoretical justification is in principle required as long as a force field is successful at reproducing data from experiment. It is nevertheless useful to try to understand the reason of the agreement (or the cause of discrepancies) by considering the relationship between the force-field building blocks (energy terms) and the underlying quantum mechanical reality.

4.1. Implicit degrees of freedom and the assumption of weak correlation

Whatever the degrees of freedom chosen to be treated explicitly within a force field, the reality behind remains quantum mechanical and involves the interaction between nuclei and electrons. Since the electronic degrees of freedom, and sometimes those of a number of nuclei, do not appear in the definition of the empirical classical potential energy function, but are still present in the underlying reality, they may be called *implicit*. The fundamental assumption (or approximation) on which empirical classical force fields are based is that the correlation between the fluctuations in these implicit degrees of freedom and the fluctuations in those which are handled explicitly can be neglected. Under this assumption, the fluctuations in the implicit degrees of freedom can be averaged out, leaving only their mean effect. This assumption is in essence a generalization of the Born–Oppenheimer principle, which allows the separation of the nuclear and electronic degrees of freedom based on the large difference between nuclear and electronic masses. Within the framework of this principle, a *mean* or *effective* potential energy function (or *potential energy surface*, PES) can be defined, which describes the interaction of the nuclei in the instantaneously averaged potential of the electron cloud. More precisely, if μ denotes the electronic degrees of freedom and i the nuclear ones, the mean potential energy describing the interaction of the nuclei, V_{nuc} $(\{\mathbf{r}_i\})$, is defined as the lowest eigenvalue of the electronic time-independent Schrödinger equation at a given configuration of the nuclei, $\{\mathbf{r}_i\}$:

$$\hat{\mathscr{H}}_\mu(\{\mathbf{r}_\mu\}, \{\mathbf{r}_i\})\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\}) = V_{nuc}(\{\mathbf{r}_i\})\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\}) \tag{4.1.1a}$$

with

$$\hat{\mathscr{H}}_\mu(\{\mathbf{r}_\mu\}, \{\mathbf{r}_i\}) = \hat{\mathscr{H}}_{tot}(\{\mathbf{r}_\mu\}, \{\mathbf{r}_i\}) - \hat{\mathrm{K}}_i(\{\mathbf{r}_i\}) \tag{4.1.1b}$$

where $\hat{\mathscr{H}}_\mu$ is equal to the total Hamiltonian of the system, $\hat{\mathscr{H}}_{tot}$, minus the kinetic energy operator $\hat{\mathrm{K}}_i$ corresponding to the nuclear degrees of freedom, and $\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\})$ is the ground state electronic wave function, which depends on $\{\mathbf{r}_i\}$ only parametrically. This treatment is valid only for an isolated system (time-independent total Hamiltonian), in which electronically excited states play no role. In Eq. 4.1.1, the assumption that the nuclei are motionless with solving the electronic problem allows for the decoupling of the $\hat{\mathrm{K}}_i$ operator from the Hamiltonian. The nuclear problem is then described by a nuclear time-independent Schrödinger equation

$$\hat{\mathscr{H}}_i(\{\mathbf{r}_i\})\Phi_i(\{\mathbf{r}_i\}) = \mathrm{E}_{tot}\Phi_i(\{\mathbf{r}_i\}) \tag{4.1.2a}$$

with

$$\hat{\mathscr{H}}_i(\{\mathbf{r}_i\}) = \mathrm{V}_{nuc}(\{\mathbf{r}_i\}) + \hat{\mathrm{K}}_i(\{\mathbf{r}_i\}) \tag{4.1.2b}$$

where the eigenvalues E_{tot} are the allowed values for the total energy of the system in its different vibrational and rotational states, and the $\Phi_i(\{\mathbf{r}_i\})$ are the corresponding nuclear wave functions. Very often, the further assumption is made that the motion of the nuclei in the mean potential $\mathrm{V}_{nuc}(\{\mathbf{r}_i\})$ can be treated classically. From a thermodynamical point of view, this approximation is normally valid for all but the lightest atoms and at high enough temperature, that is, when the classical and quantum partition functions become equivalent. When this classical treatment is adequate, using the Hellmann–Feynman theorem two equivalent formulations can be given to Eq. 4.1.1:

$$\mathrm{V}_{nuc}(\{\mathbf{r}_i\}) = <\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\})|\hat{\mathscr{H}}_\mu(\{\mathbf{r}_\mu\}, \{\mathbf{r}_i\})|\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\})>_\mu \tag{4.1.3a}$$

or

$$\frac{\partial \mathrm{V}_{nuc}(\{\mathbf{r}_i\})}{\partial \mathbf{r}_j} = - \mathbf{F}_{nuc,j}(\{\mathbf{r}_i\}) = <\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\})\left|\frac{\partial \hat{\mathscr{H}}_\mu(\{\mathbf{r}_\mu\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j}\right|\psi_\mu(\{\mathbf{r}_\mu\}; \{\mathbf{r}_i\})>_\mu \tag{4.1.3b}$$

where $< \cdots >_\mu$ denotes integration with respect to the $\{\mathbf{r}_\mu\}$ variables only and $\mathbf{F}_{nuc,j}(\{\mathbf{r}_j\})$ is the Hellmann–Feynman 'classical' force acting on nucleus j. This means that when the μ degrees of freedom are quantum mechanical, the *mean potential* (first definition) and the *potential of mean force* (second definition) are equivalent (within a constant). This is the case in an all-atom force field, where the classical interaction is described by the Born–Oppenheimer potential energy surface.

When classical degrees of freedom, m, are further removed from the interaction function by averaging (e.g. nuclei of solvent molecules or of protein side-chains), the two types of definition are no longer equivalent. Thermodynamic quantities defined in terms of an ensemble average of a microscopic (instantaneous) observable depending on the explicit degrees of freedom i will be described equivalently by an all-atom model and a lower particle resolution model (in an NVT ensemble) if the Boltzmann factors are identical, that is if

$$e^{-V_{MF}(\{\mathbf{r}_i\})/k_BT} = \int \cdots \int d\mathbf{r}_m \, e^{-V_{nuc}(\{\mathbf{r}_m\},\{\mathbf{r}_i\})/k_BT} \tag{4.1.4}$$

where $V_{MF}(\{\mathbf{r}_i\})$ is the interaction at the lower particle resolution, k_B is the Boltzmann constant and T is the (absolute) temperature. Differentiating the negative logarithm of Eq. 4.1.4 with respect to the coordinate $\mathbf{r}_j$ of an explicitly treated particle j leads to the correct statistical mechanical definition of $V_{MF}(\{\mathbf{r}_i\})$ as a *potential of mean force*:

$$\begin{aligned}\frac{\partial V_{MF}(\{\mathbf{r}_i\})}{\partial \mathbf{r}_j} &= -\mathbf{F}_{MF,j}(\{\mathbf{r}_i\}) = < -\mathbf{F}_{nuc,j}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\}) >_m \\ &= \left\langle \frac{\partial V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j} \right\rangle_m \end{aligned} \tag{4.1.5}$$

where $\mathbf{F}_{MF,j}(\{\mathbf{r}_i\})$ and $\mathbf{F}_{nuc,j}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})$ are the forces on atom j derived from the potential energies $V_{MF}(\{\mathbf{r}_i\})$ and $V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})$, and $< \cdots >_m$ denotes (Boltzmann) ensemble averaging over all possible sets of coordinates $\{\mathbf{r}_m\}$ at a given (fixed) set of coordinates $\{\mathbf{r}_i\}$. On the other hand, the derivative of the *mean potential energy*, $V_{mean}(\{\mathbf{r}_i\})$, is

$$\begin{aligned}\frac{\partial V_{mean}(\{\mathbf{r}_i\})}{\partial \mathbf{r}_j} &= \frac{\partial}{\partial \mathbf{r}_j} < V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\}) >_m \\ &= < \frac{\partial V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j} >_m \\ &\quad - \frac{1}{k_BT}\Bigg[< V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\}) \frac{\partial V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j} >_m \\ &\quad - < V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\}) >_m < \frac{\partial V_{nuc}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j} >_m \Bigg] \end{aligned} \tag{4.1.6}$$

When all the degrees of freedom are averaged out from the system, V_{MF} becomes the Helmholtz free energy of the system, A (see Eq. 4.1.4), and V_{mean} its internal energy, U. The second term in Eq. 4.1.6 can thus be interpreted as an entropic force, which arises because each set of explicit coordinates actually maps areas of different Boltzmann weighted sizes in the nuclear potential energy surface. This force can be identified by subtracting Eq. 4.1.5 from Eq. 4.1.6:

$$\begin{aligned}\frac{\partial V_{\text{entrop}}(\{\mathbf{r}_i\})}{\partial \mathbf{r}_j} &= -\mathbf{F}_{\text{entrop},j} = \frac{\partial}{\partial \mathbf{r}_j}[V_{\text{mean}}(\{\mathbf{r}_i\}) - V_{\text{MF}}(\{\mathbf{r}_i\})] \\ &= -\frac{1}{k_B T}\left[< V_{\text{nuc}}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\}) \frac{\partial V_{\text{nuc}}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j} >_m \right. \\ &\quad \left. - < V_{\text{nuc}}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\}) >_m < \frac{\partial V_{\text{nuc}}(\{\mathbf{r}_m\}, \{\mathbf{r}_i\})}{\partial \mathbf{r}_j} >_m \right] \\ &= -\frac{1}{k_B T}\left[\left\langle V_{\text{nuc}} - < V_{\text{nuc}} >_m \right\rangle_m \left\langle \frac{\partial V_{\text{nuc}}}{\partial \mathbf{r}_j} - < \frac{\partial V_{\text{nuc}}}{\partial \mathbf{r}_j} >_m \right\rangle_m \right]\end{aligned} \tag{4.1.7}$$

This entropic force is proportional to the covariance between fluctuations in the nuclear potential energy and the force derived from it, and vanishes at high temperatures. In practice, Eqs. 4.1.5–4.1.7 only give access to the derivatives of the quantities analogous to the thermodynamic quantities A, U and TS with respect to the coordinates of explicit particles, and thus define the corresponding potential energies within a constant. The estimation of absolute values would require (i) a complete sampling of conformational space and the use of Eq. 4.1.4, and (ii) a knowledge of the absolute value of V_{nuc} with respect to infinitely separated nuclei and electrons. For most purposes, absolute values are not required.

Although it may be very difficult to design in practice, a potential of mean force can in principle always be defined through Eq. 4.1.3 or Eq. 4.1.5, in the latter case within a constant, which will give the correct *thermodynamic* representation of the system (averages and fluctuations of any quantities expressed as ensemble averages of microscopic observables defined in terms of the explicit degrees of freedom i). However, it will only give an acceptable *dynamical* representation of the system if the correlation between fluctuations in the implicit and explicit degrees of freedom can be neglected (*assumption of weak correlation*). This will be the case if the two classes of degrees of freedom relax with very different timescales, that is, changes in $\{\mathbf{r}_i\}$ are slow enough and changes in $\{\mathbf{r}_m\}$ (or $\{\mathbf{r}_\mu\}$) are rapid enough, so that the m (or μ) degrees of freedom generate a full ensemble quasi-instantaneously before any change in $\{\mathbf{r}_i\}$ can take place. If this condition is satisfied, the effect of the mean force will be a good approximation to the cumulative effect of instantaneous forces. Although the approximation is reasonable for electronic degrees of freedom in many cases, this is often not the case for the classical ones (e.g. solvent nuclei). Correlation may then be introduced through forces acting on explicit degrees of freedom, which attempt to mimic the time-dependent effect of fluctuations in the implicit degrees of freedom (Langevin treatment) or by inclusion of a few additional degrees of freedom into the system (extended Lagrangian treatment). In the Langevin-type treatment [6,13], two additional forces acting on an explicit atom j are added to the mean force, a stochastic force $\mathbf{R}_j$ and a frictional force $\mathbf{D}_j$:

$$\mathbf{F}_j(t) = \mathbf{F}_{MF}(\{\mathbf{r}_i(t)\}) + \mathbf{R}_j(\{\{\mathbf{r}_i(\tau)\}, 0 < \tau < t\}) + \mathbf{D}_j(\{\mathbf{r}_i(t)\}, \dot{\mathbf{r}}_j(t)) \tag{4.1.8}$$

The simplest choice for $\mathbf{R}_j$ is a purely random force (components obeying Gaussian distributions). If $\mathbf{R}_j$ depends on $\{\mathbf{r}_i(t)\}$, correlation in space may be introduced into the stochastic force. If it also depends on previous configurations $\{\{\mathbf{r}_i(\tau)\}, 0 < \tau < t\}$, correlation in time (memory) may be introduced into the stochastic force. The frictional force is often chosen proportional to the opposite of the velocity $\dot{\mathbf{r}}_j(t)$ of atom j at time t (viscous drag), although it might be more realistic to consider also its degree of interaction with implicit degrees of freedom (e.g. solvent accessibility), depending on $\{\mathbf{r}_i(t)\}$. In the extended-Lagrangian-type treatment, classical degrees of freedom are introduced into the system and in the interaction function, which are easier to handle than the real implicit degrees of freedom, and aim at introducing fluctuations in an approximate manner. Typical examples are the approximative inclusion of fluctuations in the electronic degrees of freedom using a point charge on a spring, by letting the atomic charges fluctuate [164], or using the path-integral method. The modelling of a heat bath or pressure bath (surroundings of the system) by a single degree of freedom coupled to the system also belongs to this class of methods.

In practice, the averaging over the implicit degrees of freedom can be performed either by: (i) quantum mechanical calculation or simulation at high particle resolution, and the use of Eq. 4.1.3 or Eq. 4.1.5, respectively; (ii) analytical theories for simple systems; or (iii) an educated guess of the functional form of the potential of mean force, followed by adjustment of its parameters to reproduce experimental or quantum mechanical data. Typical examples of each method are: (i) the use of potential energy surfaces from *ab initio* molecular orbital calculations for tuning empirical force-field energies [40], the averaging of a molecular dynamics trajectory in solution over the solvent degrees of freedom [165]; (ii) the analytical continuum models used to describe solvent around a solute in a cavity, such as continuum reaction field models [15,18,166,167] or RISM equations [168,169]; and (iii) the expansion of the interaction function in a Taylor series for bonded interactions, a cosine series for torsional interactions, Coulomb plus van der Waals functions for nonbonded interactions, and possibly a solvent accessible surface area dependent term for mean-solvent effects [85], followed by parameter tuning.

Perhaps one of the most striking illustrations that force-field parameters are *effective parameters* (i.e. corresponding to a potential of mean force) is the examination of the atomic point charges used in condensed-phase force fields. To describe properly the effect of the electron cloud on the electrostatic interaction in the bulk phase, a polarization term would be required at each interaction site. Since most force fields do not contain such a term, charges used for bulk-phase simulations have to incorporate the average polarization effect and are thus considerably increased with respect to charges which would be suitable for the gas phase. For example, the dipole moment of the SPC water model [170], optimized for condensed-phase properties, is 2.27 D, 23% higher than the experimental gas-phase dipole of the water molecule, 1.85 D. Explicit inclusion of a polarizability allows one to reproduce correct condensed-phase behaviour with much lower charges [171].

4.2. Energy terms and the assumption of transferability

Under the assumption of weak correlation between implicit and explicit degrees of freedom (Sec. 4.1), a (classical) effective interaction function $V_{MF}(\{\mathbf{r}_i\})$, or loosely $V(\{\mathbf{r}_i\})$, can be defined depending solely on the explicit degrees of freedom. In the following, the explicit elementary unit will be called 'atom', whatever this unit actually is. For practical purposes, this interaction function has to be modelled in some way by an *analytical function.* If no further approximation is made, this function will, in general, be specific to a given molecular system and will depend on the coordinates of all atoms simultaneously. This dependence can be very intricate, if the required level of accuracy is high. Two distinct lines can be followed when designing this analytical function.

In the first approach, a Taylor expansion of the potential energy surface around an equilibrium conformation $\{\mathbf{r}_i^0\}$ can be performed (in Cartesian coordinates), up to the required accuracy, that is, a polynomial approximation is generated depending simultaneously on the coordinates of all atoms. The coefficients will be matrices of increasing rank,

$$V_{\text{anal}}(\{\mathbf{r}_i\}) = V(\{\mathbf{r}_i^0\}) + \sum_j^{N_{at}} \left.\frac{\partial V}{\partial \mathbf{r}_j}\right|_{\{\mathbf{r}_i^0\}} (\mathbf{r}_j - \mathbf{r}_j^0)$$

$$+ \sum_j^{N_{at}} \sum_k^{N_{at}} ({}^t\mathbf{r}_j - {}^t\mathbf{r}_j^0) \left.\frac{\partial^2 V}{\partial \mathbf{r}_j\, \partial \mathbf{r}_k}\right|_{\{\mathbf{r}_i^0\}} (\mathbf{r}_k - \mathbf{r}_k^0) + \cdots \qquad (4.2.1)$$

The eigenvectors of the matrix containing the second derivatives (Hessian matrix) can be used to define a unique, nonredundant and orthogonal basis set for the description of the system as in a spectroscopic force field. This is mathematically satisfactory, but of limited practical application since:

A. The equilibrium conformation is usually not known but something one would like to predict.

B. A system at equilibrium is generally characterized by more than one conformation.

C. Other conformations (nonequilibrium) are often also of interest – in which cases neither the Taylor expansion at $\{\mathbf{r}_i^0\}$ nor the corresponding well-defined basis set are usable.

D. The accurate description through a Taylor expansion for one configuration does not provide much insight into other parts of the configurational space, and so does not describe the physics of the system.

E. The accurate description of one molecule is useless for predictions about other molecules.

The second approach relies on the use of a sum of functionally simple analytical functions (*energy terms*) depending on selected internal coordinates, chosen on the basis of chemical intuition. This is justified because:

A. A wealth of chemical data tells us that entities such as bonds, bond angles, torsional angles and nonbonded interactions are physically meaningful, and thus the corresponding internal coordinates and distances in space appear as the natural choice in which the functional forms of the interactions are likely to adopt the most simple forms.

B. Since such internal coordinate potential energy terms are physically meaningful, they may give an appropriate description of a larger part of configurational space.

C. Since internal coordinates involve a limited number of atoms (one to four), there is a hope to obtain building blocks transferable from one molecule to another (*assumption of transferability*).

In other words, one wants to split the interaction function into a sum of functionally simple, physically meaningful (and thus insight-providing) terms, which would in addition be transferable from one molecule to another, and thus bear predictive power. These terms are called *force-field terms*. The assumption that these terms exist is similar to the one mentioned in Sec. 4.1, i.e. that for each term (e.g. bond, bond angle, etc.), the effect of the environment can be averaged out by considering an ensemble of *molecular systems* (*topologies*) and *conformations* (*geometries*). More explicitly, one would like to have an analytical expression of the form

$$V_{\text{anal}}(\{\mathbf{r}_i\}) = \sum_{\text{terms }\alpha} V^{\alpha}_{\text{anal}}(\{\mathbf{r}_j, j \in \alpha\}) \tag{4.2.2}$$

where the notation $j \in \alpha$ indicates that the atom j is involved in the force-field term α and $V_{\text{anal}}(\{\mathbf{r}_i\})$ is the analytical representation of the potential energy surface as a sum of terms α. This analytical representation and the mean force defined in Eq. 4.1.4 will give the same description of the thermodynamic properties of any molecular system if the Boltzmann factors are identical:

$$e^{-V_{MF}(\{\mathbf{r}_i\})/k_BT} = \prod_{\alpha} e^{-V^{\alpha}_{\text{anal}}(\{\mathbf{r}_j,\, j\in\alpha\})/k_BT} \tag{4.2.3}$$

If this equation is integrated with respect to the coordinates $\{\mathbf{r}_k, k\notin\beta\}$ which do not appear in a given force-field term β, and its negative logarithm is differentiated with respect to one of the coordinates $\mathbf{r}_j$, $j \in \beta$, one gets after rearrangement

$$\frac{\partial V^{\beta}_{\text{anal}}(\{\mathbf{r}_j, j\in\beta\})}{\partial \mathbf{r}_j} = < \frac{\partial V_{MF}(\{\mathbf{r}_j\}, \{\mathbf{r}_k\})}{\partial \mathbf{r}_j} >_k$$
$$- \sum_{\alpha\neq\beta, j\in\alpha} < \frac{\partial V^{\alpha}_{\text{anal}}(\{\mathbf{r}_j, j\in\alpha\}, \{\mathbf{r}_k, k\in\alpha\})}{\partial \mathbf{r}_j} >_k \tag{4.2.4}$$

where $< \cdots >_k$ denotes ensemble averaging over the coordinates $\{\mathbf{r}_k, k\notin\beta\}$. With the exception of harmonic spectroscopic force fields where the coordinates are normal mode vectors, the second term in Eq. 4.2.4 will always be present in force fields, because covalent energy terms and distance-dependent terms often act on the Cartesian coordinate of the same atom. This equation means that when the coordinate $\mathbf{r}_j$

is also involved in force-field terms α other than β, the energy term β becomes correlated to these other terms. The *geometric correlation* arising from this so-called *coordinate redundancy* may be removed by solving consistently Eq. 4.2.4 for a given molecular system. The *topological correlation* arises from the fact that the term β may be correlated to a given set of terms α for one molecular system, and to a different set for another one. This type of correlation may only be removed by considering a collection of different molecular systems. A procedure analogous to this is followed in the consistent design of force fields using *ab initio* data (Sec. 7.5). The set of conformations used there is, however, not a proper statistical ensemble, but an arbitrary selection of conformations. Unlike in the definition of the potential of mean force in Sec. 4.1 (Eq. 4.1.4), it is not guaranteed that Eq. 4.2.4 has an exact solution. In practice, one fixes functional forms for the energy terms V^{α}_{anal} and calibrates the parameters. The quality of the solution will depend on the flexibility and physical sensibleness of the selected functions.

As a consequence of this second averaging process, the parameters characterizing the force-field term α are not those found in any specific molecule, but rather characteristic of an ensemble of molecules and conformations. Thus, a force-field term is really a virtual entity incorporating the average effect of various possible environments. This means also that the parameters corresponding to a given term may be dependent on the class of compounds for which the term was calibrated (Sec. 7.6.3). When the averaging process is performed correctly and the analytical functions selected for the energy terms are sensible enough so that Eq. 4.2.4 has a good solution, the interaction function will be able to give a correct picture of the thermodynamic properties of molecular systems. It is not guaranteed, however, that a correct dynamical picture will be obtained.

4.3. Coordinate redundancy and assumption of transferability

Force fields are usually defined in terms of internal valence coordinates for the covalent interactions, and atom–atom distances for the nonbonded interactions (even if in practice the forces are calculated in Cartesian coordinates). When a molecule includes atoms with a valence higher than two, the valence internal coordinates themselves become redundant, i.e. some of them are linearly dependent from the others. The problem of *internal coordinate redundancy* can be illustrated in the case of formamide [21]. A conformation of formamide is fully specified by a set of $3N - 6 = 12$ internal coordinates. By systematic counting, one sees, however, that there are five bonds, six angles, four dihedrals and two out-of-plane coordinates, that is, 17 available internal coordinates, all likely to have an influence on the energy. Five of these are necessarily redundant. For example, the HCO, HCN and OCN angles are dependent since, due to the planarity of the carbonyl group, they must sum up to 2π. This raises the question of whether it is possible to assign specific and transferable properties to, say, a generic amide OCN angle, since it cannot vary in an independent manner from the other angles. Valence coordinate redundancy is handled differently from one force field to another, as is illustrated here for the case of ethane. Out of the nine torsional dihedrals that can be defined in ethane, only one is required to describe the relative rigid-body motion of the two methyl groups with respect to one another. There are three possibilities: (i) use a torsional coordinate which involves the six hydrogens (this is not easily generalizable to less symmetrical cases); (ii) use one of the dihedrals only (this induces asymmetry in the system and requires arbitrary choices); or (iii) accept the redundancy and calculate the nine four-body torsions with a force constant divided by nine (this may be computationally inefficient). Both choices (ii) and (iii) are found in current force fields. At last, the definition of relevant internal coordinates requires some insight into which choice may lead to the best transferable entities. For example, if the pyramidality around a nitrogen centre is maintained by three bond-angle potential energy terms, it will be difficult to get at the same time the correct bond-angle vibrational frequencies and the pyramidal inversion barrier.

Even if redundancy in the valence coordinates is avoided by limiting their number to $3N - 6$, ultimately, it will be introduced by the nonbonded interaction. Distances within a molecule are dependent on the valence coordinates, and thus the nonbonded interactions will introduce the strain effects into the valence coordinates. For example, nonbonded interaction should induce strain on the central C-C bonds in tri-tert-butyl methane so that its length is about 0.16 nm, whereas in most force fields the equilibrium bond length is about 0.152–0.153 nm.

4.4. Choices made in the averaging processes

Choices to be made with respect to the averaging process will be discussed using the example of a C-C bond. As pointed out in the previous two sections, the factors that will influence the effective length in a given conformation of a given compound are:

A. The bond *potential energy term* (a potential of mean force over an ensemble of molecules).

B. A possible *explicit dependence on topology* through different classes of C-C bonds, when different C atom types are used.

C. A possible *explicit dependence on topology and geometry* through valence coordinate cross-terms (Sec. 6.5).

D. The *implicit dependence on topology and geometry* through nonbonded strain.

For example, the DREIDING force field [69] excludes both factors B and C, which is perhaps not very accurate, but makes parametrization easy. In most force fields for biomolecules, option B is used and different classes of C-C bonds are defined, depending on the connectivity and environment of the bonded atoms. This is more accurate, allows for a specific calibration of the chemical entities required for a given purpose, and leads to a simple interaction function. In class II force fields, which are meant to be very accurate for molecules in vacuum, option C is mostly used. The inconvenience here is that many parameters are required for all but the simplest systems, the interaction function is complicated by the cross-terms, and parameters have to be calibrated all together in a consistent way. On the other hand, the interaction function is very accurate and elegant, since few atom types have to be defined (e.g. in the CFF93 force field for alkanes, only two, C and H).

5. General characteristics of the empirical interaction function

5.1. Interaction function parameters and molecular topology

An empirical interaction function, loosely called a force field, V, is defined by its functional form and the parameters that enter into its definition, i.e. its *interaction function parameters*, $\{s_i\}$. In order to express this latter dependence, the notation

$$V = V(\mathbf{q}; \{s_i\}) \tag{5.1.1}$$

can be used, where $\mathbf{q}$ is the 6N-dimensional vector defining (in any coordinate system) the configuration of the molecular system. This information is, however, not complete. In order to model a specific system, some information on the *molecular topology* is required. This arises from the fact that, in contrast to first-principles techniques, empirical force fields are based on a potential energy function that averages out the electronic degrees of freedom. This results in very different interaction regimes if different relationships between atoms at the electronic level exist. For example [21], the interaction between an Na^+ and a Cl^- ion at 1 nm in the gas phase can be calculated by solving the Schrödinger equation for the electrons of the ion pair and for two separate ions. However, this is not really required or even useful, since one can say quite safely that the potential will have a K_{el}/R dependence, where K is a constant and R is the distance between the ions. If the ions come closer and form a molecule, the Schrödinger equation can be solved again for different internuclear separations around the equilibrium distance R_{eq}. However, we know that a function like $K_b(R - R_{eq})^2$ gives a reasonable approximation to the true energy. From this example, it is clear that (i) the Schrödinger equation contains information that exceeds our needs, (ii) a correct description of relative energies is sufficient, and (iii) the analytical approximation does not bypass the quantum mechanical character of the interaction, but rather captures the essential physics from its solution. On the other hand, if the analytical description is more intuitive and computationally cheaper, it will require more information about the ions, namely, the electronic regime (bonded, nonbonded) and the parameters (K_{el}, K_b, R_{eq}) specific to the pair. It is also clear that the transition between the bonded and nonbonded regime will be a problem. Even when the interaction between two pairs of bonded atoms is described by the same functional form, the best choice of function parameters is likely to be different if the bonds are not identical. To summarize, the molecular topology information is required to decide which interaction is to be treated in the framework of which functional form and using which values for the parameters. By analogy, the only proper molecular topology information required for an *ab initio* molecular orbital calculation at a certain level of theory is the number of protons and electrons for each atom. Note that due to coordinate redundancy (Sec. 4.3), the specification of a molecular topology is in most cases not unique.

5.2. Atom types and combination rules

In a number of empirical force fields, the basic unit is the atom, which is usually considered as a charged mass point with no directionality and no internal degrees of freedom. When simulating large systems of biomolecules, hydrogen atoms are often implicitly included into the heavy atoms that are bearing them, to form so-called *united atoms.* This significantly reduces the number of degrees of freedom, since hydrogen atoms constitute about 50% of the total atom number in proteins and 30% in DNA. From a molecular dynamics point of view, this also offers the advantage of removing the high-frequency C-H bond stretching motion and enables the use of a larger timestep to integrate the equations of motion. However, since a correct modelling of the hydrogen bond becomes problematic when the donor is treated as a united atom, some force fields use a mixed method, where polar (and possibly aromatic) hydrogens are handled explicitly, whereas all the other (nonpolar) hydrogens are included into united atoms (see Table 1). The negative effects of the suppression of explicit hydrogens are the loss of dipole and quadrupole moments (this is not too serious for hydrogens linked to carbons) and a loss of steric effects (the united atoms are spherical). Finally, there are cases where the united-atom approach fails even for nonpolar hydrogens, and an explicit inclusion of all hydrogens may be required for a proper description of the system [172,173].

Common force fields usually define a limited number of *atom types* (possibly united atom). These are atoms (or groups) which are physically and chemically (i.e. with respect to their physical environment) alike. This number varies from one force field to another (e.g. 2 in CFF93/alkanes [41] and 65 in OPLS/proteins [79]). The purpose of these atom types is to facilitate the attribution of interaction function parameters to n-body interaction terms, while generating the molecular topology information for a specific system. The assumption is that the parameters s_i for an n-body interaction term between n atoms α of atom type a_α are solely determined by the types of these atoms, irrespective of their environment, i.e.

$$s_i = s_i(\{a_\alpha, \alpha = 1, \dots, n\}) \qquad (5.2.1)$$

Such rules are called *combination rules* and are an important part of the definition of a force field. Depending on the structure of the simulation program and on the type of rule, they can be weakly (i.e. easily overriden) or strongly implemented. The former possibility is to be preferred, since generation of the molecular topology information using combination rules and possible manual editing offers more flexibility [36]. One of the most well-established (and physically based) combination rules is Coulomb's law, where the magnitude of the interaction between two atoms is given by the product of the (point) charges corresponding to each atom type. Combination rules for valence terms (bond, bond angle, torsional dihedral angle and out-of-plane coordinates) are generally given in the form of tables as a function of the constituting atom types. These tables may include 'wild cards', indicating that the same parameter is to be used irrespective of the atom type of the specified atom, which reduces significantly the amount of required parameters. For van der Waals parameters, proper combination rules are still a matter of discussion (Sec. 6.6.2). Note that, in principle, the combination rules could also include atoms which are not directly participating in the interaction, but define the environment more precisely. For example, a bond type could be defined by the bonded atoms and the next covalently bonded atoms. This would, however, very rapidly increase the complexity of the force field and, to our knowledge, has never been done. Instead, when the environment of an atom is significantly modified by the type of neighbouring atoms, two different atom types are usually defined to distinguish the different environments. The use of few atom types has the advantage of simplicity and ease of parametrization. For example, if four atom types are defined for carbon (C(sp^3), C(sp^2), C(sp) and C(aromatic)), only 10 bond types have to be parametrized, but the sensitivity of the bond behaviour to the environment is low. Twenty types would surely allow one to account better for the detailed influence of the chemical environment, but this would then imply the parametrization of as much as 210 bond types, which may be a hard task.

5.3. Expression for the classical Hamiltonian

As in the quantum description of a molecular system, the classical Hamiltonian (total energy of the system) depends simultaneously on the coordinates and the momenta of all particles in the system. In a similar manner as in Hartree–Fock calculations, where the electronic Hamiltonian is approximated by a sum of one- and two-electron operators, the classical Hamiltonian can be approximated by a sum of n-body terms:

$$\begin{aligned}\mathscr{H}_{\text{class}}(\{\mathbf{q}_i,\mathbf{p}_i\}) \approx \sum_i [{}^{(1)}K(\mathbf{p}_i) + {}^{(1)}V(\mathbf{q}_i)] + \sum_i \sum_{j>i} {}^{(2)}V(\mathbf{q}_i,\mathbf{q}_j) \\ + \sum_i \sum_{j>i} \sum_{k>j} {}^{(3)}V(\mathbf{q}_i,\mathbf{q}_j,\mathbf{q}_k) + \cdots \qquad (5.3.1)\end{aligned}$$

where i, j, k, ... are indices running over the N particles constituting the system, or a subset of these, $\mathbf{q}_i$ and $\mathbf{p}_i$ are the coordinate and momentum vectors of particle i, and the (n) superscripts indicate the order of the terms. The three (single or multiple) sums in Eq. 5.3.1 correspond to the first three n-body terms of a force field, i.e. n = 1, 2, 3. The principal terms that are used in current force fields, either with a physical or a nonphysical (i.e. *ad hoc*, to perturb the system or impose restraints derived from experimental information) meaning, are listed in Table 2. The computational effort for calculating an n-body interaction term is either (i) of order O(M), M being the length of a list of possible combinations of indices entering the multiple sums of Eq. 5.3.1, if such a list is available, or (ii) of order $O(N!/(n!\cdot(N-n)!))$ if all combinations have to be calculated. Covalent interactions are typically of type (i), whereas nonbonded interactions are of type (ii). For systems of a reasonable size, N^2 will always be larger than M for any list of covalent interactions, and the bulk of computer time will be used to calculate two-body nonbonded interactions. The computation of $\mathscr{H}_{\text{class}}$ is thus essentially an $O(N^2)$ problem. Even for relatively small systems, the inclusion of three-body

Table 2 *n-Body interaction terms found in common force fields*

(n)	Subset	Type	Term
1	All atoms	P	Kinetic energy
	Charged atoms	P	Interaction with an external electric field
	Surface atoms	P	Stochastic/frictional force on a macromolecule
	Listed or all atoms	U	Atomic positional restraining
2	All-atom pairs	P	Pairwise nonbonded interaction (point charges, point charge/point dipole etc., van der Waals, solvent accessible surface area interaction)
	Bonded atoms	P	Covalent bond
	H-bonded atoms	P	H-bonding interaction (acceptor–donor)
	Listed atom pairs	U	Distance restraining
3	All-atom triples	P	Triple nonbonded interactions (expensive, seldom used)
	Atoms in bond angle	P	Covalent bond-angle bending
	Pairs of bond	P	Bond–bond cross-term
	Bond in angle	P	Bond–angle cross-term
4	Atoms in dihedrals	P	Torsional interaction, improper dihedral interaction
	H-bonded atoms	P	H-bonding (acceptor-antecedent, acceptor, hydrogen, donor)
	Pairs of angle	P	Angle–angle cross-term (around one centre)
	Atoms in dihedral	P	Bond–dihedral cross-term (central bond), angle–angle–torsion cross-term
	Atoms in dihedral	U	J-value restraining, local elevation
≥ 5	Covalent neighbours	P	Other cross-terms among bonds, angles and dihedrals
N	All atoms	P	Point polarizability
	All atoms	U	Radius of gyration unfolding force

(n): order of the term, i.e. the number of particles involved in the interaction term, N indicates all atoms; Subset: subset of atoms for which the term is calculated, either from a list or all atoms (pairs, triples, respectively); Type: physical (P) or 'unphysical' (U) term.

nonbonded terms is extremely expensive [174,175]. On the other hand, the evaluation of a single N-body term is an inexpensive problem. Examples may be the inclusion of point polarizabilities at atomic sites (when the interaction between induced dipoles is neglected) or the radius of gyration interaction that can be used to force protein unfolding in a molecular dynamics simulation [86].

6. Interaction function terms used in current force fields

In this section the most commonly used interaction function terms and corresponding combining rules are listed and briefly discussed. Only terms bearing a direct physical interpretation will be described here. For the sake of completeness it should be noted that 'unphysical' terms can be incorporated in the interaction function for the following purposes:

A. Modification of the potential energy surface in order to enhance the power of a given method to search the conformational space (Sec. 3.1.5).

B. Description of unphysical pathways linking two physical states of the system, as used for example in free energy calculations (Sec. 3.2.3), or restriction of the accessible conformational space to the neighbourhood of a given point, as used in umbrella sampling (Sec. 3.2.2).

C. Direct incorporation of experimental information in the form of constraints or (possibly time-averaged or subsystem-averaged) restraints, in order to enforce the agreement between simulation and experiment.

D. Various engineering purposes, for example restriction of the motion in selected parts of the system (position restraining energy term).

Finally, terms mimicking the potential of the mean force effect of omitted supra-atomic degrees of freedom (e.g. solvent, the side chain of protein residues) will not be discussed here.

6.1. Bond-stretching term

6.1.1. Functional forms

When simulations are performed at room temperature, and when no chemical (bond-breaking) reaction is involved, bond lengths usually remain close to their equilibrium values. The bond-stretching contribution to the potential energy can then be approximated adequately by a Taylor expansion [39]

$$E_b(\{b_i\}; \{b_i^0, {}^{(2)}k_{b,i}, {}^{(3)}k_{b,i}, \dots\}) = \sum_{\text{all bonds i}} [{}^{(2)}k_{b,i}(b_i^0 - b_i)^2 + {}^{(3)}k_{b,i}(b_i^0 - b_i)^3 + \cdots] \qquad (6.1.1.1)$$

where b^0 is the equilibrium bond length and ${}^{(n)}k_b$ is the 'force constant' corresponding to the term of power n. There is no first-order term since the derivative of the potential energy has to be zero when $b = b^0$. For example, in the MM2 force field [20,53], terms are retained till the third (cubic) power. This has the disadvantage that the potential becomes negative for high internuclear separation and, thus, an inadequate coordinate choice may cause bond dissociation. A quartic expansion is used in the MM3 [20,54] and CFF93 [22,23,41] force fields, which fixes this problem. Although the inclusion of anharmonic terms ($n > 2$) clearly improves the description of vibrational properties of molecules in the gas phase, it may not do so in other applications. When oscillations with large amplitudes are considered, when the effect of nonbonded strain on a bond length and stretching frequency are of interest, or when the breaking of a bond is required, other functional forms can be used. For example, as in the CVFF force field [47–50], a Morse-type function may be used:

$$E_{\text{Morse}}(\{b_i\}; \{b_i^0, D_i, \alpha_i\}) = \sum_{\text{all bonds i}} D_i[e^{\alpha_i(b_i^0 - b_i)} - 1]^2 \qquad (6.1.1.2)$$

where D is the well depth, b^0 is the equilibrium bond length and α is a parameter determining the width of the well. This equation already encompasses anharmonicities and provides a better description than a limited Taylor expansion around and away from the equilibrium bond length. Many other functions have been proposed [176,177], such as the Linnett, Lippincott, Rydberg and Varshni functions. Most of these have been calculated *a priori* or tailored for diatomic molecules, but at least some may be applied successfully to individual bonds in polyatomic molecules [178]. In addition to the functional form, the expansion variable may be changed, e.g. [179,180]

$$\begin{array}{ll} (b^0 - b)/b^0 & \text{Dunham} \\ b^0/(b^0 - b) & \text{Dinur/Hagler} \\ (b^0 - b)/b & \text{Simons/Parr/Finlan} \\ 2(b^0 - b)/(b^0 + b) & \text{Ogilvie} \end{array} \qquad (6.1.1.3)$$

Note that Eq. 6.1.1.1 corresponds to a Dunham expansion. The use of dissociative functions, such as the Morse function, for modelling a bond-breaking process remains, however, limited to specific systems and chemical reactions because (i) they are difficult to parametrize, and (ii) in the general case, the effect of bond breaking is not only local to a single bond and implies corresponding changes in the parameters of other covalent and nonbonded interaction terms.

In a large number of force fields (e.g. AMBER, CHARMM, GROMOS, OPLS, etc.) and especially for the simulation of large molecules or the simulation of systems in explicit solvent, the detailed formalisms mentioned above are not used. A Taylor expansion limited to the second-order (harmonic) term is assumed to be sufficient since (i) the high bond-stretching (and bond-angle bending) frequencies are weakly coupled to the rest of the system, and (ii) the low-frequency motions (conformational changes, solvent relaxation) largely determine the thermodynamic properties of the system. In other words, bond description is assumed not to be critical and the simplest function with the fewest parameters is preferred. The evaluation of the bond-stretching interaction may be made less expensive by using the quartic expression

$$E_b(\{b_i\}; \{b_i^0, k_{b,i}\}) = \sum_{\text{all bonds } i} k_{b,i}[(b_i^0)^2 - (b_i)^2]^2 \qquad (6.1.1.4)$$

which avoids a square-root operation in the calculation of the energy and force. In molecular dynamics simulations, since a proper integration of the (uninteresting) high-frequency bond-stretching vibrations requires timesteps of the order of 0.5 fs, a further (and common) time-saving technique is to constrain the bonds to their equilibrium lengths using an iterative algorithm such as SHAKE [117], which allows for the use of timesteps 4–5 times longer without substantially affecting the dynamics [181]. It has been shown, however, that the bond angles should not be constrained simultaneously. In virtually all current force fields, bonded atoms (first neighbours) are excluded from any nonbonded interactions (Sec. 6.6–6.9). This interaction would, in most cases, be unrealistically large and should already be encompassed in the bond-stretching potential energy term.

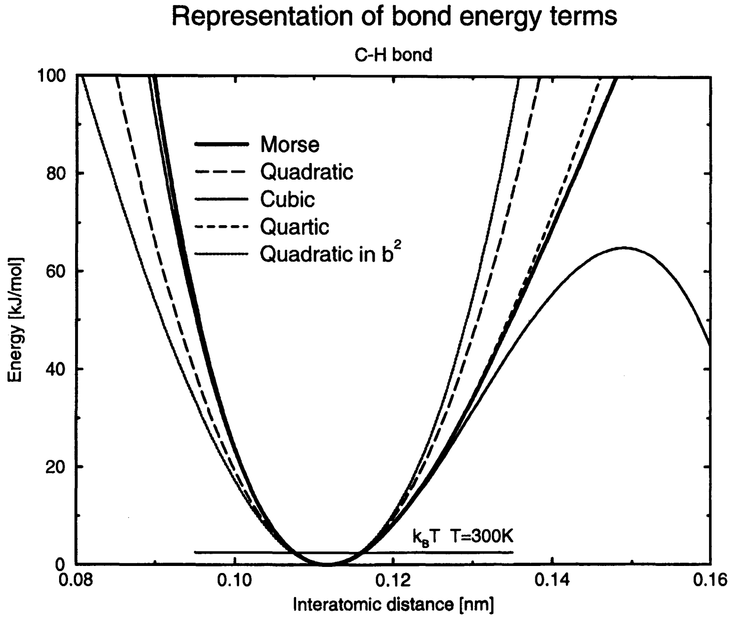


Fig. 2. Representation of bond energy terms for a C-H bond. The Morse curve (thick line) corresponds to $D = 438.2$ kJ/mol, $\alpha = 17.87$ nm^{-1} and $b^0 = 0.112$ nm, Eq. 6.1.1.2. Other curves correspond to Taylor expansions up to various powers, Eq. 6.1.1.1, and to the quartic expansion of Eq. 6.1.1.4. The expansion coefficients have been chosen to give the same curvature at the minimum for all functions. The horizontal line at $E = 2.5$ kJ/mol indicates the value of $k_B T$ at $T = 300$ K, where k_B is the Boltzmann constant.

In Fig. 2, a graphical representation of some of the energy terms described above is given for a C-H bond. The force constants have been chosen to give the same curvature at the minimum as the reference Morse function. The cubic expansion tends towards $-\infty$ for large distances (dissociative behaviour). The Morse function levels off to D at large distances, whereas all the even-power expansions grow to ∞, being smoother than the Morse curve below b^0 and steeper beyond. Except for the harmonic expansion, all functions are asymmetric around b^0 and the average bond length will not be equal to the equilibrium length b^0. At room temperature and for unstrained bonds, all functional forms are virtually equivalent for most purposes.

6.1.2. Combination rules

Combination rules for covalent bond interaction parameters are usually given in the form of a table as a function of the atoms that define the bond. An exception is the DREIDING force field [69], which uses an arithmetic combination rule

$$b_i^0(a,b) = R^0(a) + R^0(b) - 0.001 \text{ [nm]} \qquad (6.1.2.1)$$

where a and b are the atom types of the atoms forming bond i, and $R^0(a)$, $R^0(b)$ are the covalent radii corresponding to these atom types. The (harmonic) bond-stretching force constant is determined solely by the bond order.

6.2. Bond-angle bending term

6.2.1. Functional forms

Most of the considerations applying to bond-stretching terms also apply here. For small deformations around the equilibrium bond angle, a Taylor expansion can be used:

$$E_\theta(\{\theta_i\}; \{\theta_i^0, {}^{(2)}k_{\theta,i}, {}^{(3)}k_{\theta,i}, \ldots\}) = \sum_{\text{all angles i}} [{}^{(2)}k_{\theta,i}(\theta_i^0 - \theta_i)^2 + {}^{(3)}k_{\theta,i}(\theta_i^0 - \theta_i)^3 + \cdots] \qquad (6.2.1.1)$$

where θ^0 is the equilibrium angle and ${}^{(n)}k_\theta$ is the 'force constant' corresponding to the term of power n. For example, an expansion up to the fourth power is used in CFF93 [22,23,41], the second- and sixth-power terms are retained in MM2 [20,53], and MM3 [20,54] uses a full expansion up to the sixth power. An alternative potential energy term which is used in some force fields, such as the CHARMM all-atom force field for DNA [67], is the Urey-Bradley energy term

$$E_\theta(\{\theta_i\}; \{\theta_i^0, k_{\theta,i}, {}^{(1)}k_{d,i}, {}^{(2)}k_{d,i}\}) = \sum_{\text{all angles i}} [k_{\theta,i}(\theta_i^0 - \theta_i)^2 + {}^{(1)}k_{d,i}(d_i^0 - d_i) + {}^{(2)}k_{d,i}(d_i^0 - d_i)^2] \qquad (6.2.1.2)$$

where d_i is the 1,3 distance between atoms forming the extremity of the angle, d_i^0 is its equilibrium value and ${}^{(n)}k_d$ is the 'force constant' corresponding to the term of power n. If E_θ is defined to within a constant and d_i^0 is replaced by an effective distance, the linear term in Eq. 6.2.1.2 can be omitted without loss of information [182]. This function includes some anharmonicity and a coupling between the angle and the constituting bonds.

Again, in a number of force fields (e.g. AMBER, GROMOS, OPLS, etc.) dealing with large molecules or molecules in the bulk phase, only the harmonic term is retained in Eq. 6.2.1.1. A harmonic function in the angle cosine is also sometimes used [69] for computational efficiency:

$$E_\theta(\{\theta_i\}; \{\theta_i^0, k_{\theta,i}\}) = \sum_{\text{all angles i}} k_{\theta,i}(\cos\theta_i^0 - \cos\theta_i)^2 \qquad (6.2.1.3)$$

In virtually all current force fields, atoms separated by one single atom (second neighbours) are excluded from any nonbonded interactions (Secs. 6.6–6.9). This interaction would, in most cases, be unrealistically large and should already be encompassed in the bond-angle bending potential energy term.

6.2.2. Combination rules

Combination rules for bond-angle bending parameters are usually given in the form of a table as a function of the types of the atoms that define the angle. An algebraic empirical combination rule for estimating harmonic angle bending from *ab initio* results or spectroscopic force fields has, however, been proposed [183]:

$$k_{\theta,i}(a,b,c) = K\,Z(a)C(b)Z(c)\,(b^0_{ab} + b^0_{bc})^{-1}\,(\theta^0_{abc})^{-2}\exp\left(-2\frac{(b^0_{ab} - b^0_{bc})^2}{(b^0_{ab} + b^0_{bc})^2}\right) \quad (6.2.2.1)$$

where a, b and c are the atom types of the atoms forming angle i, K is a constant, Z(a), C(b) and Z(c) are parameters depending solely on the atom types, and θ^0, b^0 are equilibrium parameters.

6.3. Torsional dihedral angle term

6.3.1. Functional forms

If small oscillations around an equilibrium conformation are considered, the torsional potential energy term can, just as the bond-stretching and bond-angle bending terms, be expanded in a Taylor series. In most applications, however, when the relative energies of different conformers and the corresponding isomerization barriers are of interest, or when conformational transitions are studied by molecular dynamics, Taylor series cannot be used. In these cases, the torsional angle potential energy term needs to be 2π-periodic and symmetric at 0 and π, and can be expressed in terms of a cosine series

$$\begin{aligned} &E_\phi(\{\phi_i\};\ \{{}^{(1)}k_{\phi,i}, {}^{(2)}k_{\phi,i}, {}^{(3)}k_{\phi,i}, \ldots\}) \\ &\quad = \sum_{\text{dihedrals } i} [{}^{(1)}k_{\phi,i}(1 - \cos\phi_i) + {}^{(2)}k_{\phi,i}(1 - \cos 2\phi_i) + {}^{(3)}k_{\phi,i}(1 - \cos 3\phi_i) + \ldots] \end{aligned} \quad (6.3.1.1)$$

where ${}^{(n)}k_\phi$ is the 'force constant' corresponding to the term of order n. For example, CFF93 [22,23,41] and MM3 [20,54] use the first three terms in the expansion. The terms (of order n) are sometimes formulated slightly differently [6,64,75], as

$$|{}^{(n)}k_{\phi,i}| - {}^{(n)}k_{\phi,i}\cos n\phi_i \quad (6.3.1.2a)$$

or

$${}^{(n)}k'_{\phi,i}(1 + \cos(n\phi_i - {}^{(n)}\delta_i)) \quad \text{with } {}^{(n)}k'_{\phi,i} > 0 \text{ and } {}^{(n)}\delta_i = 0, \pi \quad (6.3.1.2b)$$

where ${}^{(n)}\delta$ in the second formulation is a phase shift, which plays the same role as the sign of ${}^{(n)}k_\phi$ in the first formulation. Since the slope of the potential has to vanish at 0 and π, the only possible values of ${}^{(n)}\delta_i$ are 0 and π. If ${}^{(n)}k_\phi$ is negative or ${}^{(n)}\delta_i$ is 0, the term has a maximum for $\phi = 0$. If ${}^{(n)}k_\phi$ is positive or ${}^{(n)}\delta_i$ is π, it has a minimum for $\phi = 0$. These latter two formulations ensure that the potential is zero at the minimum of the curve, which may not be true for Eq. 6.3.1.1. Three choices will influence the value of the force constants used and the transferability of torsional parameters from one force field to another.

A. The number of terms retained in the cosine expansion varies from one force field to another and from one dihedral type to another. Typical choices are the first three terms (e.g. CFF93) or a number of terms of selected multiplicity from one to six (e.g. CHARMM, GROMOS).

B. When two bonded atoms each have up to four covalently bound neighbour atoms, one to nine dihedrals can be defined. The summation in Eq. 6.3.1.1 need not include all these dihedral angles, but may comprise only one or a few of them (Sec. 4.3).

C. Depending on the force field, atoms separated by two other atoms (third neighbours) may be excluded from nonbonded interaction (e.g. ECEPP) or may interact with modified (e.g. GROMOS) or scaled (e.g. AMBER) van der Waals interaction parameters.

In Fig. 3, the combination of terms constituting the C-C-C-C dihedral potential energy term in the CFF93 force field [22,23,41] is illustrated. The overall energy (a) is a linear combination of terms of multiplicity one (b), two (c) and three (d). The coefficient of the first term is large and negative, which ensures that the energy is maximal in the eclipsed conformation. Due to the use of terms described in Eq. 6.3.1.1 rather than Eq. 6.3.1.2, the energy is not zero at the minimum of the curve. The two representations are otherwise equivalent. Note that curve (a) is not the energy profile for butane, since other torsional-angle-dependent terms (valence coordinate cross-terms and nonbonded interaction terms) also contribute to the overall interaction energy of the molecule.

6.3.2. Combination rules

Combination rules for torsional interaction parameters are usually given in the form of a table as a function of the atom types of the four atoms, or of the two central atoms, that define the torsional angle.

6.4. Out-of-plane coordinate distortion term

6.4.1. Functional forms

In principle, the valence terms of a force field could be entirely defined in terms of bond lengths, bond angles and torsional dihedrals, as is for instance the case in the alkane CFF93 force field [22,23,41]. There are, however, two reasons for introducing out-of-plane coordinate potential energy terms: (i) All the covalent internal coordinates mentioned till now can be expressed in terms of scalar products of vectors and there is thus no term to enforce chirality (which is just determined by the coordinates and sufficiently high isomerization barriers). Enforcing the geometry around a site by using six bond angles (tetrahedral case) or three bond angles (planar case) without including cross-terms requires an unrealistically stiff energy function. (ii) When

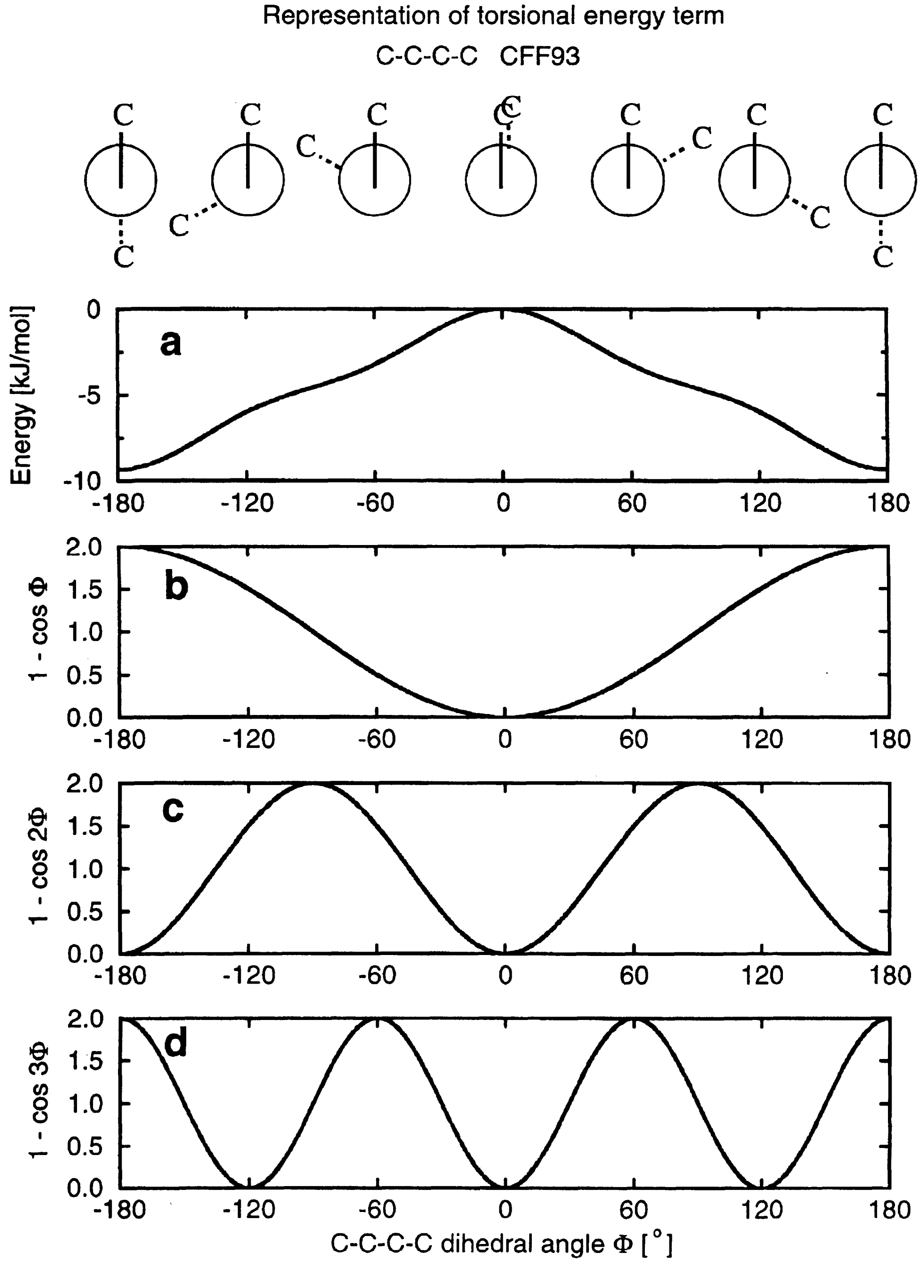


Fig. 3. Representation of the torsional dihedral angle energy term for a C-C-C-C dihedral according to the CFF93 force field, Eq. 6.3.1.1: (a) overall energy = $^{(1)}k_\phi$ (1 − cos ϕ) + $^{(2)}k_\phi$ (1 − cos 2ϕ) + $^{(3)}k_\phi$ (1 − cos 3ϕ); (b, c, d) representation of these three contributions, with $^{(1)}k_\phi$ = − 4.050, $^{(2)}k_\phi$ = 0.042 and $^{(3)}k_\phi$ = − 0.628 kJ/mol.

tetrahedral united-atom carbons, CHR_3, are used, the hydrogen is not explicitly present for the definition of an angle, but pyramidal inversion need be avoided. From a certain point of view, out-of-plane coordinate energy terms are 'unphysical', since they bias the energy in favour of one of the enantiomeric forms of the system, with the purpose of improving the description of this favoured enantiomer.

The out-of-plane coordinate energy term should describe how difficult it is to force a nonplanar geometry (trigonal site) or a nontetrahedral geometry (tetracoordinated site or CHR_3 united atom). The functional form is most often chosen to be harmonic:

$$E_\xi(\{\xi_i\};\ \{\xi_i^0, k_{\xi,i}\}) = \sum_{\text{out-of-plane coordinates } i} k_{\xi,i}(\xi_i^0 - \xi_i)^2 \tag{6.4.1.1}$$

where the summation runs over a specified set of out-of-plane coordinates. The definition of ξ is not unique [21]. Three possible choices are described in Fig. 4 for the case of a trigonal site. It can either be expressed in terms of (a) an improper dihedral angle, (b) an angle between a bond and the plane formed by the others, and (c) a pyramid height. In cases (a) and (b), three choices can be made, depending on the selected reference bond. The definition can be made unique by taking the average value over the three possibilities.

6.4.2. Combination rules

Combination rules for out-of-plane coordinate potential energy parameters are usually given in the form of a table as a function of the atom types of the four atoms, or of the two outer atoms (improper dihedral definition), that define the coordinate.

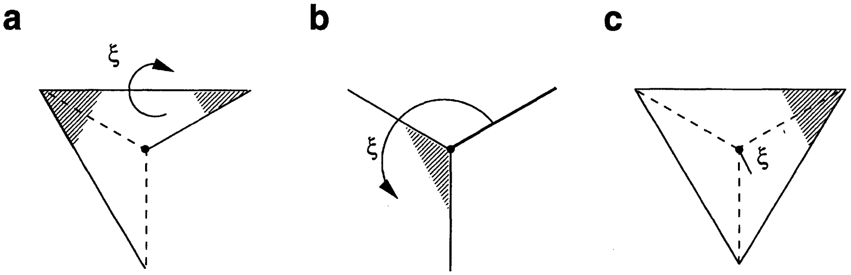


Fig. 4. Different possible definitions for the out-of-plane coordinate ξ Eq. 6.4.1.1, around a trigonal centre: (a) improper dihedral angle, i.e. the dihedral angle defined by one bond from the central atom to a peripheral atom, the vector from this peripheral atom to one of the other peripheral atoms, and the vector from this second peripheral atom to the third peripheral atom; (b) the angle between one bond from the central atom and the plane defined by the central atom and the two peripheral atoms not involved in this bond; (c) the height of the central atom above the plane defined by the three peripheral atoms. In cases (a) and (b), three choices can be made, depending on the selected reference bond.

6.5. Valence coordinate cross-terms

6.5.1. Functional forms

It has been shown that the inclusion of valence coordinate coupling terms (off-diagonal terms) significantly improves the capacity of an empirical function to reproduce trends in the energy and its first and second derivatives with respect to the atomic coordinates from *ab initio* molecular orbital calculations [22,183] and trends from experimental data in vacuum [41], see Sec. 7.5. The MM2 [20,53] and CVFF [47–50] force fields contain some of these terms, whereas the MM3 [20,54] and CFF93 [22] force fields use them in a systematic way. These terms are seldom found in force fields for the simulation of biomolecules since they (i) increase the complexity of the interaction function, (ii) lead, to some extent, to a loss of physical insight, (iii) do not allow for the use of bond constraints, and (iv) are assumed to be not very relevant to these types of problems. The commonly included off-diagonal terms are listed below (see also Table 2) and are described pictorially in Fig. 5. Interpretation of

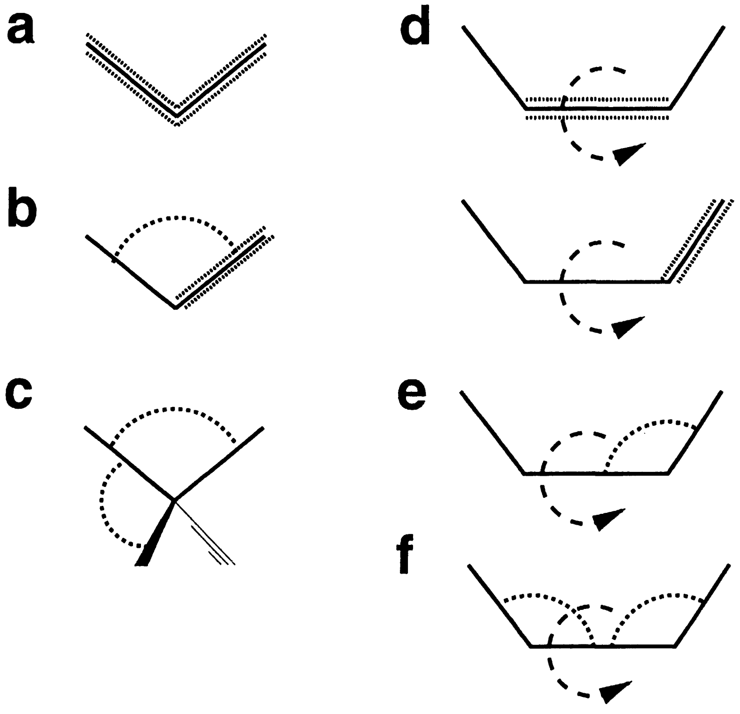


Fig. 5. Valence coordinate cross-terms commonly included in nondiagonal force fields, Eqs. 6.5.1.1–6.5.1.6: (a) bond–bond; (b) bond-angle–bond; (c) bond-angle–bond-angle; (d) torsional-angle–bond; (e) torsional-angle–bond-angle; (f) torsional-angle–bond-angle–bond-angle.

the terms with respect to the force constants is given here with reference to the force constants in the CFF93 force field for alkanes. Note that the inclusion of 1,4 nonbonded interactions (third-neighbour interaction, see Sec. 6.3.1) implicitly includes terms of type C–F.

A. *Bond–bond coupling* (bonds j sharing one common atom with bond i):

$$E_{bb'}(\{b_i,b_j\};\ \{b_i^0,b_j^0,k_{bb',ij}\}) = \sum_{\text{bonds } i}\ \sum_{\text{bonds } j>i}^{(\leq 6)} k_{bb',ij}(b_i^0 - b_i)(b_j^0 - b_j) \tag{6.5.1.1}$$

This term is present in CVFF and CFF93. Since k is positive, this term favours asymmetric bond stretching around a given site.

B. *Bond-angle–bond coupling* (two bonds j involved in the angle i):

$$E_{\theta b}(\{\theta_i,b_j\};\ \{\theta_i^0,b_j^0,k_{\theta b,ij}\}) = \sum_{\text{angles } i}\ \sum_{\text{bonds } j}^{(2)} k_{\theta b,ij}(\theta_i^0 - \theta_i)(b_j^0 - b_j) \tag{6.5.1.2}$$

This term is used in CVFF, CFF93, MM2 and MM3 to reproduce vibrational frequencies and the bond length effects in strained molecules where a bond angle is stretched or compressed. Since k is positive, bond lengthening is favoured when the bond angle is reduced.

C. *Bond-angle–bond-angle coupling* (angles j sharing one common bond with angle i):

$$E_{\theta\theta'}(\{\theta_i,\theta_j\};\ \{\theta_i^0,\theta_j^0,k_{\theta\theta',ij}\}) = \sum_{\text{angles } i}\ \sum_{\text{angles } j}^{(\leq 10)} k_{\theta\theta',ij}(\theta_i^0 - \theta_i)(\theta_j^0 - \theta_j) \tag{6.5.1.3}$$

This term is present in CVFF, CFF93 and MM3. It is used to reproduce vibrational frequencies for coupled bending modes. k may be positive or negative.

D. *Torsional-angle–bond coupling* (central bond or peripheral bonds j involved in torsion i):

$$\begin{aligned} &E_{\phi b}(\{\phi_i,b_j\};\ \{b_j^0,{}^{(1)}k_{\phi b,i},{}^{(2)}k_{\phi b,i},{}^{(3)}k_{\phi b,i},\cdots\}) \\ &= \sum_{\text{dihedrals } i}\ \sum_{\text{bonds } j}^{(1)\text{ or }(2)} (b_j^0 - b_j)\,[{}^{(1)}k_{\phi b,i}\cos\phi_i + {}^{(2)}k_{\phi b,i}\cos 2\phi_i + {}^{(3)}k_{\phi b,i}\cos 3\phi_i + \cdots] \end{aligned} \tag{6.5.1.4}$$

This term is present in CFF93 up to order 3 and in MM3 (torsion–central bond coupling only). It is used for reproducing the structures of molecules in which different conformers exhibit significant differences in bond lengths. Since ${}^{(1)}k$ is positive for the coupling to the central bond, a lengthening of this bond is favoured in the eclipsed conformations. For peripheral bonds, k is negative and small.

E. *Torsional-angle–bond-angle coupling* (two angles j involved in torsion i):

$$\begin{aligned} &E_{\phi\theta}(\{\phi_i,\theta_j\};\ \{\theta_j^0,{}^{(1)}k_{\phi\theta,i},{}^{(2)}k_{\phi\theta,i},{}^{(3)}k_{\phi\theta,i},\cdots\}) \\ &= \sum_{\text{dihedrals } i}\ \sum_{\text{angles } j}^{(2)} (\theta_j^0 - \theta_j)\,[{}^{(1)}k_{\phi\theta,i}\cos\phi_i + {}^{(2)}k_{\phi\theta,i}\cos 2\phi_i + {}^{(3)}k_{\phi\theta,i}\cos 3\phi_i + \cdots] \end{aligned} \tag{6.5.1.5}$$

This term is present in CFF93 up to order 3 and plays a similar role as the term under D.

F. *Torsional-angle–bond-angle–bond-angle coupling* (angles j and k involved in torsion i):

$$E_{\phi\theta\theta'}(\{\phi_i,\theta_j,\theta_k\};\ \{\theta_j^0,\theta_k^0,k_{\phi\theta\theta',i}\}) = \sum_{\text{dihedrals } i} k_{\phi\theta\theta',i}(\theta_j^0 - \theta_j)(\theta_k^0 - \theta_k)\cos\phi_i \qquad (6.5.1.6)$$

This term is present in CVFF and CFF93. Since k is negative, an increase in the bond angles is favoured in the eclipsed conformation.

6.6. Van der Waals interaction

6.6.1. Functional forms

It is usually assumed that the nonelectrostatic component of the interaction between nonbonded atoms can be described in the same way as the interaction between rare gas atoms, i.e. a long-range weak attraction due to induced-dipole– induced-dipole (dispersion) interaction and a short-range steep repulsion due to the overlap of the electron clouds. This type of interaction is given the generic name of van der Waals interaction. Although the features mentioned above are generally accepted, the proper functional description of van der Waals interactions is, however, still a matter of discussion [184–186]. Due to the availability of a large amount of experimental data and the absence of intermolecular forces other than van der Waals interactions, mixtures of rare gases can be used as test systems for proposed functional forms and combination rules. When the description of the many-body interaction is limited to an effective two-body interaction, accurate pairwise interaction functions can be constructed in this way. To a good approximation, these functions obey a single reduced form for all pair types from He to Xe, except at very short internuclear distances [184,187], i.e. for two atoms i and j

$$\eta_{ij} = \eta_{ij}(\varrho_{ij}) \quad \text{with } \varrho_{ij} = \frac{r_{ij}}{R_{\min}(i,j)} \text{ and } \eta_{ij} = \frac{E_{ij}}{\varepsilon(i,j)} \qquad (6.6.1.1)$$

where E_{ij} is the interaction energy, r_{ij} is the interatomic distance, $R_{\min}(i,j)$ and $\varepsilon(i,j)$ are parameters depending on the atom type of i and j, and the reduced energy $\eta_{ij}(\varrho_{ij})$ is a unique function of the reduced distance ϱ_{ij}, valid for any pair types. By convention, $R_{\min}(i,j)$ is the distance at minimum energy and $\varepsilon(i,j)$ is the corresponding energy (well depth) with respect to infinite separation. The function $\eta_{ij}(\varrho_{ij})$ thus has to satisfy

$$\eta_{ij}(\varrho_{ij} = 1) = -1, \quad \left.\frac{d\eta_{ij}}{d\varrho_{ij}}\right|_{\varrho_{ij}=1} = 0,$$
$$\lim_{\varrho_{ij}\to\infty} \eta_{ij}(\varrho_{ij}) = 0 \quad \text{and} \quad \lim_{\varrho_{ij}\to 0} \eta_{ij}(\varrho_{ij}) = +\infty \qquad (6.6.1.2)$$

Additional characteristic parameters of the interaction function are the distance at zero energy $\sigma(i,j)$, with $\sigma(i,j) < R_{min}(i,j)$, and its reduced value $\xi(i,j)$

$$\eta_{ij}(\varrho_{ij} = \xi(i,j)) = 0 \quad \text{and} \quad \sigma(i,j) = \xi(i,j)R_{min}(i,j) \tag{6.6.1.3}$$

and the equivalent harmonic force constant at the minimum $k(i,j)$ and its reduced value $\kappa(i,j)$

$$\left.\frac{d^2\eta_{ij}}{d\varrho_{ij}^2}\right|_{\varrho_{ij}=1} = \kappa(i,j) \quad \text{and} \quad k(i,j) = \kappa(i,j)/R_{min}^2(i,j) \tag{6.6.1.4}$$

Finally, the van der Waals interaction energy of the whole molecular system is given by

$$E_{vdw}(\{r_{ij}\};\ \{R_{min}(i,j),\ \varepsilon(i,j)\}) = \sum_{i}^{N_{atoms}} \sum_{j>i}^{N_{atoms}} \varepsilon(i,j)\,\eta_{ij}\left(\frac{r_{ij}}{R_{min}(i,j)}\right) \tag{6.6.1.5}$$

The following reduced functions have been proposed to describe van der Waals interactions in empirical classical force fields:

A. *n-m van der Waals function*:

$$\eta_{ij} = \frac{1}{n-m}\left[m\varrho_{ij}^{-n} - n\varrho_{ij}^{-m}\right] \tag{6.6.1.6a}$$

with

$$\xi(i,j) = (m/n)^{1/(n-m)} \quad \text{and} \quad \kappa(i,j) = nm \tag{6.6.1.6b}$$

Most current force fields use a 12-6 van der Waals function ($n = 12$, $m = 6$, also called Lennard-Jones function), where the steep repulsion is described by a $1/r_{ij}^{12}$ dependence and the dispersion by a $1/r_{ij}^6$ dependence. Three equivalent definitions can be found in the literature: either

$$E_{12\text{-}6}(\{r_{ij}\};\ \{C_{12}(i,j), C_6(i,j)\}) = \sum_{i}^{N_{atoms}} \sum_{j>i}^{N_{atoms}} C_{12}(i,j)r_{ij}^{-12} - C_6(i,j)r_{ij}^{-6} \tag{6.6.1.7a}$$

or

$$E_{12\text{-}6}(\{r_{ij}\};\ \{R_{min}(i,j),\ \varepsilon(i,j)\}) = \sum_{i}^{N_{atoms}} \sum_{j>i}^{N_{atoms}} \varepsilon(i,j)\left[\left(\frac{r_{ij}}{R_{min}(i,j)}\right)^{-12} - 2\left(\frac{r_{ij}}{R_{min}(i,j)}\right)^{-6}\right] \tag{6.6.1.7b}$$

or

$$E_{12\text{-}6}(\{r_{ij}\};\ \{\sigma(i,j),\ \varepsilon(i,j)\}) = \sum_{i}^{N_{atoms}} \sum_{j>i}^{N_{atoms}} 4\varepsilon(i,j)\left[\left(\frac{r_{ij}}{\sigma(i,j)}\right)^{-12} - \left(\frac{r_{ij}}{\sigma(i,j)}\right)^{-6}\right] \tag{6.6.1.7c}$$

The conversion between these definitions is straightforward:

$$R_{min}(i,j) = \left(\frac{2C_{12}(i,j)}{C_6(i,j)}\right)^{1/6}, \quad \sigma(i,j) = \left(\frac{C_{12}(i,j)}{C_6(i,j)}\right)^{1/6} \text{ and } \varepsilon(i,j) = \frac{C_6^2(i,j)}{4C_{12}(i,j)} \tag{6.6.1.8}$$

The 12-6 function has the advantage of being simple (few parameters, i.e. two per i, j pair) and computationally efficient (even power of r_{ij}). It has been suggested that a softer van der Waals interaction might perform better than a 12-6 form. A 9-6 van der Waals interaction (n = 9, m = 6) is used in CVFF [47–50] or CFF93 [22]:

$$E_{9\text{-}6}(\{r_{ij}\}; \{R_{\min}(i,j), \varepsilon(i,j)\}) = \sum_{i}^{N_{\text{atoms}}} \sum_{j>i}^{N_{\text{atoms}}} \varepsilon(i,j) \left[2\left(\frac{r_{ij}}{R_{\min}(i,j)}\right)^{-9} - 3\left(\frac{r_{ij}}{R_{\min}(i,j)}\right)^{-6} \right] \qquad (6.6.1.9)$$

B. *exp-m function*:

$$\eta_{ij} = \frac{1}{\zeta(i,j) - m} [m e^{\zeta(i,j)[1 - \varrho_{ij}]} - \zeta(i,j)\varrho_{ij}^{-m}] \qquad (6.6.1.10a)$$

with

$$\kappa(i,j) = \frac{m\zeta(i,j)[\zeta(i,j) - m - 1]}{\zeta(i,j) - m} \qquad (6.6.1.10b)$$

where $\zeta(i,j)$ is a dimensionless scaling parameter. The exp-6 function (m = 6) is the most used [69]:

$$E_{\text{exp-6}}(\{r_{ij}\}; \{\varepsilon(i,j), R_{\min}(i,j), \zeta(i,j)\}) = \sum_{i}^{N_{\text{atoms}}} \sum_{j>i}^{N_{\text{atoms}}} \frac{\varepsilon(i,j)}{\zeta(i,j) - 6} \left[6e^{[\zeta(i,j)(1 - r_{ij}/R_{\min}(i,j))]} - \zeta(i,j)\left(\frac{r_{ij}}{R_{\min}(i,j)}\right)^{-6} \right] \qquad (6.6.1.11)$$

When $\xi(i,j) = 13.77$ the function has the same curvature at the minimum as a Lennard-Jones function. Since this parameter can be selected for each individual pair, the function offers more flexibility. It is nevertheless less used than the Lennard-Jones function since it is computationally more expensive, and involves the calibration of three parameters per pair instead of two. Although an exponential repulsion may perform better at short distances, the exp-m function does not satisfy the last limit of Eq. 6.6.1.2 and tends towards $-\propto$ for very short distances.

C. *Double Morse function*:

$$\eta_{ij} = e^{-2\alpha(i,j)[1 - \varrho_{ij}]} - 2e^{\alpha(i,j)[1 - \varrho_{ij}]} \qquad (6.6.1.12a)$$

with

$$\xi(i,j) = 1 - \frac{\ln 2}{\alpha(i,j)} \quad \text{and} \quad \kappa(i,j) = 2\alpha^2(i,j) \qquad (6.6.1.12b)$$

Although all types of functions defined above perform similarly at distances close to the equilibrium distance, it has been suggested using both theoretical arguments and comparison to *ab initio* results [185,186] that a Morse-type function may perform better over a wider range of distances:

$$E_{\mathrm{Morse}}(\{r_{ij}\};\{\varepsilon(i,j),R_{\min}(i,j),\alpha(i,j)\})$$

$$= \sum_{i}^{N_{\mathrm{atoms}}} \sum_{j>i}^{N_{\mathrm{atoms}}} \varepsilon(i,j)\left[e^{[2\alpha(i,j)(1-r_{ij}/R_{\min}(i,j))]} - 2e^{[\alpha(i,j)(1-r_{ij}/R_{\min}(i,j))]}\right] \tag{6.6.1.13}$$

D. *n-m buffered function*:

$$\eta_{ij} = \frac{n-m}{m}\left[\frac{1+\delta}{\varrho_{ij}+\delta}\right]^{n-m}\left[\frac{1+\gamma}{\varrho_{ij}^{m}+\gamma} - \frac{m}{n-m} - 1\right] \tag{6.6.1.14}$$

A buffered 14-7 energy function has been proposed [184]:

$$E_{\mathrm{buf\text{-}n\text{-}m}}(\{r_{ij}\};\ \delta,\gamma,\{\varepsilon(i,j),R_{\min}(i,j)\})$$

$$= \sum_{i}^{N_{\mathrm{atoms}}} \sum_{j>i}^{N_{\mathrm{atoms}}} \varepsilon(i,j)\left[\frac{(1+\delta)R_{\min}(i,j)}{r_{ij}+\delta R_{\min}(i,j)}\right]^{(n-m)}\left[\frac{(1+\gamma)R_{\min}^{m}(i,j)}{r_{ij}^{m}+\gamma R_{\min}^{m}(i,j)} - 2\right] \tag{6.6.1.15}$$

where $n = 14$, $m = 7$, $\delta = 0.07$ and $\gamma = 0.12$, these parameters being obtained from a best fit to rare gas experimental data. Note that with these values of δ and γ, the minimum of $\eta_{ij}(\varrho_{ij})$ is at $(0.996;\ -1.0006)$, and thus Eq. 6.6.1.14 nearly satisfies the conditions in Eq. 6.6.1.2. The (reduced) curvature at the minimum, $\kappa(i,j)$, is in this case 79.6, and the reduced intercept, $\xi(i,j)$, is 0.89, both close to the Lennard-Jones value.

In Fig. 6 the various reduced energy functions mentioned above are displayed, with parameters corresponding to a curvature at the minimum of 72 (reduced units, the curvature of the Lennard-Jones function), except the 9-6 van der Waals function (curvature 54) and the 14-7 buffered function (curvature 79.6). As can be seen, the range of these interactions is short and they will play an essential role only for direct neighbour atoms. Due to the intrinsically small magnitude of the energies involved ($\varepsilon \approx -0.1$ kJ/mol for He-He to -2.3 kJ/mol for Xe-Xe), the divergences between the different functions above $\varrho_{ij} = 1$ are likely to affect the overall energy in a minor way in condensed-phase systems. This may of course not be true for gaseous systems. On the other hand, the steepness of the function below $\varrho_{ij} = 1$ will influence the density and compressibility of a condensed-phase system. When electrostatic effects are present, the balance between this steep repulsion and the electrostatic interaction will be the determinant part of packing forces. Since van der Waals parameters (ε, $R_{\min}$, curvature determining parameters), just as atomic charges, are effective parameters, they can be adjusted so that virtually any one of the above functional forms can give reasonable results. Of course, after such an adjustment, condensed-phase effective van der Waals parameters may not be suitable anymore to give a proper representation of the gas-phase state. Since the small energetical contributions for nearest neighbours will sum up for a large number of pairs, a proper choice for the ε and $R_{\min}$ parameters seems primordial, and combination rules (Sec. 6.6.2) should be considered carefully.

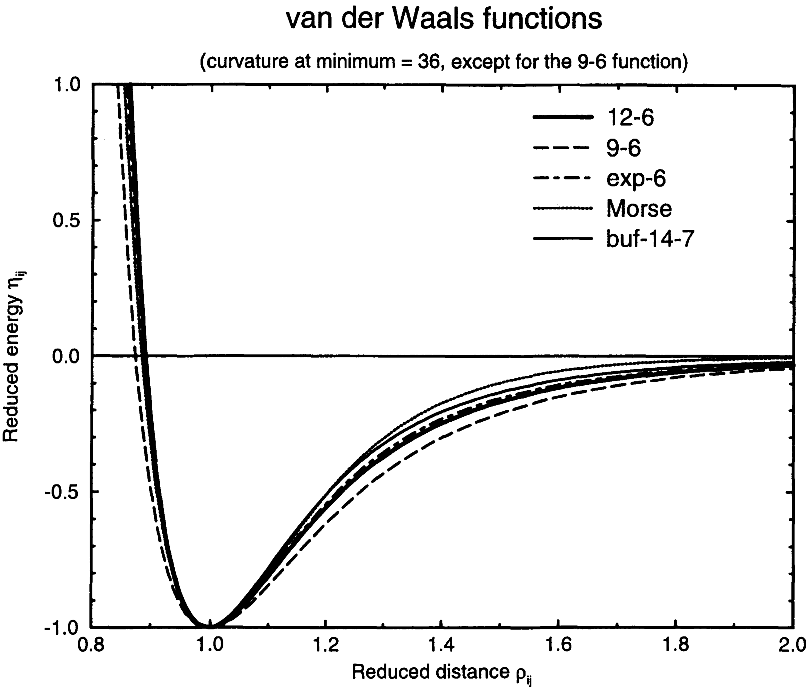


Fig. 6. Representation of the van der Waals interactions, in a reduced form, corresponding to Eq. 6.6.1.6 with $m = 6$ and $n = 12$ or $m = 6$ and $n = 9$, Eq. 6.6.1.10 with $m = 6$ and $\zeta(i, j) = 13.77$, Eq. 6.6.1.12 with $\alpha(i, j) = 6$, and Eq. 6.6.1.14 with $n = 14$, $m = 7$, $\delta = 0.07$ and $\gamma = 0.12$.

6.6.2. Combination rules

Because the definition of N atom types implies the definition of $\frac{1}{2}N(N + 1)$ van der Waals interaction parameter sets for atom pairs, most force fields use combination rules which depend on sets of N atomic parameters and can be calibrated by studying the homonuclear case [184,188]. Since the experimental energy functions for rare gases follow a single reduced form around the minimum [184,187] R_{min} or σ combination rules are interchangeable to a large extent, although often formally not equivalent.

A. *Geometric means for ε and R_{min}:*

$$R_{min}(i,j) = \sqrt{R_{min}(i,i)R_{min}(j,j)} \quad \text{and} \quad \varepsilon(i,j) = \sqrt{\varepsilon(i,i)\varepsilon(j,j)} \qquad (6.6.2.1)$$

The following two rules are equivalent to Eq. 6.6.2.1 for the case of a Lennard-Jones function and any n-m van der Waals interaction, respectively:

$$C_6(i,j) = \sqrt{C_6(i,i)C_6(j,j)} \quad \text{and} \quad C_{12}(i,j) = \sqrt{C_{12}(i,i)C_{12}(j,j)} \qquad (6.6.2.2a)$$

or

$$\sigma(i,j) = \sqrt{\sigma(i,i)\sigma(j,j)} \quad \text{and} \quad \varepsilon(i,j) = \sqrt{\varepsilon(i,i)\,\varepsilon(j,j)} \tag{6.6.2.2b}$$

B. *Geometric mean for ε and arithmetic mean for R_{min} (Lorentz–Berthelot mixing rules):* The following rules are equivalent for the case of any van der Waals interaction:

$$R_{min}(i,j) = \tfrac{1}{2}[R_{min}(i,i) + R_{min}(j,j)] \quad \text{and} \quad \varepsilon(i,j) = \sqrt{\varepsilon(i,i)\varepsilon(j,j)} \tag{6.6.2.3a}$$

or

$$\sigma(i,j) = \tfrac{1}{2}[\sigma(i,i) + \sigma(j,j)] \quad \text{and} \quad \varepsilon(i,j) = \sqrt{\varepsilon(i,i)\varepsilon(j,j)} \tag{6.6.2.3b}$$

C. *Arithmetic mean for R^6_{min} and geometric mean for εR^6_{min}:* This combination rule has been proposed recently and has been tested for rare gases [188]:

$$\begin{aligned} R_{min}(i,j) &= \left[\frac{R^6_{min}(i,i) + R^6_{min}(j,j)}{2}\right]^{1/6} \\ \varepsilon(i,j) &= \frac{1}{R^6_{min}(i,j)}\sqrt{\varepsilon(i,i)R^6_{min}(i,i)\varepsilon(j,j)R^6_{min}(j,j)} \end{aligned} \tag{6.6.2.4}$$

D. *Cubic-mean rule for R_{min} and HHG mean for ε:* This combination rule has been proposed recently and tested for rare gases [184], where the HHG mean is the harmonic mean of harmonic and geometric means:

$$\begin{aligned} R_{min}(i,j) &= \frac{R^3_{min}(i,i) + R^3_{min}(j,j)}{R^2_{min}(i,i) + R^2_{min}(j,j)} \\ \varepsilon(i,j) &= \frac{4\varepsilon(i,i)\varepsilon(j,j)}{[\varepsilon(i,i)^{1/2} + \varepsilon(j,j)^{1/2}]^2} \end{aligned} \tag{6.6.2.5}$$

E. *Slater–Kirkwood combination:* The Slater–Kirkwood expression [64,181,184] is more than a combination rule, since it also allows the estimation of van der Waals parameters from experiment:

$$\begin{aligned} C_6(i,j) &= K\frac{\alpha(i)\alpha(j)}{(\alpha(i)/N(i))^{1/2} + (\alpha(j)/N(j))^{1/2}} \\ C_{12}(i,j) &= \tfrac{1}{2}C_6(i,j)[R(i) + R(j)]^6 \end{aligned} \tag{6.6.2.6}$$

where K is a constant, α(i) is the polarizability of atom i, R(i) is its van der Waals radius and N(i) its effective number of outer shell electrons. The second rule is very similar to an arithmetic mean in R_{min}.

Other combinations have been proposed, which involve additional parameters such as polarizbility, ionization potentials or dispersion force coefficients. These are, however, not well suited for general empirical force fields, since one would like to restrict the number of parameters involved. In Fig. 7, the result of the application of

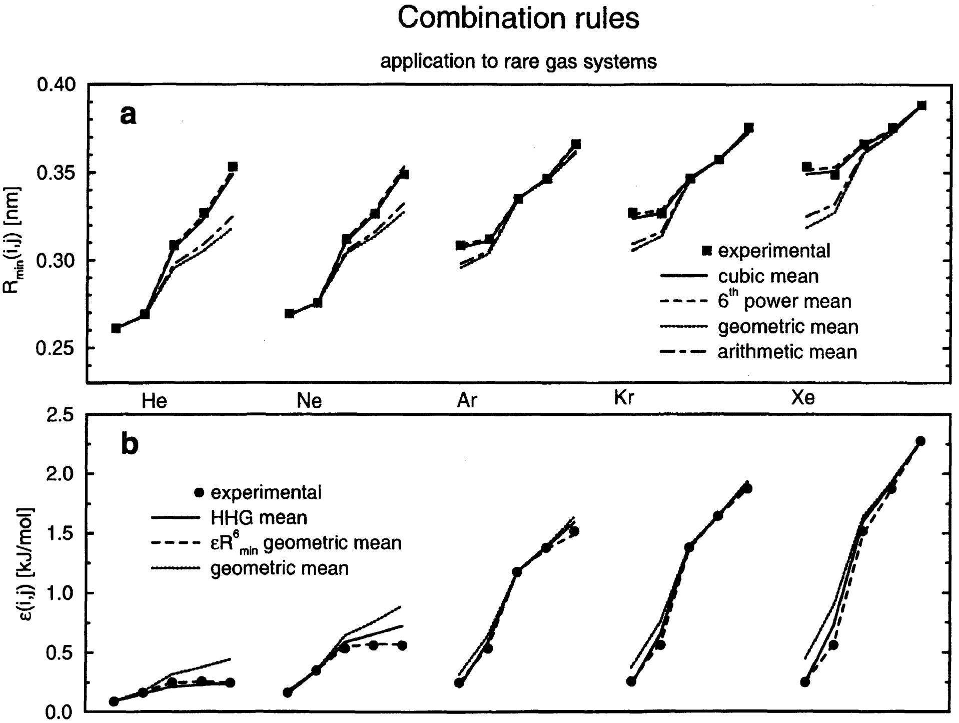


Fig. 7. R_{min} and ε combination rules applied to mixed rare gas systems. Experimental data are from Ref. 187. For each indicated gas, mixed combinations in the sequence He-Ne-Ar-Kr-Xe are indicated.

the combination rules mentioned above is reported for mixed rare gas systems and compared with experimental values [187]. It can be seen that both the geometric and the arithmetic mean rules underestimate $R_{min}(i,j)$ for systems composed of very unlike atom types (by 8–10% for the He-Xe system), the latter performing slightly better. Both other rules perform well in all cases. The best performing rule for ε(i, j) is clearly the εR^6_{min} geometric mean rule. Both the geometric mean rule and the HHG mean rule tend to overestimate the values, the latter performing slightly better. Figures 8 and 9 illustrate the importance of using a combination rule consistently within a given force field, by combining the GROMOS Br atom type with other atom types of the force field. The original GROMOS combination rule is a geometric rule in both parameters. Taking the example of interaction of Br with H, the distance at the minimum interaction energy may vary by about 10% by using different combination rules. Differences in ε vary over a range of about 0.5 kJ/mol. Note also the oscillatory behaviour of the cubic-mean rule for R_{min} when the size of the second atom type is decreased to zero.

6.7. Electrostatic interaction

In principle, a correct representation of the behaviour of the electron cloud (implicit degrees of freedom) at each atomic site would require a full multipole expansion. Very often, the expansion is truncated after the first term, i.e. a monopole approximation is used. Although models including polarizability are continuously being developed, the present discussion will be limited to the (still dominantly used) monopole interaction.

6.7.1. Functional forms

The correct treatment of electrostatic interactions is an essential but difficult problem in the design of empirical energy functions [6,189–192]. This is mainly due to their long-range nature, which causes dependence on the system size and boundary conditions, as well as high computational costs. In condensed-phase simulations, these high computational costs, together with the use of periodic boundary conditions, require approximations, which will unfortunately influence the properties of the simulated system. In most cases, the interaction is defined in terms of a pairwise

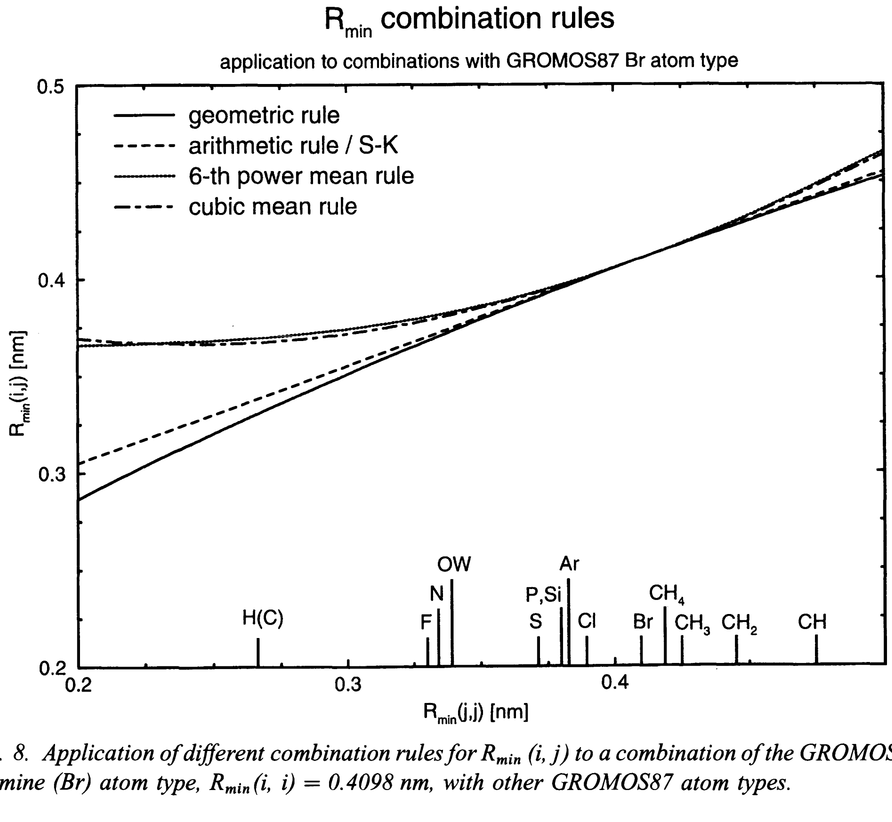


Fig. 8. Application of different combination rules for R_{min} (i, j) to a combination of the GROMOS87 bromine (Br) atom type, R_{min}(i, i) = 0.4098 nm, with other GROMOS87 atom types.

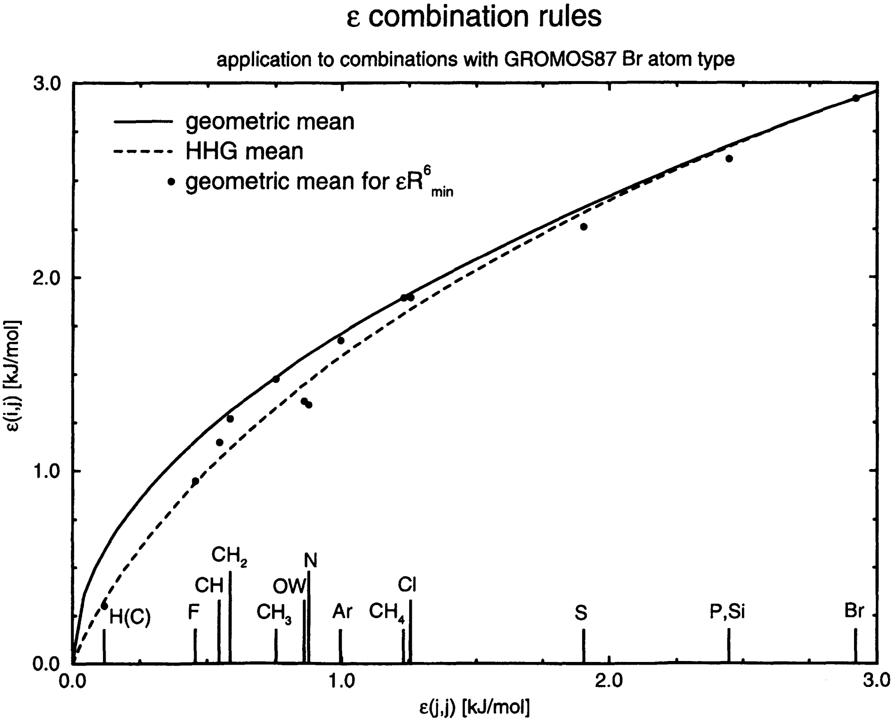


Fig. 9. Application of different combination rules for ε(i, j) to a combination of the GROMOS87 bromine (Br) atom type, ε (i, i) = 2.921 kJ/mol, with other GROMOS87 atom types.

Coulomb interaction between point (atomic or virtual site) partial charges. The effect of the polarizability of the electron cloud is assumed to be included in the interaction between these point charges in an average manner, and these charges are thus effective charges. Ideally, the interaction should be calculated by scanning all charge pairs, i.e.

$$E_{Cb}(\{r_{ij}\}; \{q_iq_j\}) = \sum_{i}^{N} \sum_{j>i}^{N} \frac{1}{4\pi\varepsilon_0\varepsilon_1} \frac{q_iq_j}{r_{ij}} \qquad (6.7.1.1)$$

where r_{ij} is the distance between charges i and j, q_iq_j is the product of the charges, ε_0 is the permittivity of vacuum, ε_1 is the relative permittivity of the medium and N is the number of atoms in the system. Equation 6.7.1.1 is, in principle, exact, but practically directly applicable only to vacuum simulations of small isolated molecules, with the aim of reproducing vacuum properties. It cannot be used in the following cases:

A. Medium- and large-scale problems: since the computational expenses grow as N^2.

B. Fixed boundary problems: if the system consists of a molecule, plus possibly some layers of solvent, surrounded by vacuum, surface tension effects will distort its properties. In the absence of dielectric screening from outside the system, the electrostatic interaction inside the system will be overestimated, and in the absence of van der Waals forces with the outside, the surface of the boundary will tend to become minimal (spherical shape). When explicit solvent molecules are present, evaporation may also occur.

C. Periodic boundary problems: if the system consists of an infinite series of replicas of a central cell (periodic boundary conditions, for crystal or solution simulations), the number of pairs in Eq. 6.7.1.1 is infinite.

A wealth of approximate treatments attempt to remedy these problems, and try to find the best compromise between efficiency and accuracy. The following list is nonexhaustive:

1. *Boundary corrections (point B):* The distortions induced at the interface to vacuum can be reduced by corrections which attempt to mimic the effect of solvent outside the boundary [193–196]: (i) short-range contacts, by the addition of a soft-wall interaction or position restraining of the atoms in the surface layer; (ii) electrostatic effects at the boundary, by the addition of dipole orientation interactions; and (iii) dynamical fluctuations, by the use of stochastic boundary condition. These boundary corrections are difficult to calibrate and often have to be reparametrized for each specific system considered.

2. *Redistribution and reduction of the charges (point B):* The distortive effect of the absence of the dielectric screening by the solvent outside the boundary can be counteracted by reducing net charges of groups of atoms to zero by a redistribution of the atomic charges [75]. This method is very *ad hoc*.

3. *Distance-dependent dielectric (point B):* The dielectric screening effect can also be mimicked by replacing ε_1 in Eq. 6.7.1.1 by an effective dielectric constant ε_{eff}, proportional to the distance between charges, i.e. $\varepsilon_{eff} = dr_{ij}$, usually with $d = 1$, 4 or 8 Å^{-1}. In this approximation, the screening effect is assumed to be proportional to the amount of bulk solvent between the charges, and thus to the distance. This method is also *ad hoc* and lacks physical meaning.

4. *Screening functions (point B):* The approach is similar to the previous one, but ε_1 is replaced by $\varepsilon_{eff}\exp(\kappa r_{ij})$, where κ is the inverse Debye screening length. The choice of an adequate ε_{eff} (constant or function of r_{ij}) is problematic and the application to heterogeneous systems is not satisfactory.

5. *Continuum methods (point B):* The system is assumed to be surrounded by a dielectric continuum of permittivity ε_2 [196]. The influence of the charge distribution in the system on the continuum outside the boundary induces a reaction field potential inside the boundary. When the shape of the boundary is highly symmetric, the interaction can be computed analytically (Born, Onsager models). In other cases, it has to be computed numerically (series expansion of the reaction field, finite difference, finite elements or boundary elements methods). The treatment of particles near the boundary is the major problem of these methods.

6. *Langevin dipoles (point B):* The solvent is modelled by a set of polarizable and rotatable dipoles on a grid, of which the average orientation is described by a Langevin-type equation [197]. The model is relatively inexpensive and seems more

realistic than a continuum approximation. It is, however, difficult to parametrize, a proper description of the interface is problematic, and the properties of the system may depend on the grid parameters.

7. *Lattice sums (point C):* These methods are based on an exact periodic treatment of the infinite system in simulations using periodic boundary conditions [57,192, 198–200]. The infinite sum over all atoms and periodic images in Eq. 6.7.1.1 can be rewritten as two finite sums over lattice (real space) and reciprocal lattice (Fourier space) vectors, plus a constant self-energy term, which can, in principle, be computed exactly. These methods are, however, complicated to implement, sometimes computationally expensive and they may enforce long-range correlations through periodicity. These are realistic in simulations of crystals, but may give rise to artefacts in bulk-phase systems, although only under special circumstances [201,202]. Lattice sum techniques include Ewald summation, particle–particle particle mesh and related methods.

8. *Minimum image convention (point C):* The interaction is only calculated between charge i of the central cell and the closest periodic image of charge j. The number of pairs is then finite, but can become large, i.e. $O(N^2)$. This convention is not used much, since all charges interacting with i belong to a volume of the same shape as the unit cell, which induces anisotropy effects.

9. *Simple spherical cutoff (points A,B,C):* The long-range correlation problems inherent to lattice sum methods (7) and the anisotropy problem inherent to the nearest image convention (8) can be reduced if the Coulomb interaction is set to zero beyond a given distance between charges, the cutoff distance R_c. The sphere of radius R_c (cutoff sphere) around a charge i has to be smaller than the unit cell, so that only nearest images are selected inside the cutoff. This method is simple to implement and allows for a significant reduction of the computational costs for large systems, since the effort is roughly $O(NR_c^3)$. Although it is a good approximation for nonpolar systems, it may, however, produce serious problems for polar systems [203,204], ionic systems [205–207] or biomolecules in solution [208–210], since the long-range Coulomb force often differs significantly from zero at the cutoff distance. This effect is illustrated in Fig. 10 by considering the radial dipole orientation correlation of water molecules around a sodium ion for different values of the cutoff radius. The main problems [57] are nonconservation of the energy in a microcanonical simulation, heating effects at the cutoff due to a nonzero force, and structural, statistical and dielectric distortions over the whole range of intermolecular distances. The following points 9a–9d describe possible corrections to the simple spherical cutoff approximation, which attempt to minimize these distortions [205,211,212].

9a. *Charge-group interaction:* Charges are grouped in terms of chemically (or intuitively) based charge groups either neutral (e.g. carbonyl groups) or bearing an integer charge (e.g. carboxylate or ammonium groups). The atom-based truncation is then replaced by a charge-group, based cutoff criterion [6,75]. For two neutral charge groups I and J, the leading term in the electrostatic interaction takes an r_{IJ}^{-3} dependence, which significantly reduces the effects of truncation. They are, however, not

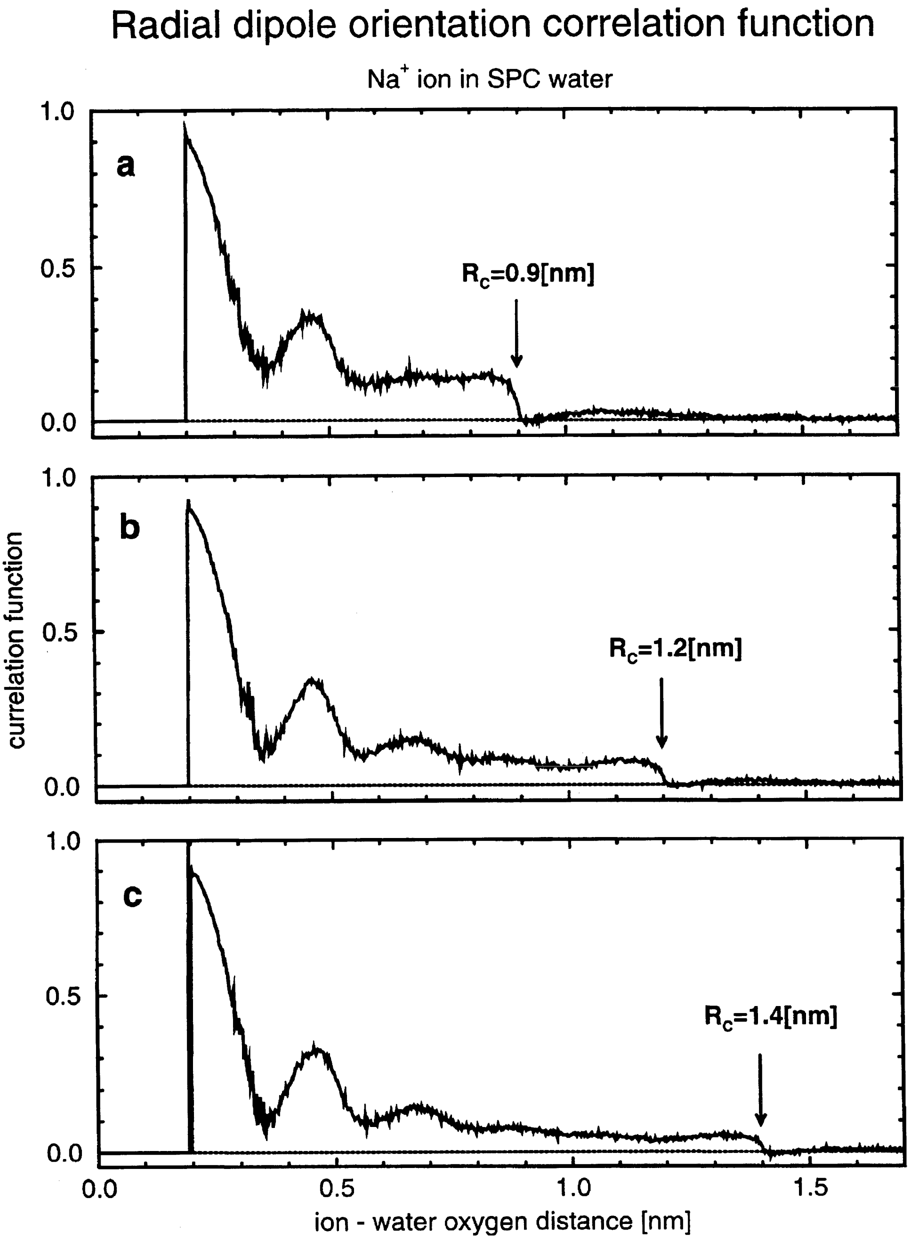


Fig. 10. Radial dipole orientation correlation of SPC water molecules around a (GROMOS87) sodium ion. The ion is at the origin, and the function is calculated as the average radial component of a unit vector along the dipole of a water molecule, for successive shells of increasing radii around the ion. Three different cutoff radii were used in the 0.7 ns simulations for the truncation of the electrostatic interaction: $R_c = 0.9$, 1.2 and 1.4 nm.

completely eliminated [203,204,206,207]. The inconvenience of the method is that it may require a modification of the original charge distribution.

9b. *Twin-range method:* In this method, a second (long-range) cutoff R_L is introduced. Interactions between charge pairs with $R_C < r_{ij} < R_L$ are calculated every n timesteps ($n > 1$, usually $\sim$ 5–10, together with the pair list update) and assumed to be constant in between [6]. If the high-frequency fluctuations in the long-range forces are negligible, the effective cutoff is increased to R_L without significant additional computational costs. In a variant, the interactions between charge groups at distances between R_C and R_L are approximated by a multipole expansion, e.g. up to quadrupole interactions [64].

9c. *Switching/shifting function:* To avoid abrupt truncation of the interaction at the cutoff radius R_C, the Coulomb interaction can be multiplied by a so-called switching function, $S(r_{ij}, R_S, R_C)$ with $R_S < R_C$, a continuous function with continuous derivative, which has the value 1 if $r_{ij} < R_S$ and 0 if $r_{ij} > R_C$ [64,211]. Energy conservation is improved, heating effects are reduced, but structural artefacts are still observed. In the special case where $R_S = 0$, the function $S(r_{ij}, R_C)$ is called a shifting function [64,213]. The inconvenience here is that the interaction is changed over the whole range of r_{ij} distances from 0 to R_C.

9d. *Reaction field correction:* The medium outside the spherical cutoff cavity may be approximated by a dielectric continuum of relative dielectric permittivity equal to that of the bulk solvent, ε_2 [166,167,214–218]. The influence of the charge distribution inside the cutoff on the continuum outside induces a reaction field potential inside the cutoff sphere. This additional interaction can be described as a correction to the Coulomb interaction term to give

$$E_{Cb+RF}(\{r_{ij}\}; \{q_iq_j\}, R_{RF}, \varepsilon_2) = -\sum_{i}^{N_{atoms}} \frac{q_i^2}{4\pi\varepsilon_0\varepsilon_1} \frac{\varepsilon_2-\varepsilon_1}{\varepsilon_2} \frac{1}{2R_{RF}}$$

$$+ \sum_{i}^{N_{atoms}} \sum_{j>i}^{N_{atoms}} \frac{q_iq_j}{4\pi\varepsilon_0\varepsilon_1} \left[\frac{1}{r_{ij}} + \frac{\varepsilon_2-\varepsilon_1}{2\varepsilon_2+\varepsilon_1} \frac{r_{ij}^2}{R_{RF}^3} - \frac{3\varepsilon_2}{2\varepsilon_2+\varepsilon_1} \frac{1}{R_{RF}} \right] \qquad (6.7.1.2)$$

where R_{RF} is in principle equal to R_C. The first summation in Eq. 6.7.1.2 corresponds to a Born term, zero when the system is neutral, constant if the total charge of the system is constant. All other interactions are defined pairwise, the last term in the double sum corresponding to the conducting boundary condition (zero potential) at the cutoff. When $\varepsilon_2 \gg \varepsilon_1$, this correction can be considered as a physically based shifting function. It makes a considerable difference whether this additional interaction is included during the simulation or as a correction afterwards [219]. This treatment is a significant improvement over a straight truncation, but not entirely correct when applied to heterogeneous systems. Its use might also require a force-field reparametrization [220].

6.7.2. Combination rules

Formally, the Coulomb law has the form of a combination rule. In the bond increment method [22,221], the charge of an atom i itself is calculated by the rule

$$q_i = \sum_{\text{1st neighbours } j}^{(4)} \delta(a(i), b(j)) \qquad (6.7.2.1)$$

where a(i) and b(j) are the atom types of i and j, respectively, and the function δ satisfies $\delta(a,b) = -\delta(b,a)$. A single bond parameter is required to evaluate all charges, and electroneutrality is always preserved.

6.8. Coupling between covalent coordinates and electrostatic interactions

Conformation-dependent charges may be used to account for the variations of electron density (shielding) at atomic sites in different conformations. Such conformation-dependent charges are usually derived from molecular orbital calculations [222].

6.9. Hydrogen-bonding term

An explicit hydrogen-bonding interaction term is sometimes added to the already present nonbonded interactions described above. Its purpose is to avoid too short hydrogen bonds due to a strong electrostatic attraction, and to allow for a specific fine tuning of hydrogen-bond distances and energies. In some force fields, the van der Waals 12-6 parameters for hydrogen-bonded atoms are reduced at the same time. For example, in CHARMM [64], the hydrogen-bond potential energy is described by a sum of four-body terms:

$$E_{hb}(\{r_{AD}, \sphericalangle(A \cdots H\text{-}D), \sphericalangle(AA\text{-}A \cdots H)\}; \{C_\gamma, C_\delta, \gamma, \delta, m, n\})$$

$$= \sum_{\substack{\text{H-bonds} \\ AA\text{-}A \cdots H\text{-}D}} \left(\frac{C_\gamma}{r_{AD}^\gamma} - \frac{C_\delta}{r_{AD}^\delta}\right) \cos^m \sphericalangle(A \cdots H\text{-}D) \cos^n \sphericalangle(AA\text{-}A \cdots H) \qquad (6.9.1)$$

where AA, A, H and D are the acceptor-antecedent, the acceptor, the hydrogen and the donor heavy atom, m depends on the type of D (m = 0, 2 or 4) and n on the type of A (n = 0 or 2). The $\cos^m$ function is zeroed if its argument is less than 90° and the $\cos^n$ function if its argument is less than 90° and $n > 0$. Normally a 12-10 function is used for the radial dependence, i.e. $\gamma = 12$ and $\delta = 10$. In other force fields (e.g. Ref. 61), only the radial dependence is retained and a two-body 12-10 function is used (i.e. $m = n = 0$ in Eq. 6.9.1). The presence of such a specific hydrogen-bonding interaction term requires some additional bookkeeping. If the structure is rigid enough, a permanent list of hydrogen-bonded groups can be defined. This list can also be automatically updated at regular intervals. If one assumes that the radial 12-10 correction can equally well be modelled by a 12-6 correction, it can be incorporated into the normal van der Waals interaction terms, as is done in GROMOS [6,75]. This requires the use of a special combination rule for 12-6 van der Waals parameters, namely

$$C_6(i,j) = \sqrt{c_6(i,t_{ij})c_6(j,t_{ij})} \quad \text{and} \quad C_{12}(i,j) = \sqrt{c_{12}(i,t_{ij})c_{12}(j,t_{ij})} \tag{6.9.2}$$

where t_{ij} determines if the interaction between i and j is polar or not. This method offers the advantage that no special bookkeeping is required for the hydrogen bonds. The inconvenience is that even if the orientation of the D-H group is not optimal for making a hydrogen bond to A, the special van der Waals parameters will be used for the A/D interaction. When no explicit angular dependence is included, the implicit dependence of the electrostatic and repulsive van der Waals nonbonded interaction upon the hydrogen-bond angle is assumed to play a similar role.

7. Force-field parametrization procedures

7.1. The basic problem

Once the degrees of freedom of the model, collectively indicated by **D**, have been selected (Sec. 2) and the functional form, **F**, combination rules, **C**, and various approximations (especially the ones dealing with nonbonded interactions), **A**, entering in the definition of the interaction function have been defined (Sec. 6), the task remains of finding the proper values of the interaction function parameters, $\{s_i, i = 1, \ldots, N_{\text{param}}\}$ [37]. These values should be adjusted to formally satisfy

$$X_\alpha^{\text{sim}}(\{s_i, i = 1, \ldots, N_{\text{param}}\}; \mathbf{D}, \mathbf{F}, \mathbf{C}, \mathbf{A}) = X_\alpha^{\text{target}}, \quad \alpha = 1, \ldots, N_{\text{obs}} \tag{7.1.1}$$

where X_α^{sim} is a simulated observable, generally depending simultaneously on all force-field parameters and on the choices mentioned above ($\mathbf{D}, \mathbf{F}, \mathbf{C}, \mathbf{A}$), and X_α^{targer} is its target value (experimental, calculated by a sufficiently accurate molecular orbital technique, or a combination of both). In most cases, the function X_α^{sim} can be calculated either from a single configuration (possibly after energy minimization and normal mode analysis), from a statistical ensemble as an ensemble average or a combination of ensemble averages (e.g. fluctuations), or from a dynamical trajectory as a time correlation function [57]. In the latter two cases, the size of the ensemble or trajectory, respectively, should be large enough so that the observable is converged, i.e. that the error bars on X_α^{sim} are sufficiently small. From a general point of view, the problems of *existence*, *stability* and *uniqueness* of a solution to the inversion problem defined by Eq. 7.1.1 is a concern [223]. The former two properties are generally assumed to be satisfied while attempting to solve the problem. In many cases, however, and when a limited number of experimental observables are considered, the uniqueness is not satisfied. Consider, for instance, the number of very different force fields developed for water or small alcohols that perform similarly well for most studied properties. From a practical point of view, the set of Eqs. 7.1.1 can only be solved in a consistent way in few cases. This is possible either for simple, few-parameter, systems (Sec. 7.4) or when *ab initio* energies and derivatives for molecules in vacuum are used as target values (Sec. 7.5). In most cases, however, one faces a number of difficult problems.

A. The computation of a single estimate of X_α^{sim}, if possible at all, is generally expensive since it requires a sufficiently long simulation for its value to be converged [7]. This is especially crucial for condensed-phase observables.

B. Many simulated observables cannot be computed until convergence within the current limit of computer power, whereas others cannot be related unambiguously to experimental observables or simply cannot be accessed experimentally. Typical examples of ambiguity are encountered in the interpretation of the heat of vaporization of liquids in the framework of an effective point charge model [224], the interpretation of crystallographic B-factors [87], or of spectroscopic measurements for parametrizing vibrational force constants (problem of zero point, possible influence of excited states). This usually limits the number of experimental observables against which a force field can be parametrized, in favourable cases, to few more than the number of parameters ($N_{obs} \gtrsim N_{param}$). The problem is then only slightly overdetermined and cross-checking, that is, reproduction of experimental observables not used in the calibration procedure, may become difficult.

C. Experimental measurements are subject to a certain uncertainty and may occasionally be erroneous, or incorrectly interpreted in terms of molecular properties. Incompatibility between experimental observables, misinterpretation in terms of simulated properties, or choice of an inadequate functional form for the interaction function may sometimes cause the absence of a solution to Eq. 7.1.1.

D. A single observable may not be sensitive to all features of the potential energy surface. Thus, a collection of several observables from very different sources should be combined, so that minimally one of them is sensitive to any force-field parameter.

E. Several (often related) observables may be determined by the same parameter or combination of parameters. Typical examples for fluids are the diffusion constant and the viscosity or the radial distribution functions and the density.

Due to the considerable size and computational costs of the problem, there is, in general, no systematic way to proceed, and parametrization procedures rely heavily on experience, judgement and intuition, with different choices made from one force field to another. A few possible ways to break down the problem into simpler ones can be found in the literature.

A. *Buildup approach:* Parameters are generated for small model compounds with one specific functional group at a time, and are assumed to be also valid for larger polyfunctional compounds. A typical example is the buildup of the OPLS force field for macromolecules (e.g. Ref. 79) from parameters for hydrocarbons [173,225,226], aromatic compounds [227], alcohols [228], sulphur compounds [229], amides [230,231] and nucleotide bases [232].

B. *Hierarchical approach:* Parameters are generated set by set, holding the previous set fixed while optimizing the next set. Typically, nonbonded parameters may be optimized against crystallographic data while holding parameters for the covalent interaction terms fixed.

Table 3 *Possible source of data for force-field parametrization or validation*

Technique	Phase	Type	Property	Parameters
Spectroscopy (IR, μ-wave)	Gas	1°	Vibrational/rotational spectra, overtone analysis	${}^{(n)}k_b$, ${}^{(n)}k_\theta$, ${}^{(2)}k_\xi$, CT
			Moments of inertia (small molecules)	b^0, θ^0
		3°	Rotational barriers and populations (estimates, small molecules)	${}^{(n)}k_\phi$
(UV, visible, μ-wave)	Solution	3°	Time-resolved fluorescence intensities, depolarization, circular dichroism	
(NMR)	Solution/ membrane	1°	Rotational barriers	${}^{(n)}k_\phi$, vdW(1,4)
		2°	Molecular structure, rotamers at equilibrium	${}^{(n)}k_\phi$, vdW(1,4)
		3°	Distances (NOE, chemical shift), orientations (J-coupling), equilibrium constants, order parameters, relaxation times, diffusion constants (translation/rotation), residence times, H/D exchange rates, etc.	
Diffraction (X-ray, neutron)	Crystal	1°	Molecular structure	b^0, θ^0
			Force-length interpolation	${}^{(2)}k_b$, ${}^{(2)}k_\theta$,
		2°	Molecular structure, crystal density, packing, lattice dynamics	vdW, q, H-bond
		3°	Electron density map, B-factors, occupancy factors	
(neutron)	Liquid/ polymers	3°	Radial distribution functions	
			Static and dynamic structure factors (polymers)	${}^{(n)}k_\phi$, vdW
Thermodynamic/ kinetic measurements	Gas	1°	Heats of formation	E
			Thermodynamic properties for rare gas mixtures	vdW, CR

Table 3 (*continued*)

Technique	Phase	Type	Property	Parameters
Thermodynamic/ kinetic measurements	Liquid/ solution	2°	Density, vapour pressure, solvation free energy, heat of vaporization, heat of mixing, partition coefficients, heat capacity, compressibility, viscosity, diffusion constant, transport properties, etc.	vdW, q
		3°	Chemical equilibrium parameters, pK_a, dielectric properties, reaction rates	
***Ab initio* and semiempirical calculations**	Gas	1°	Equilibrium geometries	b^0, θ^0
			Vibrational analysis	$^{(n)}k_b$, $^{(n)}k_\theta$, $^{(2)}k_\xi$, CT
			Conformers analysis	$^{(n)}k_\phi$
			Population analysis or fit of the electrostatic potential	q
			Van der Waals clusters (second-order perturbation or higher)	vdW
			Energy/derivatives	E, dE, d^2E
	Solution	1^0	*Idem*, supermolecule and/or reaction field approach	*Idem*

Type: 1° – primary, 2° – secondary, 3° – tertiary data, see Sec. 7.2; Parameters: parameters that can be calibrated using the corresponding data; CT: covalent coordinate cross-terms (Sec. 6.5); CR: van der Waals combination rules (Sec. 6.6.2); q: charges for Eq. 6.7.1.1; vdW: van der Waals parameters (Sec. 6.6); vdW(1,4): third-neighbour van der Waals parameters; k: force constants for Eqs. 6.1.1.1, 6.2.1.1, 6.3.1.1 and 6.4.1.1; H-bond: hydrogen-bonding interaction parameters (Sec. 6.9).

C. *Sensitivity approach:* A specific observable is brought closer to its target value by tuning the parameter it is likely to depend upon the most. This is normally done for the final fine tuning or subsequent corrections of a force field, for instance when repeated inconsistencies with experimental results are observed. A typical example is the tuning of the C_{12}(OW,OW) Lennard-Jones parameter used for the interaction with neutral carbon in GROMOS87, in order to better reproduce the solvation behaviour of proteins or peptides in solution [233,234], the solvation free energies of small hydrophobic organic molecules in water [235] and water/chloroform partition coefficients of Trp analogs [219]. Due to parameter interdependence (Sec. 7.6.1), such a tuning may cause the breakdown of other parts of the force field, and should thus be undertaken very cautiously, that is, only when considerable evidence has been gathered on different systems that the foreseen change is an improvement.

7.2. Source of data for force-field parametrization or validation

The observables that can be used to parametrize a force field can be experimental or theoretical, i.e. coming from *ab initio*, density functional or semiempirical molecular orbital calculations. Three classes can tentatively be distinguished: (i) *primary data*, i.e. data from experimental or theoretical sources that can, in principle, be interpreted directly in terms of force-field parameters; (ii) *secondary data*, i.e. data that can be compared reliably to simulation results; and (iii) *tertiary data*, i.e. data that can be compared with simulation results, but currently not accurately enough to be used for parametrization. This can be due either to convergence problems in the simulation, insufficient force-field resolution in terms of particles, difficulty of unambiguous interpretation of the experimental observable in terms of molecular properties, or too large experimental uncertainties. Table 3 summarizes various types of accessible data to parametrize (or compare) simulation results against.

7.3. Force-field parametrization using mostly experimental data

A possible scheme for the design of a force field using mostly experimental data is sketched here as an example. In the procedure, small monofunctional model compounds may be used to obtain parameters which will be used for larger polyfunctional systems.

A. Obtain the structural parameters (b^0, θ^0, ξ^0) from X-ray or neutron diffraction studies on crystals, or from spectroscopic measurements in the gas phase or on liquids (IR, Raman, NMR).

B. Obtain the corresponding force constants (k_b, k_θ ,k_ξ) from vibrational spectra in the gas phase. Possibly estimate the missing values using length-force interpolation on crystal structures, or use values from vibrational analysis on *ab initio* structures.

C. Make an initial guess at the torsional parameters (k_ϕ) from *ab initio* calculations on different conformers, from a molecular mechanics force field, or possibly from NMR measurements in solution.

D. Make an initial guess at the atomic charges (q) using results from *ab initio* calculations together with a population analysis or a fit to the electrostatic potential outside the molecule [236–239]. Note, however, that charges have different annotations in quantum calculations and in empirical force fields. In the former case, the atomic point charges (which are not observables) are tailored to approximate the electrostatic field (which is an observable) outside the molecule. In the latter case, they are effective parameters to model long-range interactions. The transfer of charges between the two techniques is thus often unreliable. A better transferability is obtained if the quantum mechanical calculation includes a reaction field correction to mimic bulk solvent, and the derived charges are constrained to reproduce the effective dipole moment of the molecule in solution.

E. Make an initial guess at the van der Waals parameters (vdW), usually from another force field or, if not available, by using the Slater–Kirkwood formula.

F. Refine simultaneously k_ϕ, q, vdW and possibly H-bond energy term parameters in order to reproduce the experimental condensed-phase properties for crystals (structure, density, packing) or liquids/solutions (density, thermodynamic observables, fluctuations, transport properties, conformer populations and isomerization barriers, radial distribution functions, thermodynamic parameters of mixing). The optimization of the intermolecular interaction may be performed in a first step using constrained covalent degrees of freedom. Ideally, this fit should be performed considering many different systems (crystals, liquids, mixed liquids) so that the extracted parameters hopefully become independent of the choice of a specific system.

Due to the computational expenses involved in getting converged values of the corresponding observables and the number of parameters to be tuned simultaneously, point F is usually the most difficult part in the design of a condensed-phase force field.

7.4. Systematic parameter optimization for simple condensed-phase systems

7.4.1. By trial and error

Most of the condensed-phase force fields (typically liquids) reported in the literature have been optimized by trial and error. An initial guess is made and parameters are subsequently varied in a more or less systematic manner, until agreement with experimental observables is reached. Since many trials have to be performed, only the essentials of the optimization process are generally reported, and the final parameter set and simulated observables are quoted (see e.g. Refs. 170, 224 and 240–242).

7.4.2. Using sensitivity analysis

Sensitivity analysis attempts to elucidate the dependence of the output of a process (in the present case, the simulated observables) on either (i) the mechanism that transforms the input into the output (the functional form of the interaction function), or (ii) the input itself (the force-field parameters). In case (i), the method relies on functional sensitivity analysis [223,243]. Except for possible *a priori* restrictions on its asymptotic behaviour, continuity and smoothness properties, no functional form is presupposed for the interaction function, which is iteratively constructed so as to reproduce experimental observables adequately. Although appealing, the method is currently limited to small molecular clusters due to its complexity and expense. In case (ii), a functional form is selected and the analysis is performed in terms of its parameters. Assuming that Eq. 7.1.1 is solved in a least-squares-fit sense, that is by minimizing an objective function, typically

$$S(\{s_i\}) = \sum_{\alpha}^{N_{obs}} W_\alpha [X_\alpha^{target} - X_\alpha^{sim}(\{s_i\})]^2 \qquad (7.4.2.1)$$

where W_α is the weight given to observable α, sensitivity coefficients can be evaluated such as

$$\frac{\partial X_\alpha^{\mathrm{sim}}}{\partial s_i}, \frac{\partial s_i}{\partial X_\alpha^{\mathrm{target}}} \text{ or } \frac{\partial X_\alpha^{\mathrm{sim}}}{\partial X_\beta^{\mathrm{target}}} \tag{7.4.2.2}$$

and used as guides for parameter tuning, and for the choice of observables to be included into the fit. The method has been applied to series of amide and carboxylic acid crystal structures in order to evaluate the sensitivities of observables upon changes in the nonbonded interaction parameters [244]. In this application, observables are single- configuration observables like lattice parameters, lattice energy, or rigid-body forces and torques on the molecules in the unit cell. Extension of the method to observables expressed as ensemble averages would require the use of the statistical perturbation formula (see Sec. 7.4.5). The sensitivity approach has also been applied to free energy calculations [245].

7.4.3. Using the weak-coupling method

A technique to automatically adjust the value of a force-field parameter to that of a given observable (e.g. an experimental liquid property) based on the weak-coupling scheme [6,246] has been applied to liquid mercury, treated as a Lennard-Jones fluid [247], and to the SPC water model [248]. In this method, the time derivative of a parameter is weakly coupled to the difference between the instantaneous value of an observable and its target (experimental) value. The technique is only applicable when a strong relationship (high sensitivity) exists between a given parameter and a corresponding observable. This relationship need not be known exactly, but has to be monotonic within the convergence interval of the parameter, i.e. a local optimum is to be found in parameter space. As a consequence, the method is applicable only to relatively simple systems, where the number of parameters is limited, and where the dominating relationships between parameters and observables are straightforward. The method is well suited for the final (usually time-consuming) step of parameter fine tuning. Multiple parameters can be refined simultaneously against the corresponding observables, for instance, Lennard-Jones repulsion parameters against pressure or density, or charges or Lennard-Jones well-depths against enthalpy of vaporization.

7.4.4. Using a search method in parameter space

Since the weak-coupling method described above can only locate a local optimum in parameter space, it will fail for systems where a good initial guess at the parameters cannot be made. Provided, however, that an objective function, $S(\{s_i\})$, can be designed to assess the quality of any trial set of parameters, $\{s_i\}$ (e.g. Eq. 7.4.2.1), and that its evaluation is reasonably cheap, some of the search techniques described in Sec. 3 (e.g. MC or MD) may be applied to search in the space of force-field parameters. They are expected to be more powerful than the weak-coupling method, since they are, in principle, able to cross barriers. When the derivatives of the objective function with respect to the parameters can be evaluated, a method related to MD may be used [26,27].

7.4.5. Using the perturbation formula

The perturbation formula [249] can be used to make extrapolations of the value an observable would take upon a given change in a force-field parameter, using a single reference ensemble. The method has been employed for SPC water in order to determine self-consistently the value of the dielectric permittivity of the continuum (parameter) in a reaction field calculation, so that the value of the simulated dielectric permittivity of the liquid (observable) equals the parameter value [250]. It was also applied for the analysis of a polarizable SPC water model in order to estimate the impact of a change in parameters (e.g. Lennard-Jones interaction parameters, charges, polarizability) on the simulated properties of the liquid [171]. In that sense, perturbation analysis can be considered as a generalization of the sensitivity analysis described in Sec. 7.4.2 to observables which are not single-configuration values but ensemble averages.

7.5. Systematic parametrization using results from ab initio calculations in vacuum

Energies and derivatives from molecular orbital calculations can also be used as target observables [223,251,252]. In this case, two features may simplify the application of Eq. 7.1.1 [41]. First, the target observables are the energies of selected conformations and their first and second derivatives with respect to the coordinates. Since these are one-configuration observables, their simulated values, X_α^{sim}, are extremely cheap to calculate for a given trial set of parameters. Second, the number of observables N_{obs} can be made much larger than the number of parameters N_{param} by increasing the number of molecules and conformations entering the fitting procedure. This turns the problem into a tractable optimization problem, well suited for the use of a systematic procedure. More precisely, an objective function S is minimized [22,23]:

$$S(\{s_i\}) = \sum_{A}^{N_{molec}} \sum_{\alpha}^{N_{conf}} W_{A,\alpha} \Bigg[{}^{(0)}W_{A,\alpha} [E_{A,\alpha} - E_{A,\alpha}^{target}]^2 + \sum_{i}^{N_{at}} {}^{(1)}W_{A,\alpha}^{i} \left[\frac{\partial E_{A,\alpha}}{\partial x_i} - \frac{\partial E_{A,\alpha}^{target}}{\partial x_i} \right]^2$$
$$+ \sum_{i}^{N_{at}} \sum_{j>i}^{N_{at}} {}^{(2)}W_{A,\alpha}^{i,j} \left[\frac{\partial^2 E_{A,\alpha}}{\partial x_i \partial x_j} - \frac{\partial^2 E_{A,\alpha}^{target}}{\partial x_i \partial x_j} \right]^2 \Bigg] \qquad (7.5.1)$$

where N_{molec} is the number of molecules of N_{at} atoms in the training set, and N_{conf} is the number of (equilibrium or distorted) conformations used for each molecule. $E_{A,\alpha} \equiv E_{A,\alpha}(\{s_i\})$ denotes the energy calculated by the force field using any trial parameter set $\{s_i\}$ (including covalent, covalent coupling and nonbonded terms) for molecule A in conformation α, and $E_{A,\alpha}^{target}$ is the corresponding quantum mechanical energy. Both energies are given relative to the lowest energy conformation of the molecule. The weights W can be adjusted to increase the impact of selected terms in the fit. Additional advantages of the procedure are that (i) *ab initio* observables, unlike experimental ones, can be generated easily for new systems, (ii) non-equilibrium (i.e. distorted) conformations are included in the parametrization set, and (iii) no model is required to interpret the observables in terms of molecular properties.

Such an approach has currently only been used systematically for alkanes. The force field for alkanes is constructed by fitting an anharmonic nondiagonal potential energy expression to the results of HF/6-31G* *ab initio* calculations [22,23]. The force field (QMFF) is successful at reproducing (at lower computational costs) the *ab initio* potential energy surface. Since it is known that *ab initio* calculations at this level of theory do not reproduce experimental results very accurately [3], this quantum mechanical force field cannot be used as such for comparisons with experimental data in the gas phase. Under the assumption that the potential energy surface from the *ab initio* calculations bears the correct trends, and that the errors in each individual term of the empirical interaction function are systematic, one may try to scale these terms using a limited number of scaling factors (five for QMFF) in order to reproduce experimental vibrational frequencies of molecules in the gas phase. Some problems and disadvantages of the method are the following:

A. Dispersion effects are generally only correctly treated at the second order of perturbation theory and not at the Hartree–Fock level of theory.

B. The selected potential energy function (including anharmonicities and cross-terms) is relatively complex and its evaluation may become expensive.

C. Parametrization has to be performed in a consistent manner over a whole class of compounds. Introduction of a new functionality, e.g. the C=C double bond, would require a full reparametrization of all terms, including those which do not include a C(sp^2) carbon atom.

D. The number of parameters is likely to increase rapidly with the number of atom types. For the alkane force field, 78 parameters define the full covalent interaction [22,23]. Inclusion of a third atom type e.g. for C(sp^2), keeping the same terms and expansions in the interaction function, would roughly multiply the number of parameters by a factor of 5.

E. The choice of the molecules in the training set, and of the selected geometries of these molecules, is arbitrary and a given choice may influence or bias the fit.

F. The nonbonded parameters are optimized solely for intramolecular interaction in small molecules. It has been shown, however, in the case of alanine dipeptide in vacuum, that the relative energy of conformers and thermodynamic properties are weakly dependent on charges [182]. On the other hand, their role in determining the solvation behaviour will be large.

7.6. Technical difficulties in the calibration of force fields

7.6.1. Parameter interdependence

Since simulated observables depend, in principle, simultaneously on all the parameters in Eq. 7.1.1, parameter optimization in force-field development can be made difficult due to correlation or anticorrelation among them. As a typical example, torsional angle parameters and third-neighbour (1,4) van der Waals interaction parameters are highly correlated and cannot be adjusted independently. When valence coordinate cross-terms are included, all covalent internal coordinate energy parameters become interdependent, and inclusion of a new functional group may require a full reoptimization of the force field. Atomic charges, van der Waals parameters and hydrogen-bonding parameters are correlated and should be adjusted consistently, in particular when a proper description of hydrogen bonds is required. Since correlated parameters cannot be optimized separately, the dimensionality and difficulty of the optimization problem is increased.

A further consequence of parameter correlation is the unclear correspondence between parameter and observable. Since usually one observable correlates with many parameters, even in the most simple case (use of primary data), this correspondence is not always straightforward. For example, the equilibrium bond angle used in a force field is the one of a virtual isolated angle, where all the effects of the neighbouring groups through nonbonded strain, valence coordinate cross-terms and internal coordinate redundancy have been averaged out into a potential of mean force (Sec. 4). It is thus not evident how such an effective parameter relates to a single bond-angle measured in a real molecule. For example, in methylcyclopropane, the effective value of a C-C-C bond angle may vary from 60° to about 120° [23]. Similarly, when bond-angle flexibility is introduced into the SPC water model [253], it is found that an equilibrium angle $\theta^0 = 109.5°$ leads to an effective average angle of about 105.4°. The parameter θ^0 has to be increased to 114° in order to get an average value of about 109.5°.

7.6.2. *Parameter dependence on degrees of freedom (**D**), functional form (**F**), combination rules (**C**) and approximations (**A**)*

In the general case, a redefinition of the degrees of freedom explicitly treated, or of the characteristics of the interaction function, should be followed by a complete reoptimization of all force-field parameters, or at least those whose effects are likely to be most strongly correlated with the effects of the changed parameters. In the following, four important choices are given on which the optimal parameters to be used will strongly depend:

A. *Treatment of the electrostatic interaction:* The optimum effective charges used in a given force field are strongly dependent on the approximations made in the functional form chosen for the electrostatic interaction (e.g. cutoff radius, reaction field and continuum dielectric constant, shifting function, distance-dependent dielectric constant, use of lattice summation) and the specific environment (explicit or implicit solvent, low or high dielectric constant). Practically, it has been shown that the reparametrization of the SPC water model is required when a reaction field term is introduced in the interaction function [220] or when polarizability effects are included [171]. Similarly, the optimal ionic nonbonded parameters, when calibrated against experimental solvation free energies, will depend on the cutoff radius used for the Coulomb interaction (see Fig. 9).

B. *Van der Waals combination rules:* When combination rules are applied to determine van der Waals interaction pair parameters, optimum atomic parameters will depend on one specific choice. Moreover, a change of the atomic parameters of one atom type will change its interaction with all other atom types.

C. *Treatment of the torsional interaction:* The choices with respect to the treatment of redundancy in torsional coordinates (e.g. one to nine torsions for ethane) will directly affect the force constants to be used. Additionally, third-neighbour nonbonded interaction may be (i) normal, (ii) scaled (AMBER), (iii) determined by a specific set of parameters, i.e. uncoupled from other nonbonded interactions (GROMOS), or (iv) absent (ECEPP). Different choices will affect the choice of a proper torsional functional form, and the corresponding force constants, if the overall torsional barriers are to be reproduced correctly.

D. *Choice of the constrained degrees of freedom:* For example, when the bond angle of the SPC water molecule is made flexible, a full reparametrization of the model is required [253].

7.6.3. *Parameter dependence on the molecule training set and calibration observables*

A force field developed for a given set of compounds using a given set of observables can only predict similar properties for related compounds, i.e. in its domain of validity. In other words, one should use it for interpolations and not for extrapolations. For example, parameters developed for linear alkanes may fail if used for cyclic alkanes. Similarly, parameters developed solely to reproduce gas-phase observables will probably fail to reproduce condensed-phase properties.

7.6.4. *Nonconvergence of important observables*

Since Eq. 7.1.1 can only be used when X_α^{sim} is a single-valued function (i.e. the observable is converged), a number of observables which are important for practical applications of a force field cannot be used directly in its parametrization procedure due to slow convergence. This problem is also encountered when, in applications of a force field, simulation lengths or system sizes grow far beyond the size or timescale that was used during its parametrization. In addition, parameters may have a weak influence on the local (short-timescale) behaviour of a molecular system, but a strong impact on the global (long-timescale) behaviour. A typical example is the effect of the parameters determining the electrostatic interaction on the simulated dielectric constant of water, which requires simulation times of the order of a nanosecond to be converged [220]. Tuning of parameters whose effect can only be detected at the limit of reachable computer power is rather difficult.

7.6.5. *Existence of conflicting requirements*

Within a given model ($\mathbf{D}, \mathbf{F}, \mathbf{C}, \mathbf{A}$), Eq. 7.1.1 may have no solution if the model is incapable of reproducing, at the same time, two observables, for whatever combination of parameters. This situation occurs, for instance, when a single set of parameters is used to describe both intermolecular and third-neighbour van der Waals interactions. The correct description of the density of condensed-phase systems is then incompatible with a correct description of torsional barriers (especially when united atoms are used). Such conflicts may be resolved by a change in the model (e.g. a different set of parameters for the two types of interactions). As another example, application of the weak-coupling method to liquid mercury modelled as a Lennard-Jones fluid [247] has shown that no set of Lennard-Jones parameters can fit simultaneously the experimental phase behaviour, the density, the heat of vaporization and the diffusion constant over the temperature range corresponding to the liquid state. Finally, in the development of the MM3 force field [54], conflicting demands appear as the impossibility to fit vibrational frequencies simultaneously with structures and heats of formations within the MM2 functional form. The conflict has been resolved by an adaptation of the functional form and an increase in the number of parameters.

7.6.6. Force-field mixing problems

Parameters are generally not transferable from one force field to another [36] and difficulties may arise while mixing force fields developed and optimized separately, as, for example, when a macromolecular solute is immersed into a solvent. The proper solvation behaviour of a macromolecule will depend strongly on the balance between solute/solute, solute/solvent and solvent/solvent interactions [12], which is not guaranteed to be correct. A potential cause for imbalance is the use of different combination rules [36]. Similar problems may arise when two solvent models from different origins are mixed. Another example is the combination of the AMBER valence force field with the OPLS nonbonded force field [79]. This combination requires removal of the AMBER hydrogen-bonding term and modifications in the 1,4 interaction handling. In general, parameters designed for explicit solvent simulations, parameters developed for vacuum simulations intending to mimic solvent environment, and parameters for proper gas-phase simulations can neither be interchanged nor combined.

7.6.7. Validation of a force field and comparison of force fields

The question of the general quality of a force field cannot be easily answered [37]. It should finally be judged by the ability of the force field to reproduce or predict experimental data. However, one should keep in mind that each force field has a range of validity determined by the systems and experimental or theoretical observables it was calibrated with. This makes fair comparison among force fields a difficult task, although some attempts can be found in the literature [254–256].

From Secs. 7.6.1–7.6.3, it should be clear that force-field parameters are by no means physical constants. Thus the direct transfer of parameters from one force field to another is a hazardous procedure. Similarly, the usefulness of *ab initio* results in vacuum is limited when designing an empirical effective interaction function for condensed-phase systems. In both cases, the transferred parameters can at most serve as an initial guess.

8. Conclusions

In the present text, some of the major issues with respect to derivation and use of empirical classical force fields have been described. Focusing mainly on models at atomic resolution, the terms that are most commonly found in the interaction energy function have been listed. From the previous discussion, it should be clear that there is no universal force field, but rather a force field best suited to a given system, a given state (phase) of the system, a given studied property and a given computer budget. The overall accuracy of a force field is limited by the crudest approximation that is made and not by the best refined part of the interaction function. This crudest approximation may occur in the energy term to which the observable is most sensitive. For condensed-phase simulations, the crudest approximation is most likely made in modelling the nonbonded interactions.

The main advantages of empirical classical force-field simulations reside in (i) the flexibility of the choice of the degrees of freedom, (ii) the limited computational costs, (iii) the ability to obtain thermodynamic and dynamical properties in addition to structural properties, whenever required, (iv) the possible inclusion of environmental effects (explicit or implicit solvent), and (v) the ability to carry out unphysical processes. The main drawbacks are (i) the non-first-principles approach, i.e. the only justification of empirical force fields resides in their ability to reproduce a large amount of experimental data, (ii) the dependence of the results on the approximations made and the choice of the force-field training set and parametrization observables, (iii) the difficulty of parametrization, and (iv) the limitations in the validity of the laws of classical mechanics, i.e. sufficiently high temperature, for all but the lightest (H, He) particles, and as long as no chemical reaction or electronically excited state is involved.

Since for many problems the use of molecular orbital methods is currently not feasible, there is nevertheless considerable interest in developing empirical force fields. Due to the constant increase of computer power, the problems that can be addressed by these techniques increase regularly in size, complexity and in terms of the volume of conformational space that can be sampled. This, almost necessarily, entails further development of the force fields themselves. New functional forms are proposed, which allow for a better energetic resolution in force fields, and systematic procedures begin to emerge for the parametrization of these functions based on both theoretical and experimental data.

References

1. Van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J., Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vols. 1 and 2, ESCOM, Leiden, 1989 ff.
2. Lipkowitz, K.B. and Boyd, D.B., Reviews in Computational Chemistry, Vols. I–VII, VCH, New York, NY, 1990 ff.

3. Hehre, W.J., Radom, L., Schleyer, P.v.R. and Pople, J.A., *Ab Initio* Molecular Orbital Theory, Wiley, New York, NY, 1986, pp. 1–548.
4. Stewart, J.J.P., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. I, VCH, New York, NY, 1990, pp. 45–118.
5. Zerner, M.C., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. II, VCH, New York, NY, 1991, pp. 313–365.
6. Van Gunsteren, W.F. and Berendsen, H.J.C., Angew. Chem., Int. Ed. Engl., 29(1990)992.
7. Van Gunsteren, W.F., Hünenberger, P.H., Mark, A.E., Smith, P.E. and Tironi, I.G., Comput. Phys. Commun., 91(1995)305.
8. Warshel, A., Computer Modeling of Chemical Reactions in Enzymes and Solutions, Wiley-Interscience, New York, NY, 1991, pp. 1–236.
9. Field, M.J., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 82–123.
10. Whitnell, R.M. and Wilson, K.R., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. IV, VCH, New York, NY, 1993, pp. 67–148.
11. Liu, H., Müller-Plathe, F. and van Gunsteren, W.F., J. Mol. Biol., 261(1996)454.
12. Van Gunsteren, W.F., Luque, F.J., Timms, D. and Torda, A.E., Annu. Rev. Biophys. Biomol. Struct., 23(1994)847.
13. Yun-Yu, S., Lu, W. and van Gunsteren, W.F., Mol. Sim., 1(1988)369.
14. Keith, T.A. and Frisch, M.J., In Smith, D.A. (Ed.) Modeling the Hydrogen Bond, American Chemical Society, Washington, DC, 1994, pp. 22–35.
15. Ángyán, J.G., J. Math. Phys., 10(1992)93.
16. Cramer, C.J. and Truhlar, D.G., Science, 256(1992)213.
17. Cramer, C.J. and Truhlar, D.G., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. VI, VCH, New York, NY, 1995, pp. 1–72.
18. Tomasi, J. and Persico, M., Chem. Rev., 94(1994)2027.
19. Müller-Plathe, F. and van Gunsteren, W.F., Macromolecules, 27(1994)6040.
20. Bowen, J.P. and Allinger, N.L., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. II, VCH, New York, NY, 1991, pp. 81–97.
21. Dinur, U. and Hagler, A.T., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. II, VCH, New York, NY, 1991, pp. 99–164.
22. Maple, J.R., Hwang, M.-J., Stockfisch, T.P., Dinur, U., Waldman, M., Ewig, C.S. and Hagler, A.T., J. Comput. Chem., 15(1994)162.
23. Maple, J.R., Hwang, M.-J., Stockfisch, T.P. and Hagler, A.T., Isr. J. Chem., 34(1994) 195.
24. Gerber, P.R., Biopolymers, 32(1992)1003.
25. Jones, D.T., Protein Sci., 3(1994)567.
26. Ulrich, P., Scott, W.R.P., van Gunsteren, W.F. and Torda, A.E., In Müller-Plathe, F. and Korosec, W. (Eds.) Annual Report 1993/1994 of the Competence Center for Computational Chemistry, ETHZ, Zürich, 1994, pp. 17–25.
27. Ulrich, P., Scott, W.R.P., van Gunsteren, W.F. and Torda, A.E., Proteins Struct. Funct. Genet., 27(1997)367.
28. Lathrop, R.H. and Smith, T.F., J. Mol. Biol., 255(1996)641.
29. Howard, A.E. and Kollman, P.A., J. Med. Chem., 31(1988)1669.
30. Leach, A.R., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules, Vol. II, VCH, New York, NY, 1991, pp. 1–55.

31. Van Gunsteren, W.F., In Lavery, R., Rivail, J.-L. and Smith, J. (Eds.) Advances in Biomolecular Simulations, Vol. 239, American Institute of Physics (AIP) Conference Proceedings, New York, NY, 1991, pp. 131–146.
32. Scheraga, H.A., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. III, VCH, New York, NY, 1992, pp. 73–142.
33. Scheraga, H.A., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 231–248.
34. Ōsawa, E. and Orville-Thomas, W.J., J. Mol. Struct., 308(1994)1–331 (The whole issue is dedicated to conformational search).
35. Van Gunsteren, W.F., Huber, T. and Torda, A.E., European Conference on Computational Chemistry (ECCC 1), American Institute of Physics Conference Proceedings, Vol. 330, AIP Press, Woodbury, NY, 1995, pp. 253–268.
36. Van Gunsteren, W.F. and Mark, A.E., Eur. J. Biochem., 204(1992)947.
37. Gelin, B.R., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 127–146.
38. Binder, K., Topics in Applied Physics, Vol. 71: The Monte Carlo Method in Condensed Matter Physics, Springer, Berlin, 1992, pp. 1–392.
39. Hagler, A.T. and Ewig, C.S., Comput. Phys. Commun., 84(1994)131.
40. Maple, J.R., Dinur, U. and Hagler, A.T., Proc. Natl. Acad. Sci. USA, 85(1988)5350.
41. Hwang, M.J., Stockfisch, T.P. and Hagler, A.T., J. Am. Chem. Soc., 116(1994)2515.
42. Lifson, S. and Warshel, A., J. Chem. Phys., 49(1968)5116.
43. Warshel, A. and Lifson, S., J. Chem. Phys., 53(1970)582.
44. Lifson, S. and Stern, P.S., J. Chem. Phys., 77(1982)4542.
45. Engelsen, S.B., Fabricius, J. and Rasmussen, K., Acta Chem. Scand., 48(1995)548.
46. Engelsen, S.B., Fabricius, J. and Rasmussen, K., Acta Chem. Scand., 48(1995)553.
47. Hagler, A.T., Lifson, S. and Dauber, P., J. Am. Chem. Soc., 101(1979)5122.
48. Hagler, A.T., Lifson, S. and Dauber, P., J. Am. Chem. Soc., 101(1979)5131.
49. Hagler, A.T., Stern, P.S., Sharon, R., Becker, J.M. and Naider, F., J. Am. Chem. Soc., 101(1979)6842.
50. Lifson, S., Hagler, A.T. and Dauber, P., J. Am. Chem. Soc., 101(1979)5111.
51. Dillen, J.L.M., J. Comput. Chem., 16(1995)565.
52. Dillen, J.L.M., J. Comput. Chem., 16(1995)610.
53. Allinger, N.L., J. Am. Chem. Soc., 99(1977)8127.
54. Allinger, N.L., Yuh, Y.H. and Lii, J.-H., J. Am. Chem. Soc., 111(1989)8551.
55. Lii, J.-H. and Allinger, N.L., J. Am. Chem. Soc., 111(1989)8566.
56. Lii, J.-H. and Allinger, N.L., J. Am. Chem. Soc., 111(1989)8576.
57. Allen, M.P. and Tildesley, D.J., Computer Simulation of Liquids, Oxford University Press, Oxford, 1987, pp. 1–385.
58. McCammon, J.A. and Harvey, S.C., Dynamics of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1987.
59. Brooks III, C.L., Karplus, M. and Pettitt, B.M., In Prigogine, I. and Rice, S. (Eds.) Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics, Wiley Series on Advances in Chemical Physics, Vol. LXXI, Wiley, New York, NY, 1988, pp. 1–259.
60. Weiner, P.K. and Kollman, P.A., J. Comput. Chem., 2(1981)287.

61. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106(1984)765.
62. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7(1986)230.
63. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham III, T.E., DeBolt, S., Ferguson, D., Seibel, G. and Kollman, P., Comput. Phys. Commun., 91(1995)1.
64. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
65. Nilsson, L. and Karplus, M., J. Comput. Chem., 7(1986)591.
66. Smith, J.C. and Karplus, M., J. Am. Chem. Soc., 114(1992)801.
67. MacKerell Jr., A.D., Wiókiewicy-Kuczera, J. and Karplus, M., J. Am. Chem. Soc., 117(1995)11946.
68. Momany, F.A. and Rone, R., J. Comput. Chem., 13(1992)888.
69. Mayo, S.L., Olafson, B.D. and Goddard III, W.A., J. Phys. Chem., 94(1990)8897.
70. Némethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S. and Scheraga, H.A., J. Phys. Chem., 96(1992)6472.
71. Levitt, M., J. Mol. Biol., 168(1983)595.
72. Levitt, M., J.,Mol. Biol., 168(1983)621.
73. Levitt, M., Hirshberg, M., Sharon, R. and Daggett, V., Comput. Phys. Commun., 91(1995)215.
74. Levitt, M., J. Mol. Biol., 82(1974)393.
75. Van Gunsteren, W.F. and Berendsen, H.J.C., Groningen Molecular Simulation (GROMOS) Library Manual, Biomos, Nijenborgh 4, Groningen, 1987.
76. Scott, W.R.P. and van Gunsteren, W.F., In Clementi, E. and Corongiu, G. (Eds.) Methods in Computational Chemistry: METECC-95, STEF, Cagliari, 1995, pp. 397–434.
77. Gerber, P.R. and Müller, K., J. Comput.-Aided Mol. Design, 9(1995)251.
78. Mohamadi, F., Richards, N.G.J., Guida, W.C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, T. and Still, W.C., J. Comput. Chem., 11(1990)440.
79. Jorgensen, W.L. and Tirado-Rives, J., J. Am. Chem. Soc., 110(1988)1657.
80. Clark, M., Cramer III, R.D. and van Opdenbosch, N., J. Comput. Chem., 10(1989)982.
81. Rappé, A.K., Casewit, C.J., Colwell, K.S., Goddard III, W.A. and Skiff, W.M., J. Am. Chem. Soc., 114(1992)10024.
82. Vedani, A., J. Comput. Chem., 9(1988)269.
83. Van Gunsteren, W.F., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 3–36.
84. Banks, J., Brower, R.C. and Ma, J., Biopolymers, 35(1995)331.
85. Fraternali, F. and van Gunsteren, W.F., J. Mol. Biol., 256(1996)939.
86. Hünenberger, P.H., Mark, A.E. and van Gunsteren, W.F., Proteins Struct. Funct. Genet., 21(1995)196.
87. Hünenberger, P.H., Mark, A.E. and van Gunsteren, W.F., J. Mol. Biol., 252(1995)492.
88. Liu, H., Müller-Plathe, F. and van Gunsteren, W.F., J. Chem. Phys., 102(1994)1722.
89. Liu, H., Müller-Plathe, F. and van Gunsteren, W.F., Chem. Eur. J., 2(1996)191.
90. Van Gunsteren, W.F., Brunne, R.M., Gros, P., van Schaik, R.C., Schiffer, C.A. and Torda, A.E., In James, T.L. and Oppenheimer, N.J. (Eds.) Methods in Enzymology: Nuclear Magnetic Resonance, Vol. 239, Academic Press, New York, NY, 1994, pp. 619–654.

91. Dammkoehler, R.A., Karasek, S.F., Berkley Shands, E.F. and Marshall, G.R., J. Comput.-Aided Mol. Design, 3(1989)3.
92. a. Judson, R.S., Jaeger, E.P. and Treasurywala, A.M., J. Mol. Struct., 308(1994)191.
    b. Judson, R.S., Jaeger, E.P., Treasurywala, A.M. and Peterson, M.L., J. Comput. Chem., 14(1993)1407.
93. DiNola, A., Roccatano, D. and Berendsen, H.J.C., Proteins Struct. Funct. Genet., 19(1994)174.
94. Torda, A.E. and van Gunsteren, W.F., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. III, VCH, New York, NY, 1992, pp. 143–172.
95. Chang, G., Guida, W.C. and Still, W.C., J. Am. Chem. Soc., 111(1989)4379.
96. Anet, F.A.L., J. Am. Chem. Soc., 112(1990)7172.
97. Saunders, M., Houk, K.N., Wu, Y.-D., Still, W.C., Lipton, M., Chang, G. and Guida, W.C., J. Am. Chem. Soc., 112(1990)1419.
98. Gibson, K.D. and Scheraga, H.A., J. Comput. Chem., 8(1987)826.
99. Gerber, P.R., Gubernator, K. and Müller, K., Helv. Chim. Acta, 71(1988)1429.
100. Ferguson, D.M. and Raber, D.J., J. Am. Chem. Soc., 111(1989)4371.
101. Saunders, M., J. Comput. Chem., 10(1989)203.
102. Bruccoleri, R.E., Haber, E. and Novotný, J., Nature, 335(1988)564.
103. Huber, T., Torda, A.E. and van Gunsteren, W.F., J. Comput.-Aided Mol. Design, 8(1994)695.
104. Ferguson, D.M., Glauser, W.A. and Raber, D.J., J. Comput. Chem., 10(1989)903.
105. Sheridan, R.P., Nilakantan, R., Dixon, J.S. and Venkataraghavan, R., J. Med. Chem., 29(1986)899.
106. Dolata, D.P., Leach, A.R. and Prout, K., J. Comput.-Aided Mol. Design, 1(1987)73.
107. Smith, G.M. and Veber, D.F., Biochem. Biophys. Res. Commun., 134(1986)907.
108. Scheek, R.M., Torda, A.E., Kemmink, J. and van Gunsteren, W.F., In Hoch, J.C. (Ed.) Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy, Plenum, New York, NY, 1991, pp. 209–217.
109. Torda, A.E., Brunne, R.M., Huber, T., Kessler, H. and van Gunsteren, W.F., J. Biomol. NMR, 3(1993)55.
110. Vásquez, M., Biopolymers, 36(1995) 53.
111. Gros, P., van Gunsteren, W.F. and Hol, W.G.J., Science, 249(1990)1149.
112. Gros, P. and van Gunsteren, W.F., Mol. Sim., 10(1993)377.
113. Schiffer, C.A., Gros, P. and van Gunsteren, W.F., Acta Crystallogr., Sect. D, 51(1995)85.
114. Van Gunsteren, W.F., Nanzer, A.P. and Torda, A.E., In Binder, K. and Ciccotti, G. (Eds.) Monte Carlo and Molecular Dynamics of Condensed Matter Systems, Proceedings of the Euroconference, 3–28 July 1995, Como, Italy, Vol. 49, SIF, Bologna, 1996, pp. 777–788.
115. Fennen, J., Torda, A.E. and van Gunsteren, W.F., J. Biomol. NMR, 6(1995)163.
116. Nanzer, A.P., Huber, T., Torda, A.E. and van Gunsteren, W.F., J. Biomol. NMR, 8(1996)285.
117. Ryckaert, J.-P., Ciccotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23(1977)327.
118. Lipton, M. and Still, W.C., J. Comput. Chem., 9(1988)343.
119. Gotō, H. and Ōsawa, E., J. Am. Chem. Soc., 111(1989)8950.
120. Kolossváry, I. and Guida, W.C., J. Comput. Chem., 14(1993)691.
121. Crippen, G.M., J. Phys. Chem., 91(1987)6341.
122. Blaney, J.M. and Dixon, J.S., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. V, VCH, New York, NY, 1994, pp. 299–335.

123. Byrne, D., Li, J., Platt, E., Robson, B. and Weiner, P., J. Comput.-Aided Mol. Design, 8(1994)67.
124. Van Schaik, R.C., van Gunsteren, W.F. and Berendsen, H.J.C., J. Comput.-Aided Mol. Design, 6(1992)97.
125. Van Schaik, R.C., Berendsen, H.J.C., Torda, A.E. and van Gunsteren, W.F., J. Mol. Biol., 234(1993)751.
126. Tufféry, P., Etchebest, C., Hazout, S. and Lavery, R., J. Comput. Chem., 14(1993)790.
127. Havel, T.F., Biopolymers, 29(1990)1565.
128. Ghose, A.K., Jaeger, A.P., Kowalczyk, P.J., Peterson, M.L. and Treasurywala, A.M., J. Comput. Chem., 14(1993)1050.
129. Schlick, T., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, Vol. III, VCH, New York, NY, 1992, pp. 1–71.
130. Moult, J. and James, M.N.G., Proteins Struct. Funct. Genet., 1(1986)146.
131. Bruccoleri, R.E. and Karplus, M., Biopolymers, 26(1987)137.
132. Li, Z. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 84(1987)6611.
133. Saunders, M., J. Am. Chem. Soc., 109(1987)3150.
134. Saunders, M., J. Comput. Chem., 12(1991)645.
135. Saunders, M. and Krause, N., J. Am. Chem. Soc., 112(1990)1791.
136. Došen-Mićovič, L., Tetrahedron, 51(1995)6789.
137. Leach, A.R., Prout, K. and Dolata, D.P., J. Comput.-Aided Mol. Design, 2(1988)107.
138. Leach, A.R., Prout, K. and Dolata, D.P., J. Comput.-Aided Mol. Design, 4(1990)271.
139. Vásquez, M. and Scheraga, H.A., Biopolymers, 24(1985)1437.
140. Vajda, S. and Delisi, C., Biopolymers, 29(1990)1755.
141. Crippen, G.M. and Havel, T.F., J. Chem. Inf. Comput. Sci., 30(1990)222.
142. Unger, R. and Moult, J., J. Mol. Biol., 231(1993)75.
143. Hartke, B., J. Phys. Chem., 97(1993)9973.
144. Gregurick, S.K., Alexander, M.H. and Hartke, B., J. Chem. Phys., 104(1996)2684.
145. Pillardy, J., Olszewski, K.A. and Piela, L., J. Phys. Chem., 96(1992)4337.
146. Piela, L., Olszewski, K.A. and Pillardy, J., J. Mol. Struct., 308(1994)229.
147. Piela, L., Kostrowicki, J. and Scheraga, H.A., J. Phys. Chem., 93(1989)3339.
148. Mark, A.E., van Gunsteren, W.F. and Berendsen, H.J.C., J. Chem. Phys., 94(1991) 3808.
149. Tsujishita, H., Moriguchi, I. and Hirono, S., J. Phys. Chem., 97(1993)4416.
150. Beutler, T.C., Mark, A.E., van Schaik, R.C., Gerber, P.R. and van Gunsteren, W.F., Chem. Phys. Lett., 222(1994)529.
151. Beutler, T.C. and van Gunsteren, W.F., J. Chem. Phys., 101(1994)1417.
152. Kirkpatrick, S., Gelatt Jr., C.D. and Vecchi, M.P., Science, 220(1983)671.
153. Wilson, S.R., Cui, W., Moskowitz, J.W. and Schmidt, K.E., Tetrahedron Lett., 29(1988)4373.
154. Brower, R.C., Vasmatzis, G., Silverman, M. and Delisi, C., Biopolymers, 33(1993)329.
155. Montcalm, T., Cui, W., Zhao, H., Guarnieri, F. and Wilson, S.R., J. Mol. Struct., 308 (1994)37.
156. Mao, B. and Friedman, A.R., Biophys. J., 58(1990)803.
157. Meza, J.C. and Martinez, M.L., J. Comput. Chem., 15(1994)127.
158. Frenkel, D., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 37–66.

159. Billeter, S.R. and van Gunsteren, W.F., Mol. Sim., 15(1995)301.
160. Van Gunsteren, W.F. and Berendsen, H.J.C., Mol. Phys., 34(1977)1311.
161. Berendsen, H.J.C. and van Gunsteren, W.F., In Ciccotti, G. and Hoover, W.G. (Eds.) Molecular Dynamics Simulation of Statistical-Mechanical Systems, Proceedings of the International School of Physics 'Enrico Fermi', Course 97, North-Holland, Amsterdam, 1986, pp. 496–519.
162. Ma, J., Hsu, D. and Straub, J.E., J. Chem. Phys., 99(1993)4024.
163. Jansen, A.P.J., Comput. Phys. Commun., 86(1995)1.
164. Rick, S.W., Stuart, S.J. and Berne, B.J., J. Chem. Phys., 101(1994)6141.
165. Tobias, D.J., Sneddon, S.F. and Brooks III, C.L., J. Mol. Biol., 216(1990)783.
166. Barker, J.A. and Watts, R.O., Mol. Phys., 26(1973)789.
167. Tironi, I.G., Sperb, R., Smith, P.E. and van Gunsteren, W.F., J. Chem. Phys., 102(1995)5451.
168. Lue, L. and Blankschtein, D., J. Phys. Chem., 96(1992)8582.
169. Perkyns, J.S. and Pettitt, B.M., Chem. Phys. Lett., 190(1992)626.
170. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F. and Hermans, J., In Pullman, B., (Ed.) Intermolecular Forces, Reidel, Dordrecht, 1981, pp. 331–342.
171. Stocker, U., Diplomarbeit, ETHZ, 1996.
172. Müller-Plathe, F., Rogers, S.C. and van Gunsteren, W.F., Macromolecules, 25 (1992)6722.
173. Kaminski, G., Duffy, E.M., Matsui, T. and Jorgensen, W.L., J. Phys. Chem., 98(1994)13077.
174. Curtiss, L.A. and Jurgens, R., J. Am. Chem. Soc., 94(1990)5509.
175. Elrod, M.J. and Saykally, R.J., Chem. Rev., 94(1994)1975.
176. Zavitsas, A.A. and Beckwith, A.L.J., J. Phys. Chem., 93(1989)5419.
177. Ermler, W.C. and Hsieh, H.C., In Dunning Jr., T.H. (Ed.) Advances in Molecular Electronic Structure Theory, Vol. 1, Calculation and Characterization of Molecular Potential Energy Surfaces, JAI Press, London, 1990, pp. 1–44.
178. Brown, F.B. and Truhlar, D.G., Chem. Phys. Lett., 113(1985)441.
179. Ogilvie, J.F., Proc. R. Soc. London A, 378(1981)287.
180. Dinur, U. and Hagler, T.A., J. Comput. Chem., 9(1994)919.
181. Van Gunsteren, W.F. and Karplus, M., Macromolecules, 15(1982)1528.
182. Pettitt, B.M. and Karplus, M., J. Am. Chem. Soc., 107(1985)1166.
183. Halgren, T.A., J. Am. Chem. Soc., 112(1990)4710.
184. Halgren, T.A., J. Am. Chem. Soc., 114(1992)7827.
185. Hart, J.R. and Rappé, A.K., J. Chem. Phys., 97(1992)1109.
186. Hart, J.R. and Rappé, A.K., J. Chem. Phys., 98(1992)2492.
187. Kestin, J., Knierim, K., Mason, E.A., Najafi, B., Ro, S.T. and Waldman, M., J. Phys. Chem. Ref. Data, 13(1984)229.
188. Waldman, M. and Hagler, A.T., J. Comput. Chem., 14(1993)1077.
189. Harvey, S.C., Proteins Struct. Funct. Genet., 5(1989)78.
190. Davis, M.E. and McCammon, J.A., Chem. Rev., 90(1990)509.
191. Berendsen, H.J.C., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 161–181.
192. Smith, P.E. and van Gunsteren, W.F., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 182–212.
193. King, G. and Warshel, A., J. Chem. Phys., 91(1989)3647.

194. Beglov, D. and Roux, B., J. Chem. Phys., 100(1994)9050.
195. Essex, J.W. and Jorgensen, W.L., J. Comput. Chem., 16(1995)951.
196. Wang, L. and Hermans, J., J. Phys. Chem., 99(1995)12001.
197. Russell, S.T. and Warshel, A., J. Mol. Biol., 185(1985)389.
198. Fincham, D., Mol. Sim., 13(1994)1.
199. Luty, B.A., Tironi, I.G. and van Gunsteren, W.F., J. Chem. Phys., 103(1995)3014.
200. Smith, P.E. and Pettitt, B.M., Comput. Phys. Commun., 91(1995)339.
201. Figueirido, F., Del Buono, G.S. and Levy, R.M., J. Chem. Phys., 103(1995)6133.
202. Luty, B.A. and van Gunsteren, W.F., J. Phys. Chem., 100(1996)2581.
203. Neumann, M., Mol. Phys., 50(1983)841.
204. Neumann, M., Steinhauser, O. and Pawley, G.S., Mol. Phys., 52(1984)97.
205. Brooks III, C.L., Pettitt, B.M. and Karplus, M., J. Chem. Phys., 83(1985)5897.
206. Brooks III, C.L., J. Chem. Phys., 86(1987)5156.
207. Madura, J.D. and Pettitt, B.M., J. Chem. Phys., 150(1988)105.
208. Schreiber, H. and Steinhauser, O., J. Mol. Biol., 228(1992)909.
209. Schreiber, H. and Steinhauser, O., Chem. Phys., 168(1992)75.
210. Schreiber, H. and Steinhauser, O., Biochemistry, 31(1992)5856.
211. Loncharich, R.J. and Brooks, B.R., Proteins Struct. Funct. Genet., 6(1989)32.
212. Steinbach, P.J. and Brooks, B.R., J. Comput. Chem., 15(1994)667.
213. Prevost, M., van Belle, D., Lippens, G. and Wodak, S., Mol. Phys., 71(1990)587.
214. Hummer, G., Soumpasis, D.M. and Neumann, M., Mol. Phys., 77(1992)769.
215. Barker, J.A., Mol. Phys., 83(1994)1057.
216. Chipot, C., Millot, C., Maigret, B. and Kollman, P.A., J. Chem. Phys., 101(1994)7953.
217. Chipot, C., Millot, C., Maigret, B. and Kollman, P.A., J. Phys. Chem., 98(1994)11362.
218. Wood, R.H., J. Chem. Phys., 103(1995)6177.
219. Daura, X., Hünenberger, P.H., Mark, A.E., Querol, E., Avilés, F.X. and van Gunsteren, W.F., J. Am. Chem. Soc., 118(1996)6285.
220. Smith, P.E. and van Gunsteren, W.F., Mol. Sim., 15(1995)233.
221. Oie, T., Maggiora, G.M., Christoffersen, R.E. and Duchamp, D.J., Int. J. Quantum Chem., Quantum Biol. Symp., 8(1981)1.
222. Dinur, U. and Hagler, T.A., J. Comput. Chem., 16(1995)154.
223. Ho, T. and Rabitz, H., J. Phys. Chem., 97(1993)13447.
224. Berendsen, H.J.C., Grigera, J.R. and Straatsma, T.P., J. Am. Chem. Soc., 91(1987)6269.
225. Jorgensen, W.L., Madura, J.D. and Swenson, C.J., J. Am. Chem. Soc., 106(1984)6638.
226. Jorgensen, W.L., Gao, J. and Ravimohan, C., J. Phys. Chem., 89(1985)3470.
227. Jorgensen, W.L. and Severance, D.L., J. Am. Chem. Soc., 112(1990)4768.
228. Jorgensen, W.L., J. Phys. Chem., 90(1986)1276.
229. Jorgensen, W.L., J. Phys. Chem., 90(1986)6379.
230. Jorgensen, W.L. and Swenson, C.J., J. Am. Chem. Soc., 107(1985)569.
231. Jorgensen, W.L. and Swenson, C.J., J. Am. Chem. Soc., 107(1985)1489.
232. Pranata, J., Wierschke, S.G. and Jorgensen, W.L., J. Am. Chem. Soc., 113(1991)2810.
233. Smith, L.J., Mark, A.E., Dobson, C.M. and van Gunsteren, W.F., Biochemistry, 34(1995)10918.
234. Daura, X., Oliva, B., Querol, E., Avilés, F.X. and Tapia, O., Proteins Struct. Funct. Genet., 25(1996)89.
235. Mark, A.E., van Helden, S.P., Smith, P.E., Janssen, L.H.M. and van Gunsteren, W.F., J. Am. Chem. Soc., 116(1994)6293.

236. Wilberg, K.B. and Rablen, P.R., J. Comput. Chem., 14(1993)1504.
237. Bachrach, S.M., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Population Analysis and Electron Densities from Quantum Mechanics, Vol. IV, VCH, New York, NY, 1994, pp. 171–227.
238. Marrone, T.J., Hartsough, D.S. and Merz Jr., K.M., J. Chem. Phys., 98(1994)1341.
239. Francl, M.M., Carey, C., Chirlian, L.E. and Gange, D.M., J. Comput. Chem., 17(1996)367.
240. Tironi, I.G. and van Gunsteren, W.F., Mol. Phys., 83(1994)381.
241. Liu, H., Müller-Plathe, F. and van Gunsteren, W.F., J. Am. Chem. Soc., 117(1995)4363.
242. Müller-Plathe, F., Mol. Sim., 18(1996)133.
243. Ho, T. and Rabitz, H., J. Chem. Phys., 89(1988)5614.
244. Thacher, T.S., Hagler, A.T. and Rabitz, H., J. Am. Chem. Soc., 113(1991)1991.
245. Wong, C.F. and Rabitz, H., J. Am. Chem. Soc., 95(1991)9628.
246. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. and Haak, J.R., J. Chem. Phys., 81(1984)3684.
247. Njo, S.L., van Gunsteren, W.F. and Müller-Plathe, F., J. Chem. Phys., 102(1995)6199.
248. Berwerger, C.D., van Gunsteren, W.F. and Müller-Plathe, F., Chem. Phys. Lett., 232(1995)429.
249. Zwanzig, R.W., J. Chem. Phys., 22(1954)1420.
250. Smith, P.E. and van Gunsteren, W.F., J. Chem. Phys., 100(1994)3169.
251. Hopfinger, A.J. and Pearlstein, R.A., J. Comput. Chem., 5(1984)486.
252. Momany, F.A. and Klimkowski, V., J. Comput. Chem., 11(1990)654.
253. Tironi, I.G., Brunne, R.M. and van Gunsteren, W.F., Chem. Phys. Lett., 250(1996)19.
254. Hall, D. and Pavitt, N., J. Comput. Chem., 5(1984)441.
255. Jorgensen, W.L., Chandrasekhar, J. and Madura, J., J. Chem. Phys., 79(1983)926.
256. Gundertofte, K., Liljefors, T., Norrby, P.-O. and Pettersson, I., J. Comput. Chem., 17(1996)429.
257. Bates, M.A. and Luckhurst, G.R., J. Chem. Phys., 104(1996)6696.

The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data

Peter Kollman[a], Richard Dixon[a], Wendy Cornell[a,*], Thomas Fox[a,], Chris Chipot[b] and Andrew Pohorille[b]**

[a] *Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94143-0446, U.S.A.*
[b] *NASA Ames Research Center, Moffet Field, CA 94035-1000, U.S.A.*

Introduction

In this chapter, we present an overview on our approach to developing a molecular mechanical model for organic and biological molecules and our opinions on what are the most important issues that go into the development of such a model. Since molecular mechanical models are more thoroughly reviewed by Hunenberger et al. [1], it is not inappropriate that we focus more on general principles and philosophy here. The main focus on new results presented here are consequences of some recent high-level *ab initio* calculations carried out by Beachy et al. [2]. This leads to a slight modification of our previously presented force field; we call this new model C96.

Development of parameters

Equation 1 represents the simplest functional form of a force field for studying molecules, in which one can vary all the degrees of freedom. The earliest force fields, which attempted to describe the structure and strain of small organic molecules, focused considerable attention on more elaborate functions of the first two terms, as well as cross terms. The modern versions of this are MM2/MM3 [3,4] and CVFF [5], which have been built with this 'top-down' philosophy.

$$
\begin{aligned}
U(\mathbf{R}) = & \sum_{\text{bonds}} K_r(r - r_{eq})^2 \quad \text{(bond)} + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \quad \text{(angle)} \\
& + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\theta - \gamma)] \quad \text{(dihedral)} \\
& + \sum_{i<j}^{\text{atoms}} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \quad \text{(van der Waals)} + \sum_{i<j}^{\text{atoms}} \frac{q_i q_j}{\varepsilon R_{ij}} \quad \text{(electrostatics)} \qquad (1)
\end{aligned}
$$

[*] Present address: Parke-Davis Pharmaceuticals, 2800 Plymouth Road, Ann Arbor, MI 48105, U.S.A.
[**] Present address: A Chem Forschung/CADD, Dr. Karl Thomae GmbH, D-88397 Biberach, Germany.

On the other hand, our approach, guided by our interest in proteins and nucleic acids, has been 'bottom-up' [6–8]. Thus, we focused on the atomic charges q_i first. Building on work by Momany [9] and Cox and Williams [10], we felt that the best, most general method to derive the atomic charges was to fit them to quantum mechanically calculated electrostatic potentials on appropriately chosen molecules or fragments. In our earlier attempt to do this, because of computational limitations in quantum mechanical calculations, we used a minimal basis set STO-3G to derive the q_i [6,7]. However, in our latest efforts [8], a 6-31G* basis set was used. This basis set has the fortunate property in that it leads to charges (dipole moments) that are enhanced over accurate gas-phase experimental values and, thus, implicitly builds in 'polarization' effects characteristic of polar molecules in aqueous solution. The fact that this basis set enhances the polarity just about the same amount as the water models TIP3P [11] and SPC [12] (where the charges are empirically adjusted to reproduce the water enthalpy of vaporization) is a fortunate fact and is key in leading to *balanced* solvent–solvent and solvent–solute interactions.

Although the 6-31G* electrostatic potential charges are well suited for intermolecular interactions, a key stumbling block in their use in a general force field is that they often are statistically ill-determined [13] for buried charges in the molecule and, in that case, can lead to a poor representation of conformational energies. The key breakthrough to solve this problem was the RESP model, developed by Bayly et al. [14]. By employing a hyperbolic restraint and multimolecule and multiconformational fitting (the latter independently noted as useful by Reynolds et al. [15]), a general and powerful method to derive 6-31G* based charges for any organic/biochemical model emerged.

Van der Waals parameters are generally dominated by the inner closed shell of electrons, and thus are fortunately far more transferable than atomic charges. Therefore, generally only one set of van der Waals parameters (radius and well depth) per atom type need be employed (with the important exception of hydrogen) [16,17]. The emergence of a general model that is empirically calibrated to fit liquid structures and enthalpies, the OPLS model [18], led us to use this approach in our force field. Although we made some adjustments and additions to that model (e.g. many different van der Waals parameters on hydrogens), our van der Waals model was very similar to OPLS and some parameters were taken from that model without modification.

Why can one not derive the van der Waals parameters for atoms using quantum mechanical calculations, as we have done for the charges? Unfortunately, such an approach is currently impractical since dispersion attraction is nonexistent at the SCF (Hartree–Fock level). Furthermore, it is very important to correctly reproduce the density of condensed-phase systems; thus, the empirical approach of OPLS is necessary at this time [19].

Continuing with the 'bottom-up' development of our force field, we come to the torsion energy term, where the V_n and γ come from either experiment or quantum mechanical calculations on small molecule models. At this point, a key conceptual difference with the 'top-down' force fields MM2/MM3 [3,4] should be stressed. Whereas MM2/MM3 often uses many terms in the Fourier series for rotation around a given bond type and attempts to reproduce the conformational energy for a collection of molecules, we have taken a minimalist approach [8]. For example, we have only a single V_3 torsional term around an X-C-C-Y bond except when X or Y are electronegative, where another term can be rationalized from electronic effects and can be derived directly using quantum mechanical calculations. This helps our model to be more easily generalized to new molecules, albeit in some cases probably at the cost of some accuracy. Exceptions to this minimalist approach are the ψ, ϕ of peptides and the χ of nucleic acids, where more terms were added to ensure as accurate as possible a reproduction of the conformational energies around these key bonds.

Finally, to ensure a reasonable representation of bond and angle terms, we use empirical data (structures and vibrational frequencies). The use of this simple harmonic model precludes high accuracy, but in our opinion such terms are of secondary importance in reproducing conformational and interaction energies in molecular recognition, proteins and nucleic acids. Thus, one does not want to compromise the simplicity and generality of the model with more complex functional forms. On the other hand, in our opinion [20,21], it is essential that bond angles are flexible for an accurate reproduction of the above properties.

Testing the model

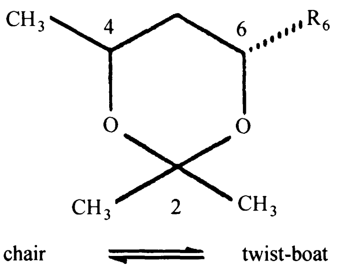
A key test of our approach was the ability to reproduce accurately liquid structures and energies and free energies of solvation. In a general sense, this is merely testing the compatibility of van der Waals parameters derived from simple liquids with ESP [22] and RESP 6-31G* [8,23] electrostatic potential based charges. The aqueous solvation free energies of a large number of molecules including substituted benzenes [22], methanol [23], hydrocarbons [8], N-methyl acetamide [23] and dimethyl sulfide [8] as well as the liquid structure and energy of methanol and N-methyl acetamide showed very good agreement with experiment. The point to emphasize is that little or no adjustment of parameters was done. Recently, Fox [24] has shown that our approach leads to a density and enthalpy of vaporization of liquid dimethyl sulfoxide (DMSO) within 2% of experiment, using RESP charges and van der Waals parameters taken without modification from the corresponding values in proteins. Liu et al. [25] have derived a united-atom DMSO model by empirically adjusting the molecular mechanical parameters to exactly reproduce the experimental density and enthalpy of vaporization. In the process, they had to make the equilibrium O-S bond length (R = 1.95 Å) significantly different from experiment (R = 1.80 Å).

The advantage of the Liu et al. approach is that a rigid united-atom model is computationally more efficient and consistent with the approach to the development of the SPC water model. Nonetheless, our all-atom approach allows the use of the correct internal geometry, is more consistent with the models of solutes, where bond angles are flexible, reproduces the dynamical properties as well as the Liu et al. model, and is easily applicable to any other liquid without parameter adjustment and with apparently little loss in accurate reproduction of the density and enthalpies of vaporization.

A test of our electrostatic model was provided by Hobza et al. [26]. Applying the highest level of *ab initio* theory practical, they calculated the 29 possible hydrogen bonding nucleic acid base–base interaction energies. They then compared these with the energies determined by various force fields and semiempirical quantum mechanical models. Encouragingly, the Cornell et al. [8] model was, on balance, closer to the *ab initio* model than any of the others, even the OPLS [27] and CHARMM23 [28] models. This was despite the fact that the Cornell et al. model simply fit the base charges with a RESP model, whereas OPLS and CHARMM23 adjusted them empirically to reproduce, among other things, hydrogen bond energies between bases or between base and water molecules.

The ability of our force field to model intramolecular (conformational) energies was provided by the studies of Rychnovsky et al. [29]. They studied a well-defined conformational equilibrium between chair and twist-boat conformers of substituted 1,3-dioxanes (see Scheme 1).

Even though high-level *ab initio* calculations reproduced the relative energies of these molecules well, MM2*/MM3* (MacroModel implementation of MM2 and MM3) and MM2/MM3 did not. However, our molecular mechanical model using RESP charges [17] had a correlation coefficient relative to these *ab initio* energies of $r^2 = 0.997$ with an average absolute error of 0.45 kcal/mol. In contrast, MM3 produced only an r^2 of 0.749 and an average error of 2.37 kcal/mol [30]. The important role of the electrostatic term, in determining these energies (the r^2 of the relative electrostatic energies with the relative total energies was 0.99), explained the superior performance of the Cornell et al. approach compared to MM3. A qualitative insight into why the electrostatic and total energies were correlated suggested that, in



Scheme 1. Chair/twist-boat conformational equilibrium of 2,2-dimethyl-trans-4-methyl-6-R-1,3-dioxane (R = substituent).

addition to a steric effect favoring the twist-boat conformer, electron withdrawing substituents favored the chair conformation because of electrostatic attraction with the oxygen along the ring. This led to the idea that a 6-CF_3 substituent would have a greater tendency than 6-CH_3 to be axial in the chair conformation, despite the smaller size of 6-CH_3. This idea was tested and supported in a joint experimental and theoretical study involving Rychnovsky's laboratory and ours [31]. Thus, even though this is one limited example, it provides encouragement that the approach described in Refs. 30 and 31 will be able to accurately represent intramolecular energies.

It is worth noting that a referee of Ref. 31 thought we were being unfair to MM3, because the V_1, V_2 and V_3 parameters had not been further optimized for this system. This is precisely the advantage of our model, where the accurate representation of the electrostatic charges with the RESP model often obviates the need for further parameter adjustment.

Non-additive and more complex models

What are the most important weaknesses in the above-described parametrizational approach and the use of Eq. 1? In our opinion, the main ones are two: the use of an effective two-body potential and the use of only atom centered charges.

$$\begin{aligned} E_{pol} &= -\frac{1}{2}\sum_{1}^{\text{atom}} \mu_i E_i^{(0)} \quad \text{(polarization)} \\ &+ \sum_{\text{ligands } j,k}^{\text{ion } i} A_{ijk} e^{\alpha_{ij}R_{ij}} e^{-\alpha_{ik}R_{ik}} e^{-\beta_{jk}R_{jk}} \quad \text{(three-body exchange)} \end{aligned} \qquad (2)$$

In the last year, we have made substantial progress in laying the foundation for the development of a complete force field including explicit nonadditive effects (adding Eq. 2 to Eq. 1). First, we have shown that such models, in contrast to additive models, lead to a good agreement with experimental solvation free energies of representative organic ions $CH_3NH_3^+$ and $CH_3CO_2^-$ without any adjustment of van der Waals parameters [32]. Secondly, we have shown that such nonadditive terms are essential, albeit nonobligatorily [33], in accurately describing cation-π interactions [34]. Thirdly, we have shown that one can equally well describe liquid CH_3OH and NMA with additive models or a nonadditive model in which the charges are uniformly reduced (by 0.88) [23]. Finally, the interaction free energy of Li^+ with hexa anisole spherand is more accurately described by nonadditive than additive molecular mechanical models [35]. In addition, considering off-center charges in electrostatic potential fit models of atoms with 'lone pairs' shows that they can often be important in leading to a very accurate description of hydrogen bond directionality [36].

Long-range electrostatic effects

To accurately describe the energy and structure of complex systems, not only are the functional form and parameters of molecular models described by Eqs. 1 and 2 important, but also the manner in which the long-range electrostatic effects are represented. The standard approach is to use a nonbonded cutoff for both electrostatic and van der Waals interactions, which seems to be a reasonable method for proteins, but appears to be a poor method to describe highly charged molecules such as nucleic acids. Darden and co-workers have shown the impressive efficiency and accuracy of the particle mesh Ewald (PME) method in representing protein crystals [37] (0.3 Å rms deviation from the observed crystal structure for bovine pancreatic trypsin inhibitor (BPTI) in a 1 ns simulation with an increase in computer time of only ~ 50% over standard cutoff methods); in collaboration with Darden, Cheatham et al. [38] have shown that the PME method leads to very accurate simulations of proteins, DNA and RNA in solution.

Recent *ab initio* calculations by Beachy et al.

In the development of the protein part of the Cornell et al. force field, torsional parameters were calibrated to reproduce, as accurately as possible, high-level *ab initio* calculations by Gould and Kollman [39] on the alanyl and glycyl dipeptides in C_{7eq}, C_{7ax}, β and α_R geometries (in glycyl dipeptide, C_{7eq} and C_{7ax} are degenerate). Since α_R is not a local minimum on the potential surface of these dipeptides, the energy of α_R was evaluated by constraining φ, ψ to a representative value of − 60.7°, − 40.7°. The final molecular mechanical energies exactly reproduced the *ab initio* relative energies for C_{7eq}, β and α_R for alanyl dipeptide and were in reasonable agreement for the other conformations.

Thus, it was rather surprising when the results of studies by Beachy et al. [2] appeared. These authors studied 10 local minima of alanyl tetrapeptide (Ace-Ala-Ala-Ala-NMe) where Ace=CH_3CO and NMe=$NH(CH_3)$, as well as the α_R geometry. Given the way the intramolecular torsional potentials were developed by Cornell et al., it was disappointing that the average difference in energy for the 10 local minima between the Cornell et al. model and the *ab initio* calculations was 2.5 kcal/mol. (Table 1). A small part of this discrepancy could be attributed to the difference in *ab initio* energies. The difference between C_{7eq} and β was 1.5 kcal/mol in Gould and Kollman [39] and 0.9 kcal/mol in the Beachy et al. [2] study. This effect could be magnified in longer peptides, so for a tetrapeptide one might attribute 1.8 kcal/mol of a discrepancy in the relative energies of a repeating β versus a repeating C_{7eq} conformation to the *ab initio* data used.

Nonetheless, one could also note that the 10 conformations chosen by Beachy et al. contained many examples of C_{7eq} and C_{7ax} conformations which are rarely found in peptides or proteins much longer than a few residues. Thus, this particular choice of conformations is somewhat unrepresentative of protein structures. Beachy et al. did study a constrained α_R conformation, given its importance in peptide and protein conformations (with φ, ψ constrained to − 52°, − 53°). They were kind enough to communicate their results on it to us, but for some reason they did not include it in their initial report [2].

In fact, the largest concern to us in the Beachy et al. study was that the Cornell et al. model found the constrained α_R only 4.0 kcal/mol above the most stable conformation #3, compared to 6.3 kcal/mol for the β-structure (conformation #1), whereas (Table 1) Beachy et al. found β more stable than constrained α_R by 6.4 kcal/mol.

We thus decided to create a model that reproduced exactly the $\alpha_R - \beta$ energy difference and to compare its performance with the Cornell et al. force field in protein molecular dynamics and conformational free energy calculations on a model peptide. We simply changed the torsional potentials around ϕ and ψ for simplicity, keeping just the same onefold and twofold Fourier components around each. This led to the model C96 presented in Table 1. Not only does it reproduce the $\beta - \alpha_R$ difference significantly better than the original Cornell et al. model, but the average error for the conformational energies goes from 2.5 to 1.6, with the largest error occurring for those conformations of relatively higher energy.

We also explored three other models in an analogous way, simply adjusting V_1 and V_2 for the ϕ, ψ torsional potentials in order to approximately reproduce the relative *ab initio* energies for conformations #1 and #3. The model labeled '88' simply has scaled the charges for the Cornell et al. force field by 0.88, which, as we have shown elsewhere [23], is an appropriate scale factor to make these charges 'gas-phase'-like; this model leads to a relatively small average error of 1.4 kcal/mol.

Instead of scaling the 6-31G* *ab initio* electrostatic potential derived charges, one could evaluate them with a basis set which more accurately represents the gas-phase dipole moments of small molecules, rather than enhancing them by ~10–20% as 6-31G* does. We have shown that a density functional based electronic structure approach does this well for small molecules, using a triple zeta plus polarization basis set [40]. We call this model DFT and the results of deriving the charges for the alanyl dipeptide using that approach and empirically altering the ϕ, ψ torsional potentials analogously to the C96 and 88 models (approximately reproducing the ΔE between conformations #1 and #3) are given in Table 1. Also reported in the table are given energies if one does the same approach using a nonadditive force field with atomic polarizabilities from Caldwell and Kollman [23].

As one can see, all the new models significantly improve the agreement with the *ab initio* calculations for the tetrapeptides, at the expense of the relative dipeptide energies. On the other hand, calibrating a model to reproduce tetrapeptide energies is probably a better, more transferable approach to proteins, provided a suitably representative set of conformations is included.

One should reemphasize, as discussed above, that the choice of a 6-31G* basis set to derive the charges was done to implicitly include polarization effects, since this basis set uniformly overestimates polarity. The use of 'gas-phase' charges may be more appropriate to compare with the *ab initio* gas-phase calculations by Beachy et al. [2] What Table 1 shows is that, whereas one can reproduce the *ab initio* data on average better with such models (88, DFT, DFTPOL), one can also adjust torsional potentials with the 6-31G* charges and create a model that compares nearly as well to *ab initio* and correctly represents the most important $\alpha_R - \beta$ energy difference. If one excludes

Table 1 *Relative energies of different models of alanyl tetrapeptide and alanyl dipeptide*

Conformation #[a]	Model (kcal/mol)					
	Ab initio[b]	Cornell et al.[c]	C96[d]	88[e]	DFT[f]	DFTPOL[g]
Tetrapeptide						
1(β)[a]	2.0	6.3	2.5	2.4	2.4	2.4
2	2.3	5.9	3.1	2.8	2.3	2.3
3	0.0	0.0	0.0	0.0	0.0	0.0
4	3.3	6.8	3.6	2.9	2.9	2.9
5	3.3	4.7	4.9	3.6	3.7	3.0
6	2.3	1.2	—	1.8	1.9	2.8
7	5.4	—	4.4	3.8	3.0	2.8
8	4.3	5.6	7.9	6.2	5.8	5.2
9	7.0	5.5	6.6	7.2	6.5	7.3
10	6.7	5.6	12.0	10.3	9.1	8.1
$\alpha_R(-52, -53)$[h]	(8.4)	(7.0)	(8.9)	(3.7)	(4.0)	(2.4)
Average deviation[i]	—	2.5	1.6	1.4	1.3	1.3
Dipeptide[b]						
Conformation[j]						
β	0.9	1.5	0.2	0.1	0.6	0.7
C_{7eq}	0.0	0.0	0.0	0.0	0.0	0.0
C_{7ax}	2.8	1.5	1.9	1.9	1.8	2.3
$\alpha_R(-60.7, -40.7)$[h]	(3.9)	(3.8)	(5.4)	(3.5)	(3.6)	(2.7)

[a] This is a β-sheet conformation; these 10 conformations are from Beachy et al. [2].

[b] See Beachy et al. [2] and unpublished results on α_R and dipeptides.

[c] Energies using the Cornell et al. force field [8].

[d] Cornell et al. force field with dihedral energies around both C-N-CT-C (ϕ) and N-C-CT-N (ψ) changed to $V_1/2 = 0.85$ kcal/mol, $\delta = 0°$ and $V_2/2 = 0.30$, $\delta = 180°$.

[e] Cornell et al. model with charges reduced by 0.88 and dihedral values changed as in footnote d with $V_1/2 = 0.55$ kcal/mol, $\delta = 0°$ and $V_2/2 = 0.30$, $\delta = 180°$.

[f] Cornell et al. model with charges changed to those derived for alanyl dipeptide in Ref. 8, but using DFT calculations with the basis set from St-Amant [40]. Torsional parameters changed as in footnote d to $V_1/2 = 0.75$, $\delta = 0°$ and $V_2/2 = 0.30$ kcal/mol, $\delta = 180°$.

[g] As in footnote d with polarization turned on using the approach in Ref. 23. $V_1/2 = 0.50$, $\delta = 0°$ and $V_2/2 = 0.30$ kcal/mol, $\delta = 180°$.

[h] ϕ, ψ constraint value used.

[i] Average magnitude of deviation from *ab initio* energy.

[j] See Gould and Kollman [39] for the definition of these conformations.

the high-energy conformations #9 and #10, the C96 model has an average difference versus *ab initio* of only 1.2 kcal/mol, comparable to those of 88, DFT and DFTPOL (1.3, 1.2 and 1.4 kcal/mol), and it represents the $\alpha_R - \beta$ energy difference much more accurately.

Given the excellent performance versus *ab initio* of nucleic acid base pairing and stacking of the Cornell et al. [8] model, as demonstrated by Hobza et al. [26], why does the model do much more poorly in representing these tetrapeptide energies? As noted by Cornell et al., intramolecular torsional energies in cases like the ϕ, ψ torsion in peptides and χ in nucleic acids are much more difficult to represent with simple, transferable torsional potentials, in contrast to the success of transferability noted above in 1,3-dioxanes. Of course, the *ab initio* data are not perfect for these systems, with the largest error probably coming from the dispersion attraction, but the average error of the relative conformational energy is likely in the range of 0.5 kcal/mol, much less than the deviations of these molecular mechanical models.

Comparing the two force fields on complex systems

Given that we have shown that ~ns trajectories of the Cornell et al. force field on ubiquitin [41] were in very good agreement with the X-ray structure and that this level of agreement was in the order Cornell et al. (PME) > Cornell et al. (cutoff) > Weiner et al. (PME) > Weiner (cutoff), we carried out molecular dynamics trajectories with the new C96 model for 600 ps with PME, for comparison with the earlier studies. As shown in Table 2, the difference between the results of the two trajectories in terms of rms movement from the crystal structure is quite small. A more detailed examination of the hydrogen bonding pattern, in terms of agreement with NMR and X-ray data, revealed that the new model led to better agreement in some areas and worse in others; all in all, the difference between the models was not significant. We also initiated free energy calculations on small peptides, determining the free energy of a $\beta \rightarrow \alpha$ a transition to a partially formed α-helix for both the Cornell et al. and C96 models.

Whereas the calculations discussed above were performed on isolated molecules, perhaps the most relevant issue is the behavior of peptides and proteins in solution. To address this issue, we used both the Cornell et al. and the C96 force fields to study

Table 2 *Rmsd from crystal structure*[a]

Model[b]		Final structure (Å)			Average structure (Å)		
		Heavy atoms	Backbone	C^α	Heavy atoms	Backbone	C^α
C96	(300 ps)[c]	1.336	0.965	0.903	0.925	0.637	0.616
Cornell et al.	(300 ps)	1.445	1.020	1.011	0.878	0.572	0.554
C96	(600 ps)	1.405	0.975	0.931	0.667	0.647	0.647
Cornell et al.	(600 ps)	1.533	1.108	1.060	1.010	0.671	0.640

[a] For details on ubiquitin simulation, see Ref. 41.
[b] Model used, see Table 1.
[c] Length of simulation.

a model oligopeptide in aqueous solution. As an example, we selected the Ac- and -NHMe terminally blocked undecamer of poly-L-leucine. This choice was motivated by its relevance to the studies of model transmembrane proteins [42]. Folding of this molecule during the transfer across the water–hexane interface was recently studied in multi-ns computer simulations using the Cornell et al. [8] model [43]. For the undecamer in the α-helical conformation in water, the free energy of unfolding the first residue from the N-terminus was determined in molecular dynamics simulations by varying discretely the ψ angle of this residue from −90° to +170°.

The simulation system consisted of a single peptide solvated by 2191 TIP4P water molecules in a box, the x, y, z dimensions of which were $41.09 \times 41.09 \times 41.09\,\text{Å}^3$. Periodic boundary conditions were applied in all three directions. The equations of motion were solved employing the Verlet algorithm with a time-step of 2.5 fs. The simulations were carried out in the (N, V, E) ensemble, with a periodic rescaling of the velocities to maintain the temperature at 300 K. Peptide–water and water–water interactions were truncated smoothly beyond 9 and 7 Å, respectively.

The free energies were evaluated as a function of the ψ angle, using the 'umbrella sampling' [44] approach. The range of values accessible to ψ was divided into four sequentially overlapping 'windows'. A harmonic restraining potential was applied to ensure that the values of ψ remained within the defined window. In addition, a biasing potential was included to yield a more uniform probability distribution of ψ in each window and, thereby, improve the accuracy of the calculation. The remaining 10 residues of poly-L-leucine were restrained to the α_R conformational region using soft harmonic potentials. The lengths of the molecular dynamics trajectories in the different windows varied between 1.4 and 3.9 ns. The total lengths of the simulations using the Cornell et al. and the C96 models were 9.3 and 9.7 ns, respectively. For each window, the probability distribution of ψ was accumulated/computed. The complete free energy profiles (or potentials of mean force) were obtained by matching the four individual curves in the overlapping regions, using the weighted histogram analysis method (WHAM) [45].

As seen in Fig. 1, the difference in stability between the folded (α_R) and the extended (β) states depends on the force field used. Whereas the Cornell et al. potential energy function favors α_R by 3.5 kcal/mol, the preference is reduced to only 1.2 kcal/mol with the C96 model. Thus, compared to the Cornell et al., the C96 stabilizes the β-conformation by 2.3 kcal/mol. This is close to the stabilization by 2.7 kcal/mol obtained for alanyl dipeptide in the gas phase (Table 1). In addition, the C96 model yields a free energy barrier 1.70 kcal/mol lower than the Cornell et al. force field. In the transition state approximation, this corresponds to an almost 20-fold increase in the average time of unfolding. This, in turn, means that the denaturation of poly-L-leucine in water will progress much slower using the Cornell et al. model than the C96 model.

An alternative approach to investigating the α-helix propensity of poly-L-leucine as a function of the force field is to calculate the free energy of breaking the first hydrogen bond along the α-helix. The distance R(O-N) between the first carbonyl oxygen and the amino group three residues away was used as the reaction coordinate. The free

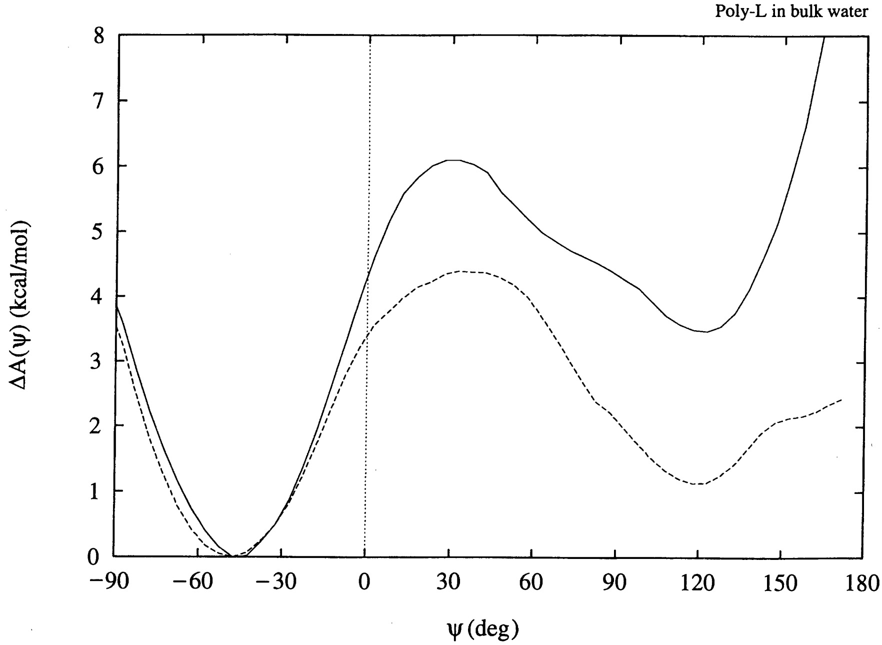


Fig. 1. Calculated free energy, plotted as a function of ψ from a ψ value characteristic of α to that characteristic of a β-conformation. Solid line: Cornell et al.; dotted line: C96 force field.

energy profile was generated from the simulations described previously by constructing the unbiased two-dimensional probability distribution P[R(O-H), ψ], and integrating the latter over ψ along the hydrogen bond distance. As can be seen from Fig. 2, the profiles generated using the two potential energy functions are markedly different. The Cornell et al. force field predicts the α-helix to be 3.5 kcal/mol lower in free energy than the extended form, whereas the same difference is only 1.2 kcal/mol with the C96 model. Using the Cornell et al. model, the extended structure is separated from the α-helical one by a very low free energy barrier, indicating that transitions from the extended structure to the α-helix should be very rapid. In contrast, the C96 model yields a barrier from the extended state that is approximately 1 kcal/mol higher. In addition, the free energy profile for this model exhibits a small minimum around 4.3 Å, absent in the Cornell et al. model, suggesting that a 3_{10}-helix might be an intermediate in coil-to-helix transitions.

A similar study on capped oligopeptides of various lengths, built from L-alanine, has been recently performed by Young and Brooks [46] using the CHARMM23 force field. The reported profiles along both ψ and R(O ··· H) are similar to those obtained here using the C96 model. Specifically, both models yield a similar relative stability of the α-helical conformation compared to the extended (β) form, and a 3_{10}-helix as an intermediate in the unfolding. The main difference is that the free energy barrier process separating the α_R and the β-states is much smaller for poly-L-alanine than for

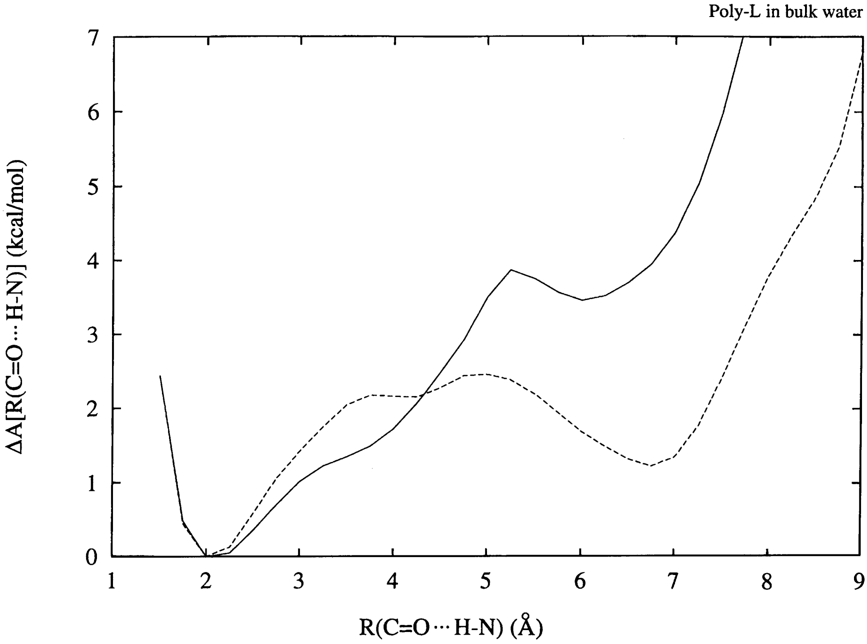


Fig. 2. Calculated free energy, plotted as a function of the hydrogen bond distance for the N-terminal hydrogen bond in the α-helix. Solid line: Cornell et al.; dotted line: C96 force field.

poly-L-leucine. In particular, it is predicted that the first residue in the undecamer of poly-L-leucine unfolds 80 times slower than in the decamer of poly-L-alanine. This is not surprising, considering that the closely packed side chains of poly-L-leucine prevent the surrounding water from disrupting the intramolecular hydrogen bonds along the backbone.

In summary, the differences in energies between the α and β energies found by Cornell et al. and C96 for the isolated peptides persist in aqueous solution. It is not clear whether simple modifications of the torsional potentials such as employed in C96 are the most appropriate way of bringing the Cornell et al. potentials in agreement with the results of Beachy et al. Furthermore, there are no experimental data to assess which force field gives results closer to reality. Thus, further experimental and theoretical studies on these systems are in order.

References

1. Hünenberger, P.H. and van Gunsteren, W.F., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 3, Kluwer, Dordrecht, 1997, pp. 3–82.
2. Beachy, M.D., Chasman, D., Murphy, R.B., Halgren, T.A. and Friesner, R.A., J. Am. Chem. Soc., submitted.

3. Burkert, U. and Allinger, N.L., Molecular Mechanics, American Chemical Society, Washington, DC, 1982.
4. Allinger, N.L., Yuh, Y.H. and Lii, J.-H., J. Am. Chem. Soc., 111(1989)8551, 8566, 8576.
5. Maple, J.R., Hwang, M.J., Stockfisch, T.P., Demur, U., Waldman, M., Ewig, C.S. and Hogle, A.T., J. Comput. Chem., 15(1994)162.
6. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106(1984)765.
7. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7(1986)230.
8. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz Jr., K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A., J. Am. Chem. Soc., 117(1995)5179.
9. Momany, F., J. Phys. Chem., 82(1978)592.
10. Cox, S.R. and Williams, D.E., J. Comput. Chem., 2(1980)304.
11. Jorgensen, W.L., Chandresekhar, J., Madura, J., Impey, R.W. and Klein, M.L., J. Chem. Phys., 79(1983)926.
12. Berendsen, H.J.C., Grigera, J.R. and Straatsma, T., J. Phys. Chem., 91(1987)6269.
13. Francl, M.M., Casey, C., Chirlian, L.E. and Gange, D.M., J. Comput. Chem., 17(1995)367.
14. a. Bayly, C.I., Cieplak, P., Cornell, W.D. and Kollman, P.A., J. Phys. Chem., 97(1993)10269.
    b. Cornell, W.D., Cieplak, P., Bayly, C.I. and Kollman, P.A., J. Am. Chem. Soc., 115(1993)9620.
15. Reynolds, C.A., Essex, J.W. and Richards, W.G., J. Am. Chem. Soc., 114(1992)9075.
16. Gough, C.A., DeBolt, S.E. and Kollman, P.A., J. Comput. Chem., 8(1992)963.
17. Veenstra, D.L., Ferguson, D.M. and Kollman, P.A., J. Comput. Chem., 8(1992)971.
18. Jorgensen, W.L. and Tirado-Rives, J., J. Am. Chem. Soc., 110(1988)1657.
19. Kaminski, G. and Jorgensen, W.L., J. Phys. Chem., in press.
20. Kollman, P.A. and Dill, K.A., J. Biomol. Struct. Dyn., 8(1991)1003.
21. Gibson, K.D. and Scheraga, H.A., J. Biomol. Struct. Dyn., 8(1991)1109.
22. Kuyper, L., Ashton, D., Merz, K.M. and Kollman, P.A., J. Phys. Chem., 95(1991)6661.
23. Caldwell, J.W. and Kollman, P.A., J. Phys. Chem., 99(1995)6208.
24. Fox, T., unpublished results.
25. Liu, H.Y., Mullerplathe, F. and van Gunsteren, W.F., J. Am. Chem. Soc., 117(1995)4313.
26. Hobza, P., Hubalek, F., Kabelac, M., Spooner, J., Mezzlik, P. and Vondrasek, J., J. Phys. Chem., 257(1996)31.
27. Pranata, J. and Jorgensen, W.L., Tetrahedron, 47(1991)2491.
28. Mackerell, A.D., Wiórkiewiecz-Kuczera, J. and Karplus, M., J. Am. Chem. Soc., 117(1995)11946.
29. Rychnovsky, S.D., Yang, G. and Powers, J.P., J. Org. Chem., 58(1993)5251.
30. Howard, A.E., Cieplak, P. and Kollman, P.A., J. Comput. Chem., 16(1995)243.
31. Cieplak, P., Howard, A.E., Powers, J., Rychnovsky, S. and Kollman, P.A., J. Org. Chem., 61(1966)3662.
32. Meng, E.C., Cieplak, P., Caldwell, J.W. and Kollman, P.A., J. Am. Chem. Soc., 116(1994)12061.
33. Chipot, C., Maigret, B., Pearlman, D.A. and Kollman, P.A., J. Am. Chem. Soc., 118(1996)2998.
34. Caldwell, J.W. and Kollman, P.A., J. Am. Chem. Soc., 117(1995)4177.
35. Sun, Y., Caldwell, J.W. and Kollman, P.A., J. Phys. Chem., 99(1995)10081.
36. Dixon, R.W. and Kollman, P.A., J. Comput. Chem., in press.

37. York, D.M., Wlodawer, A., Petersen, L. and Darden, T.A., Proc. Natl. Acad. Sci. USA, 91(1994)8715.
38. Cheatham, III, T.E., Miller, J.L., Fox, T., Darden, T.A. and Kollman, P.A., J. Am. Chem. Soc., 117(1995)4193.
39. Gould, I.R. and Kollman, P.A., J. Phys. Chem., 96(1992)9255.
40. St-Amant, A., Cornell, W.D., Halgren, T. and Kollman, P.A., J. Comput. Chem., 16(1995)1483.
41. Fox, T. and Kollman, P.A., Proteins Struct. Funct. Genet., 25(1996)315.
42. Mouritsen, O.G. and Bloom, M., Annu. Rev. Biophys. Biomol. Struct., 22(1993)145.
43. Chipot, C. and Pohorille, A., to be published.
44. Torrie, G.M. and Valleau, J.P., Chem. Phys. Lett., 28(1992)578.
45. Kumar, S., Bouzida, D., Swendsen, R.H., Kollman, P.A. and Rosenberg, J.M., J. Comput. Chem., 13(1992)1011.
46. Young, W.S. and Brooks III, C.L., J. Mol. Biol., 259(1996)560.

A separating framework for increasing the timestep in molecular dynamics

Eric Barth, Margaret Mandziuk and Tamar Schlick
Department of Chemistry and Courant Institute of Mathematical Sciences,
The Howard Hughes Medical Institute and New York University, 251 Mercer Street,
New York, NY 10012, U.S.A.

Introduction

In molecular dynamics (MD) simulations, the Newtonian equations of motion are solved numerically, and a space/time trajectory of the molecular system is obtained [1,2]. Typically, *explicit* integration algorithms are used: new positions and velocities for all atoms are computed in closed form through simple relations involving positions and velocities at previous steps. Standard explicit schemes are simple to formulate and fast to propagate, but they impose a severe restriction on the integration timestep size: Δt must resolve the most rapid vibrational mode [3]. This generally limits Δt to the femtosecond (10^{-15} s) range and the trajectory length to the nanosecond (10^{-9} s) range. This feasible simulation range is still very short relative to motions of significant biological interest.

Implicit integration algorithms [4] are widely used to increase the timestep for multiple-timescale problems where the rapid components of the motion limit numerical stability. However, implicit integrators impart two difficulties. First, the formulation is more complex, and enhanced stability is achieved at the expense of the iterative solution of a large system of nonlinear equations at each timestep. This makes the overall method computationally expensive [5]. Second, implicit integrators can *damp* the rapidly varying part of the solution; this is only suitable for problems where this component is rapidly decaying, with a negligible influence on the solution as time increases, which is not the case for biomolecules (at atomic resolution). Even in implicit symplectic methods, numerical damping can be realized due to a lower kinetic energy at large timesteps [6]. Both aspects (complexity and damping) introduce problems for the physical and computational effectiveness of implicit schemes for biomolecules [7–9]. The resulting trajectories at large timesteps must be carefully assessed by comparison with small-timestep trajectories, experiment, or enhanced-sampling simulations.

Clearly, there are two separate goals for MD simulations at large timesteps. First, if computational time were fixed per step, then certainly a larger timestep would allow the generation of trajectories spanning longer for the same computational effort. In reality, the additional cost associated with a larger integration step can still lead to an overall competitive method if the timestep is sufficiently large. The details are method-dependent. Second, larger-timestep methods can also be useful for the enhanced sampling of configuration space. Standard explicit schemes are generally restricted, even within the nanosecond range, to a relatively small region of the thermally accessible conformation space. Thus, larger-timestep methods, which can be viewed as cruder walks in conformation space, may reveal a larger range of molecular conformations and, possibly, paths of transitions among them. Of course, larger-timestep methods do not automatically lead to enhanced sampling. In practice, they might enhance configurational sampling at the expense of full dynamical detail. Therefore, the method used should be tailored to the target problem and assessed accordingly.

To evaluate current progress in this area, it is worthwhile to trace the history of MD simulations. Since the pioneering work of Rahman [10], MD has become an important tool in many areas of biophysics and biochemistry. In the 1970s, the dynamics of molecular liquids was treated by modeling molecules as rigid rotors in generalized coordinates [11–14]. The Cartesian coordinate representation for all degrees of freedom [15] soon followed with increasing interest in larger molecular systems. In the Cartesian representation, the number of degrees of freedom is increased, but the equations of motion become simpler, allowing the simulation of larger systems. Another advantage was not fully understood until later: In Cartesian coordinates, the Hamiltonian of the system is *separable* – the kinetic energy depends only on the momenta and the potential energy depends only on the coordinates. This separability makes possible the use of the second-order explicit Verlet algorithm [16]. The favorable energy preservation of Verlet has long been known and was more recently explained by its *symplecticness* [17]. (For a comprehensive discussion of symplectic integrators, see Ref. i8.) In the absence of direct experimental data for comparison, the Verlet family of methods has become the 'gold standard' for MD simulations.

With the advent of supercomputers, dynamic simulations of biological macromolecules became possible. In the pioneering work of McCammon, Gelin and Karplus [19], the small-scale motions of the protein BPTI (bovine pancreatic trypsin inhibitor, 58 residues) were followed in the Cartesian coordinates of the heavy atoms ($\sim$500). The hydrogen atoms were excluded, but their effect was incorporated implicitly via effective potentials and adjustments in the masses of the heavy atoms.

Researchers quickly realized that the total feasible length of MD simulations was severely limited by the small timestep required to resolve the bond vibrations. This led to the SHAKE algorithm [15] and the family of multiple-timestep (MTS) methods [20,21], applied to biomolecules in the pioneering work of Grubmuller et al. [22].

By constraining the bond lengths and effectively removing the most rapidly oscillating degrees of freedom, SHAKE enables timesteps two times larger compared to unconstrained methods, at a relatively small additional cost per step. MTS methods exploit the idea that the slowly varying forces can be evaluated less often than the faster components. The contribution of the slower forces to the motion can be incorporated by a Taylor expansion of the force [20,21], interpolation [23], or extrapolation [24] techniques.

System sizes of modeled biomolecules have increased steadily with the progress in computer hardware and the advent of parallel machines. Significantly, the total simulation time has increased far less dramatically [3,25]. Timescales of microseconds and milliseconds are still out of reach for macromolecules, and so the search for novel methodologies of simulating the dynamics of biomolecules continues.

In this chapter, we review current approaches for large-timestep MD and describe the progress in our normal-mode-based technique, LIN (for Langevin/Implicit integration/Normal modes), and a related method termed LN (LIN without the implicit-integration component). The separating framework of LIN solves the Langevin equations of motion in two steps: linearization and correction. The linearized equations of motion are solved numerically by an iterative technique, the cost of which is dominated by sparse-Hessian/vector products; the resulting 'harmonic' solution is then corrected by an implicit integration step, which requires minimization of a nonlinear function. LN includes LIN's linearization, but not correction, step and emerges as more competitive in terms of CPU time. We show here through applications to the model systems of alanine dipeptide and BPTI that LIN and LN become competitive methods in comparison with traditional Verlet-like algorithms, giving similar results and a computational gain, even for small systems (e.g., a dipeptide of 22 atoms).

In the next section, we briefly summarize, for a perspective, existing approaches for increasing the timestep. A description of the basic LIN framework follows, with recent algorithmic advances to improve energetic fluctuations and reduce the computational time detailed. Computational efficiency is achieved by sparse-matrix techniques, adaptive timestep selection, and fast minimization. Simulation results for alanine dipeptide and BPTI are then presented, showing good agreement with explicit-scheme simulations at 0.5 fs timesteps with respect to energetic and geometric behavior (angular distributions, rms deviations, etc.). The range of validity of the harmonic approximation is also discussed, and the performance of LN is presented. For BPTI, we demonstrate a speedup factor of 1.4 for LIN at $\Delta t = 15$ fs, and a factor of 4.38 for LN at $\Delta t = 5$ fs, in comparison with explicit-Langevin integration at $\Delta t = 0.5$ fs. Already for the dipeptide, LIN at $\Delta t = 30$ fs gives a speedup of 1.3, and LN at $\Delta t = 5$ fs gives a factor of 2.1. These speedups for small systems contrast typical results of MTS methods, which only become more competitive as the relative number of long-range (soft) forces increases. LN, in particular, is simple to implement in general packages and should yield greater speedup for larger systems. This unexpected windfall in computational performance illustrates the value of developing novel approaches (e.g., based on normal modes) that might initially appear to be not practical for macromolecules. We conclude with a brief summary and discussion of the future applications of LIN and LN.

Approaches for increasing the timestep

Methods for increasing the timestep in MD can be divided into two general types: (i) constrained and reduced-variable formulations, and (ii) separating frameworks (for the system, potential, etc.). In the first category, various SHAKE-like methods are included, as well as techniques for MD in torsion space. In the second category, we consider MTS methods and reference-system methods for splitting the equations of motion, as well as novel approaches for modeling biomolecules.

Constrained and reduced-variable formulations

In the various SHAKE-like methods [15,26–29], the equations of motion in Cartesian space are augmented by algebraic constraints via the formalism of Lagrange multipliers. In this way, the fastest degrees of freedom are frozen at their equilibrium values. Since the Hamiltonian remains separable, symplectic integrators which are explicit in the coordinates but implicit in the constraints can be used [30]. Recent advances in the mathematical treatment of the nonlinear systems arising in SHAKE make the method quite efficient [31].

The related reduced-variable formulations attempt to eliminate the fast degrees of freedom by modeling the system in a generalized coordinate system [32–34]. In torsion-angle dynamics [35,36], the polypeptide chain is treated as a chain of rigid bodies using a recursive rigid-body formulation [37]. Unfortunately, the reduced-variable formulations used in this class of methods destroy the separability of the Hamiltonian. Symplectic integrators like Verlet are *implicit* when applied to these models [18]. Therefore, in general, explicit nonsymplectic methods are used to propagate the dynamics in these approaches. This often leads to a drift in energy, especially at large timesteps [18]. In addition, due to constrained bond lengths and angles, the effective potential in torsion-angle dynamics is different from the original. For all these reasons, internal-variable dynamics is best suited for configurational searches (e.g., for structure refinement) at high temperatures where the interconfigurational barriers are effectively lowered. Recent results reinforce this [35,36].

Separating frameworks

MTS approaches for updating the slow and fast forces [20,22,38] form the first prototype of separating frameworks. These methods certainly provide speedup for systems with a clear division of timescales. For hydrocarbon systems such as fullerenes, the speedup is impressive (e.g., factors of 20–40) [39]. The speedup factor for biomolecules (e.g., 4) [40,41], however, is limited because such a clear division of timescales is lacking and the intramolecular coupling of modes is strong.

In reference-system methods, a subset of the forces or a suitable approximation to the full force is selected for which the solution is more easily obtainable, either analytically or numerically. Examples in this category are NAPA [42] and LIN [43,44]. These methods assume that the correction to the motion – due to the complementary forces – can be obtained with a much larger timestep than that associated with the reference system. A splitting of the forces into linear and nonlinear parts is the premise of LIN [43,44], described in detail below.

Table 1 *Assessment of some MD algorithms: Constrained dynamics (SHAKE and RATTLE), reduced-variable molecular dynamics (RVMD), multiple-timestep methods (MTS), LIN, and MOLDYN*

Method	Advantages	Disadvantages	Typical overall speedup
SHAKE[a] RATTLE[b]	A doubling of feasible timestep (1 → 2 fs) is possible when bonds are constrained	Angles cannot be constrained without affecting dynamics or convergence rate	~2
RVMD[c]	RVMD is useful for enhanced sampling, structure refinement, or global optimization	Overall motion is affected due to altered potential, especially increased barriers; only increased temperature or modification of potential parameters can counteract this effect	NE[d]
MTS[e]	Speedup can be achieved Vibrational spectrum can be more accurate in the high-frequency region in comparison to standard MD schemes	More frequent evaluation of the hard forces than in standard MD might be necessary (e.g., 0.25 fs) Significant speedup is achieved as the relative number of soft forces increases	4–5 [40, 46][f], 2–4 for BPTI [41]
LIN and LN[g]	Speedup can be achieved, modest for LIN, more substantial for LN These two general approaches are effective for systems without clear separation of timescales Very good agreement with small-timestep methods for LIN up to 15 fs and LN up to 5 fs has been demonstrated	Langevin approach is necessary for stability (energy drift without a heat bath) Implicit step in LIN (but not LN) is costly, due to minimization	1.4 (LIN) 4.4 (LN) for BPTI[h]
MOLDYN[i]	Large overall speedup might be obtained	Flexible substructure description is limited to propagation of a linearized, constrained system Assignment of substructures is system-dependent	NE[d]

[a] Reference 15.

[b] Reference 28.

[c] References 33–36.

[d] Not yet established; see text.

[e] References 22, 40 and 41.

[f] Speedup for MTS, LIN, and LN is given with respect to explicit schemes at 0.5 fs timesteps.

[g] Reference 43 and this paper.

[h] See this paper; greater speedups are expected for larger systems.

[i] Reference 45.

The recent MOLDYN substructuring approach of Turner et al. [45] applies multibody dynamics to molecular systems by considering a collection of rigid and flexible bodies. The motion of the atoms within these bodies is propagated via their normal-mode components, of which only the lowest frequency modes are included. The dynamics between bodies is modeled rigorously. Large overall computational gains might be possible because the number of variables is dramatically reduced (by modeling the system as a collection of large flexible bodies), and larger timesteps can be used for the flexible substructures (since the fast oscillations are absent). However, like all the novel methods above, the resulting trajectories must be carefully assessed through a comparison with all-atom, small-timestep trajectories, or experiment. It is expected that the selection of substructures and associated timesteps will influence the resulting motions significantly.

Table 1 summarizes the advantages and disadvantages of the above methods, together with effective speedup, as compared to all-atom explicit simulation. It appears that the well-known SHAKE and MTS methods certainly provide speedup at present, but separating frameworks like LIN and LN are emerging as competitive methods as well, with LN giving speedup already for small systems and both methods having the additional potential for enhanced sampling. The speedup factor of 2 for constrained dynamics usually refers to the timestep increase from 1 to 2 fs. Note that the performance of MTS schemes depends on the subdivision of forces into classes and the associated timestep combination used in each implementation (e.g., 0.25 [46] or 0.5 fs [41] for the rapid components). The speedup factors for LIN and LN are compared with 0.5 fs explicit simulations, following the same comparisons used in MTS methods [40,41]. For LN the factor of 4.4 applies to BPTI (904 atoms), and greater speedups are expected for larger systems.

It is also interesting to note that the introduction of fast multipole methods for computing the electrostatic energy *increases* the overall speedup in relation to a direct electrostatic treatment with no cutoffs, but *decreases* the relative speedup of larger-timestep methods. Although this will only be significant for very large systems, the trend can be inferred from the recent data of Zhou and Berne [46] for a 9513-atom protein: a speedup factor of 4.5 for RESPA alone compared to about 4 for the MTS method when fast multipole methods were introduced into both Verlet and RESPA. Such behavior might be relevant to other methods as well.

The LIN algorithm

Our algorithm LIN consists of linearization and correction steps, and thus combines, in theory, normal-mode (NM) techniques with implicit integration [43,44]. Let us write the collective position vector of the system as $X(t) = X_h(t) + Z(t)$. The first part of LIN solves the linearized Langevin equation for the 'harmonic' component of the motion, $X_h(t)$. The second part relies on implicit integration to compute the residual component, $Z(t)$, with a large timestep.

To describe the process formally, we start from the continuous form of the Langevin equation (in its simplest form):

$$\begin{aligned} \mathbf{M}\dot{\mathrm{V}}(t) &= -\nabla E(X(t)) - \gamma \mathbf{M} V(t) + R(t) \\ \dot{X}(t) &= V(t) \end{aligned} \tag{1}$$

The overdots denote differentiation with respect to time, V is the vector of collective velocities, **M** is the diagonal mass matrix, $\nabla E(X(t))$ denotes the gradient vector of the potential energy E, and γ is the collision parameter. The random-force vector, R, is a stationary, Gaussian process with statistical properties (mean and covariance matrix) given by

$$\langle R(t) \rangle = 0, \qquad \langle R(t) R(t')^T \rangle = 2\gamma k_B T \mathbf{M} \delta(t - t') \tag{2}$$

where k_B is Boltzmann's constant and δ is the usual Dirac symbol.

With a linear approximation to $\nabla E(X(t))$ at some reference position X_r, the system of equations for the 'harmonic' components X_h and V_h is given by

$$\begin{aligned} \mathbf{M}\dot{V}_h &= -\nabla E(X_r) - \mathbf{H}_h(X_h - X_r) - \gamma \mathbf{M} V_h + R \\ \dot{X}_h &= V_h \end{aligned} \tag{3}$$

Here $\mathbf{H}_h$ is the Hessian matrix of E at X_r, but below we discuss an approximation to $\mathbf{H}_h$, resulting from cutoffs, that is cheaper to use. System (3) can be solved by standard NM techniques [47–49]. This involves the determination of an orthogonal transformation matrix **T** that diagonalizes the mass-weighted Hessian matrix $\mathbf{H}' = \mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-1/2}$, namely

$$\mathbf{D} = \mathbf{T}\mathbf{H}'\mathbf{T}^{-1} \tag{4}$$

Eigenvalues of the diagonal matrix **D** will be denoted as λ_i. With the transformations

$$Q = \mathbf{T}\mathbf{M}^{1/2}(X - X_r) \quad \text{and} \quad F = \mathbf{T}\mathbf{M}^{-1/2}R \tag{5}$$

applied to the NM-displacement coordinates Q and random force F, system (3) is reduced to the set of decoupled, scalar differential equations

$$\begin{aligned} \dot{V}_q &= -\mathbf{D}Q - \gamma V_q + F \\ \dot{Q} &= V_q \end{aligned} \tag{6}$$

Here, the force F is a linear combination of the components of R; it also has a Gaussian distribution and autocorrelation matrix that satisfies the same properties of R(t) as shown in Eq. 2, with **I** (the $n \times n$ unit matrix) replacing **M** [43]. The initial conditions coincide with those for the original equations: $X_h(0) = X^n$ and $V_h(0) = V^n$, where the superscript n refers to the difference-equation approximations to solutions at time $n\Delta t$. The reference point, X_r, may be chosen either as the configuration of the last step, X^n, or a minimum of E near X^n (we use the former). Appropriate treatments, as discussed in Ref. 44, are essential for the random force at large timesteps ($\delta(t - t') \rightarrow \delta_{nm}/\Delta t$) to maintain thermal equilibrium. Thus, the above equations can be solved analytically for all the Q_i and associated velocities V_{qi} [43].

Once $X_h(t)$ is obtained as a solution to system (3), the residual motion component, Z(t), can be determined by solving the following set of equations [43]:

$$\begin{aligned} \mathbf{M}\dot{W}(t) &= -\nabla E(X_h + Z(t)) - \gamma \mathbf{M} W(t) + \nabla E(X_r) + \mathbf{H}_h(X_h - X_r) \\ \dot{Z}(t) &= W(t) \end{aligned} \tag{7}$$

Here W denotes the time derivative of Z, and the initial conditions for system (7) are Z(0) = 0 and W(0) = 0.

The use of the implicit-Euler scheme to discretize system (7), for example in Refs. 43 and 44, entails solution of a system of nonlinear equations, namely $\nabla\Phi(Z) = 0$, at each timestep. Following Ref. 50, this solution can be found by minimization of the 'dynamics' function Φ. Reformulating Φ in terms of X(t) rather than Z(t), we obtain

$$\Phi(X) = \tfrac{1}{2}(1 + \gamma\Delta t)(X - X_0^n)^T \mathbf{M}(X - X_0^n) + (\Delta t)^2 E(X) \tag{8}$$

where

$$X_0^n = X_h^{n+1} + \frac{(\Delta t)^2}{1 + \gamma\Delta t}\mathbf{M}^{-1}[\nabla E(X_r) + \mathbf{H}_h(X_h^{n+1} - X_r)] \tag{9}$$

Assuming that the solution X_h of the linearized system at step n + 1 is a good approximation to X, this minimization proceeds rapidly since X_h^{n+1} provides a good starting point. Furthermore, a truncated-Newton method that exploits Hessian sparsity can accelerate convergence significantly and incorporate second-derivative information [51–53]. Once X^{n+1} is found, V^{n+1} can be obtained by setting

$$V^{n+1} = V_h^{n+1} + \frac{X^{n+1} - X_h^{n+1}}{\Delta t} \tag{10}$$

The solution vectors $\{X^{n+1}, V^{n+1}\}$ provide the initial conditions for the harmonic phase of LIN at the next Δt interval.

Applications of LIN to model systems, namely butane [43] and the nucleic-acid component deoxycytidine [44], demonstrated stability at large timesteps, with activation of the high-frequency modes. The latter work also developed an appropriate treatment for the Langevin random force at large timesteps: the positional and velocity distributions were derived analytically for the decoupled oscillators on the basis of the corresponding Fokker–Planck equation [44]. However, two limitations of LIN emerged in the above studies. First, energetic fluctuations increased in LIN with the timestep, and thus LIN at large timesteps resembles more of a sampling tool than continuous dynamics. Second, computational costs are large due to the analytic normal-mode component. The precise computational cost depends on the approximation used for the Hessian in the linearized equations of motion, but certainly the $O(n^3)$ cost for a dense n × n system is *prohibitive* for macromolecules.

The work described in this contribution addresses both these issues and shows that competitiveness of LIN can be achieved at moderate timesteps and also good agreement with small-timestep dynamics. The focus on accuracy and competitiveness also leads to our new variant LN, which is far cheaper and stable at moderate timesteps.

Recent LIN progress

To accelerate LIN computations, we first developed a *numerical* approach for solving the linearized equations of motion in lieu of the analytic normal-mode procedure. This makes the first part of LIN quite cheap and the method with no correction step, LN, very competitive (see below), especially with the incorporation of efficient sparse-Hessian/vector products. To stabilize energetic fluctuations, we also replaced the implicit-Euler integrator in the second part of LIN by the symplectic implicit-midpoint (IM) scheme [54]. This substitution was also found to reduce the work in the second part of LIN (minimization). To further optimize performance, we devised an adaptive-timestep procedure to allow large timesteps when possible and force small timesteps when necessary. These components are now discussed in turn.

Numerical integration of the linearized equations

System (3) could be solved *numerically*, with timesteps $\Delta\tau \ll \Delta t$, rather than *analytically* as previously proposed [44]. The 'inner' timestep, $\Delta\tau$, required for this numerical integration is the same as for traditional MD (e.g., 0.5 or 1 fs), but *each step is cheaper than in standard* MD because updates of the energy gradient are not required at every step. After one $\mathbf{H}_h$ is evaluated (or approximated) for each *outer* LIN step, the cost of an inner integration step is dominated by *matrix/vector products* (see below). In addition, this numerical solution of the linearized equations eliminates the problem associated with the large-timestep discretization of the random forces [44] since the random force is updated every $\Delta\tau$ substep. With this new treatment of the linearized equations, LIN becomes the first multiple-timestep method to utilize implicit integration methods.

The stability of the linearized equations is assured if all vibrational modes have positive eigenvalues λ_j (corresponding to solutions $\exp(-i\lambda_j^{1/2}t)$ where $i = \sqrt{-1}$). For $\lambda < 0$, solutions diverge over large time intervals. Negative eigenvalues are generally present, but for reasonable choices of timesteps Δt and $\Delta\tau$, these instabilities appear to be mild and require no special treatment. Still, it is possible to determine, or approximate, the negative eigenvalues and the corresponding eigenvectors (e.g., by Lanczos-based techniques), *project out these imaginary frequencies*, and then solve Eq. 3 by numerical integration. In the Appendix we outline this projection method, though we did not have to resort to it.

For the explicit integration process above, we use the *second-order partitioned Runge–Kutta method* ('Lobatto IIIa,b') [55], which reduces to the velocity Verlet method when $\gamma = 0$. This yields the following iteration process for $\{X_h^{n+1}, V_h^{n+1}\}$ from initial conditions $X_h(0) = X^n$, $V_h(0) = V^n$:

$$\begin{aligned}
V_h^{i+1/2} &= V_h^i + \frac{\Delta\tau}{2}\mathbf{M}^{-1}[-\nabla E(X_r) - \mathbf{H}_h(X_h^i - X_r) - \gamma \mathbf{M} V_h^{i+1/2} + R] \\
X_h^{i+1} &= X_h^i + \Delta\tau V_h^{i+1/2} \qquad\qquad (11) \\
V_h^{i+1} &= V_h^{i+1/2} + \frac{\Delta\tau}{2}\mathbf{M}^{-1}[-\nabla E(X_r) - \mathbf{H}_h(X_h^{i+1} - X_r) - \gamma \mathbf{M} V_h^{i+1/2} + R]
\end{aligned}$$

The first equation above is implicit for $V_h^{i+1/2}$, but the linear dependency allows solution for $V_h^{i+1/2}$ in closed form. Note also the Hessian/vector products in the first and third equations. For future reference, we divide the first part of LIN (linearization) into Part Ia: $\mathbf{H}_h$ evaluation; and Part Ib: integration.

Implicit-midpoint integration

We now apply the second-order symplectic midpoint scheme [54] to the second part of LIN. Following algebraic manipulations similar to those used in Ref. 43, this implicit discretization reduces to solution of X by minimizing Φ in terms of the new variable $Y = (X + X^n)/2$. Now, instead of Eq. 8, the function Φ takes the form:

$$\Phi(Y) = 2\left(1 + \frac{\gamma\Delta t}{2}\right)(Y - Y_0^n)^T\mathbf{M}(Y - Y_0^n) + (\Delta t)^2E(Y) \tag{12}$$

where

$$Y_0^n = \frac{X_h^{n+1} + X^n}{2} + \frac{(\Delta t)^2}{4(1 + \gamma\Delta t/2)}\mathbf{M}^{-1}\left[\nabla E(X_r) + \mathbf{H}_h\left(\frac{X_h^{n+1} + X^n}{2} - X_r\right)\right] \tag{13}$$

The initial approximate minimizer, Y_0, of Φ can be set to X^n, X_h^{n+1}, or $(X_h^{n+1} + X^n)/2$ (we use the last). The new coordinate and velocity vectors for timestep n + 1 are then obtained from the relations

$$X^{n+1} = 2Y - X^n, \qquad V^{n+1} = V_h^{n+1} + 2(X^{n+1} - X_h^{n+1})/\Delta t \tag{14}$$

It is important to note that even with the symplectic integrator IM, LIN is not time-reversible due to the presence of the linearized forces which are held constant on intervals along the trajectory. Therefore, the forces in the Langevin formulation ($\gamma > 0$) are a stabilizing influence, especially at large timesteps; without these stochastic terms, the total energy will drift. The use of the diffusive regime (large γ) is one way to permit very large timesteps [56], but this is only appropriate to systems where inertial forces are relatively small.

An adaptive-timestep scheme

To further monitor energetic fluctuations, we have developed an adaptive timestep-selection subroutine heuristically. Gibson and Scheraga [33,34] used a more rigorous procedure in their torsion-angle dynamics method. Our basic idea is to resolve more accurately (with smaller timesteps than the input value) regions where significant fluctuations in energy and geometry are realized, and resolve more crudely 'smoother' regions of conformation space, where the harmonic approximation is better. Our experience suggests that large changes in the bond energy signal deterioration of the harmonic approximation [57]. Therefore, we set a certain threshold for the bond energy value for the simulated system (e.g., the mean plus five standard deviations of the bond fluctuation as obtained from a short explicit trajectory) and reduce the timestep (by one-half) if this threshold is exceeded. For subsequent steps, the original timestep is used if possible.

Economical formulation of the explicit subintegration

The cost of the explicit subintegration phase of LIN is dominated by Hessian/vector products. If $\mathbf{H}_h$ in system (3) is sparse, as in the case of cutoffs [53], efficient O(n) multiplications can be devised to treat the nonzeros only. In fact, it is reasonable in our context to employ small cutoffs for the Hessian – to approximate the harmonic motion – but to include all interactions in the correction step (via second-derivative information in the TNPACK minimization of Φ [53]).

We first formulated a sparse $\mathbf{H}_h$ by employing a 12 Å cutoff and then explored smaller cutoff values up to 4.5 Å. This small value is typically sufficient to resolve all 1–4 bonded interactions (torsions). We emphasize that full interactions are included in the correction step of LIN and, certainly, the gradient reflects all interactions. As hoped, the results for BPTI (Table 2) show no deterioration in average energies as the cutoff radius is decreased. For the dipeptide, the trends are even better for LIN at $\Delta t = 30$ fs. For illustration, we show in Fig. 1 the magnitudes of elements in the dense Hessian, $\mathbf{H}_{\text{dense}}$ (no cutoffs employed), and the difference between this matrix and the sparse $\mathbf{H}_h$ (4.5 Å cutoff). This representative pair of Hessians was evaluated in the middle of a LIN trajectory for the dipeptide, at 1.5 ns. Significantly, the entries of the difference matrix ($\mathbf{H}_{\text{dense}} - \mathbf{H}_h$) are 4 orders of magnitude smaller than the dominant entries of the dense Hessian.

The resulting savings from using a sparse matrix in the matrix/vector products are impressive. Namely, a Hessian cutoff of 4.5 Å hastens the matrix/vector product by

Table 2 *Comparison of the LIN sparse-Hessian treatment (with the range in Å indicated in the 'LIN' subscript) in the linearization part with the dense-Hessian treatment ('LIN$_{dense}$') for BPTI*

	LIN$_{12}$	LIN$_8$	LIN$_{4.5}$	LIN$_{\text{dense}}$
E	1653.52	1654.77	1655.22	1654.23
E_k	812.34	811.34	811.51	812.49
E_p	841.18	843.44	843.71	841.73
T	301.47	301.09	301.16	301.52
E_{bond}	339.26	339.82	339.95	339.39
E_{angle}	454.01	455.16	455.55	454.11
$E_{\text{U-B}}$	60.79	60.91	60.95	60.82
E_{tor}	353.62	353.61	353.52	353.76
E_{impr}	30.39	30.49	30.56	30.41
E_{vdW}	– 107.59	– 107.13	– 107.17	– 107.56
E_{elec}	– 1954.27	– 1954.39	– 1954.61	– 1954.15

The energy symbols in the first column are as follows. Total energy: E; kinetic: E_k; potential energy with respect to a local minimum (– 1664.96 kcal/mol) near the initial configuration: E_p; bond: E_{bond}; angle bending: E_{angle}; Urey-Bradley: $E_{\text{U-B}}$; torsional: E_{tor}; improper torsion: E_{impr}; van der Waals: E_{vdW}; and electrostatic: E_{elec}; all in kcal/mol. The temperature T is given in kelvin.

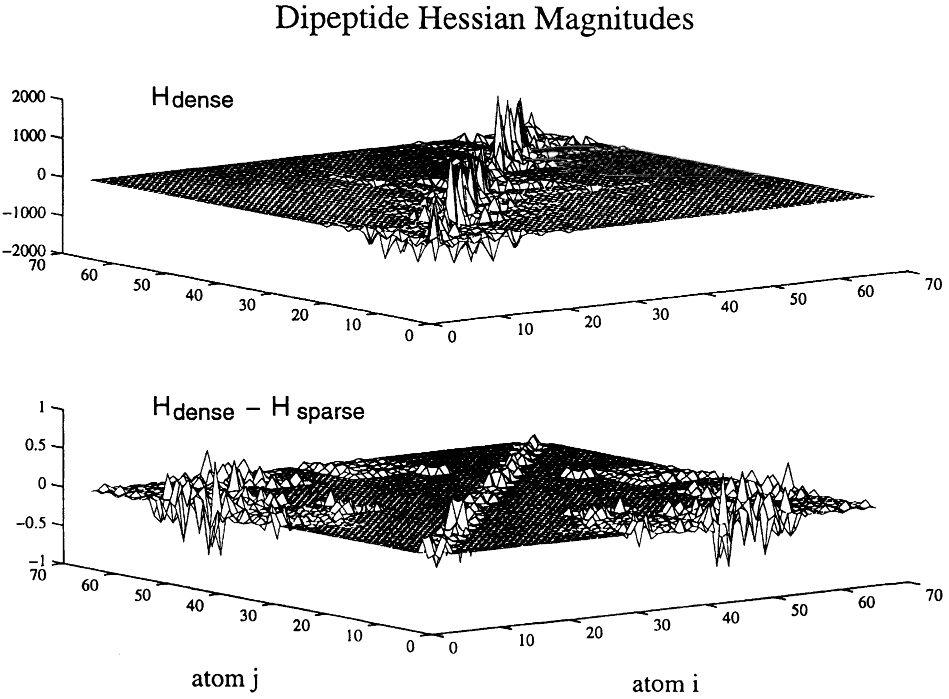


Fig. 1. Mesh plots of dense Hessian (top) and difference between the dense and the sparse Hessian (formed with 4.5 Å cutoff) (bottom) for the dipeptide model. These matrices were evaluated in the middle of the LIN trajectory, at 1.5 ns. Note the difference in scales between the two views. The maximum entry of the difference matrix is approximately 0.62, 4 orders of magnitude smaller than the dominant entries of the dense Hessian.

a factor of 19 compared to the dense Hessian for the BPTI system (2712 variables). It is of course necessary in this case to update the Hessian sparsity pattern periodically – say, every outer LIN timestep – but this updating might be done more efficiently by first computing inter-residue distances and then making atom-by-atom searches only within near residues. The actual cost of evaluating the Hessian is not very time-consuming since an efficient implementation can reuse many temporary variables calculated for the gradient (e.g., roots and powers). In fact, we found computation of the *dense* Hessian to be only a factor of 2.5 more expensive than a gradient evaluation for BPTI when the CHARMM 'slow' routines for gradient evaluations are used. In this estimate, the 'Hessian computation' actually refers to 'gradient plus Hessian computation'. With the 'fast' routines for gradient evaluation, the dense-Hessian (plus gradient) evaluation is 4.0 times more expensive than the gradient calculation alone. This factor is reduced to 2.3 and 2.0 for 12 and 8 Å cutoffs, respectively. With a 4.5 Å Hessian cutoff, the Hessian evaluation is about 1.9 times more expensive than the 'fast' gradient. In the future, it might be possible to implement in CHARMM 'fast' routines for the Hessian as well, in order to reduce these factors further.

For the value of the inner LIN timestep, $\Delta\tau$, we use 0.5 fs. The errors are reasonable in this range [41] ($\Delta\tau$ is roughly 1/20th of the fastest period) and comparisons of LIN with explicit trajectories employ 0.5 fs timesteps also. Note that the value 1 fs can also be used in both cases. Then the cost of LIN's explicit subintegration (Part Ib), as well as that of the explicit trajectory, will be reduced by one-half, but the cost of the Hessian evaluation (Part Ia) and minimization (Part II) in LIN will stay about the same, making LIN less competitive.

Comparisons and computational speedup of MTS methods are also typically reported at 0.5 fs [41,46]. The cost of Verlet will decrease by one-half with double the timestep (1 fs) when compared to MTS methods, but the performance of the RESPA scheme will depend on the timestep combination used for different classes of interactions. See Refs. 41 and 46 for recent examples.

Economical minimization

Part II of LIN, the correction step, entails numerical optimization of the nonlinear dynamics function (Eq. 12) with the truncated-Newton package TNPACK [52,58]. For the LIN simulation of BPTI, we found that the minimization subproblem is most efficiently solved using TNPACK with the *preconditioned* conjugate gradient option and the finite-differencing option for calculation of the Hessian/vector products. This implementation by a simple backward-difference scheme entails one additional gradient evaluation per conjugate gradient step [53,58], included in the counts given below. Our preconditioner is formulated from the second derivatives of the bond-length, bond-angle, dihedral-angle, improper torsion-angle and 1–4 nonbonded terms. The cost of forming and factoring this preconditioner matrix was insignificant compared to force computations due to the optimized sparse components of TNPACK, and this work certainly accelerates convergence. For BPTI, for example, approximately 11.5 gradient calculations were required per 12 fs timestep (8.5 of which were required on average for the finite-difference product). For the 15 fs timestep, an average of 14.1 gradient calculations were required per step. The counts given above were the result of more lenient minimization stopping criteria than the TNPACK default values for the final gradient norm and residual vectors [58], namely $\|g\| < 10^{-4}$ and $\|r\| < 10^{-1}$.

Simulation results and analysis

With the improvements described above, the LIN algorithm was tested on two model systems: alanine dipeptide (N-acetyl alanyl N′-methyl amide, a blocked residue of alanine with 22 atoms) and BPTI (58 residues, 904 atoms). Calculations were performed with CHARMM version 24b1, modified to include our integration and minimization algorithms [53], with the all-atom representation and parameter set 22. We used the unit dielectric constant and included *all* nonbonded interactions in the governing model. The bath temperature was set to T = 300 K, and the Langevin collision parameter was fixed at $\gamma = 50\ \text{ps}^{-1}$ [43]. For a fair comparison, the same starting position, velocity vector, and sequence of random numbers were used in the trajectories. The initial position vector for Φ minimization was chosen as $Y_0 = (X^n + X_h^{n+1})/2$. This choice generally leads to 4–8 minimization steps per substep for the dipeptide with timesteps of $\Delta t = 30$ fs, and 3–5 for BPTI with timesteps of $\Delta t = 15$ fs. All simulations were performed in serial mode on a 150 MHz R4400 SGI Indigo2 workstation.

Alanine dipeptide

We start by comparing LIN results at $\Delta t = 30$ fs with those obtained by the explicit Verlet-like Langevin integrator in CHARMM, BBK [59], at $\Delta t = 0.5$ fs. Data were collected over 3 ns, following 160 ps of equilibration, and trajectory snapshots were recorded every 120 fs. With the LIN timestep of 30 fs, only 6% of the steps are rejected (i.e., exceed the bond energy threshold of 15 kcal/mol).

In Table 3, the averages and variances of the energy components (total, kinetic, and potential) and the time-averaged properties of some internal variables are given. The results obtained with both methods are very similar. This is especially good for Langevin simulations, where *no exact trajectory exists* and representatives are

Table 3 *Averages (mean) and fluctuations (variance) for alanine dipeptide from LIN ($\Delta t = 30$ fs) and LN ($\Delta t = 5$ fs) versus explicit ($\Delta t = 0.5$ fs) Langevin trajectories over 3 ns (see the footnote to Table 2); the E_p value is given with respect to the minimum of -15.85 kcal/mol*

	Explicit		LIN		LN	
	Mean	Variance	Mean	Variance	Mean	Variance
E^a	37.9	4.8	39.6	5.4	37.8	4.8
E_k	19.7	3.4	20.1	3.6	19.6	3.4
E_p	18.3	3.3	19.5	3.8	18.2	3.3
ϕ^b	−109.1	31.2	−108.7	30.6	−111.5	31.4
ψ	119.0	47.1	118.7	47.0	123.2	47.1
$r_{C\text{-}N}{}^c$	1.340	0.028	1.340	0.028	1.339	0.028
$r_{N\text{-}C_\alpha}$	1.449	0.030	1.448	0.030	1.448	0.030
$r_{C_\alpha\text{-}C}$	1.527	0.034	1.526	0.034	1.526	0.034
$\theta_{C\text{-}N\text{-}C_\alpha}$	123.3	3.3	123.1	3.3	123.2	3.3
$\theta_{N\text{-}C_\alpha\text{-}C}$	110.6	4.2	110.6	4.2	110.5	4.2
$\theta_{C_\alpha\text{-}C\text{-}N}$	117.0	2.8	117.1	2.8	117.1	2.8

[a] Energy in kcal/mol.

[b] Angles in degrees.

[c] Bond lengths in Å.

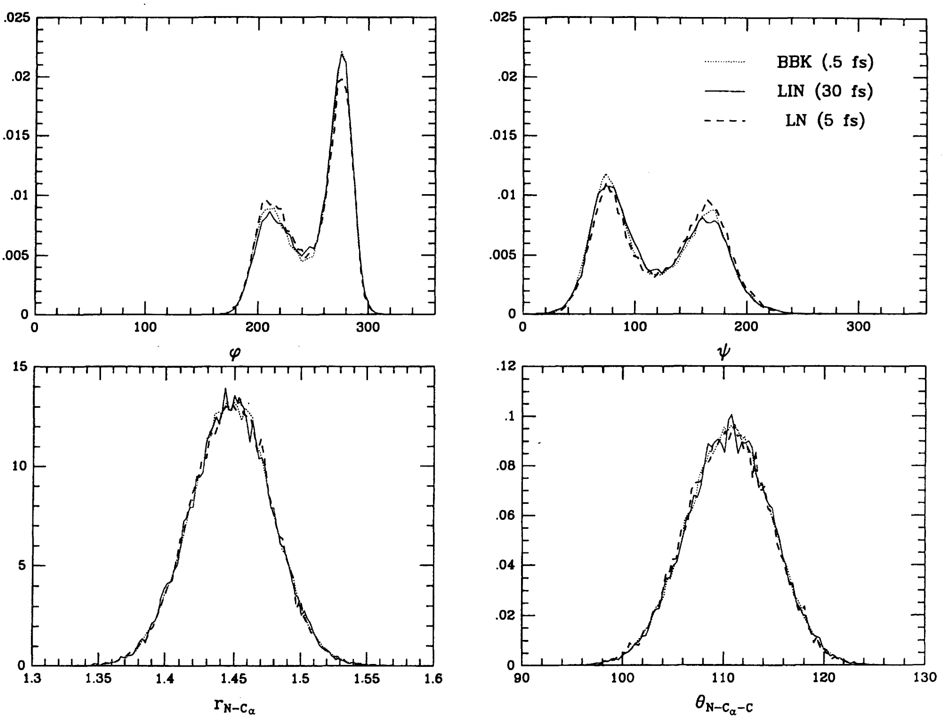


Fig. 2. Distributions from alanine dipeptide trajectories over 3 ns for one selected bond length and one bond angle, and the two dihedral angles, as obtained by LIN, $\Delta t = 30 fs$ (solid line) and LN, $\Delta t = 5 fs$ (dashed) versus explicit Langevin, $\Delta t = 0.5 fs$ (dotted).

sought [6]. The energetic fluctuations are slightly greater for LIN, due to increased energy fluctuations in the X-H bonds (where X is any non-hydrogen atom), but the global behavior is very similar.

This good agreement can be seen from Fig. 2, which compares the ensemble-generated distributions for a representative bond length ($r_{C_\alpha\text{-}N}$), a representative bond angle ($\theta_{N\text{-}C_\alpha\text{-}C}$), and the two dihedral angles (φ, ψ). We note a remarkable similarity between the LIN distributions and those of the explicit trajectory. The matching of φ and ψ distributions, in particular, indicates that the overall motion is essentially the same. For the main motion of interest here, dihedral angles, the variances from both simulations are about 31° and 47° for φ and ψ, respectively (Table 3), quite satisfactory considering that the LIN timestep is *60 times larger* and only about one-sixth the period of the dihedral-angle motion. These results suggest that at moderate timesteps we can obtain with LIN very similar trajectories to traditional MD with much smaller timesteps. The results of LN, also shown in Table 3 and Fig. 2, are discussed separately.

BPTI

As initial coordinates for BPTI, we use those described in Ref. 31, with four internal water molecules. Due to a high rejection rate for $\Delta t = 30$ fs (over 50% due to large bond-energy fluctuations), the timestep was decreased to $\Delta t = 15$ fs, where only 3% of the steps were rejected over 12 ps. With this choice of timestep, minimization proceeds rapidly (4–5 steps). While for the dipeptide we successfully used $\Delta t = 30$ fs, the more rugged potential-energy landscape for this larger, dense system appears to make the behavior more sensitive. That is, the harmonic approximation is poorer and its validity is more short-lived than for the dipeptide. Many more minima and saddle points exist for the larger system. Consequently, with too large a timestep, the correction step (minimization) can produce a distant minimum, leading to discontinuity of the smooth dynamics trajectory. The step control mechanism in our implementation of LIN is thus critical for avoiding such undesired behavior in the context of continuous dynamics (though for sampling it may be desired). It is also interesting to note that smaller timesteps for larger systems were found necessary in torsion-angle dynamics [33] and MTS methods [41].

Table 4 shows the various energy-component averages and variances obtained with BBK at 0.5 fs (the first two data columns) versus LIN at 15 fs (the central two columns). The overall energetic behavior of LIN is quite similar to the explicit trajectory, as reflected by the variance values and the small relative differences in energy. The difference in total energy is only 1%, with the largest fluctuations exhibited by the bond energy (5%).

To examine the geometric behavior, the root-mean-square (rms) fluctuations for various quantities are shown in Fig. 3: (a) total rms from the starting structure;

Table 4 *Comparison of averages and variances of various energy components and the temperature (see the footnote to Table 2) for BPTI simulations over 12 ps, LIN at $\Delta t = 15$ fs and LN ($\Delta t = 5$ fs) versus BBK with $\Delta t = 0.5$ fs*

	Explicit		LIN		LN	
	Mean	Variance	Mean	Variance	Mean	Variance
E	1632.2	30.4	1653.5	33.6	1623.7	30.8
E_k	808.3	21.8	812.3	22.1	805.7	21.9
E_p	824.0	21.9	841.2	24.7	818.0	22.1
T	300.0	8.1	301.5	8.2	299.0	8.1
E_{bond}	322.0	14.5	339.3	15.8	327.1	14.7
E_{angle}	454.3	15.4	454.0	15.7	449.3	15.3
$E_{U\text{-}B}$	60.5	3.7	60.8	3.8	59.8	3.6
E_{tor}	354.3	8.3	353.6	8.4	350.8	8.3
E_{impr}	30.4	3.8	30.4	3.9	29.8	3.8
E_{vdW}	−118.0	13.3	−107.6	13.7	−115.0	13.4
E_{elec}	−1944.5	15.2	−1954.3	15.7	−1948.9	15.0

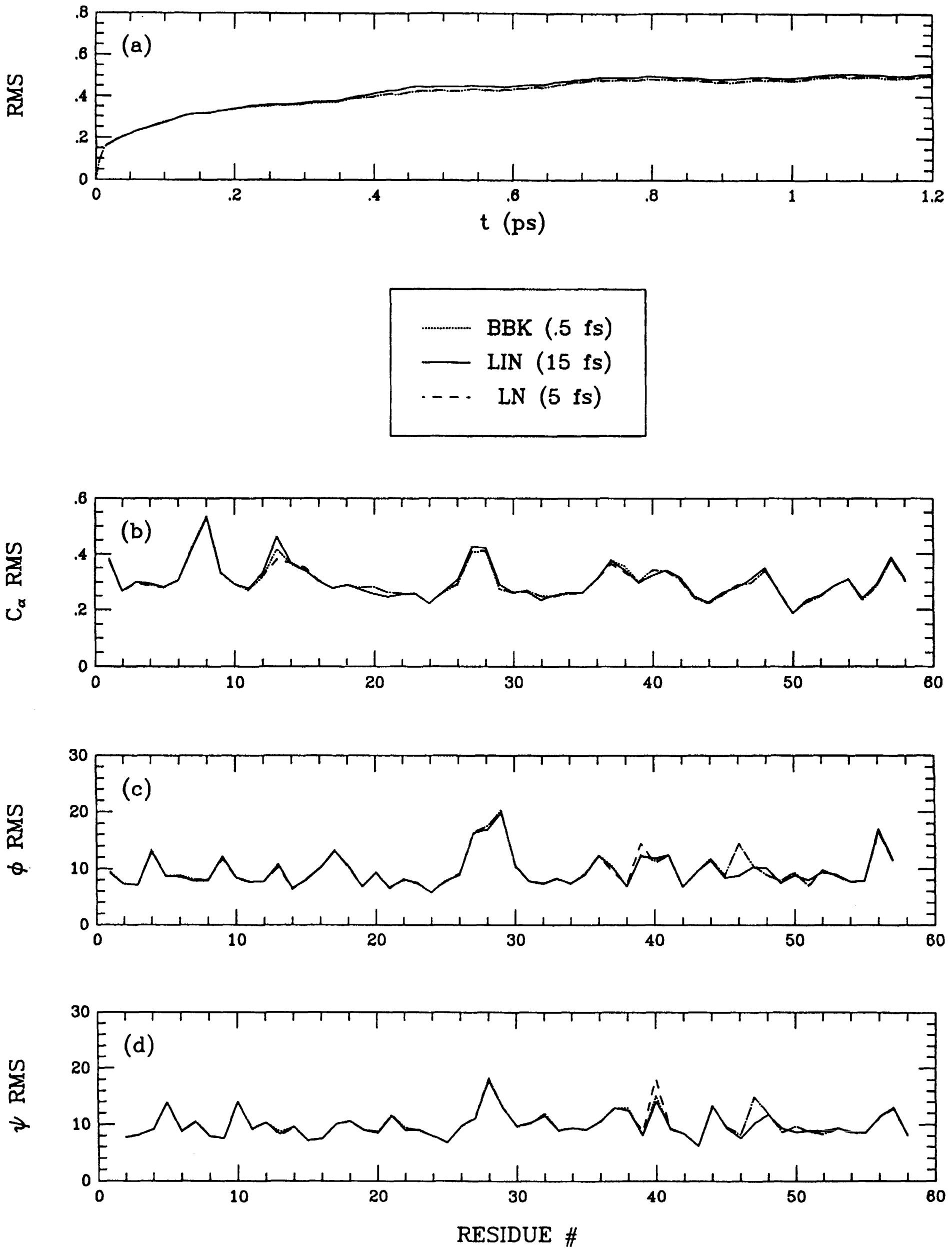


Fig. 3. Root-mean-square (rms) fluctuations for the BPTI simulations over 12 ps, with the solid line corresponding to LIN, $\Delta t = 15\,fs$, the dashed line to LN, $\Delta t = 5\,fs$, and the dotted line to the explicit trajectory, $\Delta t = 0.5\,fs$: (a) total rms from the starting point, in Å; (b) rms fluctuations of each C_α atom; (c) rms fluctuations (deg) of each ϕ angle along the backbone; (d) rms fluctuations (deg) of each ψ angle along the backbone.

(b) rms of the 58 C_α atoms of BPTI; and (c, d) rms fluctuations of the φ and ψ angles along the protein backbone. The agreement again is very good, suggesting the same mobility and a similar global pathway.

Validity of the harmonic approximation

The limit on the timestep value in LIN, when accurate reproduction of small-timestep dynamics is the goal, stems from the limited validity of the harmonic approximation. That is, anharmonicity limits the interval over which the linearized equations of motion provide a reasonable approximation to the nonlinear model. This deviation from linearity is expected to be both system and configuration dependent. In practice, we observe larger bond-length energy fluctuations and a greater computational effort in the minimization component of LIN as the timestep is increased. That is, when X_h is a poor approximation to X, the residual, Z, is 'large' in our context.

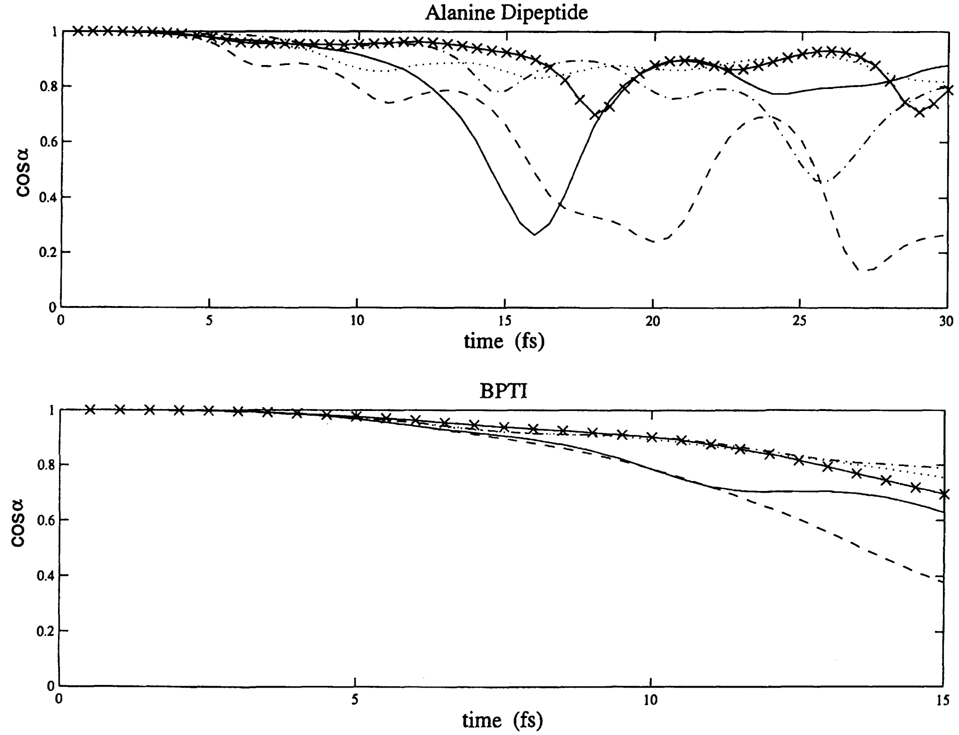
To assess the harmonic model, we plot in Fig. 4 the quantity cos α in time, where α is the angle defined between the two force vectors g(X) and $g(X_r) + \mathbf{H}_h(X - X_r)$. To obtain these views, we chose five different 15 fs intervals along the BPTI trajectory (30 fs intervals for the dipeptide), one of which (the dashed curve) was associated with a rejected step (due to high bond energies). We then computed the quantity cos α every 0.5 fs, where X is the harmonic approximation in LIN, X_r is the reference point X^n, and $\mathbf{H}_h$ is the sparse-Hessian approximation used in LIN's first part (Eq. 11). Note the different scale in the time axes for the two systems.

We see from these curves that in all the cases the harmonic approximation is very good up to 5 fs (α is very small), deteriorating in some cases after 15 fs. Thus, while at some configurational regions the harmonic approximation is quite good over the entire interval, at others very large deviations can be observed after 15 fs. This behavior justifies the use of an adaptive timestep and suggests the usefulness of this angular quantity for the analysis of trajectory behavior.

The principal difference between the dipeptide and BPTI figures is the comparative smoothness of the latter. Data show that, for the dipeptide, the quantity α can be dominated by a single oscillation. This is not the case with the larger BPTI system. Therefore, a criterion based on cos α to assess the validity of the harmonic approximation may be less useful for larger systems.

The new variant LN

To further analyze the validity range of the harmonic approximation, the degree of correction in the second part of LIN can be measured by evaluating the performance of a related method termed 'LN' (Langevin/Normal modes), which propagates the linearized Langevin Eqs. 2 and 3 with the discretization Eq. 11, as in the first part of LIN, but omits the correction phase (i.e., $Z = 0$). (An LN-type method was proposed in 1995 for molecular, not Langevin, dynamics [60].) In our implementation of LN, the adaptive-timestep selection used in LIN is not employed.


Alanine Dipeptide

BPTI

Fig. 4. The cosine of the angle in conformation space between the linearized force vector and the systematic force vector evaluated at the same point. This quantity is used to assess the validity of linearized forces along a trajectory in Part Ib of LIN. Five 30 fs trajectories of alanine dipeptide and five 15 fs trajectories of BPTI with various starting conditions are shown (see the text). The dashed lines correspond to rejected LIN steps. In all the cases, the directions of the linearized forces adequately approximate those of the systematic forces for at least 5 fs.

We first ran several exploratory simulations for BPTI with LN and LIN for the harmonic-model assessment with our sparse $\mathbf{H}_h$. We found that stable LN trajectories could be obtained only up to 10 fs timesteps. This certainly validates the need for the correction phase in LIN. At large timesteps, the correction step is crucial for producing a close trajectory to the explicit method, as illustrated in the following subsections.

By examining the trends in the various energy components before (i.e., by LN) and after (LIN) the residual correction phase for $\Delta t = 2$, 5, and 10 fs (to be detailed in Ref. 57), we found that corrections are small for 2 and 5 fs, but much larger for 10 fs. In particular, the bond-length and bond-angle energy terms reveal the largest deviations, followed by the van der Waals terms. Thus, anharmonic effects are stronger for the high-frequency terms, and our timestep criterion based on the bond-energy fluctuations can be justified by this view.

This good agreement between the LIN and LN energies up to timesteps of 5 fs immediately suggested to us that LN may be a competitive method at moderate timesteps! Indeed, with the sparse matrix/vector products, LIN's first part becomes very inexpensive compared with the second part, minimization.

To explore this intriguing possibility of using LN as a dynamics propagation scheme, we ran LN for the dipeptide and BPTI at $\Delta t = 5$ fs. The results are displayed and compared with those for LIN and BBK in Figs. 2 and 3 and Tables 3 and 4. The pattern used for LN is the dashed line.

We see from these views that the LN results are in very good agreement with those from the explicit trajectory at 0.5 fs timesteps, as well as the LIN trajectories at 15 fs (BPTI) and 30 fs (dipeptide). In fact, LN energies at the timesteps examined are, overall, in better agreement with those of BBK than LIN. Only in Fig. 3 some slight differences in behavior of the two residues can be seen in the rms plots (parts b and c). The residues near 39 and 46 are located on the protein exterior, and hence are probably more flexible. However, a difference of practical importance between LN and LIN is the much faster performance of LN. These timings are discussed next.

Computational performance

Table 5 shows the computational performance of LIN, LN, and BBK for the dipeptide and BPTI applications. Shorter simulations were used here than for the production runs. Computational speedup from our large-timestep simulations already emerges for the dipeptide. LIN shows a factor of 1.3 speedup for the dipeptide when $\Delta t = 30$ fs, and 1.4 for BPTI when $\Delta t = 15$ fs in comparison to BBK. The table shows that LIN's Part Ib (integration) is inexpensive compared to Part II (5% of the total CPU time for BPTI). This explains why LN, which has no minimization component, is very competitive. Note, however, that in LN the Hessian must be updated every 5 fs in our implementation so that the total cost of LN depends on the performance of Parts Ib and Ia (**H** updating). Indeed, already for the relatively small

Table 5 *Computational performance for Langevin dynamics, LIN and LN versus explicit integration*

	Method	Time (min)	Part Ia: H_h evaluation	Part Ib: integration	Part II: minimization	Relative time
Dipeptide	LIN, $\Delta t = 15$ fs	23.5	1.8	4.6	16.7	2.30
(300 ps)	LIN, $\Delta t = 30$ fs	16.4	1.1	4.9	10.2	1.61
	LN, $\Delta t = 5$ fs	10.2	4.9	4.9	0.0	1.00
	Exp., $\Delta t = 0.5$ fs	21.3				2.09
BPTI	LIN, $\Delta t = 12$ fs	54.3	6.2	2.6	45.2	3.19
(1.5 ps)	LIN, $\Delta t = 15$ fs	53.0	4.8	2.72	45.3	3.12
	LN, $\Delta t = 5$ fs	17.0	14.1	2.70	0.0	1.00
	Exp., $\Delta t = 0.5$ fs	74.4				4.38

For LIN and LN, the total time is slightly more than the sum for parts Ia, Ib and II.

system BPTI, LN yields a speedup factor of 4.38 in comparison to BBK with 0.5 fs timesteps for covering the same total simulation length. This speedup factor will increase with system size because the cost for gradient evaluation – dominating performance of the explicit scheme – will increase in relation to the cost of sparse-Hessian/vector products, the dominating cost of LN. For LIN, the cost of Part II must also be taken into account and is expected to be very substantial as system size increases.

Summary and Perspective

Increasing the timestep in numerical discretizations of complex, multiscale systems is a very challenging mathematical problem with applications in many areas of science and engineering. There are currently several interesting approaches for MD applications, both in the constrained-formulation and separating-framework categories. In each case, if the reproduction of detailed dynamics is required (i.e., agreement with experiment or with explicit simulations at small timesteps such as 0.5 and 1 fs), the maximum timestep is limited by the high-frequency motion and its coupling to the slower modes of the system. Unfortunately, this coupling is too strong in biomolecules to allow damping or poor resolution of the fast components [43,9,41], and some corrections are necessary for properly incorporating these contributions if larger timesteps are used [9].

In our LIN separating framework, a correction phase involving solution of a nonlinear function, closely related to the potential energy, follows a linearization phase, in which the linearized equations of motion are solved explicitly by a second-order symplectic method. The computational performance of LIN was enhanced by (i) using sparse-Hessian/vector products in the first part and (ii) using the efficient truncated-Newton minimization package TNPACK in the second part, with more lenient convergence criteria than the default values (used for structure refinement [53]). The use of a short-range $\mathbf{H}_h$ appears to be justified for resolving the harmonic motion component. Further exploration of this notion will be presented in Ref. 57. An adaptive-timestep routine was also incorporated to ensure that energetic fluctuations are controlled.

We have used 0.5 fs timesteps for solution of the linearized equations of motion because of good numerical behavior [41], and have compared the LIN results with explicit simulations at that timestep. Speedup in MTS methods is also typically reported in comparison to 0.5 fs timesteps [41,46]. It is also reasonable to use 1 fs timesteps in all cases. Although fluctuations should be larger, physical behavior should be similar, but LIN's competitiveness in terms of CPU time would decrease since only the computational time of Part Ib would decrease by a factor of one-half.

With the improvements outlined here, LIN becomes a competitive method. Speedup is modest on a single processor – a factor of 1.4 for BPTI – but this value is expected to increase with system size and, possibly, also with further improvements in our minimization scheme and adaptive-timestep control, and in the incorporation of a criterion for harmonic-model assessment.

Analysis of the results has also validated the LIN approach. For the dipeptide, LIN trajectories at 30 fs timesteps showed excellent agreement with explicit trajectories generated at $\Delta t = 0.5$ fs, in terms of both energy and geometry. For BPTI, LIN results with 15 fs timesteps gave very good agreement with respect to energetic and structural properties with the corresponding explicit simulations. We believe that 15 fs timesteps (limited by the range of validity of the harmonic approximation) with LIN are feasible for biomolecules in general. Up to this timestep regime, the global behavior of the LIN and explicit trajectories is essentially the same, with a good agreement of energy means, variances, and geometric fluctuations.

Computational competitiveness is another important issue addressed in this work. Attention to this aspect led to a delightful surprise. Our related method termed LN, which has LIN's linearization but no correction phase, demonstrates, at $\Delta t = 5$ fs, very good agreement with explicit-scheme trajectories *and* offers a far more significant computational advantage. Namely, for BPTI, LN already offers a speedup factor of 4.38 in comparison with $\Delta t = 0.5$ fs in BBK (Tables 1 and 5). Moreover, LN's competitiveness will also increase with size.

With an LN timestep of 5 fs and an inner timestep of 0.5 fs, 10 sparse-Hessian/vector products and one Hessian evaluation are required to cover each 5 fs interval. This work must be compared to 10 gradient evaluations for the explicit scheme. If the inner timestep is 1 fs in both cases, the corresponding ratio is 5 matrix/vector products and one Hessian (plus gradient) evaluation in LN versus 5 gradient evaluations in the explicit integrator. Thus, the speedup factor will depend on two ratios: the cost of a gradient evaluation versus the cost of the product of a sparse Hessian (resulting from a 4.5 Å cutoff) and a vector, and the cost of a gradient evaluation versus the cost of evaluating the gradient plus the sparse Hessian.

This unexpected windfall from LN illustrates the value of developing novel approaches that might initially appear as not practical for macromolecules (e.g., based on the computational expense of normal-mode decomposition [43,44]). Further developments regarding the scaling issues of LN, its computational performance, and applications to larger systems will form the subjects of future work. In particular, speedup issues in comparison to 1 or 2 fs timesteps will be addressed, as well as performance in the context of fast electrostatics. Possibly, the timestep in LN might be slightly increased from 5 fs if a SHAKE-like method is incorporated to monitor the bond-energy fluctuations. Other interesting mathematical issues, such as further analysis of stability and resonance issues in molecular dynamics simulations, will also be reported in future communications [57,61].

Acknowledgements

We thank Jan Hermans, Martin Karplus, and Klaus Schulten for inspiring this work. We thank Guihua Zhang for helpful discussions. Support by NIH (Parallel Research Resource in Structural Biology, award RR08102) and NSF/ARPA (HPCC award ASC-9318159) is gratefully acknowledged. T.S. is an investigator of the Howard Hughes Medical Institute.

References

1. McCammon, J.A. and Harvey, S.C., Dynamics of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, MA, 1987.
2. Allen, M.P. and Tildesley, D.J., Computer Simulation of Liquids, Oxford University Press, New York, NY, 1990.
3. Schlick, T., In Mesirov, J.P., Schulten, K. and Sumners, D.W. (Eds.) Mathematical Applications to Biomolecular Structure and Dynamics, IMA Volumes in Mathematics and its Applications, Vol. 82, Springer, New York, NY, 1996, pp. 219–224.
4. Gear, C.W., Numerical Initial Value Problems in Ordinary Differential Equations, Prentice-Hall, Englewood Cliffs, NJ, 1971.
5. Zhang, G. and Schlick, T., Mol. Phys., 84(1995)1077.
6. Mishra, B. and Schlick, T., J. Chem. Phys., 105(1996)299.
7. Schlick, T., Figueroa, S. and Mezei, M., J. Chem. Phys., 94(1991)2118.
8. Nyberg, A. and Schlick, T., Chem. Phys. Lett., 198(1992)538.
9. Schlick, T. and Peskin, C.S., J. Chem. Phys., 103(1995)9888.
10. Rahman, A., Phys. Rev. A, 136(1964)405.
11. Rahman, A. and Stillinger, F.H., J. Chem. Phys., 55(1971)3336.
12. Barojas, J., Levesque, D. and Quentrec, B., Phys. Rev. A, 7(1973)1092.
13. Rahman, A. and Stillinger, F.H., J. Chem. Phys., 60(1974)1545.
14. Ryckaert, J.P. and Bellemans, A., Chem. Phys. Lett., 30(1975)123.
15. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23(1977)327.
16. Verlet, L., Phys. Rev., 159(1967)98.
17. Ruth, R.D., IEEE Trans. Nucl. Sci., 30(1983)2669.
18. Sanz-Serna, J.M. and Calvo, M.P., Numerical Hamiltonian Problems, Chapman & Hall, London, 1994.
19. McCammon, J.A., Gelin, B.R. and Karplus, M., Nature, 267(1977)585.
20. Streett, W.B., Tildesley, D.J. and Saville, G., Mol. Phys., 35(1978)639.
21. Streett, W.B., Tildesley, D.J. and Saville, G., In Lykos, P. (Ed.) Computer Modeling of Matter, ACS Symposium Series, Vol. 86, American Chemical Society, Washington, DC, 1978, pp. 144–158.
22. Grubmuller, H., Heller, H., Windemuth, A. and Schulten, K., Mol. Sim., 6(1991)121.
23. Hale, J.M., In Lykos, P. (Ed.) Computer Modeling of Matter, ACS Symposium Series, Vol. 86, American Chemical Society, Washington, DC, 1978, pp. 172–190.
24. Scully, J.L. and Hermans, J., Mol. Sim., 11(1993)67.
25. Board Jr., J.A., Causey, J.W., Leathrum Jr., T.F., Windemuth, A. and Schulten, K., Chem. Phys. Lett., 198(1992)89.
26. Van Gunsteren, W.F. and Berendsen, H.J.C., Mol. Phys., 34(1977)1311.
27. Van Gunsteren, W.F., Mol. Phys., 40(1980)1015.
28. Andersen, H.C., J. Comput. Phys., 52(1983)24.
29. Miyamoto, S. and Kollman, P.A., J. Comput. Chem., 13(1993)952.
30. Leimkuhler, B. and Skeel, R.D., J. Comput. Phys., 112(1994)117.
31. Barth, E., Kuczera, K., Leimkuhler, B. and Skeel, R.D., J. Comput. Chem., 16(1995)1192.

32. Mazur, A.K. and Abagyan, R.A., J. Biomol. Struct. Dyn., 6(1989)815.
33. Gibson, K.D. and Scheraga, H., J. Comput. Chem., 11(1990)468.
34. Gibson, K.D. and Scheraga, H., J. Comput. Chem., 11(1990)487.
35. Rice, L.M. and Brünger, A.T., Proteins Struct. Funct. Genet., 19(1994)277.
36. Turner, J., Weiner, P., Robson, B., Venugopal, R., Schubele III, H. and Singh, R., J. Comput. Chem., 16(1995)1271.
37. Bae, D.-S. and Haug, E.J., Mech. Struct. Mach., 15(1987)359.
38. Tuckerman, M.E. and Berne, B.J., J. Comput. Chem., 95(1992)8362.
39. Procacci, P. and Berne, B.J., J. Chem. Phys., 101(1994)2421.
40. Humphreys, D.E., Friesner, R.A. and Berne, B.J., J. Phys. Chem., 98(1994)6885.
41. Watanabe, M. and Karplus, M., J. Phys. Chem., 99(1995)5680.
42. Tuckerman, M.E., Martyna, G.J. and Berne, B.J., J. Chem. Phys., 93(1990)1287.
43. Zhang, G. and Schlick, T., J. Comput. Chem., 14(1993)1212.
44. Zhang, G. and Schlick, T., J. Chem. Phys., 101(1994)4995.
45. Turner, J.D., Weiner, P.K., Chun, H.M., Lupi, V., Gallion, S. and Singh, U.C., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 535–555.
46. Zhou, R. and Berne, B.J., J. Chem. Phys., 103(1995)9444.
47. Harrison, R.W., Biopolymers, 23(1984)2943.
48. Levy, R.M., Srinivasan, A.R., Olson, W.K. and McCammon, J.A., Biopolymers, 23(1984)1099.
49. Brooks, B.R. and Karplus, M., Proc. Natl. Acad. Sci. USA, 82(1985)4995.
50. Peskin, C.S. and Schlick, T., Commun. Pure Appl. Math., 42(1989)1001.
51. Schlick, T. and Overton, M.L., J. Comput. Chem., 8(1987)1025.
52. Schlick, T. and Fogelson, A., ACM Trans. Math. Software, 14(1992)46.
53. Derreumaux, P., Zhang, G., Brooks, B. and Schlick, T., J. Comput. Chem., 15(1994)532.
54. Mandziuk, M. and Schlick, T., Chem. Phys. Lett., 237(1995)525.
55. Hairer, E. and Wanner, G., Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, Springer Series in Computational Mathematics, Vol. 14, Springer, New York, NY, 1991.
56. Grønbech-Jensen, N. and Doniach, S., J. Comput. Chem., 15(1994)997.
57. Barth, E., Mandziuk, M. and Schlick, T., in preparation.
58. Schlick, T. and Fogelson, A., ACM Trans. Math. Software, 14(1992)71.
59. Brünger, A., Brooks, C.B. and Karplus, M., Chem. Phys. Lett., 105(1982)495.
60. Askar, A., Space, B. and Rabitz, H., J. Phys. Chem., 99(1995)7330.
61. Schlick, T., Barth, E. and Mandziuk, M., Annu. Rev. Biophys. Biomol. Struct., 26(1997), to appear.
62. Hao, M. and Harvey, S.C., Biopolymers, 32(1992)1393.
63. O'Neil, J.O. and Szyld, D.B., SIAM J. Sci. Stat. Comput., 11(1990)811.
64. Duff, I.S., Erisman, A.M. and Reid, J.K., Direct Methods for Sparse Matrices, Oxford University Press, New York, NY, 1986.
65. Schlick, T., SIAM J. Sci. Stat. Comput., 14(1993)424.
66. Golub, G.H. and van Loan, C.F., Matrix Computations, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1986.
67. Zlatev, Z., Computational Methods for General Sparse Matrices, Kluwer Academic Publishers, Dordrecht, 1991.
68. Parlett, B.N., The Symmetric Eigenvalue Problem, Prentice-Hall Series in Computational Mathematics, Prentice-Hall, Englewood Cliffs, NJ, 1980.

Appendix. Projection method

We discuss a projection method for isolating the negative portion of the spectrum that emerged in our LIN studies. In general, this method can be used to treat any portion of the spectrum (i.e., the few slowest modes) to obtain a Hessian of desired eigenstructure. For the projection method, we rewrite the unitary transformations $\mathbf{H}' = \mathbf{T}^{\mathrm{T}}\mathbf{D}\mathbf{T}$ by splitting $\mathbf{T}$ and $\mathbf{D}$ into two parts corresponding to the negative (*first* k entries) and nonnegative eigenvalues of $\mathbf{H}'$:

$\mathbf{D}_-$ and $\mathbf{D}_+$ are the $k \times k$ and $(n-k) \times (n-k)$ submatrices of $\mathbf{D}$, and

T_- and $\mathbf{T}_+$ are the $k \times n$ and $(n-k) \times n$ submatrices of $\mathbf{T}$.

The mass-weighted Hessian, $\mathbf{H}' = \mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-1/2}$, can then be expressed as

$$\mathbf{H}' = \mathbf{T}^{\mathrm{T}}\mathbf{D}\mathbf{T} = [\mathbf{T}_- \ \ \mathbf{T}_+]^{\mathrm{T}}\begin{bmatrix}\mathbf{D}_- & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_+\end{bmatrix}[\mathbf{T}_- \ \ \mathbf{T}_+] = \mathbf{T}_-^{\mathrm{T}}\mathbf{D}_-\mathbf{T}_- + \mathbf{T}_+^{\mathrm{T}}\mathbf{D}_+\mathbf{T}_+ \quad \text{(A1)}$$

For simplicity, we now omit the prime superscripts of mass weighting. Our matrix

$$\mathbf{H}_+ \equiv \mathbf{H} - \mathbf{T}_-^{\mathrm{T}}\mathbf{D}_-\mathbf{T}_- \quad \text{(A2)}$$

has only nonnegative eigenvalues, and we can solve Eq. 3 with $\mathbf{H}_+$ replacing $\mathbf{H}_h$ by explicit integration at timesteps $\Delta\tau = 0.5$ or 1 fs, with the random force R updated at every iteration according to Eq. 2.

It now remains to determine $\mathbf{T}_-$. One strategy for approximating the negative eigenvalues of a large symmetric matrix is based on a block-diagonal approximation for $\mathbf{H}_h$. This form can be obtained by exploiting the molecular topology or using a partitioning scheme [62]. Alternatively, sparse molecular preconditioners, as used in our minimization [53], have a nearly band structure and can be subjected to reordering to yield block-diagonal or banded forms [63,64]. Our experience in truncated-Newton minimization for molecular systems in CHARMM [53], reordering schemes, and sparse modified factorizations [65] will be valuable here. Such sparse structures can be exploited to perform parallel decompositions for the blocks or to apply block-Lanczos or Givens decomposition techniques [66,67]. More generally, preconditioned Lanczos techniques (e.g., Davidson-type) are attractive for large systems [68] since they are iterative and cheap per step, like related conjugate gradient methods [66]. Givens orthogonalization techniques are especially efficient for banded systems, as small rotation matrices are repeatedly applied to obtain a tridiagonal form. Roundoff problems in Lanczos techniques can be monitored: the appearance of spurious multiple copies of eigenvalues can be *detected* (e.g., program lanczos by Cullum and Willoughby, available from netlib) or *avoided* by partial reorthogonalization (program laso by Scott in netlib). Dan Sorensen's package, ARPACK, a variant of the Arnoldi/Lanczos method based on an implicit restarting technique, has the special advantage of requiring only matrix/vector products and no dense similarity transformations. It finds k eigenvalues with user-specified properties in limited storage (of order 2kn).

Reduced variable molecular dynamics

James Turner[a], Paul K. Weiner[a], Barry Robson[b], Ravi Venugopal[c], Harry Schubele III[c] and Ramen Singh[c]

[a]Amdyn Systems Inc., 28 Tower Street, Somerville, MA 02143, U.S.A. [b]Proteus Molecular Design Ltd., Proteus House, Lyme Green Business Park, Macclesfield, Cheshire SK11 0JL, U.K. [c]Dynacs Engineering Company Inc., 28870 U.S. Highway 19 North, Suite 405, Clearwater, FL 34621, U.S.A.

Introduction

Computer simulations of molecular systems provide invaluable insights for understanding the structure–function relationships pertinent to the discovery of new molecules with desired properties, such as new pharmaceutical drugs. The use of molecular simulations and the increasing performance of modern computers makes it now possible to study the precise physicochemical nature of protein–ligand interactions, protein engineering, solvation phenomena, and to characterize the thermodynamical properties of complex systems with many thousands of atoms [1,2]. Even with the availability of high-performance computers, many problems of practical interest are so computationally challenging that new solution methods are required for formulating and solving the resulting mathematical models. This paper explores the capabilities of recursive dynamics methods for reducing the computational effort required for studying complex molecular systems. A numerical example is presented that demonstrates the application of the basic recursive algorithms.

Molecular dynamics (MD) is one of the established simulation techniques in the prediction, analysis, and design of complex molecules [1,2]. These techniques typically apply the Newtonian laws (F = ma) to the motion of all of the atoms in the system, where the forces have been defined as a function of atom type, bond type, dihedral type, and interatomic distances. There are many software packages that implement these techniques very efficiently on computers ranging from workstations to large parallel supercomputers. However, these software tools typically analyze the atomic motions in terms of particle-based equations of motion.

There are fundamental difficulties associated with handling large MD problems with particle-based methods, that cannot be overcome by attention to the details of hardware and software implementation. The first difficulty is due to the N^2 nonbonded interactions (assuming no cutoff criteria), where N is the number of atoms, that occur in the energy function because every atom can experience the forces of every other atom via long-range interaction effects. Even with the use of arbitrarily imposed cutoff distances, the evaluation of the nonbonded terms can take over 90% of the computer time involved in each energy function evaluation.

The second difficulty relates to the step size limitation in the integration step of the dynamics algorithm. In order to accurately integrate the equations of motion in atomistic models, it is necessary to sample adequately the period of the highest frequency motion in the system. The high-frequency bond stretching motion that occurs in all molecules requires the use of a small step size (i.e., 0.5–1.0 fs) in the simulation. This limitation prevents very large simulations from being run, even on supercomputers, for periods much greater than hundreds of picoseconds. Unfortunately, this is much too short a time scale to study most processes of biological interest or to study rapidly occurring but infrequent events, such as conformational barrier crossings. Even for the study of noncovalent protein–ligand interactions, the simulations may be too short to adequately sample conformational space and therefore to accurately calculate the free energy of interaction.

The third difficulty relates to the presence of multiple minima in the (3N − 6)-dimensional conformational space that defines the potential energy surface. These minima can cause the simulations to become locally trapped or delayed. Even if one had enormous amounts of supercomputer time, it is difficult to ensure adequate sampling to estimate thermodynamic properties accurately, such as free energy, or even to ensure that the structures reached include all of the biologically relevant ones. Because of the time required for each simulation, it would be difficult to repeat the simulations for many other modified molecules. The additional simulations are important for an understanding of the specificity of the biological process being studied.

If the equations of motion for a particular problem are too difficult to analyze with available computer resources, one is forced to consider approximate techniques for extracting the useful chemical design information. The researcher tries to establish the simplest mathematical models that can be used, while retaining predictive capabilities for providing insights into the biologically interesting behaviors. One approach for solving this problem consists of applying non-Newtonian dynamics. These methods are non-Newtonian in the sense that the Newtonian laws are highly modified or Newtonian motion is in some sense interrupted [3]. These methods include the use of increased dimensionality to avoid entrapment in three-dimensional x, y, z Euclidean space [4,5]. By transforming the normal energy/force functions and extending the simulations into four dimensions, barriers, which exist due to movements in true Euclidean space, could be tunneled through and a lower minimum found. The appropriate direction of tunneling is controlled by 'target functions' which express heuristic information concerning which direction the global minimum must lie in. The target function can also represent specific experimental data obtained from the protein of interest by experimental work [6].

Other methods, such as Rush dynamics, modify the basic simulation equations in the familiar three-dimensional space. It has been shown, for at least small problems, that these methods can more efficiently search conformational space than the traditional MD methods [3,7]. There is, however, still a need to reduce a very large problem down to a simpler one even when applying these methods.

A better approach is to significantly reduce the number of degrees of freedom in the simulation (while retaining the quantitative aspects of the simulation). Then each simulation takes less time and one can run many different simulations. A class of dynamics algorithms, called reduced variable dynamics, makes this reduction in time possible. Since there are many fewer degrees of freedom, phase space can be more efficiently searched. The following sections will describe the current available methods.

Traditional constrained dynamics approaches

The traditional approach for eliminating uninteresting degrees of freedom (DOF) and high-frequency motions is through the use of defined constraints. The constraints are introduced into the equations of motion by adding a force-like term. Typically, this term consists of the gradient of a position-dependent constraint equation times an unknown Lagrange multiplier (LM). The resulting atomistic constrained equation of motion (EOM) follows as

$$m_j \ddot{R}_j = F_j(R_1, \dots, R_N) + \sum_{r=1}^{N_c} \nabla_j C_r(R_1, \dots, R_N) \lambda_r$$

subject to

$$\nabla_j C_r(R_1, \dots, R_N) \dot{R}_j = 0, \quad j = 1, \dots, N_p, \quad r = 1, \dots, N_c$$

where ∇_j denotes the gradient with respect to the jth atom position coordinates, $C_r(*)$ denotes the functional form of the rth constraint function, and λ_r denotes the rth LM. The solution for the LM is obtained by differentiating the gradient of the constraint equation and introducing the EOM. Straightforward manipulations lead to an LM solution defined by a linear matrix–vector algebraic equation of dimension $(N_c \times N_c)$, where N_c denotes the number of constraints. As N_c becomes large, the computational burden associated with the inversion of the constraint matrix rapidly becomes prohibitive (e.g., $O(N^3)$). This approach retains all of the atomic degrees of freedom and does not exploit the inner connection topology between molecules implied by the imposition of the constraints.

There are two alternative methods available for solving this problem. The first method involves the construction of a set of generalized coordinates, leading to

$$M(q)\ddot{q} = F(q, \dot{q})$$

where q denotes the $(N \times 1)$ generalized coordinate vector, $M(q)$ denotes the $(N \times N)$ mass matrix, and $F(*)$ denotes the $(N \times 1)$ force vector. The method is completely general and has the added benefit of eliminating the constraint variables completely in the problem formulation. However, it requires the inversion of an $(N \times N)$ matrix $M(q)$, which is an $O(N^3)$ algorithm. Typically, $N \ll N_c$; nevertheless, as problem sizes continue to increase even these approaches reach practical limits for the time required to solve for the system accelerations.

Generalized coordinates have been used in molecular dynamics for many years, although the applications have been previously limited to short polymers [8]. General-purpose formulations have not appeared previously for several reasons. First, the classical analytical methods, such as Lagrange's, involve the formulation of kinetic and potential energy expressions, which require computing potentially thousands of time-varying first- and second-order partial derivatives [9]. For systems consisting of N degrees of freedom, the equations of motion are formulated and solved by computing N time-varying first- and N x N second-order partial derivatives with respect to the generalized coordinates at each time step in the integration process. This procedure is well defined but rapidly becomes unwieldy even for relatively low-order problems (i.e., ≥ 30 independent variables). In addition, there is still the problem of inverting an (N x N) mass matrix.

Mazur and Abagyan [10,11] have developed a modeling approach using a Lagrangian-based internal coordinate model for generating the equations of motion. The accelerations are computed by inverting the mass matrix for all the degrees of freedom at each time step in the integration process. An advantage of this approach is that the constraint equations are analytically eliminated from the problem. The major disadvantage of their approach is that the time-varying mass matrix must be computed and inverted at each time step. This approach is useful for small problems, but does not scale up well for large problems because the computational effort required for inverting a large mass matrix ($O(N^3)$) dominates the effort required for the force-field calculations.

In addition, the methodology has some limitations [12], including restrictions on intermolecular connectivity and torsion angle definitions. In a latter article [12], these limitations were eliminated, but only a Monte Carlo technique was presented and there were no updates for the formulation of the equations of motion to account for the additional degrees of freedom. This update would involve a major modification of the original formulation and would still not remove the fundamental problem of requiring a large matrix inversion.

Rudnicki et al. [13] present an algorithm for computing the pseudorotation dynamics of a furanose ring. By retaining internal vibrational behaviors, their approach goes beyond the limitations of a rigid-body model. However, the authors did not present a methodology for coupling their pseudorotation model to the underlying rigid-body motion of the ring, nor did they discuss how this model could be coupled to other bodies in a simulation. This must be carried out in such a way that the resulting equations of motion scale up efficiently for large systems. This method is most closely related to the generalized coordinate method, because the constraints are eliminated by a Lagrangian algorithmic approach. Their approach is useful for generating simplified dynamics models for subcomponents of more complex systems.

Other methods involve calculating approximate solutions to the constrained equations of motion. These methods use iterative approaches to approximately solve for the Lagrange multipliers. For simple bond constraints, these algorithms typically only need a few iterations to converge. The most popular method of this type, SHAKE [14,15], is easy to implement and scales as O(N) as the number of constraints increases. This method has been very popular for macromolecules. An alternative method, RATTLE [16], is based on the velocity version of the Verlet algorithm. Like SHAKE, RATTLE is an iterative algorithm. However, adding any other types of constraints, even bond angle constraints, can greatly slow the convergence of SHAKE and limit the maximum step size [17].

Recent work [9,18] has resulted in extensions of the SHAKE algorithm that allow for internal coordinate constraints. The newer methods no longer have the convergence problems associated with the original SHAKE algorithm. These methods are also iterative and straightforward to implement. However, the maximum time step size used in the papers describing these methods is still limited to 3 fs. The systems used to test the methods were small (< 50 atoms) and it is not known how the methods would perform for larger systems with many thousands of constraints.

A new dynamics algorithm coupling implicit integration and normal mode techniques has recently been developed [9]. This algorithm gains a factor of 10 × over other explicit integration schemes. The integration technique introduces no damping and is stable for step sizes as large as 50 fs. This new integration scheme is very interesting and could be useful even with the method described in this paper. The method, however, requires a linearized model. As shown below, dynamics equations between coupled bodies must have nonlinear terms to allow the simulation to be valid for arbitrarily large displacements and rapid motions. In very dense systems, such as explicit solvation studies, the linearized model could be a reasonable approximation. In large protein simulations with no or implicit solvent, such as might be used in folding studies, one wants large and rapid motions to occur so that interesting low-energy structures can be quickly located. These simulations would require the presence of the nonlinear dynamics terms.

The shortcomings of these iterative methods are (i) they are not exact, (ii) they are still limited to a relatively small step size, or (iii) they do not scale up as O(N). The proposed approach is based on a recursive generalized coordinate formulation for the equations of motion and can be generalized to handle both particles and rigid-body constraints in a unified framework. The proposed algorithm maintains all internal constraints with no approximations. This new approach leads to a reduced variable molecular dynamics simulation (RVMD) technique. The RVMD will prove to be quite powerful because there are far too many variables, even with all bonds constrained, for macromolecular simulations to explore more than a small region of phase space. However, once the interesting events are located using RVMD, unconstrained simulations can be carried out to calculate the thermodynamic properties of interest.

It is a practical necessity to be able to greatly reduce the number of variables in a macromolecular system, so that qualitative aspects of its behavior, such as reaction pathways, can be studied and interesting events located. In several papers [17,20], it has been noted that the use of constraints beyond bond length constraints can affect calculated properties, such as conformational interconversion rates. As a result, the effects of the constraints must be closely examined. The impact of imposing various constraints can be assessed by comparing the results of the reduced variable simulations with unconstrained simulations (when possible), as well as a comparison with any available experimental data.

This paper describes an extension to previously developed constraint techniques applied by Turner et al. [21–24] and developed by Singh et al. [25–27]. These enhanced constraint methods will enable the study of large computational chemistry problems that cannot be easily handled with current constrained molecular dynamics methods. These methods are based on an O(N) solution to the constrained equations of motion. The benefits of this approach are that (1) the system constraints are solved exactly at each time step, (2) the solution algorithm is noniterative, (3) the algorithm is recursive and scales as O(N), (4) the algorithm is numerically stable, (5) the algorithm is highly amenable to parallel processing, and (6) potentially greater integration step sizes are possible. It is anticipated that application of this methodology can potentially provide a 10–100-fold improvement in the speed of a large molecular trajectory as compared with the time required to run a conventional atomistic unconstrained simulation. It is anticipated that the RVMD methodology will provide an enabling capacity for pursuing the drug discovery process for large molecular problems.

General constrained dynamics formulation

Rigid-body constraints

For rigid bodies the constraint equations are generalized as follows [28,29]:

$$C_j(R_1, \theta_1, \ldots, R_N, \theta_N) = \alpha_j(t), \quad j = 1, \ldots, N_c$$

where R_j denotes the (3×1) jth body reference point position vector locating the jth body relative to inertial space and θ_j denotes the vector of the jth body kinematic variables used for describing the orientation of the jth body relative to the inertial frame. The constraint rates follow as

$$\sum_{r=1}^{N} \left(\frac{\partial C_j}{\partial R_r} \dot{R}_r + \frac{\partial C_j}{\partial \theta_r} \dot{\theta}_r \right) = \dot{\alpha}_j, \quad j = 1, \ldots, N_c$$

leading to a constrained EOM of the form

$$M_j \ddot{q}_j = F_j + b_j^T \Lambda, \quad \sum_{\sigma=1}^{N_b} b_\sigma (\dot{R}_\sigma \, \dot{\theta}_\sigma)^T = \dot{\alpha}_j$$

where

$$b_\sigma = \begin{bmatrix} \nabla_{R_1} & & \nabla_{\theta_1} \\ & \cdots & \\ \nabla_{R_N} & & \nabla_{\theta_N} \end{bmatrix}_\sigma C_\sigma$$

M_j denotes the (6×6) jth rigid-body mass matrix, F_i denotes the (6×1) jth body force, torque, and rotating frame kinematic effects vector, and $q_{,i} = d^2(R_{xj}, R_{yj}, R_{zj}, \theta_{xj}, \theta_{yj}, \theta_{zj})/dt^2$ denotes the acceleration vector for the jth body translation and orientation. Other sets of orientation parameters can be used for characterizing the rotational motion of rigid bodies (e.g., Euler parameters, etc.). The $(N_c \times 6)$ constraint matrix, b_σ, is generalized for both translational and rotational components. Typically, the constraints are defined at the interconnection hinges between contiguous bodies. For example, if only one rotational DOF is allowed at a joint, then two rotational and three translational constraints must be defined and the constraint equation above is a (5×1) vector. The solution for Λ can be expressed in a functional form that is identical to the form obtained for the particle formulation [28].

Recursive generalized coordinate methodology

The modeling problem is naturally divided into two parts: (i) the formulation and solution process for the equations of motion, and (ii) the development of mathematical models for the constraint functions to be supported by the RVMD algorithm. The generation of equations of motion for interconnected systems defined by generalized coordinates requires special consideration. Large matrices can occur either because of the need to solve the constraint equations, given above with a large Lagrange multiplier constraint matrix [30,31], or the need to solve the accelerations at the system level. The large matrices arise because conventional approaches attempt to obtain the solutions in a single operation, such as matrix inversion, for the desired unknowns.

Recursive techniques reformulate the solution process to eliminate these large matrices by using body-level operations. The body-level operations lead to implicitly defined sets of equations that are noniteratively solved. Small matrices arise in the solution process because only local body-level models are considered at any one time. The recursive process leads to a multistep algorithm. An added benefit of recursive formulations is that they allow the user greater freedom in setting up new constraint models.

The recursive algorithm greatly reduces the computational complexity involved in solving large-order linear equations of the form

$$Ax = b$$

They work by using noniterative recursive equations, which involve many small matrices, to generate the solution

$$x = A^{-1}b$$

without forming A or A^{-1} explicitly. An example of a recursive algorithm [23,24,32] has been run comparing two methods for solving the constraints between multiple rigid bodies. The $O(N^3)$ generalized coordinate algorithm inverted a large matrix in a single step. The $O(N)$ recursive algorithm used a series of steps that only required the inversion of (5×5) matrices. Each body has six degrees of freedom. There are five constraints between each interconnected body. For 14 bodies, the recursive algorithm is 1.9-fold faster than the generalized coordinate method. For 350 bodies, the recursive algorithm is 2523-fold faster. In this case, the generalized coordinate method had to invert a (1745 × 1745) matrix.

Current modeling approach

The equations of motion are modeled using N generalized coordinates [33–37] $q = \{q_1, \ldots, q_N\}$. In this model all of the individual body equations of motion are collected together to create a system-level equation of motion. The solution for the system-level equation then treats all interaction effects simultaneously throughout the system. This approach leads to large matrix equations, which are required to balance the interaction effects everywhere in the system. This approach is presented first since it is conceptually easier to understand the algorithmic strategies required to dealing with systems described by free and constrained degrees of freedom. Then this approach is generalized to restructure the algorithm strategies to be recursive in nature. The recursive approach eliminates the need for building large matrix equations, leading to significant computational savings.

The system-level equation of motion is described by

$$M(q)\ddot{q} = f(q, \dot{q}, t) + C(q, t)^T \lambda$$

subject to the $(N_c \times 1)$ vector constraint equation

$$C(q, t)\dot{q} = b(t)$$

The constraint equation can be factored into free and constrained parts as follows:

$$A_f \dot{q}_f + B_c \dot{q}_c = b \tag{1}$$

where A_f is the $(N_c \times N)$ free degrees-of-freedom transformation matrix, q_f is the $(N \times 1)$ vector of free degrees of freedom, B_c is the $(N_c \times N_c)$ constrained degrees-of-freedom transformation matrix, and q_c is the $(N_c \times 1)$ vector of constrained degrees of freedom.

To eliminate the constrained degrees of freedom, we need a kinematic expression for the constrained rates as a function of the free degrees of freedom. The required equation is obtained by solving Eq. 1 for q_c as follows:

$$\dot{q}_c = D\dot{q}_f + Z \tag{2}$$

where

$$D = -[B_c]^{-1}A_f, \quad Z = [B_c]^{-1}b$$

and the inverse matrix is assumed to be well defined.

The equation of motion can be transformed to eliminate the constrained degrees of freedom and the Lagrange multiplier by using Eq. 2 to define the following coordinate transformation:

$$\dot{q} = \begin{pmatrix} \dot{q}_f \\ \dot{q}_c \end{pmatrix} = W\dot{q}_f + Y \tag{3}$$

where

$$W = \begin{bmatrix} I_p \\ D \end{bmatrix}, \quad Y = \begin{pmatrix} 0 \\ Z \end{pmatrix}$$

Differentiating Eq. 3 with respect to time leads to

$$\ddot{q} = \dot{W}\dot{q}_f + W\ddot{q}_f + \dot{Y} \tag{4}$$

Introducing Eq. 4 into the equation of motion and multiplying the resulting equation by W^T leads to the free degree-of-freedom equation of motion

$$\ddot{q}_f = (W^T M W)^{-1} W^T (f_{ext} + C^T\lambda - M[\dot{W}\dot{q}_f + \dot{Y}])$$

The solution process is completed by observing that the *Lagrange multiplier* term above vanishes because the following product is identically zero

$$W^T C^T = [I_p D^T] \begin{bmatrix} A_f^T \\ B_c^T \end{bmatrix} = A_f^T [I - (B_c^T)^{-1} B_c^T] = A_f^T [0] = 0$$

leading to the final form for the free degree-of-freedom equation of motion:

$$\ddot{q}_f = (W^T M W)^{-1} W^T (f_{ext} - M[\dot{W}\dot{q}_f + \dot{Y}]) \tag{5}$$

where all constraint terms have been eliminated.

Recursive solution approach

The recursive approach begins by writing the individual body equations of motion as

$$M_j(q_j)\ddot{q}_j = f_j(q_j, \dot{q}_j, t) + C_j(q_1, \ldots, q_{Nb})^T \lambda$$

where the constraint matrix C_j selects a subset of the system-level Lagrange multipliers. For bodies connected in a tree structure, the recursive algorithms begin with bodies at the end of branches of the tree structure. This leads to equations of the form

$$M_{j+1}(q_{j+1})\ddot{q}_{j+1} = F_{j+1}(q_{j+1}, \dot{q}_{j+1}, t) + C^T_{j+1,j}\bar{\Phi}_j\lambda_j$$

$$M_j(q_j)\ddot{q}_j = F_j(q_j, \dot{q}_j, t) + C^T_{j,j+1}\bar{\Phi}_j\lambda_j + C^T_{j,j-1}\bar{\Phi}_{j-1}\lambda_{j-1}$$

where the first equation has one Lagrange multiplier for the constraint to the j body and the second equation has two Lagrange multipliers for the constraints to the j and j − 1 bodies. The constraint equation for the first body can be shown to be

$$\phi_j \dot{f}_j + \bar{\phi}_j \dot{c}_j = C_{j+1,j}\dot{q}_{j+1} + C_{j,j+1}\dot{q}_j$$

$$\phi_j^T \phi_j = I_f, \quad \bar{\phi}_j^T \bar{\phi}_j = I_c, \quad \bar{\phi}_j^T \phi_j = 0$$

where f_j denotes the free degrees of freedom at the joint connecting the jth and (j + 1)th bodies, c_j denotes the constrained degrees of freedom at the joint connecting the jth and (j + 1)th bodies, and ϕ_j and $\bar{\phi}_j$ denote selection operators for the free and constrained degrees of freedom.

Solving the constraint equation for q_{j+1} leads to

$$\dot{q}_{j+1} = C^{-1}_{j+1,j}(\phi_j \dot{f}_j + \bar{\phi}_j \dot{c}_j - C_{j,j+1}\dot{q}_j)$$

where the motion rate for the last body in the branch is now described in terms of the rates at the joint and the rate for the jth body. This equation is said to be implicit because of the dependence on the jth body rate.

Differentiating the constraint equation with respect to time and solving for q_{j+1} leads to

$$\ddot{q}_{j+1} = C^{-1}_{j+1,j}\phi_j \ddot{f}_j + h_{j+1}(\ddot{q}_j)$$

where h_{j+1} contains the derivative terms not displayed. Introducing the equation above into the equation of motion for the body at the end of the branch and multiplying the resulting equation by

$$\phi_j^T C^{-T}_{j+1,j}$$

leads to

$$\phi_j^T C^{-T}_{j+1,j} M_{j+1} C^{-1}_{j+1,j}(\phi_j \ddot{f}_j + h_{j+1}) = \phi_j^T C^{-T}_{j+1,j}(F_{j+1} + C^T_{j+1,j}\bar{\phi}_j \lambda_j)$$

or

$$\ddot{f}_j = [\phi_j^T C^{-T}_{j+1,j} M_{j+1} C^{-1}_{j+1,j}\phi_j]^{-1}(-h_{j+1}(\ddot{q}_j) + \phi_j^T C^{-T}_{j+1,j} F_{j+1})$$

where the Lagrange multipliers have vanished because of the orthogonality of ϕ_j and $\bar{\phi}_j$. The solution for the Lagrange multiplier is obtained by solving the constraint equation for the constraint rates as follows:

$$\dot{c}_j = \bar{\phi}^T(C_{j+1,j}\dot{q}_{j+1} + C_{j,j+1}\dot{q}_j)$$

Differentiating this equation with respect to time leads to

$$\ddot{c}_j = \bar{\phi}^T(C_{j+1,j}\ddot{q}_{j+1} + p_j(\ddot{q}_j)) \tag{6}$$

where p_j is implicitly dependent on the q_j body acceleration. Next, solving the constrained equation of motion for the q_{j+1} acceleration one obtains

$$\ddot{q}_{j+1} = M^{-1}_{j+1}(F_{j+1} + C^T_{j+1}\bar{\phi}_j \lambda_j)$$

and introducing the equation above into Eq. 6 the solution for the Lagrange multiplier follows as

$$\lambda_j = [\bar{\phi}_j^T C_{j+1} M_{j+1}^{-1} C_{j+1,j}^T \bar{\phi}_j]^{-1} (\ddot{c}_j - \bar{\phi}_j^T C_{j+1,j} M_{j+1}^{-1} F_{j+1} - \bar{\phi}_j^T P_j) \qquad (7)$$

This completes the first stage of the recursive algorithm. The equation of motion for the joint degrees of freedom is implicitly dependent on the q_j body acceleration. Next, the equation of motion for the jth body is processed. The first step is to introduce Eq. 7 into the jth body equation of motion, leading to

$$\bar{M}_j \ddot{q}_j = \bar{F}_j + C_{j,j-1}^T \bar{\phi}_{j-1} \lambda_{j-1}$$

where the mass matrix and the force vector have been modified. This equation has the same form as the first equation solved. Following the same procedure, we obtain equations of the form

$$\ddot{f}_{j-1} = [\phi_{j-1}^T C_{j,j-1}^{-T} M_j C_{j,j-1}^{-1} \phi_{j-1}]^{-1} (-h_j(\ddot{q}_{j-1}) + \phi_{j-1}^T C_{j,j-1}^{-T} F_j - \bar{\phi}_{j-1}^T P_{j-1})$$

and

$$\lambda_{j-1} = [\bar{\phi}_{j-1}^T C_j M_j^{-1} C_{j,j-1}^T \bar{\phi}_{j-1}]^{-1} (\ddot{c}_{j-1} - \bar{\phi}_{j-1}^T C_{j,j-1} M_j^{-1} F_j) \qquad (8)$$

The process is repeated until a body is reached which is not constrained by only one other body. After introducing the previously calculated Lagrange multiplier, the transformed equation of motion can be shown to be

$$\bar{M}_x \ddot{q}_x = \bar{F}_x$$

The solution for this equation follows as

$$\ddot{q}_x = \bar{M}_x^{-1} \bar{F}_x$$

which is possible because this equation is referenced to the inertial frame, which has an assumed acceleration of zero. By reversing the order of solution just described, the implicit dependence in each of the equations can now be defined. The key point of the recursive solution process is that small matrices are used at each step.

Nonlinear effects

Nonlinear motion models allow simulations to be valid for arbitrarily large displacements and rapid motions. Unlike linear models, nonlinear models can lead to fundamentally different types of solutions. For example, when one considers the equations governing the angular momentum of a rotating body, the solution for even unforced models is radically different. In the linear model below, when no torques act the angular momentum is constant, which implies a constant rotation rate as seen in a body-fixed frame. On the other hand, the nonlinear model has time-varying momentum components. Indeed, the time-varying terms lead to nonconstant rotation rates. As a result, the orientation of the body in inertial space can be characterized as a general tumbling motion. The evolving motions quickly become uncorrelated.

Linear model:

$$d\vec{P}/dt = \vec{0}, \quad \vec{P} = \vec{I}\cdot\vec{\omega}$$

Nonlinear model:

$$d\vec{P}/dt + \omega \times \vec{P} = \vec{0}$$

The conservation laws are given by

Kinetic energy:

$$T = (P_1^2/I_1 + P_2^2/I_2 + P_3^2/I_3)/2$$

Angular momentum:

$$\vec{P}\cdot\vec{P} = P_1^2 + P_2^2 + P_3^2$$

In both the linear and nonlinear cases, the kinetic energy and angular momentum are conserved. The resulting motions, however, are very different. The nonlinear model allows for coupling between axes that cannot exist in linear models. Without the nonlinear terms, one cannot be sure that the resulting motions are accurately predicted, unless the rates are extremely small.

The nonlinear effects arise from two sources. First, dynamic reference frames that move with each body are used to simplify the equations of motion. A significant advantage of this approach is that the rigid-body mass properties are constant. Constraints are also easier to describe in terms of the dynamic reference frames. The use of dynamic reference frames leads to nonlinear $\omega \times (*)$ terms in the translational velocity and acceleration, and the rotational angular momentum vector equations. These terms must be retained to account for rapid motions of the rigid bodies.

Second, when large changes in the orientation of rigid bodies with respect to inertial space are possible, the direction cosine matrices describing the large rotational displacements must retain all of the nonlinear products of sine and cosine terms. These effects are particularly important in tree structures, where many products of direction cosine matrices may be required to orient bodies with respect to a reference body's orientation.

Methods

The initial conformations of alanine dipeptide were obtained by model building, using AMBER+ [38] using the AMBER 3 force field [39,40]. The ϕ and ψ angles, shown in Fig. 1, were set at the desired initial value and a constrained energy minimization was performed. This was followed by gradually heating (3–5 ps with a 1 fs step size) to the specified temperature (300 or 600 K) using a dynamics simulation with SHAKE.

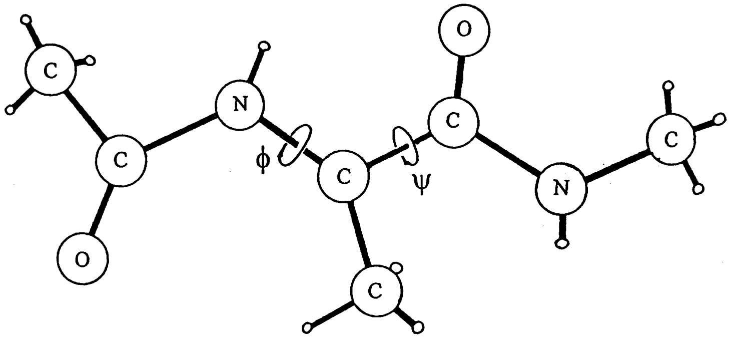


Fig. 1. Alanine dipeptide.

The resulting coordinates, particle velocities, rigid group definitions, and a list of the allowed degrees of freedom were then passed to the RVMD program. This program, as shown in Fig. 2, is completely independent from the molecular mechanics software. This enables the reduced variable method to be easily attached to any molecular mechanics or quantum mechanics code. The programs communicate via common blocks and an interface routine. The RVMD program takes care of all of the

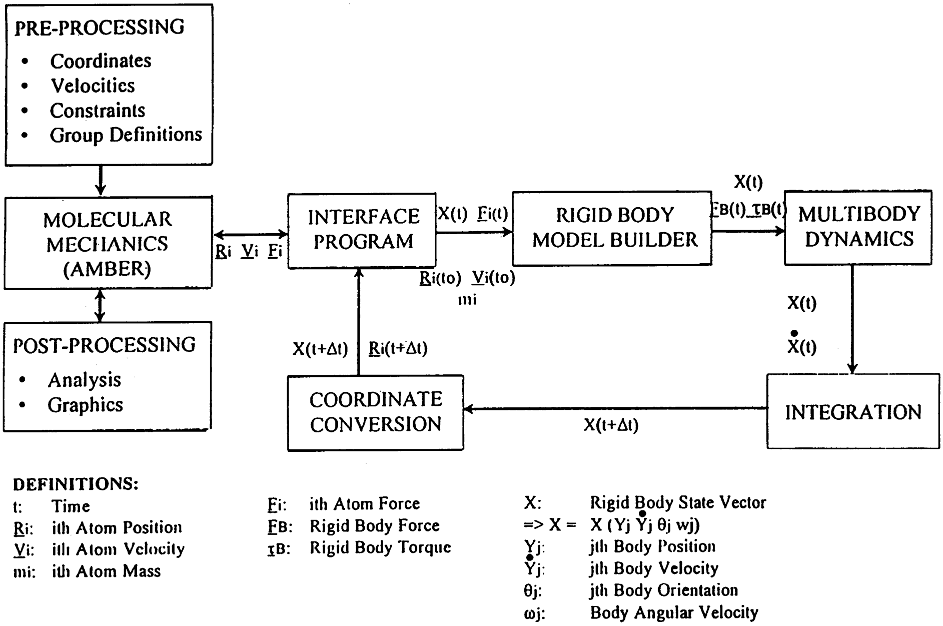


Fig. 2. Software flow chart.

coordinate transformations and the molecular mechanics program only needs to have routines that compute the forces in terms of Cartesian coordinates and does not have to be modified if the body definitions are changed.

The RVMD program calculates equivalent rigid-body masses, first mass moment vectors, and inertia tensors. From the particle position, velocity, and mass data, the linear and angular momentum are computed. After the rigid-body mass data are computed, a model can be developed for the inertial linear and angular momentum as follows:

$$\begin{bmatrix} \sum_{r=1}^{N_b} m_r V_r \\ \sum_{r=1}^{N_b} m_r R_r \times V_r \end{bmatrix} = \begin{vmatrix} I & \tilde{R} \\ 0 & I \end{vmatrix} \begin{vmatrix} J_k & \tilde{S}_k \\ -\tilde{S}_k & M_k \end{vmatrix} \begin{pmatrix} \omega_k \\ v_k \end{pmatrix}$$

where J_r denotes the rth bodies inertia tensor, S_r denotes the first mass moment, the tilde denotes that the first mass moment is expressed in terms of a (3×3) matrix equivalent of the vector cross product, M_r denotes the mass of the rth body, and the R vector locates the rigid-body reference point relative to the inertial frame. The only unknowns in this equation are ω_k and v_k. The solution for these variables is obtained by inverting the matrices. The initial rates were further processed to account for the constraints in the system. The constraints are defined by specifying the degrees of freedom at each joint between the bodies. For example, if a body was attached with a single rotational degrees of freedom, then three translational constraints and two rotational constraints were defined. This process consists of projecting the adjacent body rotational and translational rates along the specified joint axes.

Constant-energy and constant-temperature simulations were then run. In addition, annealing runs where the temperature was heated and then cooled were made in an attempt to locate the global minimum. For the later runs, it was necessary to implement a rigid-body temperature scaling algorithm. The all-atom AMBER simulation used a one-point Verlet integrator and the RVMD simulation used a four-point Runge–Kutta integrator. The accuracy of this integrator is extremely high. Constant-energy simulations of 100000 points had less than a 10% fluctuation in the total energy with no restarts. Simulations are currently being undertaken with a two-point integrator and a restart procedure. It is anticipated that this will provide similar accuracy with only 50% of the function evaluations.

Temperature scaling

The temperature was scaled by developing a model for the system kinetic energy as follows:

$$K = \frac{1}{2} \sum_{r=1}^{N_b} \begin{bmatrix} \omega_r \\ v_f \end{bmatrix}^T \begin{vmatrix} J_r & \tilde{S}_r \\ -\tilde{S}_r & M_r \end{vmatrix} \begin{bmatrix} \omega_r \\ v_r \end{bmatrix}$$

where J_r denotes the rth bodies inertia tensor, S_r denotes the first mass moment, the tilde denotes that the first mass moment is expressed in terms of a (3×3) matrix equivalent of the vector cross product, and M_r denotes the mass of the rth body. The angular velocities and linear velocities are linear functions of the degrees of freedom that are integrated by the equations of motion. Accordingly, the kinetic energy and the temperature can be adjusted by computing a scale factor

$$\rho = (K/K_{\text{desired}})^{1/2}$$

where K denotes the currently computed kinetic energy and K_{desired} denotes the desired kinetic energy. This scale factor is used to modify the velocity terms in the equation of motion. Because the kinetic energy is linearly related to the temperature, scaling the kinetic energy is equivalent to temperature scaling.

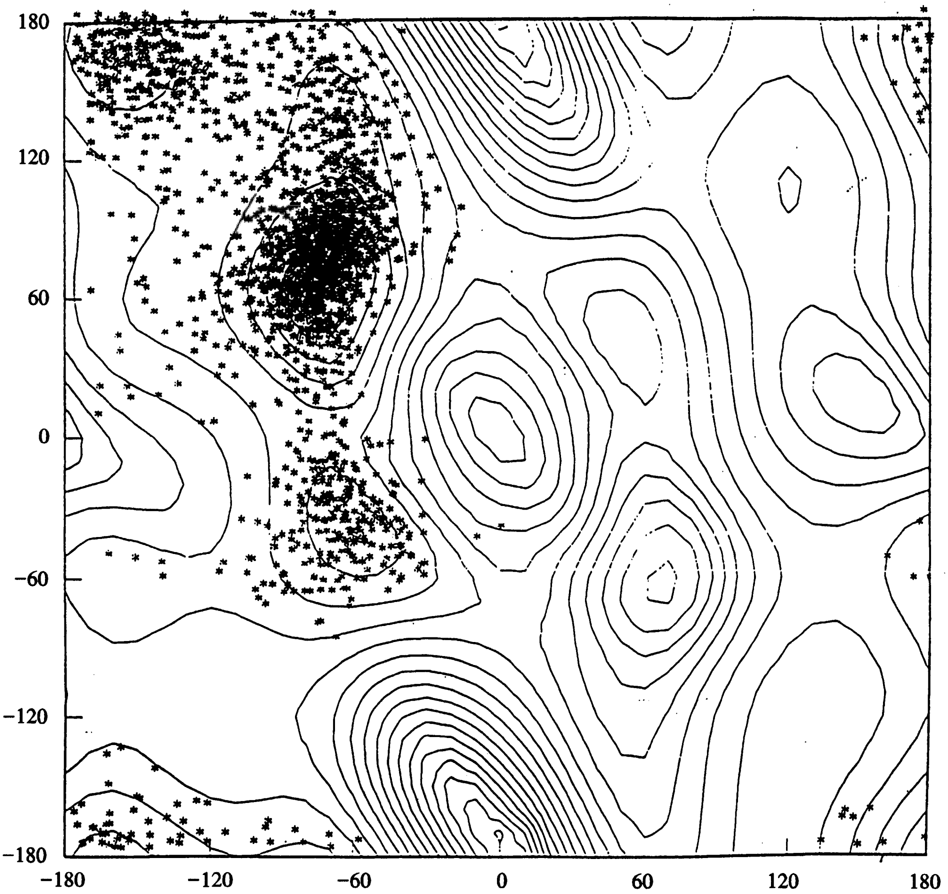


Fig. 3. 200 ps AMBER simulation with SHAKE bond constraints. Starting value of $\phi = -60$ and $\psi = 60$.

Results

Figures 3 and 4 show the results from all-atom AMBER simulations of alanine dipeptide with SHAKE bond constraints (22 atoms with 39 degrees of freedom). These runs were carried out at constant energy and only varied according to their starting conformation and length of simulation. In both of these runs, the system was heated to 600 K over 5 ps and the runs were made with a 1 fs step size. All of the following plots are made on ϕ/ψ contours generated with all other degrees of freedom minimized at each point.

Figures 5–9 give the results for the rigid-body simulations with a total of five degrees of freedom (three methyl torsional angles and ϕ and ψ dihedral angles with

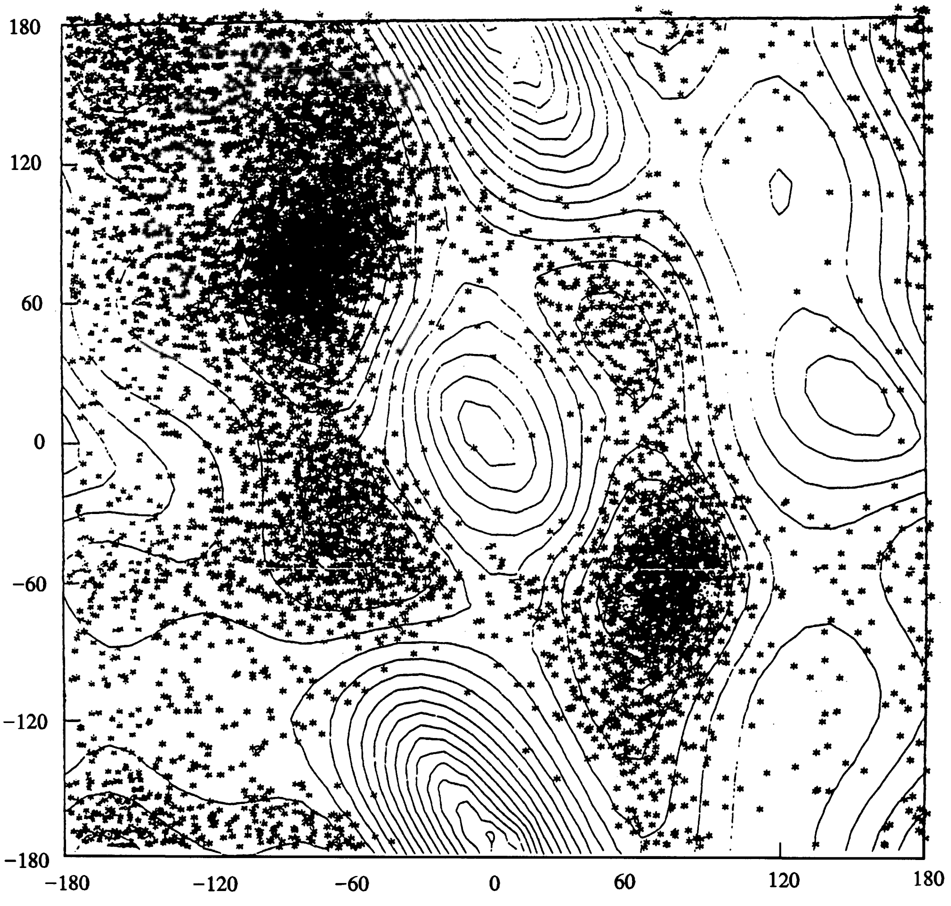


Fig. 4. 1600 ps AMBER simulation with SHAKE bond constraints. Starting value of ϕ = — 60 and ψ = 60.

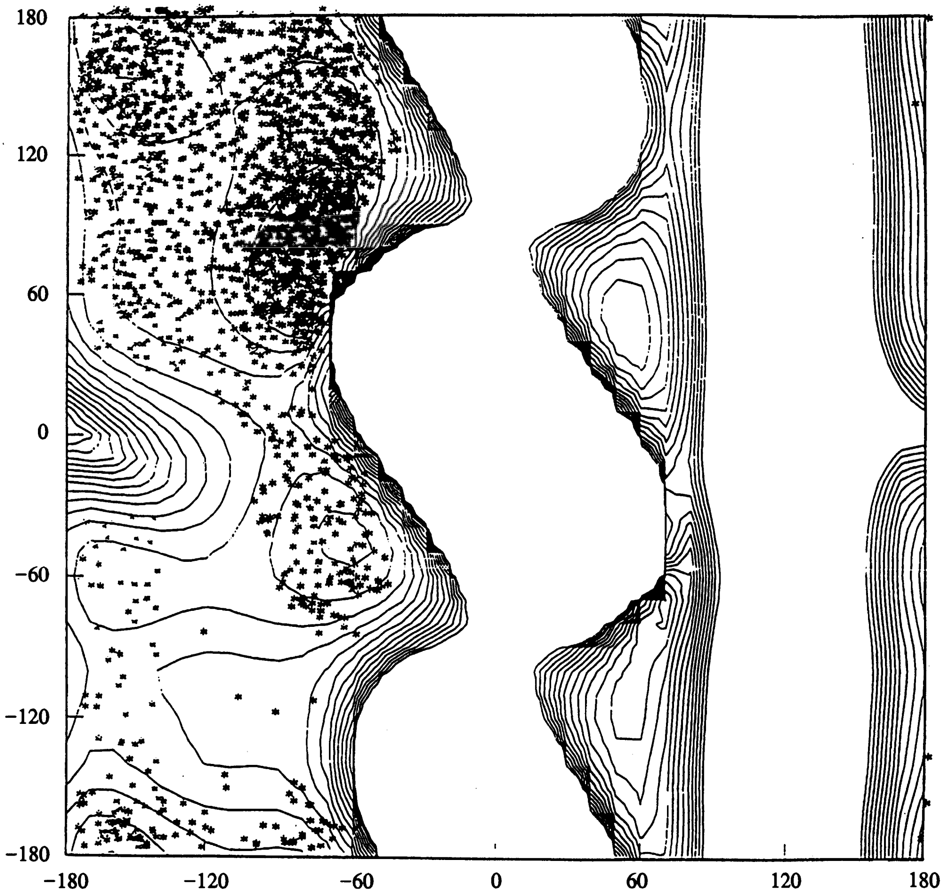


Fig. 5. 400 ps multibody simulation with a 2 fs step size. Starting value of $\phi = -60$ and $\psi = 60$.

a total of six rigid bodies). The runs are plotted on ϕ/ψ contours generated with rigid-body rotations about the ϕ/ψ torsional angles. These simulations are run at either constant energy (Figs. 5–7) or constant temperature (Figs. 8 and 9). The run in Fig. 5 uses 300 K AMBER velocities as a starting point to compute the rigid-body velocities. The runs in Figs. 6–9 use the same initial velocities as used in the run in Fig. 5, but with each component multiplied by the square root of 2 (doubling the amount of KE). Both the ability of the multibody simulation to search conformational space and to locate the global minimum using annealing algorithms were investigated. The figure captions give the relevant details of each simulation, including starting point, time step, length of simulation, and type of simulation (all atom or rigid body).

Runs made at 300 K with the AMBER SHAKE algorithm did not move out of the starting point minimum at $\phi = -60$ and $\psi = 60$. At 600 K, Fig. 3 shows that the simulation can explore a few local minima, but the structure still stays relatively near the starting conformation after 200 ps. Figure 4 explores the effect of a long simulation, 1.6 ns. This simulation explores all four low-energy minima and samples other reasonably low-energy portions of the surface. It should be noted that the molecule does remain trapped for long periods near the low-energy regions. This is a common phenomenon of molecular dynamics simulations.

Figure 5 shows that the 'normal temperature' (rigid-body velocities are derived from the 300 K all-atom velocities) rigid-body simulation appears to sample phase space as well as the 600 K all-atom simulation (Fig. 3). It should be noted that the

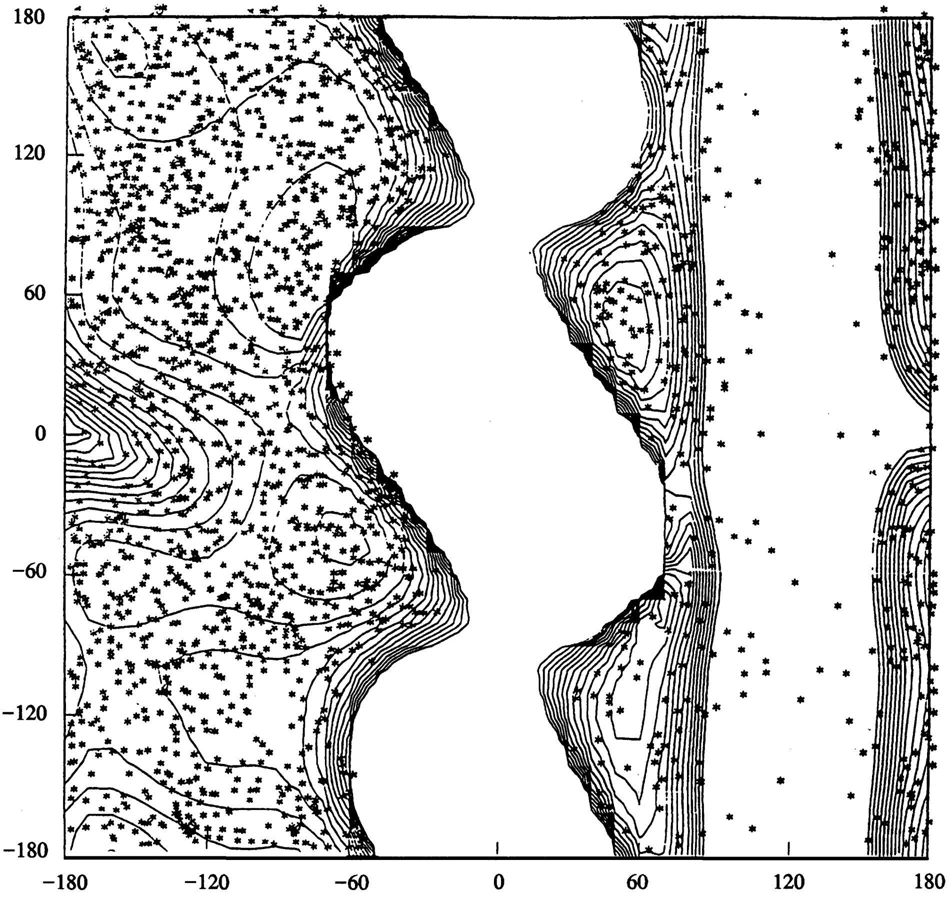


Fig. 6. 800 ps multibody simulation with a 4 fs step size. Starting value of $\phi = -60$ and $\psi = 60$.

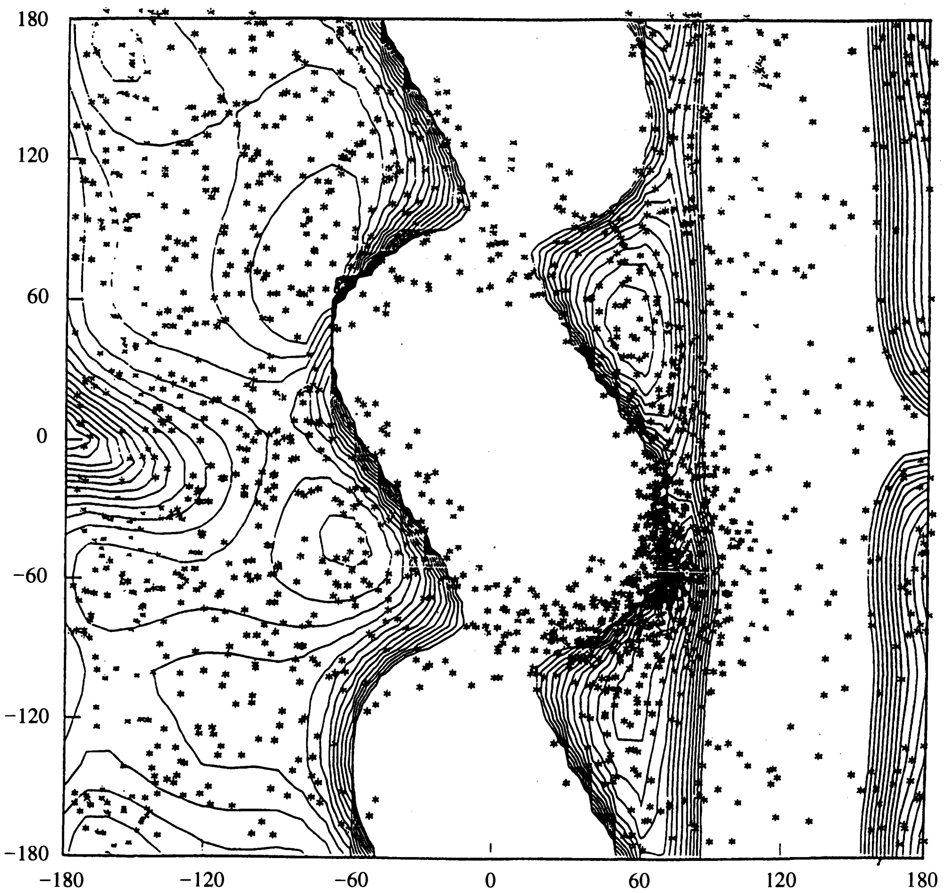


Fig. 7. 800 ps multibody simulation with a 4 fs step size. Starting value of $\phi = 60$ and $\psi = 120$.

contours for this rigid map are steeper than those of the flexible map in Fig. 3 and that it is more difficult for the simulation to move around the rigid map.

By doubling the kinetic energy or the temperature of this simulation, Figs. 6 and 7 show that the accessible regions of phase space are very well sampled independent of the starting point or the step size (2 or 4 fs). The distribution of points is very uniform and the simulation does not appear to get trapped in a local minimum for long periods of time. The molecule in Fig. 7 had to surmount barriers of 5–10 kcal to go from the minimum at $\phi = 60$ and $\psi = -120$ to the portion of the surface including the global minimum near $\phi = -60$ and $\psi = 60$. Every 100th point was saved during these runs. An analysis of the ϕ/ψ values showed frequent changes of 60° or more between points saved. This showed that energy is easily able to flow between the bodies.

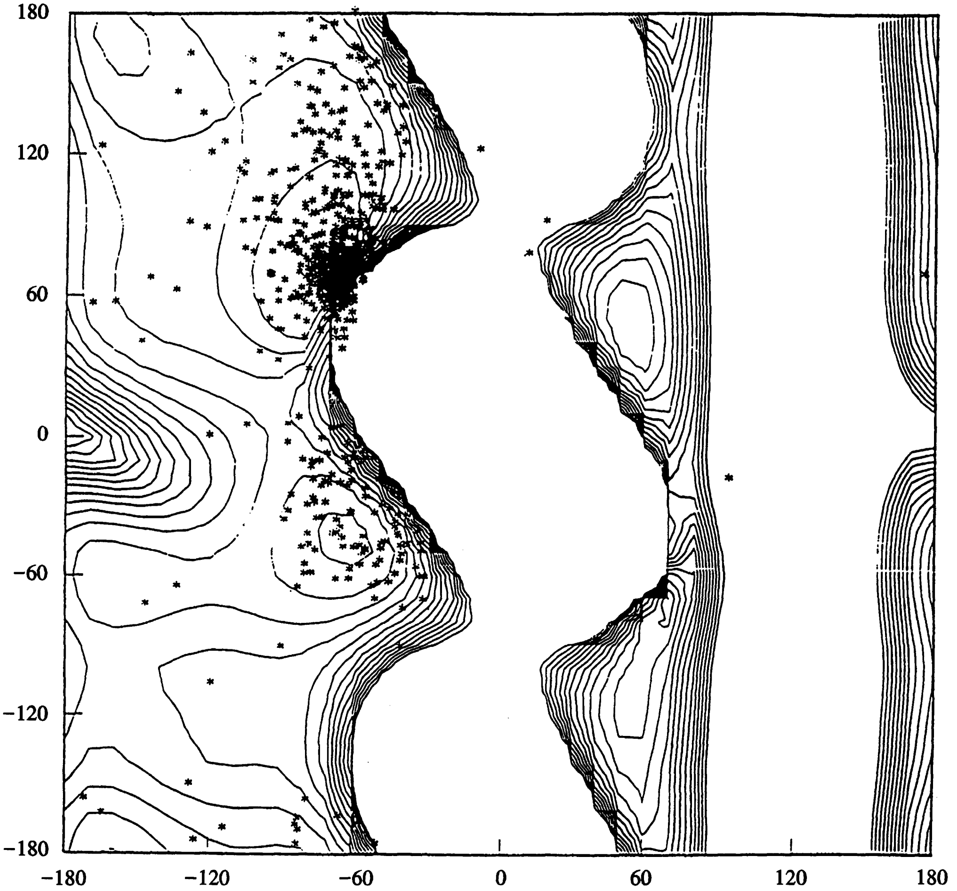


Fig. 8. 67 400-point annealing multibody run with a 4 fs step size. Starting value of ϕ = − 120 and ψ = 120.

Figures 8 and 9 show the results of heating and then cooling alanine dipeptide. All protocols were carried out over approximately a 300 ps time period. The simulations were stopped once they had been cooled sufficiently low that further movement on the energy surface was unlikely. Only the simulation carried out in Fig. 9 seemed to depend on the cooling protocol. If cooling began while the molecule was over the center strip, the simulation converged to the minimum near $\phi = 60$ and $\psi = -60$. If cooling began over the other regions, the simulation tended to converge to the global minimum near $\phi = -60$ and $\psi = 60$.

One question that might be asked is whether or not a multibody run is equivalent to an all-atom molecular mechanics run at a greatly raised temperature. This question is raised since the formula for temperature has a factor in the denominator that is equal

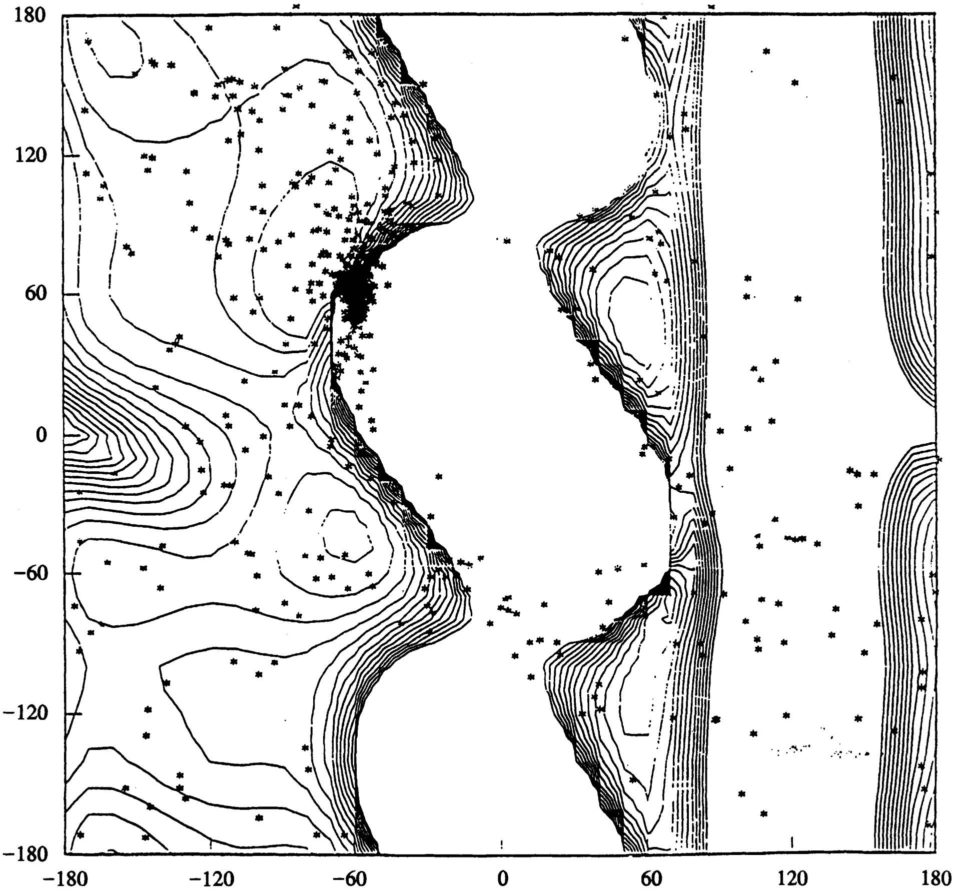


Fig. 9. 67 400-point annealing multibody run with a 4 fs step size. Starting value of $\phi = 60$ and $\psi = -120$.

to the number of degrees of freedom. Thus, one would expect that a simulation with only a few degrees of freedom would have a much higher temperature than with all degrees of freedom.

To address this question, the following computer experiment was performed. A 300 K, 800 ps, all-atom AMBER run with only SHAKE constraints was made for n-butane. This is a system with only one significant degree of freedom (rotation about the middle two carbon atoms). A step size of 1 fs was used. This run only had 0 or 1 dihedral transitions, depending on the starting point. Figure 10 shows a typical plot of the torsion angle versus the dynamics step.

The velocities of all of the atoms were taken from the starting point of this run (which had been equilibrated at 300 K) and used in the multibody program.

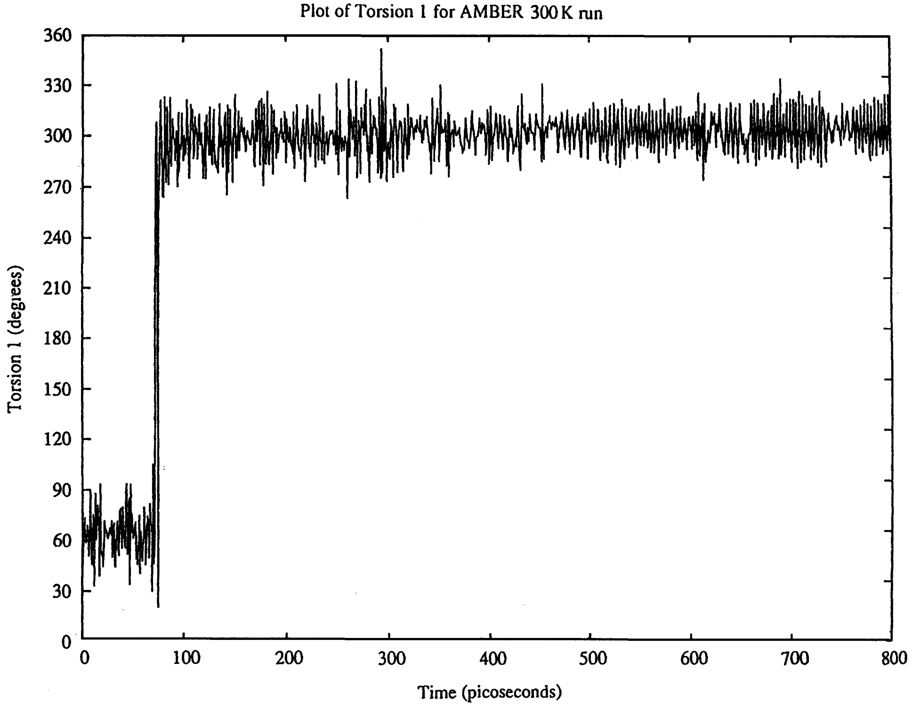


Fig. 10. Torsion 1 versus time for 300 K AMBER simulation.

These atom velocities were then mapped onto the body velocities. The body velocities were used to start the multibody simulation, which was run for 800 ps with a step size of 6 fs. The same total kinetic energy was used in both simulations. This simulation, as shown in Fig. 11, had many transitions.

Another run was made with the all-atom AMBER + SHAKE algorithm. This time the atoms were heated to 2400 K (the necessary temperature required to match the theoretical temperature of the multibody run). This run, as shown in Fig. 12, had many more dihedral transitions than the multibody run. However, many points were in higher energy regions than those in the multibody run.

This is graphically shown in Figs. 13 and 14. Figure 13 shows that the multibody simulation samples the dihedral values in a manner similar to that of an all-atom 300 K AMBER simulation. Figure 14 shows that the 2400 K all-atom AMBER simulation samples the high-energy states much more frequently and thus has a flatter distribution of dihedral values.

Therefore, the heated all-atom AMBER run and the multibody run are not equivalent. The multibody run avoids the problem of energy transfer between the high- and low-frequency modes by freezing the group geometries and only adding the

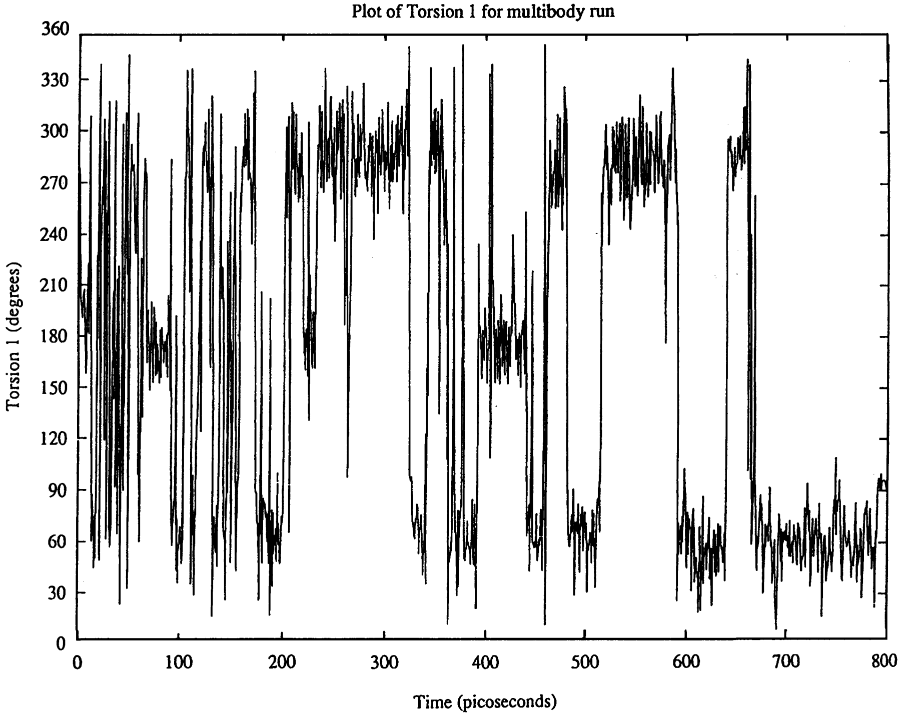


Fig. 11. Torsion 1 versus time for multibody simulation.

energy to the low-frequency modes that the room temperature all-atom AMBER run would have.

Two important issues that arise when comparing this method to more traditional all-atom methods are the speedup in the total simulation time and the amount of time required for convergence of the method. The first issue of speedup is only partially addressed in this study. The speedup in the use of multibody technology results from (i) the use of a much simpler energy surface that can be searched more quickly because there are many fewer local minima [12] (many fewer degrees of freedom), (ii) the increase in step size possible because of the reduction in high-frequency motion (it should be noted that each step, even if taken with the same step size as the all-atom simulation, is taken using dihedral internal coordinates; this allows a much larger motion of the atoms than a step using Cartesian coordinates with all atoms moving independently), and (iii) the decrease in time for computing the energy function that arises because a multipole expansion can be used to compute the interactions between the rigid bodies.

The reduction in the number of degrees of freedom in this problem is approximately eight (39 degrees of freedom in the SHAKE all-atom simulation and five degrees of freedom in the reduced variable simulation). The speedup from step size because of the

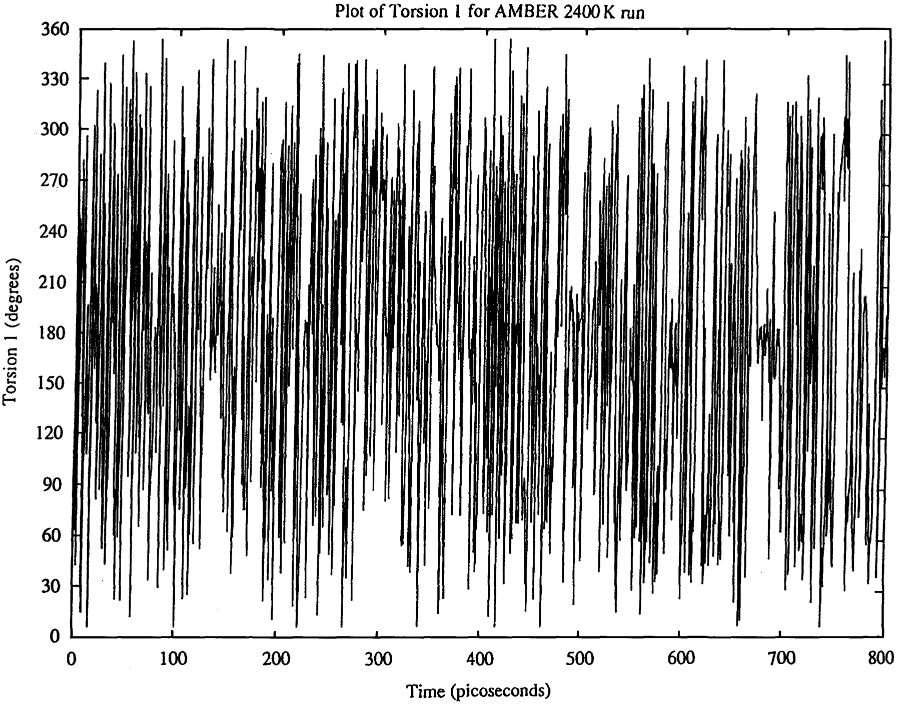


Fig. 12. Torsion 1 versus time for 2400 K AMBER simulation.

reduction in high-frequency motion is a factor of 2–4 (1 fs versus 4 fs). While step sizes as large as 2 fs may be carried out using the SHAKE algorithm, simulations at high temperatures and those for large molecules with starting points far from stable low-energy configurations may require a step size nearer 1 fs for stability of the simulation. This gives a speedup of 16–32. Large molecules will have possibilities for greater reductions in the degrees of freedom because of the possible groupings of larger numbers of atoms (i.e., the atoms in a helix).

There is also the possibility of reduction in the time for the energy function evaluation through the use of multipole expansions. Previous work [23,24] has shown a 44 × reduction in the time required to evaluate the electrostatic interaction energy between two helices using fourth-order multipole expansions. In an actual problem requiring the use of a mixture of methods for calculating short- and long-range interactions, this factor will decrease but should be at least a factor of 2. Combining all of these factors gives a conservative estimate of 10–100 times possible speedup in simulation time using this methodology.

Of course, in a speedup comparison, one must also take into account the overhead of the multibody portion of the code. There is no measurable cost for communication between the molecular mechanics and multibody code since a single program is used

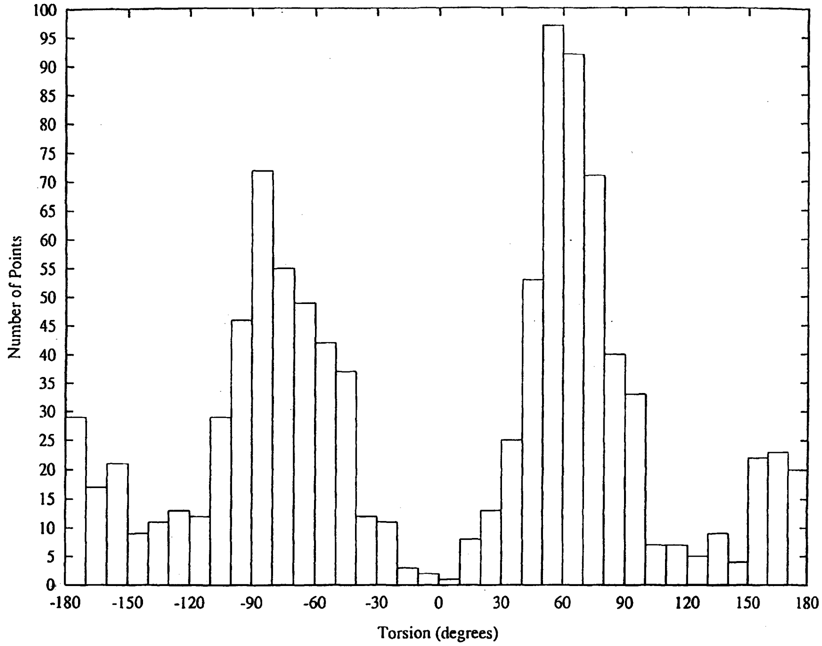


Fig. 13. Histogram of torsion values in multibody simulation.

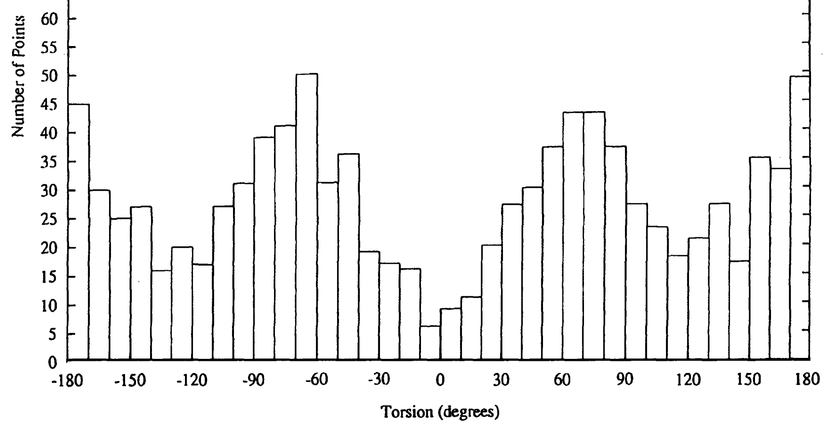


Fig. 14. Histogram of torsion values in 2400 K AMBER simulation.

and information is passed through common blocks. The portion of time required for the multibody software overhead rapidly decreases as the time for the energy function evaluation increases (i.e., for larger problems). For small problems, such as the one here, the total multibody overhead time compared to the total time for the simulation is 75–90% of the total time. This proportion decreases to 10–30% for simulations of moderate-sized molecules (approximately 500 atoms and 100 rigid bodies).

Conclusions

The RVMD software appears to be stable for step sizes larger than those permitted in a conventional all-atom simulation and can be used to reduce easily the number of degrees of freedom in a problem. Since energy easily flows from one body to another without getting trapped in small-amplitude high-frequency motion, the molecule can more easily sample many low-energy regions of phase space. Larger problems are being run to test the scaling properties of the method and integrators that require fewer energy evaluations are being investigated. The method appears to be a promising computational tool for drug design.

Acknowledgements

Two of the authors (J.T. and P.K.W.) wish to thank Proteus for its generous hospitality and support for these studies, which were carried out at its headquarters in Macclesfield, U.K. All of the authors wish to thank Jin Li, Chris Murray, and Eric Platt for helpful scientific discussions, Brian Petch and Richard Walker for computational assistance, and Peter Brady for general guidance. They are also grateful to John Pool for inviting them to England and making this work possible.

References

1. Brooks III, C.L., Karplus, M. and Pettitt, B.M., Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, Wiley, New York, NY, 1988.
2. Van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulations of Biomolecular Systems, Vol. 2, ESCOM, Leiden, 1994.
3. Byrne, D., Li, J., Platt, E., Robson, B. and Weiner, P., J. Comput.-Aided Mol. Design, 8(1994)67.
4. Robson, B., 'Computer-aided biomolecular engineering', Cyber 205 Newsletter, Winter, 1985, p. 5.
5. Robson, B., In Darbre, A. (Ed.) Practical Protein Chemistry – A Handbook, Wiley, New York, NY, 1986, pp. 567–607.
6. Robson, B., In Hadzi, D. and Jerman-Blazic, B. (Eds.) QSAR in Drug Design and Toxicology, Elsevier, Amsterdam, 1987, pp. 239–245.

7. Li, J., Platt, E., Waskowycz, B., Cotterill, R.M.J. and Robson, B., Biophys. Chem., 43(1992)221.
8. Pear, M.R. and Weiner, J.J., J. Chem. Phys., 71(1979)212.
9. Ryckaert, J.P., Mol. Phys., 55(1985)549.
10. Mazur, A.K. and Abagyan, R.A., J. Biomol. Struct. Dyn., 6(1989)815.
11. Abagyan, R.A. and Mazur, A.K., J. Biomol. Struct. Dyn., 6(1989)833.
12. Abagyan, R.A., Totrov, M. and Kuznetsov, D., J. Comput. Chem., 15(1994)488.
13. Rudnicki, W.R., Lesyng, B. and Harvey, S.C., Biopolymers, 34(1994)383.
14. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23(1977)327.
15. Van Gunsteren, W.F. and Berendsen, H.J.C., Mol. Phys., 34(1977)1311.
16. Andersen, J.C., J. Comput. Phys., 52(1983)24.
17. Van Gunsteren, W.F. and Karplus, M., Macromolecules, 15(1982)1528.
18. Tobias, D.J. and Brooks III, C.L., J. Chem. Phys., 89(1988)5115.
19. Zhang, G. and Schlick, T., J. Comput. Chem., 14(1993)1212.
20. Van Gunsteren, W.F., Mol. Phys., 40(1980)1015.
21. Turner, J.D., Chun, H.M., Weiner, P., Gallion, S. and Singh, C., 'Order (n) multibody dynamics', Conference on Research Perspectives in Structural Biology and Chemistry, Hilton Head, SC, 27–30 January 1991.
22. Turner, J.D., Chun, H.M., Weiner, P., Gallion, S. and Singh, C., 'Order (n) multibody dynamics', 7th International Congress of Quantum Chemistry, Menton, France, 2–5 July 1991.
23. Turner, J.D., Chun, H., Lupi, V., Weiner, P., Gallion, S. and Singh, C., Chem. Design Automation News, 7(12)(1992)34.
24. Turner, J.D., Chun, H., Lupi, V., Weiner, P., Gallion, S. and Singh, C., Chem. Design Automation News, 8(1)(1993)16.
25. Singh, R.P., VanderVoort, R.J. and Likens, P.W., TREETOPS User's Manual, Dynacs Engineering Company, 1990.
26. Singh, R.P. and Ravi, V., Code Generator User's Document, Internal Report 920625-1, Dynacs Engineering Company, 1992.
27. Venugopal, R. and Kumar, M., Proceedings of the 5th NASA/NSF/DOD Workshop on Computational Control, 15 February 1993.
28. Whittaker, E.T., A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, 4th ed., Cambridge University Press, Cambridge, 1961.
29. Meirovitch, L., Methods of Analytical Dynamics, McGraw-Hill, New York, NY, 1970.
30. Frisch, H.P., Chun, H.M. and Turner, J.D., NDISCOS – Users and Programmers Manual, Photon Research Technical Report, December 1992.
31. Chun, H.M., Turner, J.D. and Frisch, H.P., AAS/AIAA Astrodynamics Specialists Conference, Durango, CO, August 1991.
32. Chun, H.M., Turner, J.D. and Frisch, H.P., 'Recursive multibody formulations for robotic applications with harmonic drives', Presented to International Conference on Dynamics of Flexible Structures in Space, Cranfield, U.K., 15–18 May 1990.
33. Kane, R.R. and Wang, C.R., J. Siam, 13(1965)2.
34. Kane, T.R. and Levinson, D.A., J. Guidance Control, 3(1980)2.
35. Singh, R.P. and Likins, P.W., Automatic Control Conference, San Francisco, CA, June 1983.
36. Kane, T.R. and Levinson, D.A., Dynamics: Theory and Applications, McGraw-Hill, New York, NY, 1985.

37. Turner, T., Weiner, P., Robson, B., Venugopal, R., Schubele III, H. and Singh, R., J. Comput. Chem., 16(1271)10.
38. AMBER+ is a fully vectorized molecular mechanics code by Singh, U.C., Ramnarayan, K., Weiner, P.K. and Kollman, P. This code is distributed by Amber Systems Inc.
39. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, G., Alagona, G., Profeta, S. and Weiner, P.K., J. Am. Chem. Soc., 106(1985)765.
40. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7(1986)230.

Gaussian shape methods

J.A. Grant[a] and B.T. Pickup[b]

[a] *Zeneca Pharmaceuticals, Mereside, Macclesfield, Cheshire SK10 4TF, U.K.*
[b] *Centre for Molecular Materials, Department of Chemistry, The University of Sheffield, Sheffield S3 7HF, U.K.*

Introduction

The aim of this article is to give an overview of recent developments in Gaussian 'shape methods', with an emphasis on their advantages in biological and macromolecular applications.

The term 'shape methods' refers to a whole range of techniques in which molecules are represented as atom-centered overlapping hard spheres. The applications of such techniques include the simple computation of a molecular volume and surface area, or the construction of elaborate protein surfaces. The origins of these ideas lie in the chemist's early appreciation of steric factors, i.e. many molecular properties rely on the fact that atoms in molecules have relatively impenetrable cores. These effects have been explained by the later discovery of the quantum mechanical nature of electronic structure. Excluded-volume effects arise because of the high energy required when wave functions from neighboring molecules begin to overlap.

Many physical properties depend, to some extent, on excluded volume or shape effects (including changes in shape). For example, the volume and surface area are needed to construct the equation of state for a liquid comprising rigid polyatomic molecules, and in understanding the molecular packing of liquids and crystals. The phase behavior of complex fluids is thought to arise partly from shape effects; for example, rod-like molecules exhibit nematic, smectic and other complex phases. Simple models of hydration describe the free energy of solvation in terms of atomic contributions that are proportional to surface area or volume. A central idea in drug design is the concept of 'lock and key', in which the binding of a ligand to a macromolecular receptor is closely related to the 'goodness of fit' into the active site. The binding of a ligand is a matter of finding an appropriate local minimum on a potential energy binding surface. The potential energy surface in question is normally a classical potential energy surface which is of parametrized form. The actual physical process of ligand binding is much more complicated; the real potential energy surface should be a quantum mechanical one, and indeed computed properties should be a statistical mechanical average.

This article is particularly concerned with two aspects of shape methods, namely lock-and-key matching and key comparison. Key comparison involves the comparison between shapes of putative ligands. Because molecules that give rise to similar pharmacological responses often show a degree of molecular shape similarity, the characterization of shape, and the evaluation of shape similarity among molecules (keys), is an important tool for rational drug design. Lock-and-key matching refers to our own implementation of the rigid-body geometric docking problem. This procedure attempts to determine the optimal surface complementarity of a molecular complex using very efficient Gaussian technology. This methodology is a precursor to a more complete solution of the ligand binding problem, in which some aspects of the free energy of association are modelled.

A goal of both of these techniques is to act as a low-resolution but rapid computational screen of large three-dimensional databases of molecules, to identify small numbers of potential ligands for receptors, both in the case when the receptor structure is completely unknown (key comparison) or when some information about receptor structure is available (lock-and-key matching).

This article presents a Gaussian description of shape which allows an analytical description. That is to say, that volume, area, shape assessment and matching algorithms are given by continuous and smooth functions. The properties of the Gaussian function lead to very efficient algorithms. The low computational overhead in obtaining nuclear-coordinate derivatives enables us to use the apparatus of optimization theory to compare the shapes of ligands, and dock molecules geometrically into their receptor sites.

The recent comprehensive treatise on molecular shape by Mezey [1] gives the following criteria required of an ideal description of molecular shape, i.e. that it

1. is based on the physical properties of the molecule,
2. describes the full, three-dimensional shape of the molecule,
3. leads to numerical shape characterization, such as a numerical shape code,
4. is easily computable, leading to computer-based molecular shape analysis,
5. is reproducible and objective,
6. provides tools for the evaluation of shape similarity, and
7. provides tools for the evaluation of shape complementarity.

This chapter attempts to highlight how a simple Gaussian description of molecular shape [2,3] fulfills these requirements. We do not attempt to describe the numerous applications of the Gaussian function in chemistry, because we have reviewed the relevant applications elsewhere [2,4]. However, we note that the work described here begins from the ideas introduced into quantum chemistry by Boys [5], in which Gaussian functions represent atomic orbitals. The more general description of atoms as spherical Gaussians has been used by other authors to construct *different* models of molecular shape [6–10] from those described in this chapter. Central to these has been the very innovative work of Good and Richards [11,12].

Gaussians and their properties

The primary purpose of this section of the article is to introduce Gaussians and their properties. In order to retain a practical orientation, however, we shall do this from the view of molecular shape. It is necessary, therefore, to introduce 'hard-sphere' shape techniques as a preamble.

We consider an 'atom' A with coordinates $\mathbf{R}_A = (X_A, Y_A, Z_A)$ which is a center for the hard-sphere density

$$\rho^{hs} = \begin{cases} 1, & 0 \leq r_A \leq \sigma_A \\ 0, & \sigma_A < r_A \end{cases} \tag{1}$$

where the local radial coordinate is defined by the equation

$$r_A^2 = (x - X_A)^2 + (y - Y_A)^2 + (z - Z_A)^2 \tag{2}$$

and (x, y, z) are coordinates of a point specified in some global coordinate system. It is obvious that the three-dimensional integral

$$\begin{aligned} V_A &= \int d\mathbf{r}\, \rho_A^{hs} \\ &= 4\pi \int_0^\infty dr_A\, r_A^2\, \rho_A^{hs} \\ &= \tfrac{4}{3}\, \pi\sigma_A^3 \end{aligned} \tag{3}$$

gives the volume of atom A which has radius σ_A. In this article we shall use a convention for integration in which an integral over three-dimensional space implies integration over all space, even when no integration limits are specified. This convention is widely used in quantum chemistry. We also define a spherical Gaussian density on atom A:

$$\rho_A^g = p_A \exp(-\alpha_A r_A^2) \tag{4}$$

where p_A is a 'height' factor, and the exponent

$$\alpha_A = \kappa_A/\sigma_A^2 \tag{5}$$

where κ_A is dimensionless. The Gaussian and hard-sphere densities are depicted in Fig. 1. The factors p_A, κ_A are regarded as parameters which can be adjusted appropriately to give an atomic volume which agrees with the hard-sphere equivalent. Hence,

$$\begin{aligned} V_A &= \int d\mathbf{r}\, \rho_A^g \\ &= 4\pi p_A \int_0^\infty dr_A\, r_A^2 \exp(-\alpha r_A^2) \\ &= p_A \left(\frac{\pi}{\alpha}\right)^{3/2} \\ &= p_A \left(\frac{\pi}{\kappa_A}\right)^{3/2} \sigma_A^3 \end{aligned} \tag{6}$$

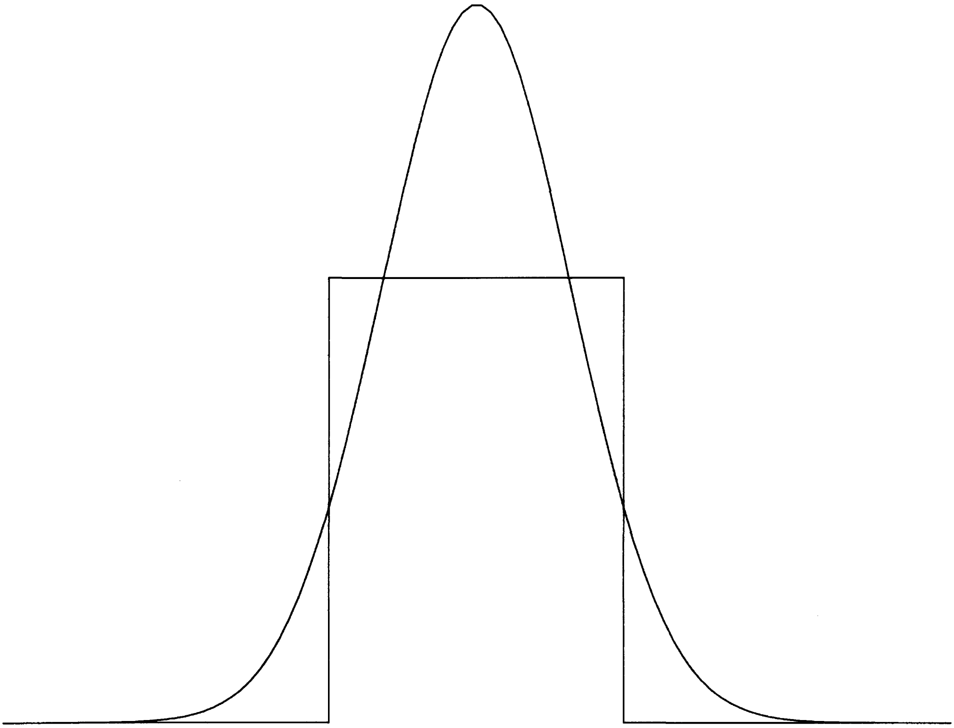


Fig. 1. A cross-section through hard-sphere (ρ_A^{hs}) and Gaussian densities (ρ_A^g) defined for center A.

It follows that the adjustable parameters must satisfy the constraint

$$p_A \left(\frac{\pi}{\kappa_A}\right)^{3/2} = \frac{4\pi}{3} \tag{7}$$

Evidently, we are free to vary the Gaussian height, in which case κ_A is fixed by Eq. 7, or *vice versa.* The philosophy surrounding the introduction of a 'soft' sphere is chiefly concerned with its technical advantages. We emphasize that ρ_A^g is not a probability density unless $p_A = 1$, but we prefer to retain the possibility of nonunit heights. ρ_A^g is best regarded as a device for the computation of molecular volumes, areas and other associated shape-dependent factors. The purpose of the rest of this section is to outline the technical advantages of Gaussians. These technical advantages rely on the following important features: (i) ease of integration and (ii) the Gaussian product theorem. The ease of Gaussian integration depends on the standard integral

$$\begin{aligned} I_n &= \int_{-\infty}^{\infty} dx\, x^{2n} \exp(-\alpha x^2) \\ &= \frac{(2n-1)!!}{(2\alpha)^n} \left(\frac{\pi}{\alpha}\right)^{1/2} \end{aligned} \tag{8}$$

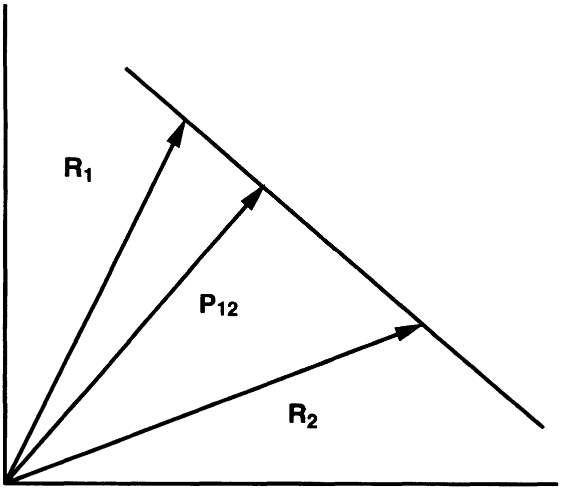


Fig. 2. The coalescence center, P_{12}, for a pair of Gaussians at centers R_1 and R_2 lies along the line between the two centers at a position determined by the relative exponents.

where the double factorial symbol represents just the product of odd integers down to unity, and $((-1)!! = 1)$. From Eq. 8 one can obtain the volume integral 6 and moment integrals, such as

$$\int \mathrm{d}\mathbf{r}\, x^{l} y^{m} z^{n} \rho_A^{g} \tag{9}$$

If we consider two Gaussians with exponents α_i $(i = 1, 2)$ centered at $\mathbf{R}_i$, the product satisfies

$$\exp(-\alpha_1 r_1^2)\exp(-\alpha_2 r_2^2) = K_{12}\exp(-\alpha_{12} r_{12}^2) \tag{10}$$

where

$$\alpha_{12} = \alpha_1 + \alpha_2 \tag{11}$$

and $\mathbf{r}_{12} = \mathbf{r} - \mathbf{P}_{12}$, with the product center

$$\mathbf{P}_{12} = \frac{\alpha_1}{\alpha_{12}}\mathbf{R}_1 + \frac{\alpha_2}{\alpha_{12}}\mathbf{R}_2 \tag{12}$$

along the line between $\mathbf{R}_1$ and $\mathbf{R}_2$ in the ratio of the exponents (see Fig. 2). The constant

$$K_{12} = \exp\left(\frac{-\alpha_1\alpha_2 R_{12}^2}{\alpha_{12}}\right) \tag{13}$$

Equation 10 implies that a product of two Gaussians is also a Gaussian, but centered at the coalescence point $\mathbf{P}_{12}$ defined in Eq. 12. It can be shown by induction that a product of n spherical Gaussians, each centered at $\mathrm{R_i}$, can be written as

$$\prod_{i=1}^{n} \exp(-\alpha_i r_i^2) = K_{12\cdots n} \exp(-\alpha_{12\cdots n} r_{12\cdots n}^2) \tag{14}$$

where

$$\alpha_{12\cdots n} = \sum_{i=1}^{n} \alpha_i$$

and

$$\mathbf{r}_{12\cdots n} = \mathbf{r} - \mathbf{P}_{12\cdots n}$$

with the coalescence center

$$\mathbf{P}_{12\cdots n} = \frac{1}{\alpha_{12\cdots n}} \sum_{i=1}^{n} \alpha_i \mathbf{R}_i \tag{15}$$

and the coalescence constant

$$K_{12\cdots n} = \exp\left\{ -\frac{1}{\alpha_{12\cdots n}} \sum_{i>j} \alpha_i \alpha_j R_{ij}^2 \right\} \tag{16}$$

The importance of Eqs. 10–16 lies in the fact that a multiple Gaussian product is a single Gaussian centered at a coalescence point. It follows that multiple Gaussian products can be integrated to give analytical expressions for moment integrals, such as Eq. 9.

We now turn to the representation of a molecule, $\mathcal{M}$, as a set of overlapping hard spheres (see Fig. 3). It is obvious that we must write the volume as a series

$$V_{\mathcal{M}} = \sum_{A} V_A - \sum_{A>B} V_{A\cap B} + \sum_{A>B>C} V_{A\cap B\cap C} - \cdots \tag{17}$$

involving the volumes of individual atoms, together with corrections to allow for pair and higher overlaps. We want to write the molecular volumes as an integral (equivalent to Eq. 3) over a molecular density written as

$$\rho_{\mathcal{M}}^{hs} = \sum_{A} \rho_A^{hs} - \sum_{A>B} \rho_{A\cap B}^{hs} + \sum_{A>B>C} \rho_{A\cap B\cap C}^{hs} - \cdots \tag{18}$$

in which the individual terms are densities achieving unit values only inside the designated region, for example

$$\rho_{A_1\cap A_2\cdots\cap A_n}^{hs} = \begin{cases} 1 & \text{inside } A_1 \cap A_2 \cdots \cap A_n \\ 0 & \text{elsewhere} \end{cases} \tag{19}$$

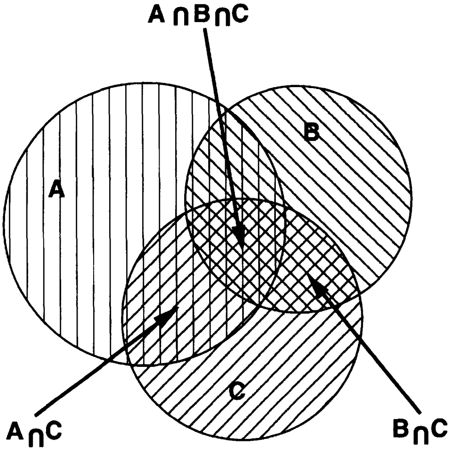


Fig. 3. A molecule envisaged as a set of overlapping hard-spheres.

The unit density functions in Eq. 1 satisfy an obvious multiplicative property

$$\rho^{hs}_{A_1 \cap A_2 \cdots \cap A_n} = \prod_{i=1}^{n} \rho^{hs}_{A_i} \tag{20}$$

i.e. the product is only nonzero in the intersection region. It follows that Eq. 18 can be rewritten through all orders in product form as

$$\rho^{hs}_{\mathcal{M}} = 1 - \prod_{i=1}^{n} (1 - \rho^{hs}_{A_i}) \tag{21}$$

This expression includes corrections for intersections through all orders. The intersection volume is just the integral of the intersection density Eq. 20:

$$V^{hs}_{A_1 \cap A_2 \cdots \cap A_n} = \int \mathbf{dr}\, \rho^{hs}_{A_1 \cap A_2 \cdots \cap A_n} \tag{22}$$

The integration of expressions such as Eq. 22 is extremely difficult. Analytical formulae exist for low-order intersections, but these are computationally expensive, as are their nuclear-coordinate derivatives.

The generalization of Eqs. 18–22 in the Gaussian context is simple. We repeat all of the hard-sphere formulae, but replacing hard-sphere densities by their Gaussian equivalents. Hence, we write

$$\rho^{g}_{\mathcal{M}} = \sum_{A} \rho^{g}_{A} - \sum_{A>B} \rho^{g}_{A \cap B} + \sum_{A>B>C} \rho^{g}_{A \cap B \cap C} - \cdots \tag{23}$$

with a corresponding expression like Eq. 17 for the molecular volume. The intersection densities in Eq. 23 are defined by analogy with Eq. 20 as

$$\rho^{g}_{A_1 \cap A_2 \cdots \cap A_n} = \prod_{i=1}^{n} \rho^{g}_{A_i} \tag{24}$$

It may be argued that Eq. 24 does not have the property exhibited in Eq. 19 because the Gaussians do not represent uniform shape-density distributions. We can see, however, that, by virtue of the Gaussian coalescence theorem, Eq. 14, the intersection density in Eq. 24 is itself a spherical Gaussian centered in the region where the intersection is to be found. We write the full Gaussian density as

$$\rho^g_{\mathcal{M}} = 1 - \prod_{i=1}^{n} (1 - \rho^g_{A_i}) \tag{25}$$

and the Gaussian intersection volume as

$$V^g_{A_1 \cap A_2 \cdots \cap A_n} = \int \mathrm{d}\mathbf{r}\, \rho^g_{A_1 \cap A_2 \cdots \cap A_n} \tag{26}$$

Equation 26 can be integrated analytically using the Gaussian coalescence theorem 14. It is not possible to integrate Eq. 25 analytically without expanding the product in terms of intersection densities. The total volume then has to be constructed using the series in Eq. 17.

Areas can be derived by the simple expedient of differentiation with respect to sphere radii. Hence, referring to Eq. 3, the surface area of a sphere is

$$A_A = \frac{\partial V_A}{\partial \sigma_A} \tag{27}$$

We extend this formula to the Gaussian case by writing

$$\begin{aligned} A_{\mathcal{M}} &= \sum_{i=1}^{n} \frac{\partial V_{\mathcal{M}}}{\partial \sigma_{A_i}} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \sigma_{A_i}} \int \mathrm{d}\mathbf{r}\, \rho^g_{\mathcal{M}} \end{aligned} \tag{28}$$

The analytical formulae for the area, $A_{\mathcal{M}}$, have been presented by us previously [2]. Such formulae can be understood by the observation that

$$\frac{\partial V_{\mathcal{M}}}{\partial \sigma_{A_i}} \tag{29}$$

is the volume of an infinitesimal shell created by increasing the radius of a single Gaussian 'atom' from σ_{A_i} to $\sigma_{A_i} + \mathrm{d}\sigma_{A_i}$. Equation 29 gives the area contribution from atom A_i including the appropriate allowance for intersections. The partial derivatives can be viewed therefore as 'area' contributions arising from individual atoms. The sum of such contributions is the total molecular area. It should be stressed that, in the hard-sphere domain, the measure of surface area can be interpreted as the amount of 'paint' required to color an object, i.e. an actual surface exists. Gaussian areas tell you how much 'paint' is required for an analogous object, but unfortunately not where to put it. That is, Gaussians have no surface, although one can readily and efficiently display contour iso-surfaces [13]. Equation 28 is a mathematical device which enables accurate and efficient computations. For any practical optimization method in which molecular shape is part of the function being minimized or maximized, the positional gradient, and preferably the Hessian of the molecular shape, is required. The simplicity of function 26 for the terms that comprise the total Gaussian volume ensures that the coordinate derivatives are almost as simple. We have presented in Ref. 2 formulae for the first and second coordinate derivatives for the general n-fold volume intersection. These expressions give rise to an extremely efficient algorithm for computing derivatives, because the derivatives differ from the volume term only by simple linear factors involving a vector difference between the atom centers and the coalescence center.

Results

Calculations on model systems [2] established that a suitable Gaussian height parameter (see Eq. 7) to reproduce hard-sphere volumes is $p = 2.70$, while a slightly smaller Gaussian height value of $p = 2.60$ is better suited for the computation of area. However, the differences with respect to hard-sphere quantities are small when either one of these Gaussian height parameters is chosen to compute both volumes and areas. The computation of Gaussian surface area and volume for real molecules with arbitrary numbers of atoms requires a simple algorithm to reduce the combinatorial number of terms that appears in summations defining the volume 17. We have therefore described a trivial algorithm [2], based on a neighbor list approach, and retaining only summations up to sixth order in Eq. 17. Adopting such an algorithm we have computed the Gaussian volumes and areas for many hundreds of small molecules (of potential pharmaceutical interest) in Zeneca databases, and have found only small differences ($\approx$1–2%) with respect to conventional hard-sphere methods. Figure 4 illustrates the agreement between Gaussian and hard-sphere volumes for the subset of small molecules found in the Cambridge Structural Database (CSD) [14,15], in which the term 'drug' or 'activity' appears in the text qualifier field. This subset contains $\approx$ 4000 molecules, and we find an average percentage error difference of 0.6% between the Gaussian volume and the hard-sphere volume. The analogous error for the area measurement is 2.0%. These calculations each use a Gaussian height parameter of $p = 2.70$, which is not optimal for area. Tables 1 and 2 illustrate the computational and numerical performance of the Gaussian method applied to a number of proteins. The coordinates of these proteins were taken from the Brookhaven Protein Data Bank [16]. The CPU times were obtained using an Indigo R3000 (spec 92(fp) 24.2), and include the analytical computation of the first nuclear-coordinate derivative, but do not include the time required to compute the interatomic pairwise distances necessary for the algorithm. The point of separating these parts of the algorithm is that neighbor distances are usually available in precomputed form, in for example molecular mechanics/dynamics packages, in which the Gaussian shape model could be introduced as part of a simple shape-based solvation model. It can be

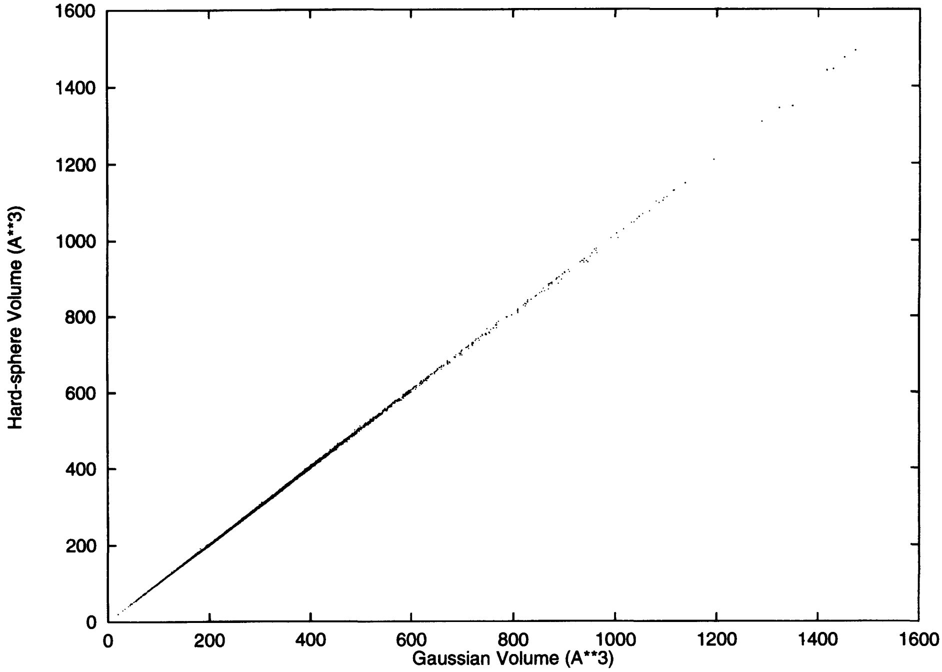


Fig. 4. Comparison of Gaussian and hard-sphere volumes.

seen that the Gaussian area and volume computation is approximately linear with respect to the number of protein atoms (N). We find that the computation of the first (or second) position gradients adds a negligible overhead relative to the computation of Gaussian volumes or areas, whereas we would expect the computation of hard-sphere derivatives for molecules with thousands of atoms to be extraordinarily

Table 1 *Comparison between Gaussian and hard-sphere volumes for a few proteins*

Protein (Brookhaven entry)	Number of residues	Gaussian volume (Å^3)	Hard-sphere volume (Å^3)	Percentage difference	CPU time (s)
1crn	46	3735	3737	0.0	0.32
2ins	100	8779	8801	0.3	0.79
5cyt	103	9070	9060	0.1	0.67
2rhe	114	9346	9358	0.1	0.73
1lz1	130	11638	11628	0.1	0.94
3fxn	138	12149	12180	0.3	0.99
3app	323	26467	26449	0.1	2.63

The CPU timings refer to the Gaussian method and include the computation of the first nuclear-coordinate derivative and the shape multipoles.

Table 2 *Comparison between Gaussian and hard-sphere areas for a few proteins*

Protein (Brookhaven entry)	Number of residues	Gaussian area (Å^2)	Hard-sphere area (Å^2)	Percentage difference	CPU time (s)
1crn	46	4222	4288	1.5	0.33
2ins	100	9821	9907	0.9	0.79
5cyt	103	10371	10528	1.5	0.67
2rhe	114	10738	10858	1.1	0.74
1lz1	130	13194	13281	0.7	0.93
3fxn	138	13805	13957	1.1	0.98
3app	323	30112	30456	1.1	2.67

The CPU timings refer to the Gaussian method and include the computation of the first nuclear-coordinate derivative and the shape multipoles.

expensive. The neighbor distance computation is roughly quadratic in N, although it should be noted that the algorithm does not require the computation of any square roots, because the method actually utilizes the square of the distance between atoms.

In comparison with hard-sphere results, Gaussian areas are less 'accurate' than volumes. It should be noted, however, that hard-sphere results are only one criterion of success for the Gaussian methodology, and probably not a very good one. The hard-sphere representation of molecular shape is not well founded in physics (or chemistry), or at least no better founded than its Gaussian equivalent. If one is looking for a methodology for the prediction of the solvation free energies of proteins, for example, based on molecular areas, then the Gaussian area is just as good as the hard-sphere area. Errors due to other factors are much more important than the precise definition of 'area' to be used. The use of continuous functions for shape representation enables the use of many mathematical techniques and provides new opportunities for the extension of shape technology simply because of the ease with which formulae can be generated by integration or differentiation. One possibility along these lines is the computation of a pointwise curvature tensor; another is the Fourier transformation method presented in the next section.

Shape characterization

The Gaussian methodology we have described leads to a very useful method for the characterization of shape in terms of a set of moment averages which can be computed analytically with trivial cost. In elementary physics one discusses the distribution of charge in a system using the concept of electrostatic moments, charges (monopoles), dipoles, quadrupoles, etc., and these quantities are well understood in elementary chemistry. The electric dipole moment, for example, is defined as

$$\mathbf{p} = \int d\mathbf{r}\, \mathbf{r}\rho^{\text{elec}}(\mathbf{r}) \tag{30}$$

where ρ^{elec} is the electrostatic charge density. The connection between the dipole moment and the charge distribution of a molecule is well known. We will introduce 'shape' multipoles based on our Gaussian density. In doing this we are doing no more than introducing moment averages like 30 which can be used as simple indices for shape comparison. The Gaussian shape analogue of Eq. 30 is the first moment, $\mathbf{S}^{(1)}$, which in component form is

$$S^{(1)}_{\alpha} = \frac{1}{V}\int \mathbf{dr}\, r_{\alpha}\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{31}$$

where we have normalized the integral with the *Gaussian* volume,

$$V = \int \mathbf{dr}\ \rho^{g}_{\mathcal{M}}\ (\mathbf{r}) \tag{32}$$

and where the subscript α represents the Cartesian direction (x, y, z). The first observation one can make is that the 'zeroth' moment, V, in Eq. 32 is invariant to a change in origin for the coordinate system. The first moment translates in a simple manner, viz. changing to a coordinate origin $(X, Y, Z) = \mathbf{R}$,

$$\left.\begin{aligned} x' &= x - X \\ y' &= y - Y \\ z' &= z - Z \end{aligned}\right\} \mathbf{r}' = \mathbf{r} - \mathbf{R} \tag{33}$$

so that

$$\mathbf{S}^{(1)'} = \frac{1}{V}\int \mathbf{dr}'\, \mathbf{r}'\ \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{34}$$

$$= \frac{1}{V}\int \mathbf{dr}\, \mathbf{r}\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) - \frac{1}{V}\mathbf{R}\int \mathbf{dr}\, \rho^{g}_{\mathcal{M}}(\mathbf{r})$$

$$= \mathbf{S}^{(1)} - \mathbf{R} \tag{35}$$

It follows that one can choose an origin like

$$\mathbf{R} = \mathbf{S}^{(1)}, \Rightarrow \mathbf{S}^{(1)'} = \mathbf{0} \tag{36}$$

that is a 'centroid' for the molecule, which makes the first moment vanish. It is now possible to define a shape quadrupole, $\mathbf{S}^{(2)}$, which in component form is

$$S^{(2)}_{\alpha\beta} = \frac{1}{V}\int \mathbf{dr}\, r_{\alpha}\, r_{\beta}\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{37}$$

This is a second-rank symmetric tensor. We can choose an axis system, called the principal axis system, such that Eq. 37 is diagonal, i.e.

$$S^{(2)}_{\alpha\beta} = \delta_{\alpha\beta}\,\varepsilon^{2}_{\alpha} \tag{38}$$

Strictly speaking, Eq. 37 is not necessarily positive definite because of the definition of $\rho^{g}_{\mathcal{M}}(\mathbf{r})$ (which can go locally negative), but in any practical case we have not seen this happen. The existence of a centroid and a principal axis system for the shape multipoles provides a natural coordinate system for the alignment of molecules. The three eigenvalues in Eq. 38 give a simple set of shape indices which assess the ellipticity of the molecule. The higher order multipoles, $\mathbf{S}^{(n)}$,

$$S^{(n)}_{\alpha_1\alpha_2\cdots\alpha_n} = \frac{1}{V}\int d\mathbf{r}\, r_{\alpha_1} r_{\alpha_2} \cdots r_{\alpha_n}\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{39}$$

are rank-n symmetric tensors. We can calculate octopolar (n = 3) and higher order shape tensors as a way of providing a more and more detailed assessment of the shape as n increases.

The shape multipoles defined above provide a set of simple indices that can be used as shape comparators. We can also use the shape multipoles as a way of producing a coarse-grained representation of molecular shape. This can be achieved by considering the Fourier transform of the molecular density:

$$\rho^{g}_{\mathcal{M}}(\mathbf{k}) = \int d\mathbf{r} \exp(i\mathbf{k}\cdot\mathbf{r})\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{40}$$

and its moment expansion in terms of the plane wave in powers of $\mathbf{k}$:

$$\rho^{g}_{\mathcal{M}}(\mathbf{k}) = V \sum_{n=0}^{\infty} \frac{(i\mathbf{k})^{n}}{n!} \cdot \mathbf{S}^{(n)} \tag{41}$$

where

$$V = \int d\mathbf{r}\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{42}$$

is the molecular volume, and the moments

$$\mathbf{S}^{(n)} = \frac{1}{V}\int d\mathbf{r}\, \mathbf{r}^{n}\, \rho^{g}_{\mathcal{M}}(\mathbf{r}) \tag{43}$$

are Cartesian tensors which are symmetric in their indices. The dot product in Eq. 41 implies a full scalar contraction with the $\mathbf{k}^{n}$ vectors. Hence, using the Einstein summation convention (in which repeated indices are summed over),

$$\rho^{g}_{\mathcal{M}}(\mathbf{k}) = V\left\{1 + ik_{\alpha}S^{(1)}_{\alpha} + \frac{i^{2}}{2!}k_{\alpha}k_{\beta}S^{(2)}_{\alpha\beta} + \frac{i^{3}}{3!}k_{\alpha}k_{\beta}k_{\gamma}S^{(3)}_{\alpha\beta\gamma} + \cdots\right\} \tag{44}$$

where the Greek indices α, β, γ, etc. stand for the Cartesian directions x, y, z.

Equation 44 becomes

$$\rho^{g}_{\mathcal{M}}(\mathbf{k}) = V\left\{1 - \frac{k_{\alpha}k_{\beta}}{2!}S^{(2)}_{\alpha\beta} - \frac{i}{3!}k_{\alpha}k_{\beta}k_{\gamma}S^{(3)}_{\alpha\beta\gamma} + \cdots\right\} \tag{45}$$

which we can rewrite by assuming that the terms in the series above $S^{(2)}$ are geometric. Hence

$$\rho^{g}_{\mathcal{M}}(\mathbf{k}) = V\exp\left(-\frac{k_\alpha k_\beta}{2!}S^{(2)}_{\alpha\beta}\right)f(\mathbf{k}) \tag{46}$$

where we can define

$$f(\mathbf{k}) = 1 + \frac{k_\alpha k_\beta k_\gamma}{3!}f_{\alpha\beta\gamma} + \frac{k_\alpha k_\beta k_\gamma k_\delta}{4!}f_{\alpha\beta\gamma\delta} + \cdots \tag{47}$$

The purpose of Eq. 47 is to 'correct' the expansion of Eq. 46 in powers of **k** so that it agrees term by term with Eq. 45. It can be shown by examining powers of **k** that

$$f(\mathbf{k}) = 1 - \frac{ik_\alpha k_\beta k_\gamma}{3!}S^{(3)}_{\alpha\beta\gamma} + \frac{k_\alpha k_\beta k_\gamma k_\delta}{4!}\left[S^{(4)}_{\alpha\beta\gamma\delta} - S^{(2)}_{\alpha\beta}S^{(2)}_{\gamma\delta} - S^{(2)}_{\alpha\gamma}S^{(2)}_{\beta\delta} - S^{(2)}_{\alpha\delta}S^{(2)}_{\beta\gamma}\right] + \cdots \tag{48}$$

The fourth-order term in the square brackets in Eq. 48 contains lower order terms which correct for the overcounting for $S^{(2)}$ terms implied by the exponential in Eq. 46. We now consider what happens when we truncate the expansion 46 and reconstitute the direct space representation by an inverse Fourier transform, i.e.

$$\rho^{g}_{\mathcal{M}}(\mathbf{r}) = \frac{1}{8\pi^3}\int d\mathbf{k}\exp(-i\mathbf{k}\cdot\mathbf{r})\,\rho^{g}_{\mathcal{M}}(\mathbf{k}) \tag{49}$$

Thus, taking

$$\rho^{g[2]}_{\mathcal{M}}(\mathbf{k}) = V\exp\left(-\frac{k_\alpha k_\beta}{2}S^{(2)}_{\alpha\beta}\right) \tag{50}$$

where the superscript [2] implies that Eq. 50 is only correct through second order, and substituting into Eq. 49 we obtain

$$\begin{aligned}\rho^{g[2]}_{\mathcal{M}}(\mathbf{r}) &= \frac{V}{8\pi^3}\frac{(2\pi)^{3/2}}{(\det S^{(2)})^{1/2}}\exp(-r_\alpha(S^{(2)})^{-1}_{\alpha\beta}r_\beta)\\ &= \frac{V}{(2\pi)^{3/2}}(\det S^{(2)})^{-1/2}\exp(-r_\alpha(S^{(2)})^{-1}_{\alpha\beta}r_\beta)\end{aligned} \tag{51}$$

It is easy to confirm that (using Eq. 51)

$$\int d\mathbf{r}\,\rho^{g[2]}_{\mathcal{M}}(\mathbf{r}) = V, \quad \int d\mathbf{r}\,\mathbf{r}^2\,\rho^{g[2]}_{\mathcal{M}}(\mathbf{r}) = S^{(2)} \tag{52}$$

so that Eq. 51 gives a simple Gaussian representation of the molecular shape density which ensures that we regain the correct zeroth and second moments, as it should.

Adopting the rotated coordinate system, Eq. 51 becomes

$$\rho^{g[2]}_{\mathcal{M}}(\mathbf{r}) = \frac{V(\det S^{(2)})^{-1/2}}{(2\pi)^{3/2}} \exp\left\{ - \left(\frac{x^2}{\varepsilon_{xx}^2} + \frac{y^2}{\varepsilon_{yy}^2} + \frac{z^2}{\varepsilon_{zz}^2} \right) \right\} \tag{53}$$

which shows that the 'through second order' representation of the *molecule* is a single elliptical Gaussian with principal diameters ε_x, ε_y, ε_z. This is the real justification for the *ansatz* in Eq. 46 in which $S^{(2)}$ terms are summed geometrically through all orders. It is possible to produce a more detailed description of the molecular shape by adding to Eq. 51 the next term (third-order) coming from the expansion of f (**k**), viz.

$$\rho^{g[3]}_{\mathcal{M}}(\mathbf{r}) = \left\{ 1 - \frac{1}{3!} S^{(3)}_{\alpha\beta\gamma} \nabla_\alpha \nabla_\beta \nabla_\gamma \right\} \rho^{g[2]}_{\mathcal{M}}(\mathbf{r}) \tag{54}$$

The 'through third order' term in Eq. 54 and higher order variants generate ellipsoids with 'ears', i.e. embellished by a higher order angular dependence. Equations such as 50 and 54 give an objective tool for coarse-grained representations of molecular shape, in which the whole molecule can be represented by a single ellipsoid or 'eared' ellipsoid. Alternatively, one can introduce distributed ellipsoids to represent groups of atoms, such as methyl or amino acid side-chains, and backbones in a protein.

Results

The shape-multipole method that we have outlined provides a simple way of encapsulating the shapes of molecules in terms of a small number of numerical values. One can calculate centroids and shape multipoles at negligible cost; for example, the computation of the protein volumes in Table 1 also includes the CPU time required to compute the shape multipoles. The centroids and quadrupolar axes can be used as a simple way of aligning sets of complex molecules. It is also possible to screen databases of molecules using a similarity function based on the components of the shape multipoles. Such an approach resembles the suggestion of Sudarsanam et al. [17], in which molecules are represented as ellipsoids computed from the coordinates of atom centers. The shape multipoles, however, describe complex geometric molecular shapes more precisely, taking into account the true extent of atoms in molecules. As an illustration, Table 3 gives the shape multipoles for the arbitrarily chosen molecule with the CSD code CEYYAG, and the 15 molecules extracted from the previously described subset of the CSD, ranked according to possessing the most similar shape-multipoles to CEYYAG. For brevity, the table only gives the largest components of the octopole, and not all 10 unique values. The structure of CEYYAG, and some of the molecules determined to be similar are given in Fig. 5. CEYYAG is a simple sulfonamide with antibacterial activity, with an approximate 'V' shape largely conferred by the stereochemistry of the central sulfur atom. A number of the molecules determined to be similar are also sulfonamides, but some are not, although from the diagram it can be seen that all possess a similar 'V'-shaped motif. In this approach we use a simple rms-type function to make a quantitative comparison of the molecular shapes. We have yet to establish the optimal way for carrying out such comparisons. A drawback of this approach as a low-resolution search for potential inhibitor lead compounds is that it relies on a global molecular shape description. Obviously, such a technique cannot distinguish between those regions of a molecule directly involved in interactions with the receptor and those that are not. It is possible to distribute the shape multipoles (and the related area multipoles) to provide localized descriptions of molecular shape, although this in turn complicates the comparison procedure between different molecules. However, such searches are not limited to attempting to identify lead compounds. One strategy for exploiting *de novo* ligand design [18] is to look for 'spacers' or molecular frameworks that seek to position functional groups according to some pharmacophoric model. We are currently attempting to design a method incorporating multipole searches of the CSD to find a set of *similar* spacers to aid the ligand design process. One advantage of utilizing the CSD is that, in the case of flexible spacers (or 'hinges'), it is likely that the crystal structure conformation of the spacer will also be accessible if this molecular framework is transferred into a potential ligand. Another application of the shape multipoles is to investigate their relationship to the analogous electrostatic moments. The idea of using the connection between shape and charge distribution tensors has been proposed by Silverman and Platt [19,20] as part of a 3D-QSAR procedure, although such a technique could also be modified for database searching.

Visualization [13] of the iso-contours of the Gaussian density function in Eq. 54 gives rise to surfaces that closely resemble the smooth molecular surface. The analytical Fourier transform approach that we have outlined gives a way of coarse-graining the Gaussian density representation, in a similar way to the numerical decomposition of the molecular surface into spherical harmonics [21]. Such low-resolution versions of protein surfaces and density volumes may prove useful in searching the Protein Data Bank to recognize tertiary-structure motifs, or in conjunction with some of the surface comparison and docking methods presented in the next section. There are a number of other potential applications for low-resolution models of complex biomolecular systems, and these have been very well reviewed by Carson [22] in his presentation of surface decomposition using wavelet multiresolution analysis.

In our approach one can replace the atom-based densities for functional groups, residues or whole molecules by elliptical Gaussian or 'decorated' ellipsoids. This method of approximating shape ensures that the individual components, be they parts or whole molecules, have correct shape multipoles up to prescribed orders. For visual representation this coarse-graining may have important advantages, since the number of Gaussians needed to represent a protein, for example, would be drastically reduced. One might only require two Gaussians per residue, depending on the degree of exactitude required. The conservation of multipoles referred to above is reminiscent of Stone's [23,24] 'distributed multipole' approach to molecular interactions.

Table 3 *Shape multipoles computed for the molecule CEYYAG (CSD code) and the next 15 most similar molecules found in a subset of the CSD*

CSD code	V^g	Q_{xx}	Q_{yy}	Q_{zz}	Ω_{xxx}	Ω_{xxy}	Ω_{xyz}
	(Å^3)	(Å^2)	(Å^2)	(Å^2)	(Å^3)	(Å^3)	(Å^3)
CEYYAG	252.61	8.52	3.72	2.16	3.21	– 7.06	– 1.47
CPROMZ	236.58	8.92	4.54	1.39	6.18	– 6.99	– 1.39
DPHPZL	260.35	8.88	3.70	2.07	4.84	– 4.98	1.03
YIBFEU	221.57	7.94	3.38	2.06	4.87	– 7.07	0.40
SOBXUC	243.81	9.11	3.82	1.89	0.83	– 7.55	– 0.45
DUKXAI	239.79	9.61	3.92	1.68	6.00	– 7.04	– 1.44
MXPEAC	245.73	7.96	4.55	1.67	7.14	– 5.71	– 0.40
SUTHAZ	220.29	8.17	2.97	2.48	2.51	– 8.37	0.03
JADDIB	238.34	9.21	3.33	1.62	3.72	– 6.59	0.31
YADBOU	225.07	8.29	3.39	1.40	5.22	– 3.97	– 0.94
KIXFOM	249.72	7.89	4.29	2.06	6.82	– 3.88	– 1.53
SLFNMF02	252.48	9.57	4.45	2.08	– 4.77	– 9.23	– 1.72
BILSEU	209.03	7.99	3.13	2.20	0.73	– 8.01	– 0.84
BGIFUL	260.76	9.20	3.96	1.97	1.65	– 3.12	– 2.85
GAHGOL	223.38	8.23	3.62	1.82	7.98	– 2.54	– 1.31
BARGEG	297.57	7.44	3.94	2.55	1.89	– 5.21	– 0.78

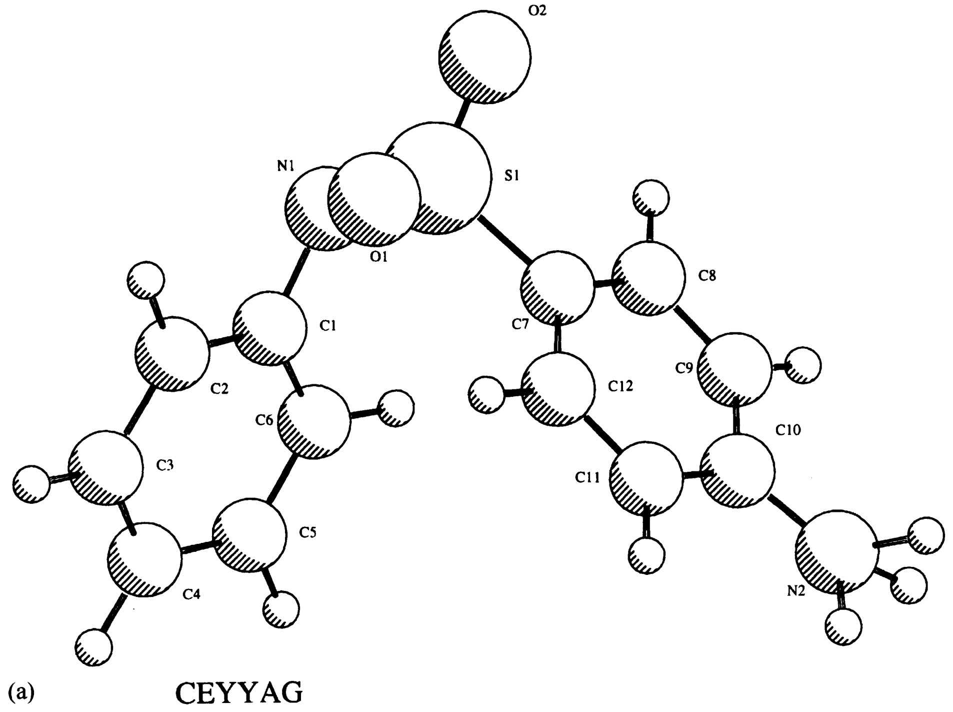


Fig. 5. Three-dimensional structure of CEYYAG and some similar molecules as determined by shape multipoles.

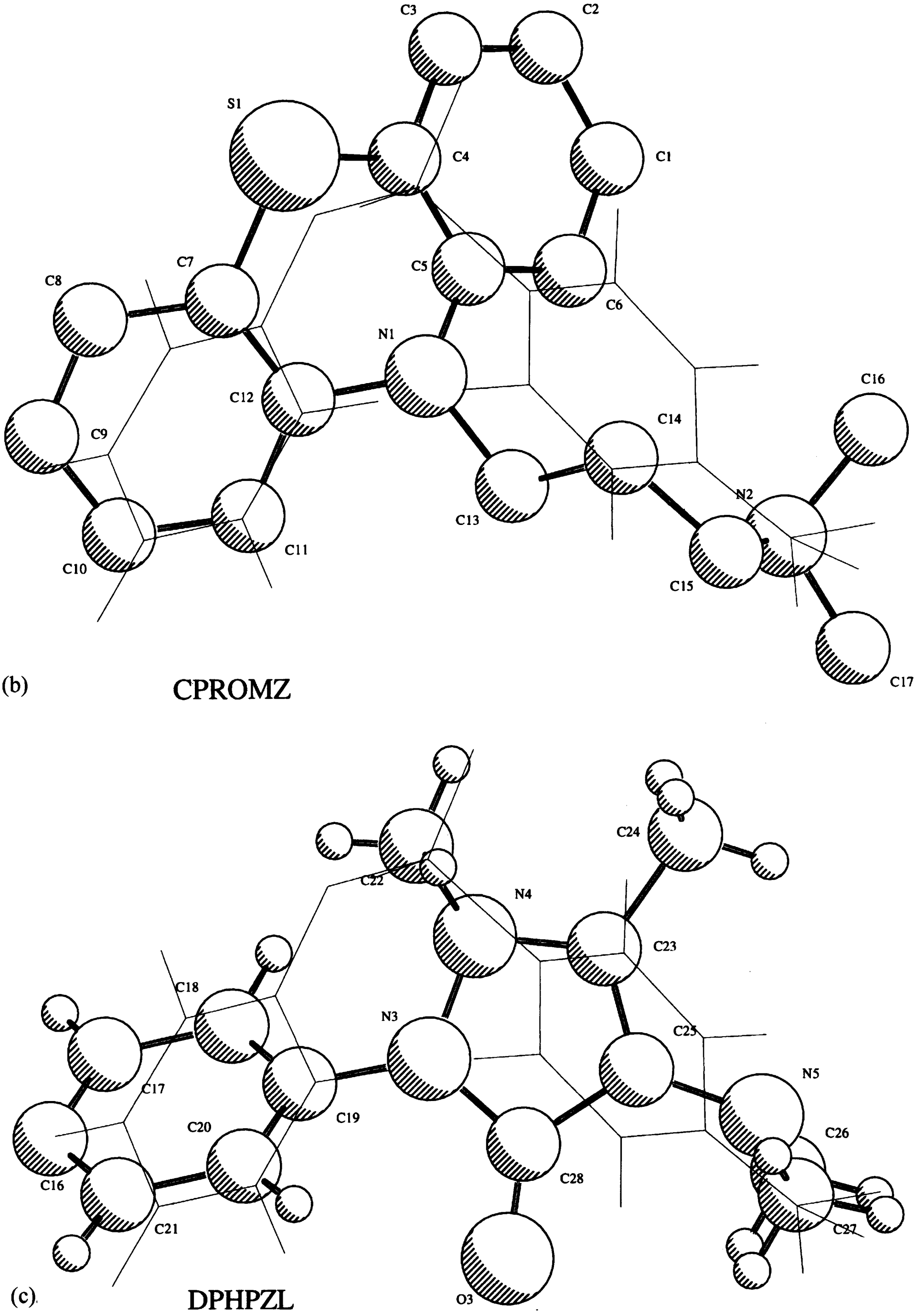


Fig. 5. (continued).

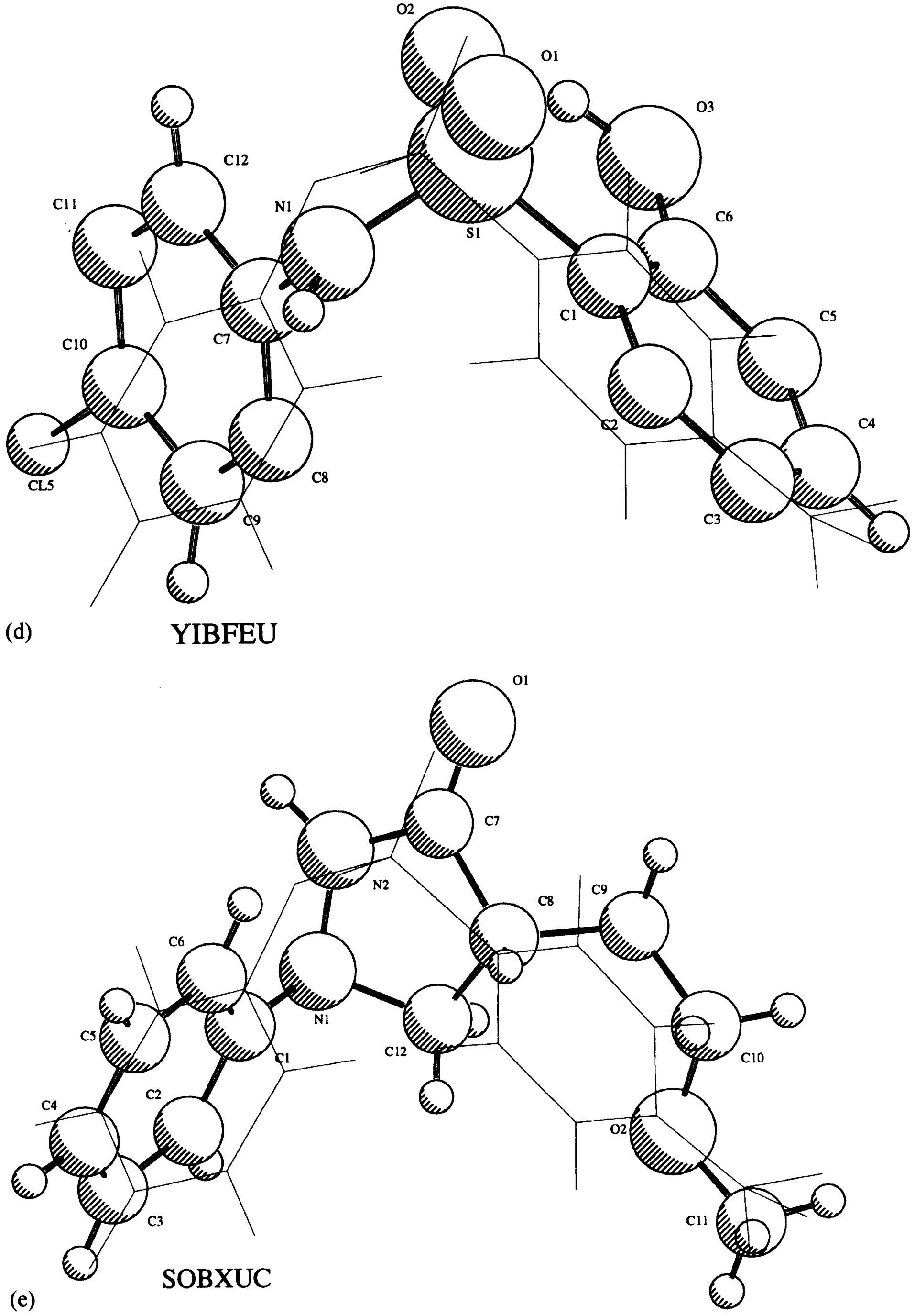


Fig. 5. (continued).

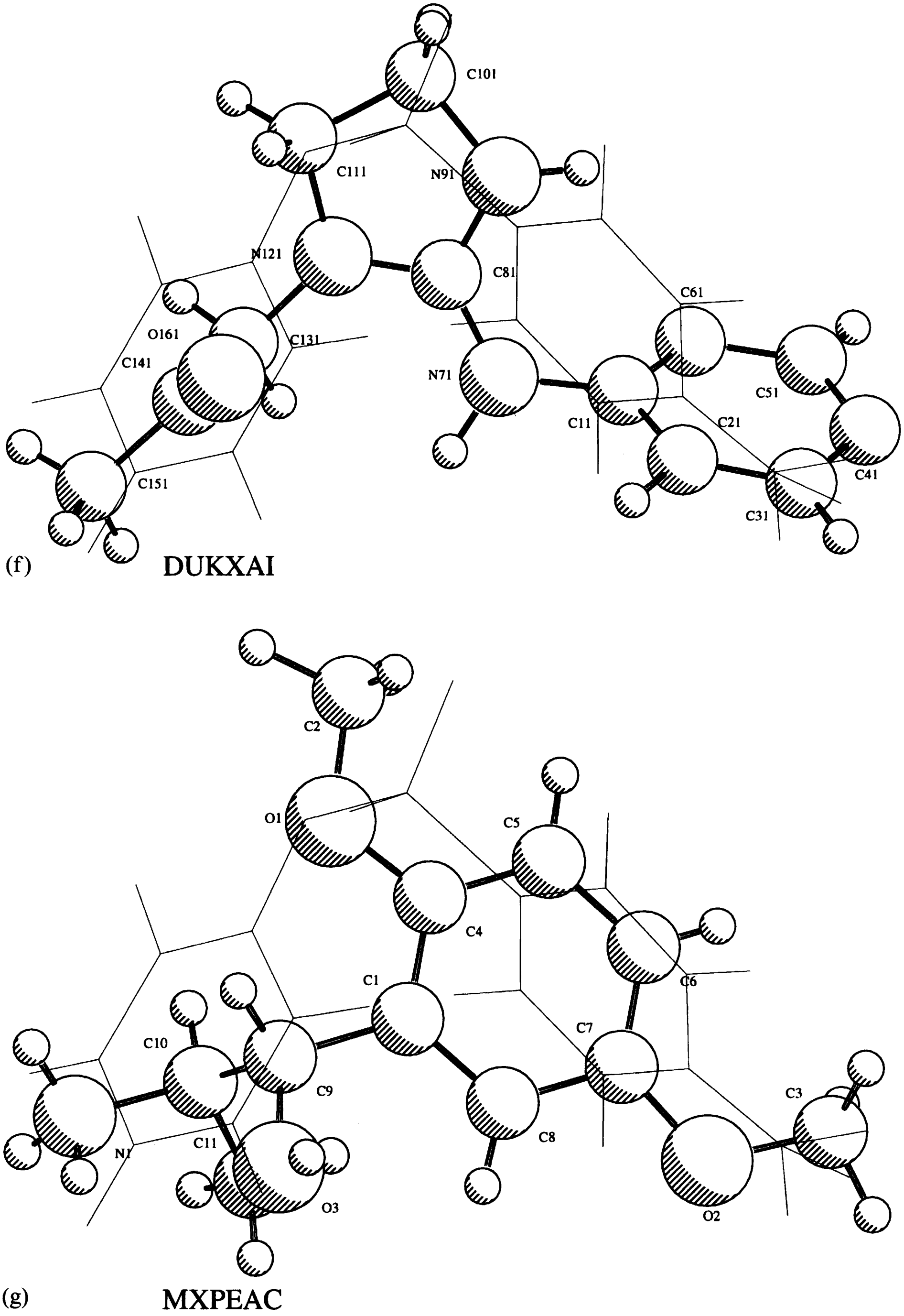


Fig. 5. (continued).

Docking

We are interested in two related docking procedures. External docking is the classical docking procedure, which is widely understood and involves the alignment of two molecular shapes without overlap. Hence, we seek to rotate and translate rigidly a guest molecule or substrate (B), with respect to a host molecule (A), such that the surfaces are in maximal contact. Internal docking, on the other hand, involves the same rigid motions of B, but with maximal overlap. The simplest aspect of internal docking involves the shape comparison of a series of molecules of similar size. In this instance one is essentially producing maximal overlaps. A more complicated case involves molecules of dissimilar sizes in which one is interested merely in comparing two surface sections [25]. The problem of external docking is well studied, and recent detailed reviews can be found [26–30]. Perhaps the most successful of the molecular docking methods is the set-theoretic method of Kuntz and co-workers [31–33]. This method is a clique-based method which matches nodes between graphs representing the protein and the ligand. However, an alternative viewpoint is to treat the docking problem as a search in a rigid-body Cartesian coordinate system [34,35]. This is a more direct simulation of the molecular recognition process, and is readily extended to introduce ligand flexibility. In hard-sphere terms the molecular docking problem is intuitively obvious, but how does one implement it in the context of the Gaussian representation? The answer lies in the ease with which the molecular intersection volume, V_{AB}, and the molecular intersection area, A_{AB}, are computed with Gaussians. We propose a model in which external docking seeks to maximize A_{AB} and to minimize V_{AB}. Internal docking seeks a maximum in both quantities.

Let us consider two Gaussians centered at R_A and R_B. The Gaussians have exponents

$$\alpha = \frac{\kappa}{\sigma_A^2}, \quad \beta = \frac{\kappa}{\sigma_B^2} \tag{55}$$

so that the two Gaussians represent atoms of radius σ_A, σ_B, respectively. The intersection volume, V_{AB}, can be calculated using Eqs. 26 and 8 as

$$V_{AB} = V_{AB}(0)\exp\left(-\xi R_{AB}^2\right) \tag{56}$$

where R_{AB} is the interatomic distance, and the parameter

$$\xi = \frac{\alpha\beta}{\alpha + \beta} \tag{57}$$

The other constant in Eq. 56 is given by

$$V_{AB}(0) = p_A p_B \left(\frac{\pi}{\alpha + \beta}\right)^{3/2} \tag{58}$$

We now define a quantity which is related to the intersection area, namely

$$\mathrm{B_{AB}} = \left(\sigma_\mathrm{A} \frac{\partial}{\partial\sigma_\mathrm{A}} + \sigma_\mathrm{B} \frac{\partial}{\partial\sigma_\mathrm{B}}\right) \mathrm{V_{AB}} \tag{59}$$

This quantity has the same (volume) dimensions as $\mathrm{V_{AB}}$. The Gaussian docking function is constructed using a function

$$\mathrm{F_{AB}} = \mathrm{B_{AB}} - \lambda \mathrm{V_{AB}} \tag{60}$$

where the constant λ will be fixed so as to confer desirable properties on the docking function $\mathrm{F_{AB}}$. Hence, we ensure that $\mathrm{F_{AB}}$ is a maximum when the intersphere distance $\mathrm{R_{AB}}$ achieves the correct value for hard-sphere docking (see Fig. 6). Hence

$$\frac{\mathrm{dF_{AB}}}{\mathrm{dR^2_{AB}}} = 0 \quad \text{at } \mathrm{R^2_{AB}} = \mathrm{D^2_{AB}} = (\sigma_\mathrm{A} \pm \sigma_\mathrm{B})^2 \tag{61}$$

The positive sign in Eq. 61 is appropriate for external docking, whilst the negative sign is for internal docking. In this manner we can produce the Gaussian version of hard-sphere docking. Some algebra leads to

$$\mathrm{F_{AB}} = 2\mathrm{V_{AB}}(0)\,[1 + \xi(\mathrm{R^2_{AB}} - \mathrm{D^2_{AB}})]\,\exp(-\xi \mathrm{R^2_{AB}}) \tag{62}$$

It is more useful in practice to use a normalized docking function,

$$\begin{aligned}\mathrm{N_{AB}} &= \frac{\mathrm{F_{AB}}}{\mathrm{F^{max}_{AB}}} \\ &= [1 + \xi(\mathrm{R^2_{AB}} - \mathrm{D^2_{AB}})]\exp(-\xi(\mathrm{R^2_{AB}} - \mathrm{D^2_{AB}}))\end{aligned} \tag{63}$$

which has a maximum

$$\mathrm{N^{max}_{AB}} = 1 \quad \text{at } \mathrm{R^2_{AB}} = \mathrm{D^2_{AB}} \tag{64}$$

and a minimum

$$\mathrm{N^{min}_{AB}} = (1 - \xi \mathrm{D^2_{AB}})\exp(\xi \mathrm{D^2_{AB}}) \tag{65}$$

at

$$\mathrm{R_{AB}} = 0 \tag{66}$$

The quantity $\mathrm{N_{AB}}$ can be thought of as a function which evaluates the sphere–sphere contacts for docking. The sharpness of the maximum can be controlled by adjusting the parameter κ, since

$$\xi = \frac{\kappa}{\sigma^2_\mathrm{A} + \sigma^2_\mathrm{B}} \tag{67}$$

A large κ gives a sharp maximum and, as we shall see, more hard-sphere-like behavior. For $\mathrm{D_{AB}} = 0$ a large κ gives a very deep minimum at $\mathrm{R_{AB}} = 0$.

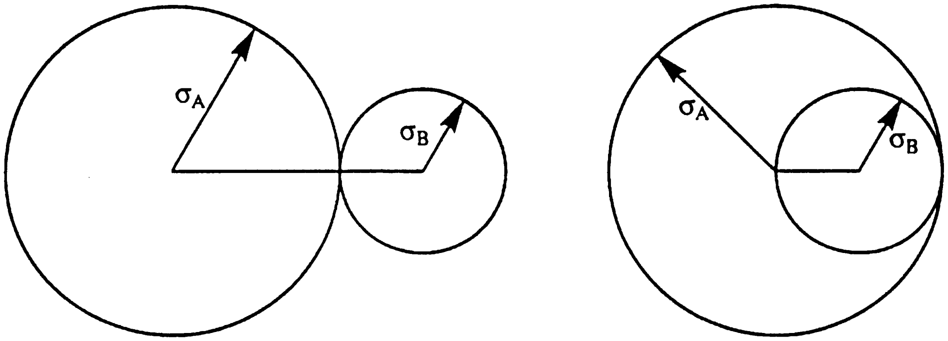


Fig. 6. External (A) and internal (B) docking for two spheres of radii σ_A and σ_B.

How can we define a Gaussian docking function for a pair of molecules such as a protein–ligand complex? This is simply answered, since one can define a pair-docking function

$$N = \sum_{A \in \mu_A} \sum_{B \in \mu_B} N_{AB} \tag{68}$$

which sums over contact atoms from molecules μ_A and μ_B. The function 68 can be maximized by rigid rotations and translations of μ_B. The size of the maximum gives a pair contact number for the docked molecules. In this way we achieve a formula embracing both external and internal docking. The optimization of Eq. 68 with respect to the rigid motions of B can be achieved using an analytical determination of the Cartesian first and second derivatives. The rotational motions are computed using a quaternionic representation, so that the Cartesian derivatives must be transformed into this representation. The parameter κ acts as a natural annealing parameter controlling the shape of the surface. Another simple approach to shape comparison and internal docking has been utilized [4] by simply maximizing the intermolecular intersection volume V_{AB}.

Results

We have coded a local optimization procedure based on the method described in the previous section. In what follows we review some preliminary results obtained investigating some protein–ligand interactions. Table 4 shows the results for the external docking of a number of protein–ligand systems. All the results are obtained with a value of $\kappa = 3.0$ and from starting geometries derived from experimental X-ray crystallographic data. The protein is represented by *all* atoms belonging to each residue that contains at least one atom within 10 Å of the ligand, in the experimentally determined binding orientation. The number of protein atoms and ligand atoms is given in Table 4. The primary intention here is to check that our procedure is in broad agreement with experiment. This is confirmed by looking at the rms deviations

Table 4 *Accuracy of the Gaussian docking method with respect to crystallographic ligand orientations*

PDB code	Description	Natoms[a]	Rms (Å)
2GBP [38]	Periplasmic binding protein/β-D-glucose	539/13	0.51
3DFR [39]	Dihydrofolate reductase/NADPH	562/33	0.44
3DFR [39]	Dihydrofolate reductase/methotrexate	629/48	0.28
6RSA [40]	Ribonuclease A/uridine-vanadate	867/31	1.29
1GST [41]	Mu class glutathione S-transferase/glutathione	595/20	0.36
2CCP [42]	Cytochrome c peroxidase/iron-heme	827/43	0.45
1STP [43]	Streptavidin/biotin	428/16	0.73
1HVR [44]	HIV-protease/cyclic urea (dupont xk263)	762/46	0.13

[a] Number of protein/ligand atoms.

between optimized and experimental structures reported in Table 4. We have also conducted theoretical experiments in which ligands are started in random positions. We find that there is no problem associated with finding local maxima, that ligands can easily move distances of 8–10 Å, and that it is easy to achieve convergence in the rms of the gradient down to 10^{-6} Å^2. The relatively large motions of the ligand during rigid-body optimization from some random search point are in part because the Gaussian docking function 'softens' the 'surface' of the interacting molecule, allowing for a certain degree of penetration of the two interacting surfaces. This fuzziness allows for both a degree of conformational change of the interacting molecules, and does not require that the surfaces make a perfect fit, as is required for hard-sphere surfaces. Clearly, this behavior is governed by the value of the κ parameter, i.e. a smaller κ gives a softer surface. In this respect our analytical algorithm performs in a similar way to the purely numerical soft-docking algorithm of Jiang and Kim [35]. In general, the values N_{opt} provide a good discrimination between local maxima, and it is rare for maxima to be found which are larger than the true experimental ones. Nonetheless, we expect that it will be important to analyze the structures corresponding to the best maxima in the Gaussian docking function, with a more physical method such as the Poisson–Boltzmann estimation of binding energies. It is also easy to visualize the docked structures using the calculated N_{AB} values for N_{opt} to color atom pairs in contact. No attempt has been made, as yet, to find an optimum algorithm, either for global searching of the rigid-body coordinate space or for improving the efficiency of the local optimization. For example, global optimization techniques such as simulated annealing [36] or the diffusion equation approach [9,37] in which the potential surface can be deformed, in our case by varying the κ parameter, may be very suitable. We expect that the performance of the local optimization can be improved by using a neighbor list approach to minimize the amount of work. The current algorithm uses all intermolecular atom pairs for the relatively large systems studied, which, given the exponential nature of the docking function, is probably unnecessary. Even given these restrictions, the local optimizations described in Table 4 require only a few seconds of CPU resource on an average workstation. We have also used our technique for systems based on protein–protein interactions, cyclodextrin complexes and also for DNA binding, and we obtain similar results to those described for the protein–ligand systems.

An important feature of the docking method described in the previous section is that, by simply changing the sign appearing in the constant D_{AB} defined in Eq. 6, the method can be switched from evaluating surface complementarity to surface similarity. Thus, it is trivial to reuse the code to compute the results of Table 4, to align molecules based on the similarity of their shapes. In this respect, the new unified docking procedure presented in this section is an extension of our old approach presented in Ref. 4 (which is equivalent to setting $\lambda = 0$ in Eq. 60). The advantage of this new approach is that it is now possible to identify surface similarities arising from molecules of differing sizes such as proteins with their small molecule mimics. We have used the new method for internal docking to predict the relative orientation of ligand series binding to the proteins thrombin, thermolysin and HIV-protease. The accuracy of these predictions, with respect to the experimentally observed relative orientations, is at least as accurate as those reported using our original shape matching method [4] ($\lambda = 0$), and in certain cases there is some improvement. We have also carried out some *preliminary* calculations attempting to generate the relative orientation of turkey ovomucoid inhibitor (TOMI, a 56-residue peptide) and a difluoroketone inhibitor (DFKi) bound to the enzyme elastase. These calculations were originally carried out using the very elegant hard-sphere method of Masek et al. [25]. The results that we have obtained so far are similar to those obtained by Masek et al. We find that it is not possible to generate the experimental alignment using only volume intersection ($\lambda = 0$), but successfully find a maximum that corresponds to the experimental alignment using the new internal docking method. The computational performance of the Gaussian algorithm is much improved, as expected, relative to the original hard-sphere method. We need to establish if the global value of the Gaussian docking function corresponds to the experimental alignment by placing the DFKi inhibitor at random start points on (or close to) the surface of the TOMI protein.

Conclusions

We have described a Gaussian variant of hard-sphere shape techniques which has the advantage that all expressions for quantities are analytically determined. We have shown how to compute volumes and areas, and have indicated how one can characterize molecular shapes using multipolar averages. The multipole shape tensors give a coarse-grained representation of the shape density based on the whole molecule or on fragments within it. We have also extended the Gaussian method to include a description of surface similarity and complementarity.

Acknowledgements

We thank Dr. Jaroslaw Kostrowicki and Dr. Anthony Nicholls for very valuable discussions concerning the use of Gaussian functions to represent the unit step function. We also thank Dave Cosgrove, Maria Gallardo, Peter Kenny, Philip Jewsbury, Christine Kitchen, Brian Masek, Jeff Morris, Dave Timms and Tony Wilkinson for their help and advice in assembling the work presented in this chapter.

References

1. Mezey, P.G., Shape in Chemistry, VCH, New York, NY, 1993.
2. Grant, J.A. and Pickup, B.T., J. Phys. Chem., 99(1995)3503.
3. Grant, J.A. and Pickup, B.T., J. Phys. Chem., 100(1996)2456.
4. Grant, J.A., Gallardo, M.A. and Pickup, B.T., J. Comput. Chem., 17(1996)1653.
5. Boys, S.F., Proc. R. Soc. London, Ser. A, 200(1950)542.
6. Diamond, R., Acta Crystallogr., Sect. A, 27(1971)436.
7. Marshall, G.R. and Barry, C.D., Abstr. Am. Crystallogr. Assoc., Honolulu, HI, 1979.
8. Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A. and Dunn, D.A., Computer-Assisted Drug Design, ACS Symposium Series, Vol. 112, American Chemical Society, Washington, DC, 1979, pp. 205–226.
9. Kostrowicki, J., Piela, L., Cherayil, B.J. and Scheraga, H.A., J. Phys. Chem., 95(1991)4113.
10. Kearsley, S.K., Tetrahedron Comput. Methodol., 3(1990)615.
11. Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32(1992)188.
12. Good, A.C. and Richards, W.G., J. Chem. Inf. Comput. Sci., 33(1993)112.
13. Grant, J.A. and Nicholls, A., in preparation.
14. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M. and Watson, D.G., J. Chem. Inf. Comput. Sci., 31(1991)187.
15. Allen, F.H. and Kennard, O., Chem. Design Autom. News, 8(1993)31.
16. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112(1977)535.
17. Sudarsanam, S., Duke, G., March, C.J. and Srinivasan, S., J. Comput.-Aided Mol. Design, 6(1992)223.
18. Cosgrove, D.A. and Kenny, P.W., J. Mol. Graph., 14(1996)1.
19. Platt, D.E. and Silverman, B.D., J. Comput. Chem., 17(1996)358.
20. Silverman, B.D. and Platt, D.E., J. Med. Chem., 39(1996)2129.
21. Leicester, S.E., Finney, J.L. and Bywater, R.P., J. Mol. Graph., 6(1988)104.
22. Carson, M., J. Comput.-Aided Mol. Design, 10(1996)273.
23. Stone, A.J., Chem. Phys. Lett., 83(1981)233.
24. Stone, A.J. and Alderton, M., Mol. Phys., 56(1985)1047.
25. Masek, B.B., Merchant, A. and Matthew, J.B., Proteins, 17(1993)193.
26. Cherfils, J. and Janin, J., Curr. Opin. Struct. Biol., 3(1993)265.
27. Lybrand, T., Curr. Opin. Struct. Biol., 5(1995)224.
28. Masek, B., Molecular Similarity in Drug Design, Blackie Academic, London, 1995.
29. Bamborough, P. and Cohen, F.E., Curr. Opin. Struct. Biol., 6(1996)236.
30. Strynadka, N.C.J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B.K., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. and James, M.N.G., Nat. Struct. Biol., 3(1996)233.
31. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161(1982)269.

32. Roe, D.C. and Kuntz, I.D., Pharm. News, 2(1995)13.
33. Shoichet, B.K. and Kuntz, I.D., Chem. Biol., 3(1996)151.
34. Wodak, S.J. and Janin, J., J. Mol. Biol., 124(1978)323.
35. Jiang, F. and Kim, S., J. Mol. Biol., 219(1991)79.
36. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., Science, 220(1983)671.
37. Kostrowicki, J. and Scheraga, H.A., J. Phys. Chem., 96(1992)7442.
38. Vyas, N.K., Vyas, M.N. and Quiocho, F.A., Science, 242(1988)1290.
39. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257(1982)13650.
40. Borah, B., Chen, C., Egan, W., Miller, M., Wlodawer, A. and Cohen, J.S., Biochemistry, 24(1985)2058.
41. Ji, X., Zhang, P., Armstrong, R.N. and Gilliland, G.L., Biochemistry, 31(1992)10169.
42. Wang, J., Mauro, J.M., Edwards, S.L., Oatley, S.J., Fishel, L.A. and Ashford, V.A., Biochemistry, 29(1990)7160.
43. Weber, P.C., Ohlendorf, D.H., Wendoloski, J.J. and Salemme, F.R., Science, 243(1989)85.
44. Lam, P.Y.S., Jadhav, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C., Weber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-Viitanen, S., Science, 263(1994)380.

Systematic procedure for the development of accurate QM/MM model Hamiltonians

Mark A. Cunningham and Paul A. Bash
Center for Mechanistic Biology and Biotechnology, Argonne National Laboratory, Argonne, IL 60439, U.S.A.

Introduction

Enzymes play a crucial role in the biochemical machinery, serving as extraordinarily specific, highly efficient and regulatable catalysts of the fundamental chemical reactions that occur in living organisms. Understanding the mechanisms by which enzymes achieve their remarkable catalytic abilities has been a long-standing goal of biochemists, and significant progress has been made since 1878 when Fredrich Wilhelm Kühne first coined the word 'enzyme' (from the Greek *en* in + *zyme* leaven) to emphasize that there was some agent in yeast cells and not the yeast itself that was responsible for fermentation. Biochemists have since deduced that enzymes are proteins, composed of sequences of amino acids, and have developed powerful experimental techniques for determining the precise amino acid sequences that define proteins. Moreover, it is possible to modify individual residues through site-directed mutagenesis techniques to help identify the key functional groups of the enzyme. Additionally, X-ray crystallography has advanced to the point where it is now feasible to determine the three-dimensional structure of enzymes, providing us with the ability to visualize the active site and confirm Emil Fischer's 1894 hypothesis that the specificity of an enzyme for a particular substrate is due to their geometrically complementary structures.

This lock-and-key hypothesis is borne out in the binding of the substrate malate in the active site of the enzyme malate dehydrogenase (MDH), which interconverts malate and oxaloacetate as part of the citric acid cycle. The active site of MDH has two arginine residues, Arg^{81} and Arg^{153}, that form salt bridges with the carboxylate oxygen atoms of the malate substrate (Fig. 1) and orient it optimally for the subsequent transfer of a proton to the nearby histidine residue (His^{177}) and transfer of a hydride ion to the cofactor nicotinamide adenine dinucleotide (NAD^+). A third arginine residue, Arg^{87}, forms hydrogen bonds with both the O2 oxygen atom (which donates the proton) and one of the carboxylate oxygens of the malate substrate, further locking the substrate into place.

Despite tremendous progress toward the goal of understanding the detailed mechanisms of catalysis, a number of issues remain difficult to address with current experimental methodology, because the critical catalytic steps of molecular recognition and capture of the substrate and the ensuing making and breaking of bonds take place on exceedingly short time scales. Crystal structures are insightful, but yield an essentially static picture of the enzyme–substrate system. The measurement of Michaelis–Menten types of rate constants cannot always differentiate between competing mechanisms. Site-directed mutagenesis techniques do afford the ability to modify individual residues and to study the effects of mutations on structure and catalytic activity, but the results are not always definitive. This situation has motivated our efforts to develop numerical tools which can be used to address some of these critical issues of molecular recognition and catalytic processes in enzymes that are difficult to analyze with current biophysical and biochemical experimental methods.

Given that the Schrödinger equation provides an appropriate framework for describing the quantum mechanical behavior of atoms and molecules, one might hope that the study of enzyme-catalyzed reactions might prove to be a reasonably straightforward exercise. Sometime in the next century, that will undoubtedly be the case. For now, however, there are significant restraints imposed by both the sophistication of the algorithms which produce approximate solutions of the Schrödinger equation and the computing power available, even in supercomputing environments. High-level *ab initio* quantum mechanical calculations based on Hartree–Fock or density functional methods are capable of generating enthalpies of formation, for example, which agree with experimental measurements to within 4–8 kJ/mol [1]. Unfortunately, such calculations are quite limited in practice. A single energy calculation in a system containing a dozen atoms may require tens of hours to complete on even a large supercomputer. Computations of entire enzyme–substrate complexes composed of thousands of atoms are clearly outside the realm of feasibility for these methods.

As an alternative, one might consider using one of the so-called semiempirical quantum mechanical methods [2]. These techniques rely upon a parametrization of various functionals to approximate solutions to the Schrödinger equation and are orders of magnitude faster than the *ab initio* quantum methods. Unfortunately, even these methods are too computationally intensive to be applicable to the study of complete enzyme–substrate systems. If we are willing to dispense with the constraint of simulating the entire enzyme, it is possible to build approximate models of the active site by replacing the active-site residues with small-molecule analogues and orienting them according to the crystal structure [3–5].

A method that is capable of providing a dynamical simulation of large systems like the enzyme–substrate complex is the so-called molecular mechanics model [6]. In this method, the atoms are treated classically. Chemical bonds are represented by force constants that define the bond lengths and angles between adjacent bonds. Atoms that are not bound interact electrostatically and via a Lennard-Jones potential. The charge distributions and parameters that define the model are determined by calibrating to the known structures and infrared spectra of small molecules in the gas phase and measured thermodynamic properties in the condensed phase. The molecular mechanics model provides a rough approximation to solutions of the Schrödinger equation but, in practice, provides realistic simulations of enzyme systems.

One issue that can be addressed with molecular mechanics methods is the formation of the Michaelis complex, which can be considered to be the point in the reaction sequence when the substrate has been captured and oriented by the enzyme. Hydrogen bonds stabilizing the complex have formed but no covalent bonds have been reorganized. Figure 1 is a depiction of the Michaelis complex of malate and MDH and was obtained from a dynamical simulation using molecular mechanics methods: a so-called molecular dynamics calculation [7]. The structure of the Michaelis complex for malate and MDH cannot be determined experimentally: the reaction proceeds too quickly. By fortuitous circumstance, it happens that MDH binds citrate [8] into a stable complex, permitting X-ray crystallographers to define its structure. The structure displayed in Fig. 1 was obtained by placing a malate substrate into a conformation analogous to the one occupied by citrate in the experimentally defined crystal structure and then allowing the protein and malate substrate to equilibrate into a minimum-energy configuration. As a measure of the ability of the molecular mechanics method to realistically reproduce the actual protein environment, we can compute the differences in position between atoms in the X-ray crystal structure and those determined by the numerical model. The root-mean-squared difference summed over α-carbon atoms is 0.35 Å; for all atoms, the figure rises to 0.89 Å. These values are quite reasonable and we can conclude that the numerically derived structure of the Michaelis complex illustrated in Fig. 1 is a realistic representation of the actual Michaelis complex of the MDH:malate:NAD$^+$ system.

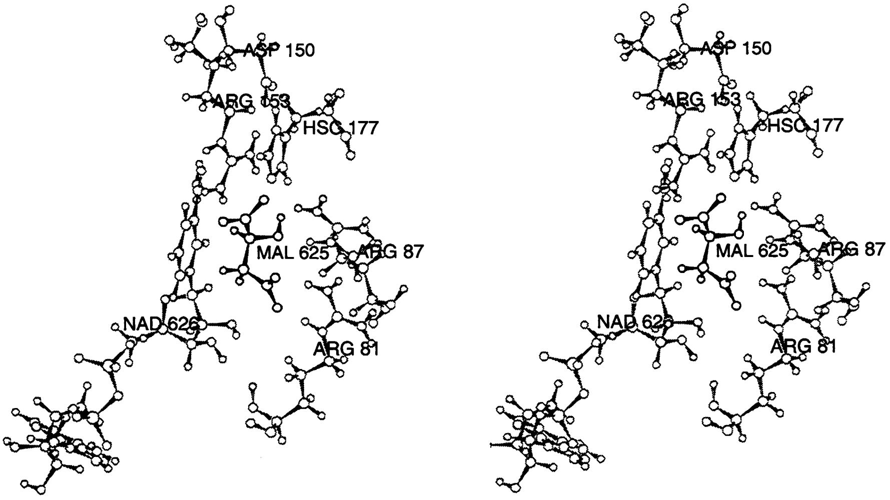


Fig. 1. Stereoview of the active site of MDH, with the substrate malate and cofactor NAD$^+$.

QM/MM method

Unfortunately, the molecular mechanics model cannot adequately describe the subsequent processes of breaking and forming chemical bonds. A more accurate description of the inherently quantum phenomena associated with bond formation is necessary. Rather than resort to a cluster-type approach to address the issue of electronic structure during the catalysis process, we note first that the molecular mechanics method does provide a realistic description of the dynamics of the protein as a whole. If we further postulate that the quantum chemical activity is confined to a small region near the active site, then it is possible to construct a hybrid model in which only a relatively few atoms are described quantum mechanically and the remainder are treated with molecular mechanics [9–11]. This approach has the advantage that the entire enzyme is present and the effects of the enzyme environment on the reaction are included implicitly, unlike the cluster models in which enzyme environmental effects are more crudely approximated. Unfortunately, even with the reduction in the size of the problem to be treated quantum mechanically from thousands of atoms to tens of atoms, it is not possible to incorporate any of the high-level *ab initio* quantum methods and retain any hope of running dynamical simulations. Consequently, we are constrained to use one of the semiempirical quantum models.

While this is, on the face of it, not as satisfying a situation as one might hope, it is still possible to generate reasonable simulations with the semiempirical models. In Table 1, we list a small sampling of values from the so-called G2 test set, which compares computed enthalpies of formation with experimental measurements [1]. The G2 method is a composite theory based on the Hartree–Fock 6-311 G(d, p) basis

Table 1 *Errors in estimation of enthalpies of formation*

Molecule	ΔH^f_0 (kJ/mol)	|G2 – expt.|	|B3LYP – expt.|	|AM1 – expt.|
Methane	– 74.8 ± 0.4	2.9	6.7	37.6
Ethane	– 84.0 ± 0.4	2.1	2.5	10.9
Benzene	82.3 ± 0.8	16.3	18.8	9.2
Formic acid	– 378.3 ± 0.4	8.4	3.8	28.8
Acetic acid	– 432.2 ± 1.7	6.3	10.9	1.7
Methylamine	– 23.0 ± 0.4	0.0	13.4	7.9
Trimethylamine	– 23.8 ± 0.8	5.9	0.8	16.7
Acetaldehyde	– 165.9 ± 0.4	5.4	1.3	7.9
Pyridine	140.4 ± 0.8	9.2	0.8	10.9

G2 is a composite theory based on the Hartree–Fock 6-311 G(d,p) basis set and several basis extensions. Electron correlation is incorporated by Møller–Plesset perturbation theory and quadratic configuration interaction. B3LYP is a density functional method based on Becke's three-parameter functional and the gradient-corrected correlation functional of Lee, Yang and Parr. AM1 is the semiempirical model of Dewar et al.

set and several basis extensions. Electron correlation is incorporated by means of Møller–Plesset perturbation theory and quadratic configuration interaction. At the present time, it represents the state-of-the-art in high-level *ab initio* quantum methods. While the root-mean-square variance for the G2 method over the entire 148-member test suite was 5.1 kJ/mol, there were instances where the deviation was as large as 34 kJ/mol. The variance of the density functional method B3LYP was 10.2 kJ/mol, with a maximum deviation of 84 kJ/mol. We have not computed AM1 enthalpies for all the members of the G2 test set, but a quick inspection of Table 1 indicates that the variance for the AM1 method will be significantly larger than that of either of the two *ab initio* methods. On the other hand, we are not interested in reproducing the entire slate of chemical reactions for all known elements. Rather, we are interested in a specific few interactions which may occur in a particular enzyme–substrate system. Because the AM1 model is defined by a set of empirically determined parameters, we might anticipate that suitable adjustments to these parameters will produce a model Hamiltonian which can be quite accurate for a limited set of interactions. Using the MDH:malate:NAD$^+$ system as an example, we shall lay out a systematic procedure for developing accurate model Hamiltonians suitable for studying enzyme–substrate systems.

Calibration of the QM model Hamiltonian

The first problem we must confront is a specification of the important aspects of the enzyme system in question. Figure 2 is a schematic depiction of the active site of the MDH:malate:NAD$^+$ complex, with the important atoms labeled. We note that the

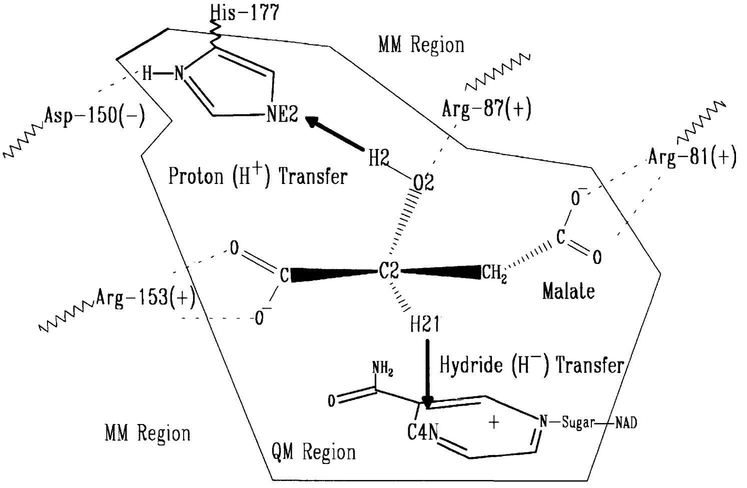


Fig. 2. Schematic drawing of the active site of the MDH:malate:NAD$^+$ complex.

crystal structure and other biochemical data [12,13] indicate that a proton is transferred from the O2 oxygen atom of the malate substrate to the NE2 nitrogen atom in the imidazole ring of the His^{177} residue in MDH and a hydride ion is transferred from the C2 carbon atom in malate to the C4N carbon atom in the nicotinamide ring of the NAD^+ cofactor. Consequently, in addition to requiring that the quantum Hamiltonian provide a reasonable description of the structures of the key elements, we want to ensure that both the proton and hydride transfer reactions are accurately represented. To do so, we will need to examine analogous reactions in small-molecule systems, for which experimental data are available and for which high-level *ab initio* quantum calculations can be performed.

Focusing for the moment on the proton transfer reaction, we note that the proton is derived from a hydroxyl group on the malate substrate. We chose methanol as an analogue for the proton donor; it has been well studied experimentally and can be numerically modeled to high order. Because the nitrogen atom which serves as the proton acceptor is a member of the imidazole ring, it is wiser to use the imidazole ring as an acceptor analogue rather than trying to abstract the system to a smaller entity such as NH_2. This entails a larger computational problem, but is still one which can be reasonably addressed.

The next step is to fit the parameters of the AM1 model Hamiltonian to reproduce the target data, which we will choose as the experimental enthalpies of formation of the reactants, methanol and imidazole, and the products, methoxide and imidazolium. We also include experimental dipole moments for methanol and imidazole, theoretical dipole moments for methoxide and imidazolium (from HF/6-31G(d) calculations) and structural information obtained from high-level (MP2/6-31G(d)) *ab initio* quantum calculations: bond lengths, angles and dihedral angles. The nonlinear optimization problem associated with this step has been outlined by Dewar and Thiel [14]. Basically, we seek the set of parameters $\mathbf{x} = (x_1, \ldots, x_N)$ defining the AM1 model Hamiltonian (listed in Table 2) that minimize the following scalar function:

$$\sigma(\mathbf{x}) = \sum w_i | y_i(\mathbf{x}) - y_i^0 | \qquad (1)$$

where the w_i are the relative weights of the $i = 1, \ldots, M$ terms. The $y_i(\mathbf{x})$ are the results of AM1 calculations with the parameter set $\mathbf{x}$ and the y_i^0 are the target values: experimental observables and results of *ab initio* quantum calculations. We have implemented an approach like that of Rossi and Truhlar [15] and utilize a genetic algorithm to ptimize the Hamiltonian parameters. The genetic algorithm we employed is similar to one described by Goldberg [16], but represents the variables as real numbers instead of bit patterns and provides a uniform distribution of crossover points instead of one or two [17]. Additionally, this implementation relies upon a steady-state algorithm for population replacement [18].

To address the proton transfer reaction, we used the following set of target values [19]:

1. Experimental enthalpies of formation, with a relative weight of 1.
2. Dipole moments, with a relative weight of 30.

Table 2 *AM1 model Hamiltonian parameters*

Parameter	AM1				AM1-SSP			
	H	C	N	O	H	C	N	O
U_{ss}	– 11.396427	– 52.028658	– 71.860000	– 97.830000	– 10.690510	– 53.180427	– 75.073641	– 104.113886
U_{pp}		– 39.614239	– 57.167581	– 78.262380		– 38.767876	– 57.341356	– 78.275023
β_s	– 6.173887	– 15.715783	– 20.299110	– 29.272773	– 6.422484	– 16.242199	– 20.380228	– 33.109001
β_p		– 7.719283	– 18.238666	– 29.272773		– 7.272735	– 16.121743	– 27.876218
α	2.882324	2.648274	2.947286	4.455371	2.893675	2.809818	3.248052	4.235416
K_1	0.122796	0.011355	0.025251	0.280962	0.125194	0.011703	0.026569	0.293314
K_2	0.005090	0.045924	0.028953	0.081430	0.005108	0.042409	0.028418	0.078530
K_3	– 0.018336	– 0.020061	– 0.005806		– 0.018117	– 0.021220	– 0.004842	
K_4		– 0.001260				– 0.001600		
M_1	1.200000	1.600000	1.500000	0.847918	1.108805	1.415666	1.457421	0.907351
M_2	1.800000	1.850000	2.100000	1.445071	1.799327	2.033559	2.067424	1.655841
M_3	2.100000	2.050000	2.400000		2.254171	1.832265	2.455736	
M_4		2.650000				2.408368		
ζ_s	1.188078	1.808665	2.315410	3.108032				
ζ_p		1.685116	2.157940	2.524039				
L_1	5.000000	5.000000	5.000000	5.000000				
L_2	5.000000	5.000000	5.000000	7.000000				
L_3	2.000000	5.000000	2.000000					
L_4		5.000000						

All parameters are in units of eV. The Slater exponents ζ_s and ζ_p were not optimized; neither were the L parameters. These values are the same for both AM1 and AM1-SSP parameter sets.

3. Internal coordinates taken from an *ab initio* quantum calculation to define the structures. We used a Hartree–Fock method with the 6-31G(d) basis set and MP2 correlation correction. Bond lengths and dihedral angles were given relative weights of 1 and angles were given relative weights of 5.

The genetic algorithm was initialized with the standard AM1 parameter set (there were 42 independent parameters optimized by the fitting procedure). A population of 300 chromosomes was used, and initial values were selected from a random Gaussian distribution with a standard deviation of 0.1, centered on the standard AM1 parameter values. Crossover and mutation probabilities were chosen to be 0.7 and 0.01, respectively. The algorithm was run for 15 000 generations, with 1% of the population selected for crossover in each generation. Optimized parameters for the system-specific parametrization (AM1-SSP) are listed in Table 2.

Some results of the genetic algorithm fit are listed in Table 3. The rather extensive geometry comparisons are omitted from Table 3 for the sake of brevity, but the overall comparisons are quite good. The optimized geometries differ from their target values for bonds by 0.011 ± 0.008 Å, the computed angles differ by 1.06 ± 0.97° and dihedrals differ by 0.19 ± 0.14°. All of the AM1-SSP enthalpies of formation agree with the experimental values to within 3 kJ/mol; dipole moments agree to within 0.3 D. Additionally, if we look at the overall enthalpy of reaction for the transfer of a proton from methanol to imidazole, $\Delta\Delta H_f^0 = \Delta H_f^0(\text{products}) - \Delta H_f^0(\text{reactants})$, we find that the experimental value is 656.7 kJ/mol and the AM1-SSP value is 661.3 kJ/mol. The value from the standard AM1 formulation is 681.5 kJ/mol. The optimized AM1-SSP value agrees to within 5 kJ/mol, which is the same level of accuracy obtained with very computationally intensive G2 calculations in the G2 test set. While this optimized AM1-SSP model Hamiltonian would not fare well if tested in a broad range of problems, at least for the case of the proton transfer reaction between malate and His^{177}, we have a fair amount of confidence that the AM1-SSP model will produce accurate results.

The development of a complete model Hamiltonian for an enzyme–substrate system will require including in the above procedure any other probable or possible reactions which are part of the catalytic process. In the case of MDH, we need to add information about the hydride transfer reaction in a way similar to what we did for the

Table 3 *Design targets for the proton transfer reaction*

Molecule	ΔH_0^f (kJ/mol)			$\lvert\mu\rvert$ (D)		
	Target	AM1	AM1-SSP	Target	AM1	AM1-SSP
Methanol	− 201.9	− 234.2	− 201.2	1.70	1.62	1.97
Methoxide	− 138.8	− 161.0	− 136.2	2.16	1.38	2.09
Imidazole	146.3	212.2	145.5	3.80	3.60	3.69
Imidazolium	739.9	820.6	741.7	1.74	1.63	1.76

Experimental enthalpies of formation are taken from Ref. 23. Experimental dipole moments are obtained from Ref. 24.

proton transfer reaction. It is instructive to sketch out a possible approach to handling the hydride transfer reaction. The hydride ion transferred from the malate substrate to NAD$^+$ is originally attached to the C2 carbon atom of malate and opposite the O2 oxygen, which serves as the proton donor. A reasonable choice for an analogue of the proton donor would be either methoxide or methanol, with the methyl group serving to donate the hydride ion. The hydride acceptor in the MDH system is the cofactor NAD$^+$, specifically the C4N carbon of the nicotinamide ring. The principal problem here is to determine a reasonable analogue to the complete NAD$^+$ molecule. One could choose to work solely with the pyridine ring (it is probably not prudent to use an analogue which does not include the ring structure). It is likely, however, that the carboxyamide group which is attached to the neighboring C3N carbon will affect the reaction to some degree. Unfortunately, experimental data for 1,4-dihydronicotinamide are sparse, which means that some crucial constraints are not available for the fitting procedure. It is, of course, possible to use high-level *ab initio* quantum methods such as G2 or B3LYP to supply the missing information but, as we can see in Table 1, even these methods are not as reliable as one might hope. It is important to find some place to touch base with experiment. The most cautious approach might then be to use experimental data from pyridine and *ab initio* quantum calculations in 1,4-dihydronicotinamide to guide the parameter fitting process.

QM/MM interactions

The next issue we must address concerns the interactions between the atoms described with a quantum mechanical potential, in what is termed the QM partition of the model, and those described classically through molecular mechanics, or in the MM partition. The hybrid model includes both electrostatic and van der Waals interactions between atoms in the two partitions. Explicitly, the model incorporates the following terms:

1. QM electron–MM partial charge electrostatic potential. Atoms in the MM partition have no electrons; their effective charges are positioned at the atom center.
2. QM nucleus–MM partial charge electrostatic potential.
3. QM/MM van der Waals potential, which models the electronic repulsion and dispersion properties that are missing because the MM atoms have no electrons.

In the case where an atom in the QM partition is bonded to an atom in the MM partition, as would be the case if we chose to draw the QM boundary between the α-carbon of a protein residue and the β-carbon of its side chain, the model incorporates fictitious 'link' atoms that serve to terminate the QM electron density along the bond. The link atoms have no interactions with atoms in the MM partition but do contribute to the energy and forces felt by atoms in the QM partition. There are no adjustable parameters for the link atoms. Neither of the QM/MM electrostatic potential terms contain any free parameters but the van der Waals interactions must be calibrated to realistically represent the forces on atoms in the QM partition due to atoms in the MM partition.

There are no experimental data to guide this parametrization effort. Instead, we rely on high-level *ab initio* quantum calculations of a water molecule interacting with the small-molecule analogues. Hartree–Fock calculations using the 6-31 G(d) basis set were performed, optimizing the individual structures of water and the analogue molecules. A series of HF/6-31 G(d) optimizations was then performed with a single degree of freedom: the distance between the water molecule and a specific target atom in the analogue molecule. We display the relevant orientations in Fig. 3 for the analogue molecules used in the proton transfer reaction. The 6-31 G(d) basis set has been shown to generate reasonable structures, but the interaction energies for neutral molecules are systematically underestimated [20]. We have used a scale factor of 1.16 to compensate for this bias on all of the interactions between neutral reactants [21].

The set of calculations performed to establish the *ab initio* target values was then repeated with the hybrid model. Geometries of the small analogue molecules were optimized with the AM1-SSP quantum model Hamiltonian. The water molecules were treated with molecular mechanics and we used the TIP3P model of Jorgensen et al. [22]. The van der Waals parameters for atoms in the small-molecule analogues

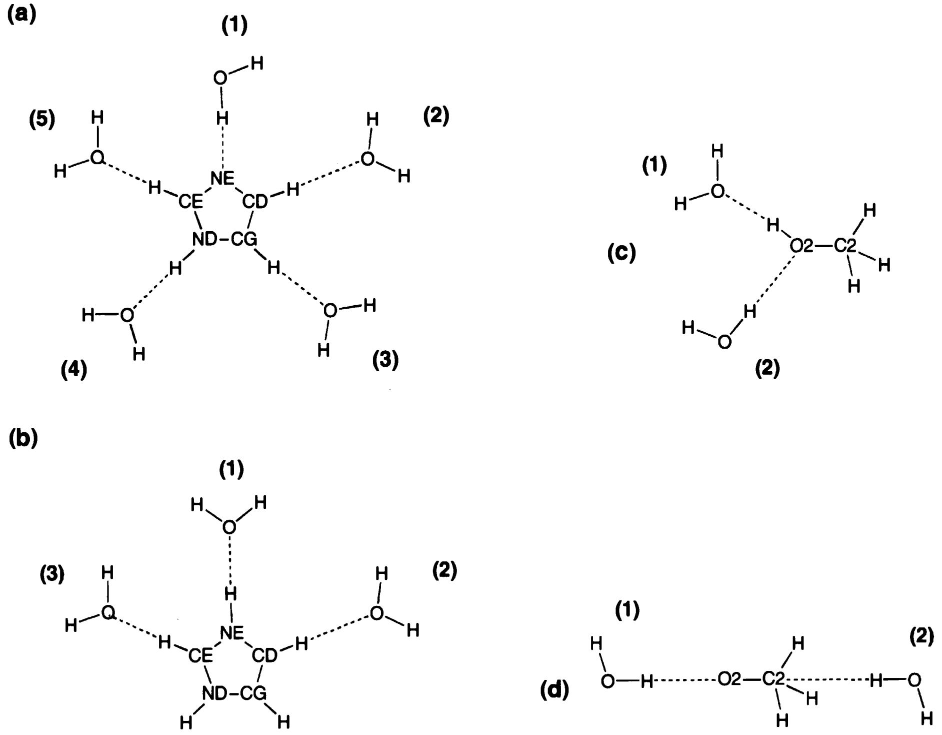


Fig. 3. Schematic of the microsolvation procedure: (a) imidazole; (b) imidazolium; (c) methanol; (d) methoxide.

were then adjusted to reproduce both the interaction energies and the optimal intermolecular distances obtained from the Hartree–Fock calculations. A summary of these results is presented in Table 4. We have also plotted the interaction energies obtained from the AM1-SSP quantum calculations against those obtained from the Hartree–Fock method in Fig. 4, which reflects the 0.7 kJ/mol root-mean-square deviation for interaction energies. For optimum distances, we find a root-mean-square deviation of 0.06 Å.

Proton transfer in solution

The pieces are now in place to perform simulations with the calibrated Hamiltonian parameters. Before investigating the reaction mechanism in the context of the enzyme, however, we can touch base with experiment one last time by examining the proton transfer reaction between methanol and imidazole in solution. The free energy change ΔG in solution can be obtained from the experimental pK_a values of the reactants by means of the following relation:

$$\Delta G = -2.3RT[pK_a(\text{imidazole}) - pK_a(\text{methanol})] \tag{2}$$

The pK_a value of methanol [25] is 15.5 and that of imidazole [26] is 6.05, yielding an experimental value of 53.6 kJ/mol for ΔG. We compute the free energy change in solution utilizing a free energy perturbation method [27].

Table 4 *Microsolvation results*

Molecule	Orientation	Atom	HF/6-31 g(d)		Hybrid method	
			d(O-X) Å	E (kJ/mol)	d(O-X) Å	E (kJ/mol)
Imidazole	1	NE	3.09	– 30.4	2.83	– 30.6
	2	CD	3.75	– 4.1	3.70	– 3.6
	3	CG	3.60	– 10.0	3.66	– 8.4
	4	ND	3.16	– 27.8	2.96	– 26.3
	5	CE	3.54	– 10.2	3.57	– 10.5
Imidazolium	1	NE/ND	2.94	– 66.7	2.85	– 65.6
	2	CD/CG	3.22	– 39.9	3.18	– 41.5
	3	CE	3.12	– 51.4	3.31	– 46.9
Methanol	1	O2	2.99	– 22.5	2.60	– 28.4
	2	O2	2.99	– 23.8	2.72	– 21.6
Methoxide	1	O2	2.67	– 87.7	2.47	– 89.1
	2	C2	2.44	– 28.8	2.25	– 29.1

Orientations of the molecules correspond to those in Fig. 3. The columns labeled d(O-X) represent the distances between the oxygen in the water molecule and the heavy atom attacked in the small molecule.

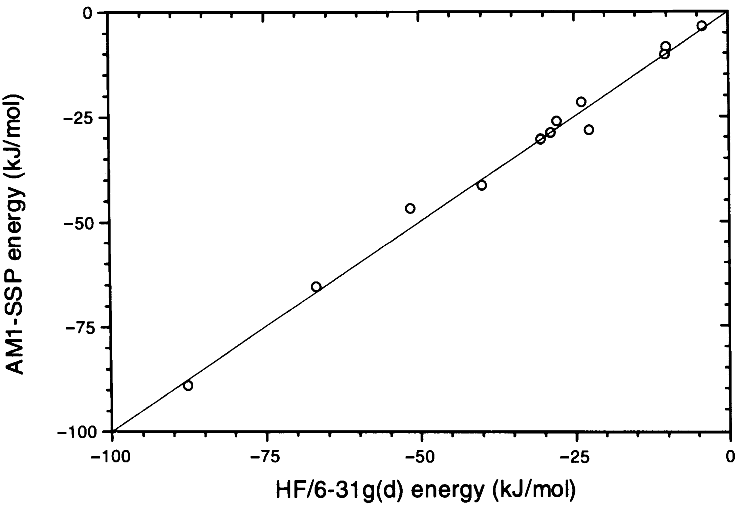


Fig. 4. Calibration of QM/MM interactions. Hybrid method interaction energies are shown plotted against HF/6-31 g(d) values. The solid line is drawn to guide the eye.

In this formalism, the system is characterized by a Hamiltonian $H(\mathbf{p}, \mathbf{q}, \lambda)$ that is a function of the coordinates $\mathbf{q}$ and conjugate momenta $\mathbf{p}$ and a multidimensional coupling parameter λ. The parameter λ serves to define a pathway between two states A and B. We define a sequence of discrete states represented by values λ_i, $i = 1, \ldots, N$, which transforms state A into state B via a series of suitably small steps. The free energy difference between two adjacent states is given by the following relation:

$$\Delta G(\lambda_i \rightarrow \lambda_{i+1}) = -RT \ln\langle \exp\{[H(\mathbf{p}, \mathbf{q}, \lambda_{i+1}) - H(\mathbf{p}, \mathbf{q}, \lambda_i)]/RT\}\rangle \quad (3)$$

where the term enclosed in angle brackets $\langle\,\rangle$ represents an ensemble average. The total free energy change from state A to state B is just the sum over all the intermediate steps, as given below:

$$\Delta G(A \rightarrow B) = \sum_i \Delta G(\lambda_i \rightarrow \lambda_{i+1}) \quad (4)$$

In computing the free energies, we use the ergodic hypothesis and assume that a time-averaged sampling over the structures as they evolve dynamically is equivalent to the actual ensemble average over all possible configurations. An estimate of the computational error that arises due to our discrete method (employed in Eqs. 3 and 4) can be obtained by computing the free energy change for the inverse reaction, proceeding from state B to state A. That is, at each state λ_i, we compute the free energy change for both the forward $\lambda_i \rightarrow \lambda_{i+1}$ and backward $\lambda_i \rightarrow \lambda_{i-1}$ directions.

To simulate the proton transfer reaction between methanol and imidazole, we immersed the solute molecules in an 18 Å radius ball of TIP3P water. This produced a model consisting of 2388 atoms: the 15 atoms of methanol and imidazole that are to be treated quantum mechanically and 791 water molecules in the MM partition. A deformable, stochastic boundary condition [28,29] was enforced on atoms in the region from 16 to 18 Å. The initial configuration of methanol and imidazole was transformed into methoxide and imidazolium by moving the proton from the O2 oxygen atom in methanol to the NE nitrogen atom in imidazole in a series of 0.05 Å steps. The distances between the proton and the two heavy atoms were constrained at each step, along the path shown in Fig. 5. Because the free energy is a thermodynamic state variable, the difference in free energy between states A and B is path-independent. Each intermediate state was equilibrated for 20 ps and data were collected for 10 ps using 1 fs molecular dynamics time steps. The computed free energy changes were 51.0 kJ/mol in the forward direction and 49.8 kJ/mol in the backward direction, which compare quite favorably with the experimental value of 53.6 kJ/mol. The free energy profile for the forward path is depicted in Fig. 6. The rather jagged patch in the center of the path is an artifact of the path we chose and is not physically significant; only the end point values are meaningful.

We note that the solvating water molecules have a significant influence on the reaction. In the gas phase, the experimental free energy change was 656.7 kJ/mol. In solution, there is a dramatic energy stabilization of the charges on the product species. It will be interesting to compare this result with the equivalent reaction in the enzyme environment.

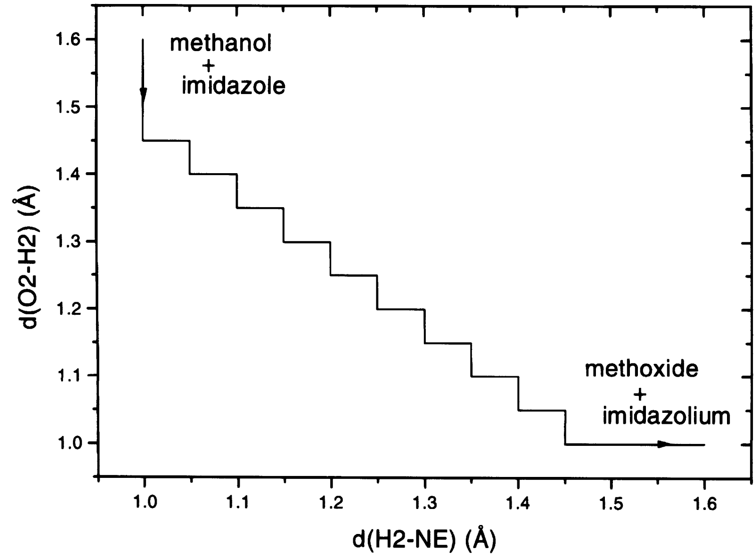


Fig. 5. Path taken during proton transfer reaction. The distances between the proton and the heavy atoms were constrained at each intermediate step.

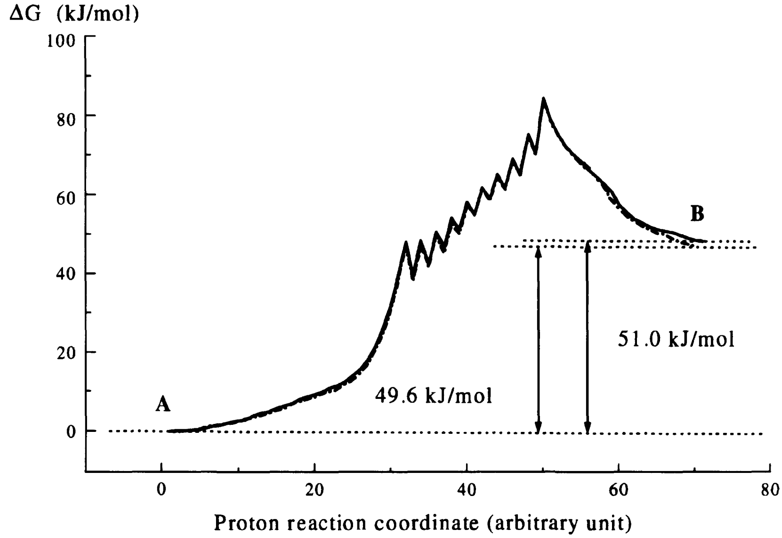


Fig. 6. Free energy profile of the proton transfer between methanol and imidazole. The pathway parameter λ represents a sequence of 70 intermediate states in which the distances between the proton and the two heavy atoms (O2 oxygen of methanol and NE nitrogen of imidazole) were varied in increments of 0.05 Å. The forward calculation is represented by a solid line; the backward by a dashed line.

Transferability of the QM/MM parameters

One last concern that we should address before beginning our studies of the reaction mechanism in the enzyme is the transferability of the QM/MM parameters. These parameters were established by the microsolvation procedure described above, defining the van der Waals parameters according to interactions with water molecules. The proton transfer reaction between methanol and imidazole in solution was well described by the QM/MM method but, in the protein, the interactions with residues in the active site will not always be with oxygen atoms. In particular, nitrogen atoms in the guanidinium groups of active-site arginine residues will play a key role in stabilizing the malate substrate. As a final check on the model Hamiltonian, we should examine some of the key interactions between atoms in the QM partition and side chains in the protein.

We have performed an extensive study of these interactions in small model systems [7] and found that, without further refinement, the parameters obtained through the microsolvation process can adequately describe the QM/MM interactions. In Fig. 7, we illustrate one example of these studies, in which a methyl-guanidinium (representing an arginine residue) interacts with an acetate ion (representing one of the carboxylate groups of malate). The methyl-guanidinium was placed in the MM partition

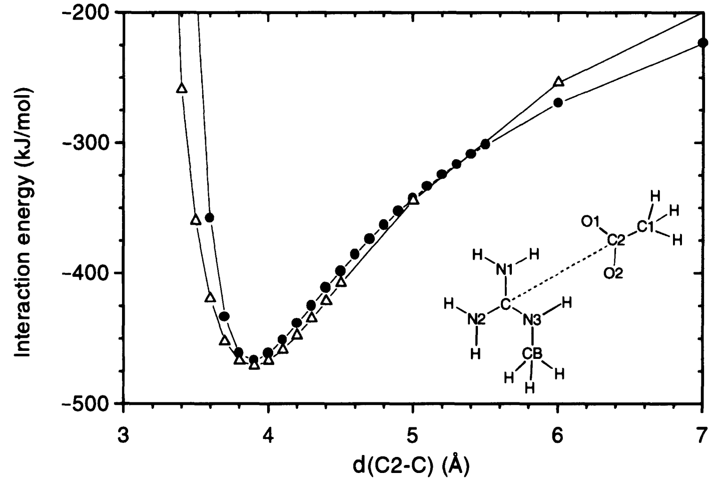


Fig. 7. Interaction energy of methyl-guanidinium and acetate. (●) QM/MM energies; (△) results of DFT calculations.

and the acetate in the QM partition. An analysis of the entire system was also performed with a density functional model where, again, the reactant structures were optimized independently and then were translated with one degree of freedom to obtain the interaction energy as a function of distance. Figure 7 is representative of the agreement we found; energies were reproduced within 20 kJ/mol and minimum-interaction distances to within 0.3 Å.

Reaction mechanism in the enzyme

With the calibration of the semiempirical Hamiltonian and van der Waals interaction parameters, we are now prepared to run simulations in the enzyme environment. We constructed a model of the enzyme by considering all of the amino acid residues within an 18 Å radius of the C2 carbon atom of malate, the NAD$^+$ cofactor and 39 water molecules deduced from the crystal structure. Another 105 water molecules were added by superimposing a 20 Å ball of TIP3P water and then removing all TIP3P molecules within 3.1 Å of non-hydrogen protein atoms, substrate, cofactor or crystal water molecules. Finally, another 27 water molecules were added from a resolvation procedure like that just described, but after 40 ps of molecular dynamics calculations with all atoms fixed except water molecules. Using solely a molecular mechanics description for all atoms, the system was heated from 0 K to 300 K over an interval of 20 ps with atom velocities assigned from a Gaussian distribution every 2 ps in 30 K increments. The system was then equilibrated for another 80 ps, followed by 40 ps of data collection to define the Michaelis complex illustrated in Fig. 1. At this point, the QM/MM calculations were initiated and 20 ps of equilibration was performed, followed by 20 ps of data collection.

One concern about the QM/MM method is whether the dynamical behavior of atoms in the QM partition is accurately depicted. We can compute the root-mean-square deviation between structures and find that the difference between the Michaelis complex defined by the MM simulations and the equilibrated structure defined by the QM/MM calculations is 0.16 Å for α-carbon atoms and 0.39 Å for all atoms. We conclude that, when properly calibrated, the QM/MM method provides a realistic dynamical model of the enzyme–substrate system.

To explore the reaction mechanism of the MDH:malate:NAD$^+$ system, we could employ the free energy perturbation method that we utilized in the proton transfer study between methanol and imidazole. A somewhat less computationally intensive alternative is to examine the minimum-energy surface of the reaction. We started with the QM/MM equilibrated structure defined above and annealed it from 300 K to 0 K over 20 ps while constraining the H2 proton and H21 hydride to be equidistant (1.3 Å) from the donor and acceptor atoms. We then minimized the energy of the resulting configuration for 5000 steps, maintaining the constraints on the proton and hydride positions. The distances between the O2 oxygen atom of the malate substrate and the H2 proton, between the H2 proton and the NE2 nitrogen atom of His^{177} of MDH, between the C2 carbon atom of the malate substrate and the H21 hydride ion, and between the H21 hydride ion and the C4N carbon atom of NAD$^+$ were varied on a four-dimensional grid with 0.2 Å spacing. An energy minimization of 1000 steps was performed at each grid point, resulting in 675 separate minimum-energy values in the four-dimensional space.

We produced a minimum-energy surface in the following manner. For each of two degrees of freedom, the distances between (i) the proton and (ii) the hydride ion and

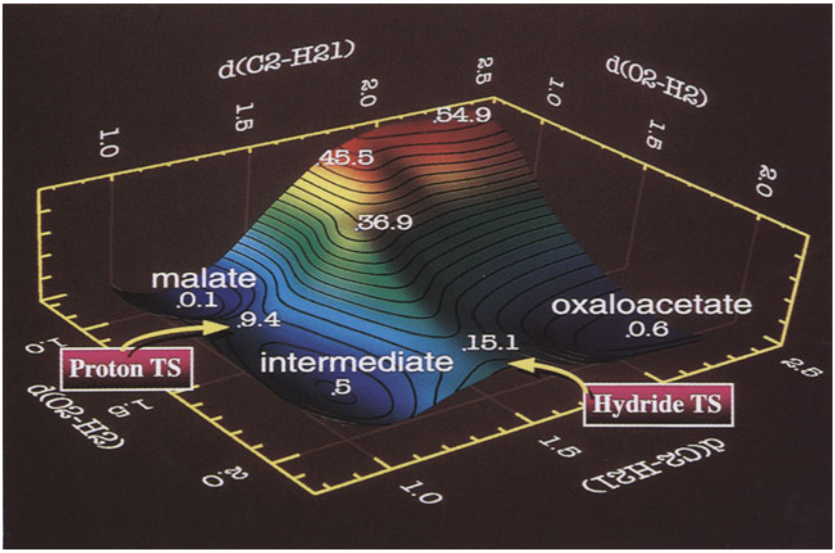


Fig. 8. Minimum-energy surface of the enzyme-catalyzed reaction. Reprinted by permission of Biochemistry. Copyright 1997 American Chemical Society.

the malate substrate, the minimum-energy value was sought from the possible values of the other two degrees of freedom. A plot of this reaction surface is depicted in Fig. 8. What is striking about the minimum-energy surface is the large barrier facing an initial hydride transfer. The minimum-energy pathway clearly indicates that the

the malate substrate, the minimum-energy value was sought from the possible values of the other two degrees of freedom. A plot of this reaction surface is depicted in Fig. 8.

proton transfer occurs first, followed by the hydride transfer. This is somewhat surprising due to the fact that the intermediate state after the proton transfer has taken place will have a net charge of − 3e on the substrate. In the gas phase, we might have expected the hydride reaction to proceed first, to minimize the charge separation in the intermediate state. As we saw in the study of the proton transfer in solution, however, solvation effects can be quite large. The environment of the enzyme clearly provides some stabilizing effects on the charged intermediate state [7]. We can see how important these solvation effects are by the following analysis. For each of the minimum-energy states defined by the surface in Fig. 8, we performed a single-point energy calculation in which all the charges on MM atoms were set to zero, thereby removing the principal solvation effects from the calculated energies. We plot these data in Fig. 9 and note that there is now a large barrier opposing the initial proton

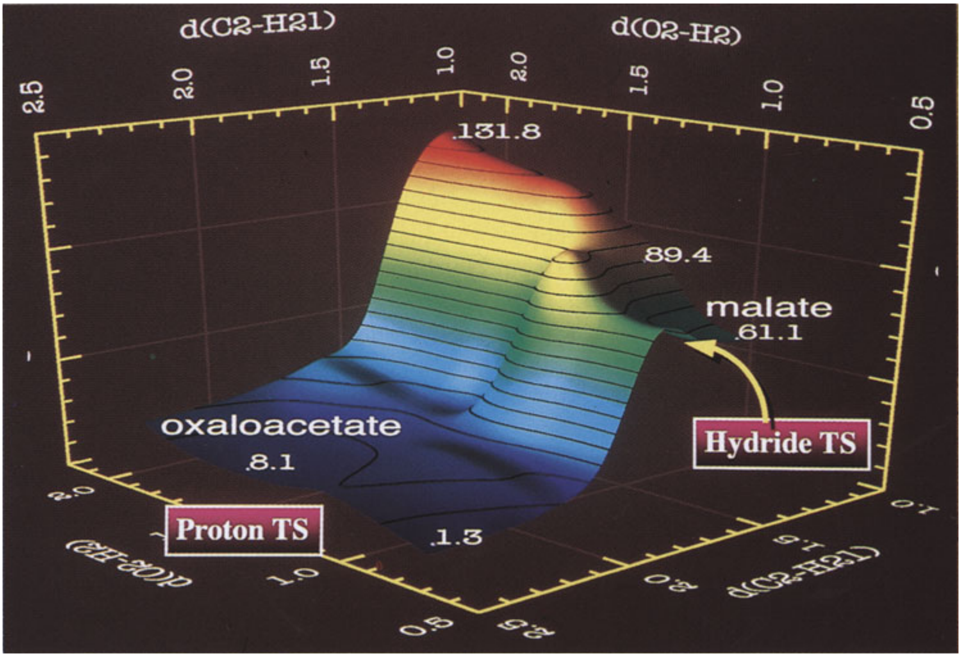


Fig. 9. ‘Gas-phase’ reaction surface. Single-point energy calculations were performed for the minimum energy states defined by the surface in Fig. 8 with the MM charges set to zero. Note that this figure is reoriented with respect to Fig. 8. Reprinted by permission of Biochemistry. Copyright 1997 American Chemical Society.

transfer. Without the solvation effects due to the enzyme, the hydride transfer would occur first. Consequently, we can see directly in the MDH:malate:NAD$^+$ system that the catalytic properties of the enzyme are not solely due to the proper orientation of the substrate and proximity of the reacting element. Solvation effects due to other residues present in the active site can have significant impact on the reaction mechanism.

Conclusions

We have demonstrated that realistic simulations of enzyme–substrate systems are possible with currently available computing resources. The key elements in our method are the use of a quantum mechanical description of atoms in the active site of the enzyme, calibration of the semiempirical quantum Hamiltonian, calibration of the interactions between atoms in the QM partition and those in the MM partition and, of course, good experimental data on the crystal structure of the enzyme. Without the quantum description of atoms in the active site, we would be unable to treat the important bond formation events that define the catalytic process. Unfortunately, the limitations of present algorithms and computing resources require that we use a semiempirical quantum method, but we have demonstrated that it is possible to calibrate the method against experimental data and produce results which rival those of the best *ab initio* quantum approaches, albeit in a limited set of circumstances. Furthermore, by calibrating the interactions of atoms in the QM partition with those in the MM partition, it is possible to produce realistic dynamics calculations. In essence, one can turn on the quantum description of the atoms without affecting the dynamics of the system as a whole. In this way, we can directly compute the effects of the protein matrix on the reaction mechanism, without resorting to any *ad hoc* schemes for estimating the influence of active-site residues. We note that recent experiments by John Burgner's group at Purdue are consistent with our proposed mechanism, lending some credence to our belief that the simulations produce a realistic description of the enzyme–substrate system.

Finally, for the case of the MDH:malate:NAD$^+$ system, it appears as though the enzyme produces an environment much like the solvating environment of aqueous solution, providing for the stabilization of charged intermediate states. Additionally, following the lock-and-key hypothesis, the active-site residues serve to orient the malate substrate into a configuration that is optimal for the subsequent chemical events. We are now in the process of extending this effort to consider other aspects of the MDH system, such as substrate specificity and the changes in activity brought about by modification of important subgroups. These are questions of prime importance to the complete understanding of enzyme systems and are questions that can be addressed by numerical simulation, working in concert with careful experimentation.

Acknowledgements

This work was supported in part by the U.S. Department of Energy, Office of Health and Environmental Research, under Contract W-31-109-Eng-38.

References

1. Curtiss, L.A., Raghavachari, K., Redfern, P.C. and Pople, J.A., J. Chem. Phys., 106(1997)1063.
2. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107(1985)3092
3. Wilkie, J. and Williams, I.H., J. Am. Chem. Soc., 114(1992)5422.
4. Andrés, J., Moliner, V. and Safont, V.S., J. Chem. Soc., Faraday Trans., 90(1994)1703.
5. Ranganathan, S. and Gready, J.E., J. Chem. Soc., Faraday Trans., 90(1994)2047.
6. Brooks, C.L., Karplus, M. and Pettit, B.M., Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics, Advances in Chemical Physics, Vol. LXXI, Wiley, New York, NY, 1988.
7. Cunningham, M.A., Ho, L.L., Gillilan, R.E. and Bash, P.A., Biochemistry, 36(1997)4800.
8. Hall, M.D. and Banaszak, L.J., J. Mol. Biol., 232(1993)213.
9. Warshel, A. and Levitt, M., J Mol. Biol., 103(1976)227.
10. Singh, U.C. and Kollman, P.A., J. Comput. Chem., 7(1986)718.
11. Field, M.J., Bash, P.A. and Karplus, M., J. Comput. Chem., 11(1990)700.
12. Parker, D.M., Lodola, A. and Holbrook, J.J., Biochem. J., 173(1978)959.
13. Lodola, A., Shore, J.D., Parker, D.M. and Holbrook, J.J., Biochem. J., 175(1978)987.
14. Dewar, M.J.S. and Thiel, W., J. Am. Chem. Soc., 99(1977)4899.
15. Rossi, I. and Truhlar, D.G., Chem. Phys. Lett., 233(1995)231.
16. Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
17. Syswerda, G. and Schaffer, J., Proceedings of the 3rd International Conference on Genetic Algorithms, Morgan Kaufmann, New York, NY, 1989, pp. 2–8.
18. Whitley, D. and Kauth, J., Rocky Mountain Conference on Artificial Intelligence, Morgan Kaufmann, New York, NY, 1988, pp. 118–125.
19. Bash, P.A., Ho, L.L., MacKerell Jr., A.D., Levine, D. and Hallstrom, P., Proc. Natl. Acad. Sci. USA, 93(1996)3698.
20. Hehre, W.J., Radom, L., Schleyer, P. and Pople, J.A., Ab Initio Molecular Orbital Theory, Wiley, New York, NY, 1988.
21. MacKerell Jr., A.D. and Karplus, M., J. Phys. Chem., 95(1991)10559.
22. Jorgensen, W.L., Chandrasekar, J., Madura, J., Impey, R.W. and Klein, M.L., J. Chem. Phys., 79(1983)926.
23. Lias, S.G., Bartmess, J.E., Liebman, J.F., Holmes, J.L., Levin, R.D. and Mallard, W.G., J. Phys. Chem. Ref. Data, 17(Suppl. 1)(1988)1–872.
24. McClellan, A.L., Tables of Experimental Dipole Moments, Vol. 2, Rahara Enterprises, El Cerrito, CA, 1974.
25. Maskill, H., The Physical Basis of Organic Chemistry, Oxford University Press, New York, NY, 1989.
26. Dawson, R.M.C., Elliot, D.C., Elliot, W.H. and Jones, K.M., Data for Biochemical Research, 3rd Ed., Oxford University Press, New York, NY, 1986.
27. Brooks, C.L. and Karplus, M., J. Mol. Biol., 208(1989)159.
28. Van Gunsteren, W.F. and Berendsen, H.J.C., Mol. Phys., 34(1977)1311.
29. Brooks, C.L. and Brunger, A., Biopolymers, 24(1985)843.

Part II
Electrostatics and solvation

Modeling protonation equilibria in biomolecules

Michael K. Gilson
*Center for Advanced Research in Biotechnology, National Institute of Standards and Technology,
9600 Gudelsky Drive, Rockville, MD 20850-3479, U.S.A.*

Introduction: The importance of protonation equilibria

pH is one of the fundamental physiological variables. In fact, serious illness is usually associated with deviations of blood pH only a few tenths from the normal value, 7.4. It is very unlikely that the effects of abnormal pH result from the destructive chemical action of protons or hydroxide; if this were the case, a blood pH of 6.8, say, would not be dangerous. It is more likely that the sensitivity of organisms to pH results chiefly from the influence of pH upon the stability and reactivity of the many biomolecules that possess acidic and basic chemical groups. If this is correct, then the ill effects of abnormal pH values should become substantial for pH changes that correspond to free energy changes equal to the thermal energy, ~RT. Given that $\Delta\Delta G = \Delta pH(RT/\log e) = \Delta pH(RT/0.43)$ [1], the critical pH change should be about 0.4. In fact, this defines quite well the range of blood pH values that are tolerable to humans.

Biomolecules would be far less sensitive to pH if they lacked chemical groups that titrated near pH 7. For example, the stability of a protein might well be fixed between pH's 5 and 9 if it lacked histidines, free cysteines, and an N-terminal amine group. Why, then, should natural selection have generated biomolecules that are exquisitely sensitive to pH? Part of the explanation may have to do with the importance of proton transfers in enzyme catalysis. In order for a chemical group in an enzyme to be an efficient general acid or base, it must be poised to surrender or abstract a proton. It must also revert readily to its initial protonation state in order to regenerate the active enzyme. In order for a group to meet these requirements, its pK_a must be poised near the ambient pH. Therefore, catalysis will necessarily be sensitive to pH.

The dependence of structure and function upon pH poses important challenges to those who wish to understand or engineer the properties of biomolecules. Optimally, one would like to be able to compute accurately the fractional charge of important ionizable groups as a function of molecular conformation and of pH. This capability would be of enormous value in elucidating the mechanisms of enzymes, and in predicting conformational properties such as the stability of folded proteins and of noncovalent complexes. Tautomer equilibria, correlations among the protonation states of multiple groups, and the kinetics of proton transfer will also be important in some cases.

This chapter presents an overview of the efforts to model protonation equilibria accurately. It presents brief discussions of the observed pK$_a$'s of ionizable groups in proteins, and of the theory of multiple-site ionization equilibria. It then sketches several computational approaches to the problem, focusing upon use of the Poisson–Boltzmann (PB) model for electrostatic interactions to compute protonation equilibria in proteins. Applications of the PB method of computing pK$_a$'s to the mechanism of acetylcholinesterase, and to the prediction of the pH-dependence of protein stability are then presented. Finally, promising areas for future research are suggested.

Observed pK$_a$ shifts in proteins, and the 'Null' model

Before presenting a discussion of the methods for predicting pK$_a$'s in proteins, it is worth examining the rapidly growing body of measured pK$_a$'s. The histogram of Fig. 1 shows that the pK$_a$'s of ionizable groups in proteins are often close to the pK$_a$'s of chemically similar compounds – 'model compounds' – in solution [2,3]. The distribution of pK$_a$ shifts, relative to model-compound values, is centered at 0, and most shifts are less than 1 pK$_a$ unit in magnitude. However, there is considerable scatter: some outliers are shifted by as much as 3 and even 4 pK$_a$ units. Not surprisingly, residues that are thoroughly solvated at the surface of the protein usually show only small pK$_a$ shifts [3], and the largest shifts are associated with groups that are at least partially sequestered from solvent. For example, the most remote outlier in the histogram is the pK$_a$ of a lysine side chain artificially placed in the hydrophobic interior of staphylococcal nuclease [4]. Its large pK$_a$ shift, measured indirectly through the thermodynamic linkage of its titration to the unfolding equilibrium of the protein, appears to result almost entirely from desolvation, rather than from

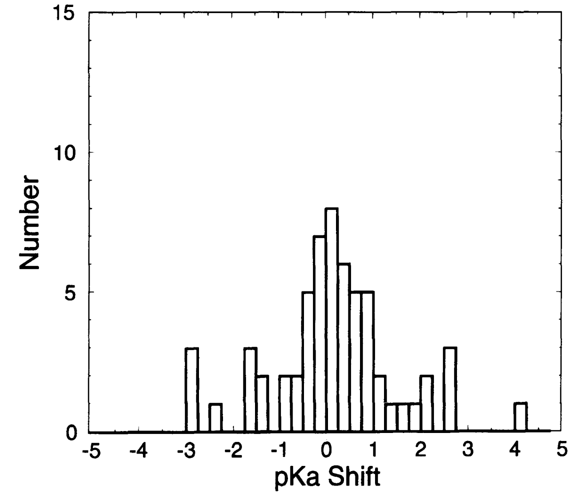


Fig. 1. Histogram of pK$_a$ shifts for 60 ionizable groups in several globular proteins.

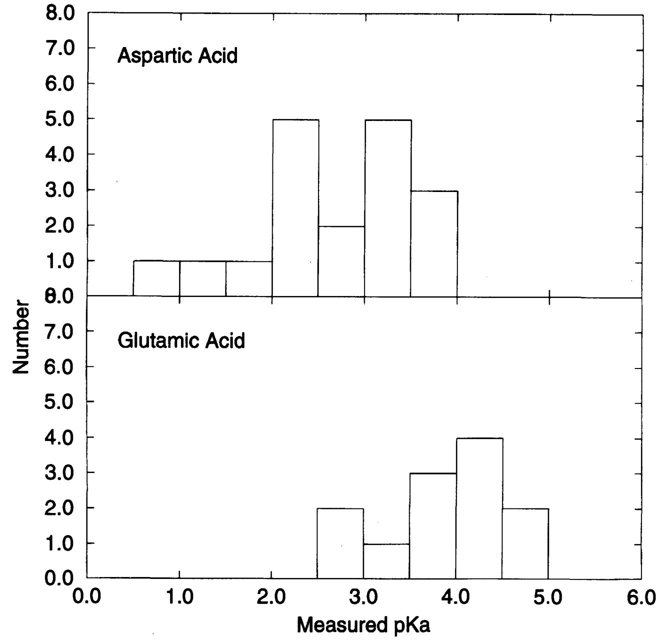


Fig. 2. Top: distribution of measured pK_a*'s for 19 aspartic acid side chains; bottom: distribution of measured* pK_a*'s for 13 glutamic acid side chains.*

interactions with polar or ionized neighbors. Buried ionizable groups introduced into T4 lysozyme [5] and myoglobin [6] also have large pK_a shifts. On the other hand, not all desolvated ionizable groups have large pK_a shifts; for example, the pK_a of the partly desolvated catalytic His of chymotrypsin is about 7 [7,8].

Because the pK_a's of most ionizable groups in proteins are not shifted much relative to model compounds, the simplest reasonable approach to estimating these pK_a's is to assume that they equal the model-compound values. Accordingly, it has been argued that computational models should, at the least, be more accurate than this trivially simple approximation, which has been termed the 'Null' model [3]. This approximation is frequently used in setting up molecular dynamics simulations of biomolecules, with the additional constraint that each ionizable group must be either fully protonated or fully deprotonated.

Further analysis of the measured pK_a's we have so far assembled yields an unexpected observation: aspartic side chains tend to be significantly more acidic than glutamic side chains. The mean pK_a of all 19 aspartic side chains is 2.7, but the mean pK_a of the 13 glutamic acids is 4.0. The distributions of the pK_a's of these groups are compared in the histograms of Fig. 2. As is shown in Table 1, this result holds not only for the pooled data, but also for each separate protein. The pK_a's of ethanoic (acetic) acid, propanoic acid, and butanoic acid differ by less than 0.1 pK_a unit [9]. This

Table 1 *Mean experimental pK$_a$'s of Asp and Glu*

Protein	Asp		Glu	
	N	Mean	N	Mean
HEWL	7	2.6	2	4.6
RNASE A	5	3.1	5	3.8
BPTI	2	3.2	2	3.8
OMTKY3	2	2.5	3	4.0
CHYMO	2	2.0	0	na
T4	1	2.0	0	na
RNASE T1	0	na	1	4.3
Cumul	19	2.7	13	4.0

N: number of measured pK$_a$'s in each protein. (Most data are from NMR studies.) Mean: average pK$_a$ of group; Cumul: cumulative means; HEWL: hen egg-white lysozyme [3,87]; RNASE A: ribonuclease A [88] (see also the discussion in [89]); BPTI: bovine pancreatic trypsin inhibitor [90–92]. OMTKY3: turkey ovomucoid third domain [93,94]. CHYMO: chymotrypsin [7,8]; T4: T4 lysozyme [95]; RNASE T1: ribonuclease T1 [96,97].

suggests that it is the protein environment of aspartic side chains that causes them to be more acidic than glutamic side chains, rather than something intrinsic to the group. This issue is revisited later in this chapter. For now, it appears that the investigator confronted with an aspartic acid residue of uncertain pK$_a$ would do best to assume a pK$_a$ of about 2.7 rather than the customary 4.0. More generally, it is becoming possible to establish a more sophisticated Null model, in which the pK$_a$ of every group of a certain type, say aspartic acid, is set to the mean pK$_a$ observed for groups of that type in proteins. However, for the data we have examined to date, aspartic acid is the only group whose mean pK$_a$ in proteins deviates significantly from the model-compound pK$_a$ [10].

Theory of multiple-site protonation equilibria

Modeling protonation equilibria is particularly challenging because of the large number of ionizable groups in most proteins, and the long range of the electrostatic interactions among these groups. This section discusses the theory of linked protonation equilibria, also known as multiple-site titration.

Because each of the N ionizable groups in a protein may exist in two chemically different protonation states (more for groups with different tautomers), the number of possible ionization states for a protein is at least 2^N. Computing the influence of protonation equilibria upon a protein involves, at least in principle, accounting for all of these states. For example, the fractional ionization of a titratable group, i, is given by a Boltzmann average of its charge over all ionization states, α, of the protein:

$$\langle x_i \rangle = \frac{\sum_{\alpha=0}^{2^N-1} x_\alpha(i) \exp\left[-\frac{\mu_\alpha^o - \mu_0^o - q_\alpha \mu_{H^+}^o}{RT} \right]}{\sum_{\alpha=0}^{2^N-1} \exp\left[-\frac{\mu_\alpha^o - \mu_0^o - q_\alpha \mu_{H^+}^o}{RT} \right]} \tag{1}$$

Here $x_\alpha(i)$ is 0 when group i is not ionized in state α, and 1 if it is ionized; μ_α^o is the standard chemical potential of the protein in ionization state α; μ_0^o is the standard chemical potential of an arbitrarily selected reference ionization state indexed by 0; q_α is the number of protons gained by the protein on going from the reference state, 0, to state α; and R and T are the gas constant and the absolute temperature, respectively. The thermodynamic averages of quantities other than charge take similar forms.

Processes whose equilibrium constants depend upon pH are coupled to protonation equilibria; conversely, a process is coupled to protonation if it perturbs the energetics of protonation of one or more ionizable groups. Protein denaturation and the binding of ligands by proteins are two important processes that are often coupled to protonation. Such couplings may be analyzed by considering the free energy change for the process – say protein folding – in the reference protonation state 0 (ΔG_0^o), and then adjusting for the change in the populations of protonation states that occurs upon folding (see Fig. 3). This adjustment may be written in terms of a binding polynomial, Σ(pH), for the multiple proton-binding sites [1,11–13]. The standard free energy change, ΔG^o for the complete process is

$$\Delta G^o = \Delta G_0^o + \Delta\Delta G_{ion}^o(pH) \tag{2}$$

where

$$\Delta\Delta G_{ion}^o(pH) = \Delta G_{ion,2}^o(pH) - \Delta G_{ion,1}^o(pH) = -RT\ln\frac{\Sigma_2(pH)}{\Sigma_1(pH)}$$

$$\Sigma_j(pH) \equiv \sum_{\alpha=0}^{2^N-1} K_{\alpha,j} 10^{-pHq_\alpha}$$

$$K_{\alpha,j} = \left(\frac{[P_\alpha]}{[P_0][H^+]^{q_\alpha}}\right)_{equilibrium} = \exp\left[-\frac{\mu_{\alpha,j}^o - \mu_{0,j}^o - q_\alpha \mu_{H^+}^o}{RT} \right] \tag{3}$$

Here $K_{\alpha,j}$ is the equilibrium constant connecting the protein in its reference state, P_0, and in its ionization state, P_α; the subscript j = 1,2 indicates values before and after the folding reaction, respectively; and square brackets indicate concentration in standard units.

The equations presented here demonstrate that the problem of computing the pH-dependence of the properties of proteins can be divided into two parts: (i) computing the equilibrium constants, K_α, that connect the various ionization states; and (ii) using the many values of K_α to compute thermodynamic quantities of interest. Neither problem would be difficult if the pK_a's of individual ionizable groups in proteins equaled those of chemically similar groups in solution, and if the groups did

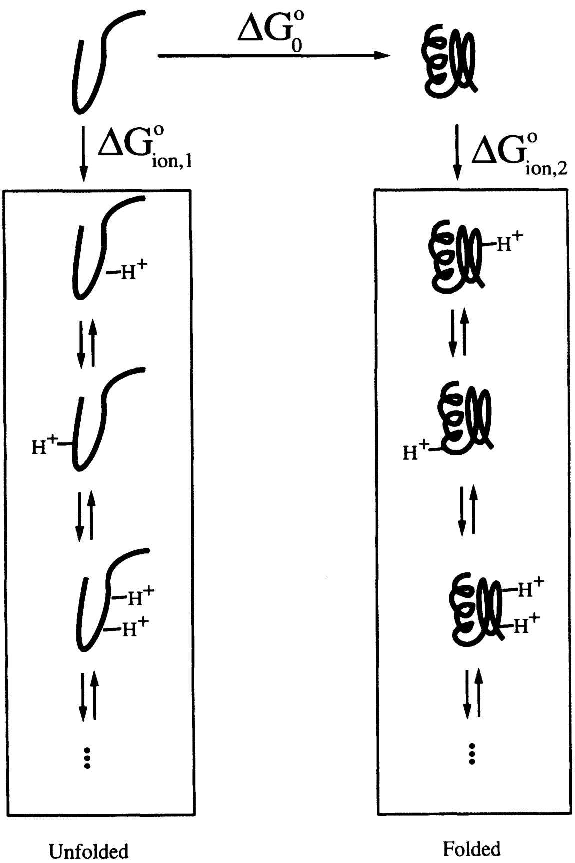


Fig. 3. Linkage of protein folding with protonation. ΔG_0^o: change in standard chemical potential of the protein upon folding, in the reference protonation state; $\Delta G_{ion,1}^o$: change in standard chemical potential of unfolded protein when the unfolded protein in the reference protonation state is allowed to equilibrate with buffer at some pH; $\Delta G_{ion,2}^o$: same as $\Delta G_{ion,1}^o$, but for folded protein. Protons have been left out of the stoichiometry for brevity.

not interact. Then the equilibrium constants K_α could be computed trivially from the pK_a's of the individual groups. What makes the problem interesting is that the protein environment does perturb the pK_a of each group, and that the groups do interact with each other.

Formulating the problem

In general, the energy of changing the protonation state of an ionizable group in a protein is determined by its gas-phase proton affinity, and by interactions with solvent, with other parts of protein, and with other solutes. Computing the work of protonation *ab initio* is a daunting challenge. The problem is made more tractable when one starts with the work of protonating a chemically similar group in solution, and computes only the perturbations that result from the protein environment. Most methods in use or under investigation today take this approach. The fact that pK_a shifts in proteins are usually small implies that these perturbations are usually modest.

Computational methods for computing the effect of the protein environment upon the energetics of protonation may be constructed in a number of different ways. Here, these methods are divided into detailed simulations that account, at least in principle, for nonlinear responses to changes in protonation states; and methods that assume additivity of the energy contributions to the overall chemical potentials of a molecule in a given ionization state.

Detailed simulations

Ideally, one would compute the effect of the protein environment by free energy simulations that explicitly include many solvent molecules and that allow for the conformational fluctuations of the biomolecules. In such approaches, the free energies cannot be said to be purely electrostatic, because changes in electrostatic interactions associated with ionization are coupled to other components of the molecular Hamiltonian. (For a discussion of related issues, see Refs. 14–16.) It might be expected that such approaches would provide highly accurate results, because they represent the systems in great detail. In fact, it is difficult to assess the accuracy of this approach. One report of pK_a calculations by free energy simulation suggests that the accuracy of the method is about $\pm 3\,pK_a$ units [17]. This is less than the accuracy of the Null model, but the report focuses upon a small number of relatively difficult cases, so it is not possible to draw broad conclusions about accuracy.

Convergence of the calculations is also a concern: free energy simulations for a glutamic acid in ribonuclease T1 display large differences in the ionization energies computed by forward and backward integration [18]. The greatest difference, 20 kJ/mol (3.4 pK_a units), is observed in the reference calculation for acetic acid in water. Test calculations on sodium in water suggest that these differences might be diminished if long-range interactions were included [19]. The results of pK_a calculations by free energy simulations that do include long-range electrostatic interactions are promising, but there still appear to be convergence problems: pK_a's from free energies averaged over eight short (22 ps) molecular dynamics (MD) trajectories differ by 0.1–2.9 pK_a units from the pK_a's computed from a single long (110 ps) trajectory [19].

In summary, free energy simulations represent an elegant approach to computing pK_a's. Improvements in computational procedures and computer speed continue to increase their utility. At present, however, they are still subject to convergence problems. Also, because they are time-consuming, it is difficult to accumulate enough comparisons with experiment to assess their validity.

Treating multiple-site titration by assuming additivity

Because free energy simulations are very time-consuming, it is difficult to use them to compute more than a few of the equilibrium constants, K_α, that are required for computing pH-dependent properties. The problem of rapidly computing the differences in chemical potential among many ionization states of a protein may be solved efficiently if it is assumed that the response of the system to ionization is linear. This assumption permits the construction of a symmetric matrix, $\| G_{ij} \|$, of interaction free energies among the ionizable groups. It is necessary to define a reference ionization state, 0, for the whole system; here, this will be the state in which all groups are neutral, but any state is valid. Then each diagonal term (i = j) is the difference between the work of ionizing i in the protein with all other groups neutral, and the work of ionizing the same group in solution. Each off-diagonal term (i ≠ j) is the additional free energy contribution when groups i and j are ionized simultaneously in the protein. Then the equilibrium constant connecting an arbitrary ionization state α to the reference state is

$$K_\alpha = \exp\left[-\beta \sum_{i=1}^{N} x_\alpha(i)\left(G_{ii} + \sum_{j>i}^{N} x_\alpha(j)G_{ij}\right)\right] \prod_{i=1}^{N} K_{ai}^{-x_\alpha(i)z(i)} \tag{5}$$

where $\beta \equiv (RT)^{-1}$, z(i) is 1 for bases and −1 for acids, and K_{ai} is the equilibrium constant for ionizing the model compound corresponding to group i. The other terms are defined above. Thus, the assumption of additivity makes it possible to compute each of the $2^N - 1$ equilibrium constants, K_α, from the far fewer $(N(N + 1)/2)$ matrix elements in $\| G_{ij} \|$. The remainder of this chapter is restricted to algorithms based upon this approach.

1. Simulation-based method for computing the matrix of interactions

First-order response theory can be used to extract the terms in the matrix $\| G_{ij} \|$ from detailed simulations [20]. The response theory would be exactly applicable if the fluctuations in electrostatic potential were normally distributed. This approach has been used to compute the diagonal terms of the interaction matrix for lysozyme [21]. The approach appears to be quite promising, but to date it does not seem to have overcome the same problems of convergence and accuracy associated with free energy simulations.

2. Electrostatic models for computing the matrix of interactions

Currently, the matrix elements are most frequently computed from solutions of the linearized Poisson–Boltzmann (LPB) equation [22]. This approach implicitly assumes that the perturbations of pK_a's in proteins are dominated by electrostatic interactions. The reliability of this approximation has been debated [3,23–25]. Empirically, the success of electrostatic models at predicting pK_a's (see below) suggests it is reasonably good.

The electrostatic approach to computing pK_a shifts in biomolecules dates at least to the smeared-charge model, which was published in 1924 [26], long before the structure of any protein had been solved to atomic resolution. The smeared-charge model treats the protein as a giant spherical Born ion [27], with the net charge of all its ionizable groups smeared out over its surface. In 1957, still without the benefit of fast computers or detailed information on the structure of any protein, Tanford and Kirkwood added detail, modeling individual ionizable groups as discrete charges in arbitrarily chosen locations near the surface of a spherical protein of low dielectric constant [28,29]. The solvent was treated as a high dielectric continuum, and the influence of ionic strength was incorporated by allowing the dissolved electrolyte to redistribute in response to the electrostatic potential of the protein in accord with the LPB equation. When protein structures and detailed information on the pK_a's of individual groups became available, the equations of the Tanford–Kirkwood model were used to estimate interactions among ionizable groups, now including the experimentally determined distance between each pair of groups [30,31]. Interestingly, it was discovered that the accuracy of the results could be increased by weakening the interactions of a group with others by a factor related to its accessibility to solvent. The resulting modified Tanford–Kirkwood (MTK) model was relatively successful at reproducing experimental data [32].

However, the MTK model remained restricted to a spherical representation of the protein. Also, like its predecessors, it did not yield the diagonal elements of the matrix $\|G_{ij}\|$. That is, pK_a's were assumed to be perturbed only by interactions among ionizable groups; the influence of neutral dipolar groups was neglected, as were desolvation effects. The work of moving an ionizable group from the high-dielectric solvent into the low-dielectric interior of the protein [33,34] can lead to large pK_a shifts, such as the ~ 4 unit shift of the lysine artificially introduced into the hydrophobic interior of staphylococcal nuclease [3,4].

The problem of treating the shape of the protein accurately was addressed by Warwicker and Watson, who solved the Poisson equation for a protein by the method of finite differences [35]. This approach was then used to solve the LPB equation [36], and the finite-difference method was soon used to compute a pK_a shift in the active site of subtilisin due to mutations of charged groups about 10 Å away [37,38]. Charge–charge interactions, G_{ij}, were computed as the interaction of the charges of one group with the potential generated by the other. The interactions of charged groups with neutral dipolar groups in the protein may be computed in the same way [39]. These interactions, dubbed 'background' terms [13], contribute to the diagonal terms, G_{ii}, in the matrix of interactions.

However, these advances did not yet offer a way of computing the electrostatic work of transferring a polar group from bulk solvent to the low-dielectric interior of a protein – the 'Born' term. This problem was solved in 1988 [40]. Briefly, the electrostatic potential, ϕ_k, at each charged or partially charged atom k is computed for the group in bulk solvent, and then in the interior of the fully discharged protein (see Fig. 4). The difference in the total electrostatic energy, $\frac{1}{2}\sum_k q_k(\phi_k^{\text{protein}} - \phi_k^{\text{solvent}})$, is the work of transfer. The calculation can be arranged so that the singularity – or near-singularity – in the potential at a charge cancels when the difference is taken. In certain cases, a single finite-difference solution of the LPB equation can be used to

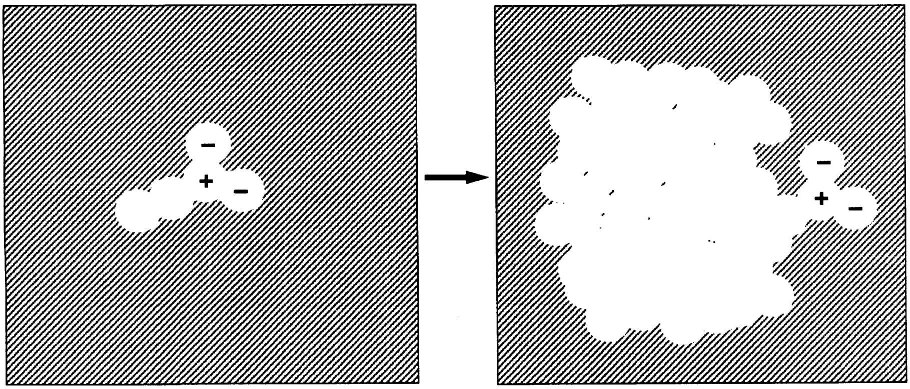


Fig. 4. Diagram of the process for computing the work of transfer of a polar side chain from solvent to a hypothetically neutral protein, the 'Born' term.

compute the work of transfer [41]. These methods for computing electrostatic energies in the context of the LPB equation were used in 1990 to establish a complete matrix of energy terms, and thus to compute the pK_a's of most of the ionizable groups in hen egg white lysozyme [13]. Because this approach and its variants are now in common use, it is worth summarizing here. This discussion focuses upon use of the finite-difference method for solving the LPB equation, but similar approaches can be set up with other numerical methods.

In its simplest form, the method treats ionization as the addition of a ±1e charge to one atom of an ionizable group. This might be the N^{ζ} of a lysine, the C^{γ} of an aspartic acid, and so forth. This atom will be termed the ionization site. The LPB equation is solved twice for each group, for a total of 2N calculations. In each calculation, the only source is a unit charge at the ionization site. In the first calculation, the group is in the dielectric and ionic environment defined by the protein and solvent. The boundary between protein and solvent is typically the molecular surface (contact + reentrant) defined by Richards [42]. The atomic radii that define the boundary are equal or similar to standard van der Waals radii. This first calculation gives information about the reaction field produced by the solvent at the charge itself, and also about the interaction of the charge with charges at any other location in the protein. In the second calculation, the group has been artificially removed from the protein, and is surrounded by solvent. This calculation gives information about the reaction field produced by the solvent at the charge when the group is fully solvated. It is assumed that the pK_a of the group in this state equals the model-compound value; the change in the pK_a of the group when it is moved from bulk solvent to the protein is directly related to the change in the computed ionization energy for the group in the solvent versus that in the protein. The calculations just described are used as follows to compute the elements of the interaction matrix, $\|G_{ij}\|$.

The off-diagonal terms that represent group–group interactions are

$$G_{ij} = z_i z_j \Phi_{ij} \tag{5}$$

where Φ_{ij} is the electrostatic potential at the ionization site of group j due to a unit charge at the ionization site of group i. The potentials are those computed in the protein. The principle of reciprocity implies that $\Phi_{ij} = \Phi_{ji}$, so the matrix is symmetric.

The diagonal terms that represent the work of ionizing each group i in the otherwise un-ionized protein, relative to the work of ionizing it in solution, are given by

$$G_{ii} = \frac{1}{2}(\Phi_{ii} - \Psi_{ii}) + z_i \sum_{k=1}^{M_{prot}} \Phi_{ik} q_k - z_i \sum_{k=1}^{M_i} \Psi_{ik} q_k \tag{6}$$

where Ψ_{ik} is the potential at the atom site of group i due to the charge at atom k, when the group is completely surrounded by solvent; q_k is the partial charge on atom k; the first sum extends over all M_{prot} partially charged atoms of the protein, such as main-chain amides, dipolar side chains, and the neutral forms of ionizable groups; and the second sum extends only over the M_i partially charged atoms of ionizable group i. The first part of Eq. 6 is the Born term. Note that the singularities (or near-singularities for the finite-difference method) in Φ_{ii} and Ψ_{ii} cancel when their difference is taken. The last two terms in Eq. 6 yield the change in the 'background' energy when the group is inserted into the protein. The atomic charges, q_k, are typically drawn from an empirical force field, such as those used in molecular dynamics simulations.

More detailed representations of the charge redistributions that result from ionization may be accommodated within this scheme, at the expense of increasing the number of LPB calculations [43–45]. The cost is still not prohibitive, however, especially when efficiency is increased by the use of 'focusing' methods, which use very fine finite-difference grids for calculating short-range interactions, but use coarser grids for long-range interactions for which accuracy is less critical [3,44]. For a protein of a few hundred residues, all the calculations can normally be completed in a matter of hours on a Silicon Graphics workstation with an R4400 CPU.

3. *Computing properties with the energy matrix*

These methods provide the entries in the matrix of electrostatic interactions, $\|G_{ij}\|$. As discussed above, the matrix can then be used to compute the relative chemical potentials of the various ionization states of the system, and thus the dependence of properties upon pH (see Eq. 1). However, this step can be challenging if many states are possible, as is frequently the case. A simple approach to this multiple-site titration problem, first used by Tanford and Kirkwood [28], involves assuming that each ionizable group 'feels' the average potential of each other group. For example, near pH 4, a histidine may feel the influence of several partially ionized carboxylic acids. This mean-field approximation [46] allows for an efficient iterative method that yields the fractional charge of each group [28]. However, it becomes highly inaccurate in cases where the protonation states of two groups correlate with each other [1,13,44].

For example, two aspartic acids that are close together will tend not to be ionized at the same moment. At some pH, they will both be 50% ionized, but their repulsion will be smaller than that predicted by the mean-field approximation, for when one is ionized, the other probably will not be.

This problem has been addressed in several ways. First, for systems of moderate size, it is possible to treat all ionization states explicitly. It may also be possible to show that certain groups are essentially fixed in an ionized or neutral form in the entire pH range of interest [13]. Computationally fixing the state of such groups reduces the number of different ionization states by a factor of 2 for each fixed group. Another approach uses the mean-field approximation to treat interactions among groups that interact weakly, but enumerates all the ionization states of each cluster of strongly interacting groups [1,44,47]. Such methods are convenient for globular proteins with up to several hundred ionizable groups [1], and typically run in fractions of a second to a few minutes. They are inadequate for systems in which the ionizations of many groups are tightly coupled, such as the photosynthetic reaction center of *Rhodopseudomonas sphaeroides*, for then the clusters become intractably large. Such systems may be dealt with by Monte Carlo methods [44,48,49]. For typical globular proteins, these are slower than the cluster methods mentioned above, but they are the method of choice for large systems of strongly coupled groups.

Accuracy of the PB method for computing pK$_a$'s

The number of protein pK$_a$'s that have been measured experimentally is substantial, and it is increasing rapidly. Therefore, fairly extensive assessments of the reliability of models for predicting pK$_a$'s can be carried out. The accuracy of the PB model depends on the details of its implementation and upon parameters. For example, the results are expected to depend upon the treatment of tightly bound solvent molecules and of hydrogen atoms, whose coordinates are not determined crystallographically. Important parameters include the dielectric constant of the protein, and the atomic charges and radii; the latter determine the position of the boundary between the low-dielectric interior and the high-dielectric solvent. It is therefore not surprising that the apparent accuracy of the model varies from one study to another.

In 1994, it was observed that few, if any, previous calculations were more accurate than the Null model, the assumption that proteins do not perturb pK$_a$'s from their model-compound values (see above). However, calculations based upon a high protein dielectric constant did consistently beat the Null model [3], as illustrated in Fig. 5, a histogram of errors in pK$_a$'s computed with a protein dielectric constant of 20. These calculations also beat, by a smaller margin, the more sophisticated Null model, in which the pK$_a$ of every group of a certain type is set to the mean value observed for it in proteins [10]. These results are somewhat surprising, because theoretical considerations suggest that the dielectric constant of the protein interior is usually much lower, perhaps 2–4 [50–52]. Some computational studies have yielded higher values for the dielectric constant of a protein [53–55], but this is because they

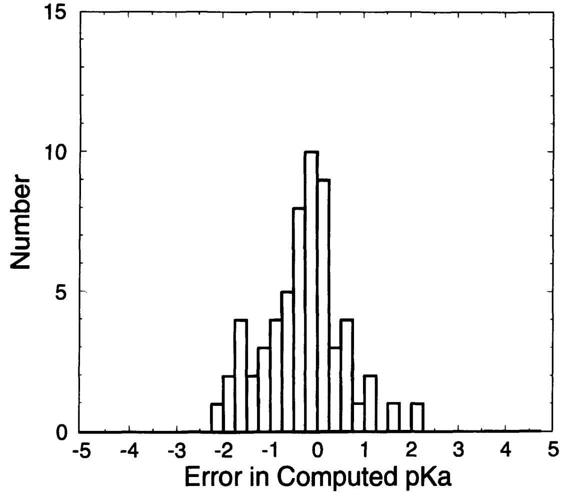


Fig. 5. Histogram of errors in pK$_a$'s computed with a protein dielectric constant of 20, for the same groups as in Fig. 1.

consider the motions of either solvent molecules or ionized groups to contribute to the dipole moment fluctuations of the protein, and thus to its dielectric constant. Including solvent molecules would appear to be inappropriate for the present application, because the influence of solvent is already explicitly included in the LPB model. It is less clear whether one should include the fluctuations of the dipole moment of the protein due to motion of, for example, a lysine side chain. However, sample calculations suggest that the large fluctuations in dipole moment associated with motions of the ionized group are not directly relevant to the screening of charge–charge interactions within the protein [3,56].

Why the use of a seemingly unrealistically high dielectric constant for the protein should yield accurate pK$_a$'s remains uncertain. Detailed case studies should be enlightening. For now, it is worth discussing some possible explanations:

1. A high dielectric constant may compensate for inadequacies in the atomic charges and radii used in the calculations. It has, in fact, been shown that modeling ionization as the addition of a unit charge to a single atom of a neutral group leads to a model of the ionized group that is unrealistically difficult to desolvate [10]. When used with a protein dielectric constant of 4, this ionization model therefore creates a systematic bias in favor of the neutral forms of both acids and bases. Switching to a more realistic model of the ionized form [57] eliminates this systematic error, and improves the accuracy of pK$_a$ calculations carried out with a protein dielectric constant of 4; however, the accuracy is still not as good as that with a high dielectric constant, or even as the Null model [10].

2. The high dielectric constant might in some way compensate for the use of single crystal conformations of proteins in the calculations instead of multiple

solution conformations. In fact, pK$_a$'s computed with a low protein dielectric constant and averaged over sets of conformations generated by nuclear magnetic resonance (NMR) studies tend to be more accurate than those based upon single crystal conformations; however, the results are still less accurate than the Null model [10].

3. The conformation of a salt-bridge is coupled to the protonation states of the acid and base that form it. The conformation observed in a crystal structure solved near neutral pH does not reflect the tendency of the bridge to fall apart when the pH is near the pK$_a$ of either group. Therefore, pK$_a$'s computed with the crystal structure may overestimate the actual influence of each group upon the pK$_a$ of the other. In fact, it might be argued that the PB model is more accurate than it should be, given that it uses protein conformations appropriate to one pH to compute protonation transitions at quite different pH's. A recent study shows that accounting for the conformational flexibility of ionizable side chains in the PB model improves the accuracy of computed pK$_a$'s when a low dielectric constant is assumed for the protein [58]. Although this form of conformational flexibility is not a dielectric relaxation, an artificially high protein dielectric constant could compensate for its neglect.

4. The treatment of tightly bound solvent in the PB model may be inadequate. Even solvent molecules detected by crystallography are often neglected in computations of pK$_a$'s [3]. This avoids the need to propose and test rules for which solvent molecules should be included, and how they should be oriented. On the other hand, leaving out these molecules may lead to inaccuracy [44]. It is conceivable that these inaccuracies are to some degree compensated by the use of a high dielectric constant for the protein.

5. The use of a high dielectric constant for the protein may suppress 'noise' in the calculated pK$_a$'s without fully suppressing the 'signal'. One source of noise is uncertainty in the atomic coordinates. Small changes in the positions of atoms may have significant effects upon computed pK$_a$'s when a low protein dielectric constant is assumed. These changes are diminished when the protein dielectric constant is set to a high value. Model calculations show that, even if the true dielectric constant is 4, use of a dielectric constant of 20 can suppress the errors associated with uncertainty in the distance between the groups, without introducing large new errors [59]. This is because the energy of ionization in the protein usually includes a positive desolvation (Born) term, and compensatory, stabilizing background and interaction terms [44]: raising the protein dielectric constant reduces the destabilizing Born term and the stabilizing interaction terms in parallel.

Applications of the PB method of computing pK$_a$'s

Although the basis for the success of the PB model with a protein dielectric constant of 20 is currently uncertain, the method is applicable to real systems because it is predictive. This section highlights two previously described applications.

Acetylcholinesterase and chymotrypsin

The enzyme acetylcholinesterase (AChE) rapidly hydrolyzes the neurotransmitter acetylcholine in species from insects to humans. AChE possesses an active-site triad remarkably similar to that of chymotrypsin, and the chemical mechanism by which it hydrolyzes esters is thought to be similar to that by which chymotrypsin hydrolyzes amides [60–64]. The active-site triads in both enzymes consist of a catalytic serine, a histidine presumed to act as a general base, and a carboxylic acid hydrogen-bonded to the histidine. Figure 6 demonstrates the structural similarity of the active-site triads of chymotrypsin (4CHA) [65,66] and *Torpedo californica* AChE (TcAChE; 1ACE) [63]. Kinetic studies of AChE indicate that a group with a pK_a of ~6.3 must be neutral at the start of catalysis; similarly, in chymotrypsin a group with a pK_a of about 7 must be neutral. These groups are believed to be the catalytic histidines [7,8,67–70].

However, straightforward application of the PB model to both systems yields quite different results; the computed pK_a of the catalytic histidine of chymotrypsin is about 7 [3], in excellent agreement with experiment, but the computed pK_a for the catalytic histidine of AChE is 9.3, 3 pK_a units too high [2]. It seems odd that the PB model should fail in the case of AChE, given that it works so well for the similar triad of chymotrypsin. Furthermore, the magnitude of the error, 3 pK_a units, is much larger than the errors usually associated with the method (see Fig. 5). It is therefore worth seeking an explanation for the discrepancy.

Further study shows that the computed pK_a of the histidine in AChE is driven up by two nearby glutamic residues (see Fig. 6): when the closest glutamic acid is forced to remain neutral in the calculations, the pK_a of the catalytic histidine drops to 7.3, and when both glutamic acids are forced to remain neutral, the computed pK_a of the catalytic histidine becomes 6.1. No groups comparable to these glutamic acids are

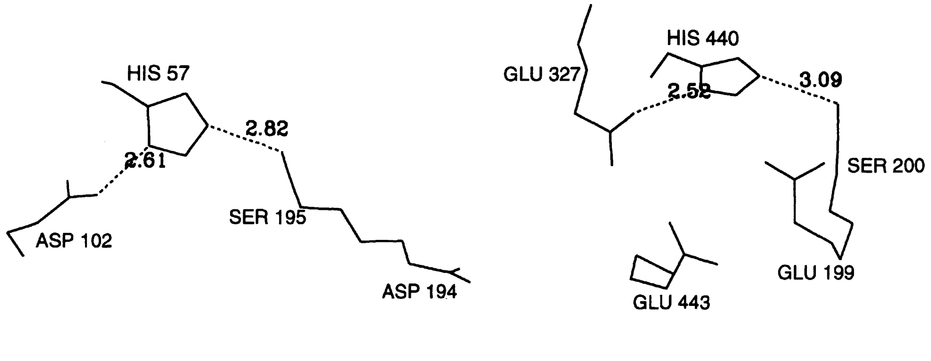


Fig. 6. Comparison of the active-site triads of chymotrypsin (A) and TcAChE (B) (distances are in Å).

found in chymotrypsin. The calculations thus yield the intuitively reasonable result that two nearby negative charges make a histidine much more basic. However, it is still necessary to reconcile this view with the experimental result that the pK_a of the histidine of AChE is essentially normal. In fact, a reasonable and even enlightening explanation can be found.

The physiological substrate of AChE is cationic, as are most of its competitive inhibitors. Crystallographic studies show that the cationic moieties of inhibitors bind essentially at the surface of the glutamic acid closest to the catalytic histidine, Glu^{199} [71]. When the pK_a calculations are repeated with the cationic ligand tacrine at its experimentally determined location, the pK_a of the catalytic histidine falls to about 6.5 [2]. Similarly, the computed pK_a is 7.4 when a sodium ion is modeled at the cation-binding site [2]. These results suggest that the original pK_a calculations for AChE are in error because they neglect the large perturbing influence of a monovalent cation bound at the surface of Glu^{199}, near the catalytic histidine.

This conclusion proves to be consistent with kinetic studies of the enzyme. Because the histidine must be neutral for catalysis to occur, and because a bound cation seems necessary to neutralize the histidine at neutral pH, the efficiency of catalysis should, under certain circumstances, depend upon the concentration of cations in solution. This is seen experimentally. Under conditions where the rate-limiting step is deacylation – which occurs after cleavage of the cationic moiety of the substrate – catalysis is in fact accelerated by the cations [69,72–74].

The calculations also offer an explanation for the puzzling observation that the pH-dependence of the catalysis is unchanged when Glu^{199} is replaced by a neutral residue [62]. The original pK_a calculations suggest that eliminating this negative group would make the histidine much less basic. However, if eliminating this negative charge *also* essentially eliminates the binding of a cation, the net charge in the vicinity of the histidine will be unchanged by the mutation. Therefore, the pK_a of the histidine should not change much, as observed experimentally.

Thus, the initially discrepant pK_a calculations lead to reasonable explanations for possibly puzzling experimental results.

The pH-dependence of protein stability

The linkage of protonation with other equilibria means that pH alters the apparent equilibrium constants of many reactions, such as those associated with the noncovalent association of biomolecules and the denaturation of proteins [1,11,12,75]. This section describes efforts to predict, from structural information, the pH-dependence of the stability of folded proteins.

The pH-dependence of the free energy of denaturation can be isolated in a single term that depends logarithmically upon the ratio of the proton-binding polynomial for the native and denatured states of the protein. This is shown in Eqs. 2 and 3, and conceptually in Fig. 3. As outlined above, the PB model can be used to compute the equilibrium constants in the binding polynomial for proteins of known structure.

However, there is no computational procedure for predicting the protonation energetics of a denatured protein. In the absence of titration data for the denatured state, it is usual to assume that the pK$_a$'s of unfolded proteins equal model compound pK$_a$'s [3,75].

With this assumption, the PB model reproduces fairly well the overall shapes of curves of stability versus pH for ribonuclease A, hen egg-white lysozyme, barnase, and T4 lysozyme, as shown in Figs. 7–11, which have been presented previously [3]. However, the results are not always satisfying quantitatively. For example, the computed increase in the stability of barnase from pH 3 to 6 is markedly overestimated by the calculations [3]. Such errors result in part from inaccuracies in the pK$_a$'s calculated for the native protein [3]. However, there are also other sources of error that have to do with the character of the denatured protein. As recently emphasized, different denaturants, such as heat, acid, and concentrated urea, create different denatured states [76]. Such effects can lead to discrepancies between the directly measured variation in protein stability with pH, and the variation in stability that is predicted from titration curves in native and denatured states [3]. Recent studies provide more information on the persistence of shifted pK$_a$'s in denatured proteins and in protein fragments [76–78].

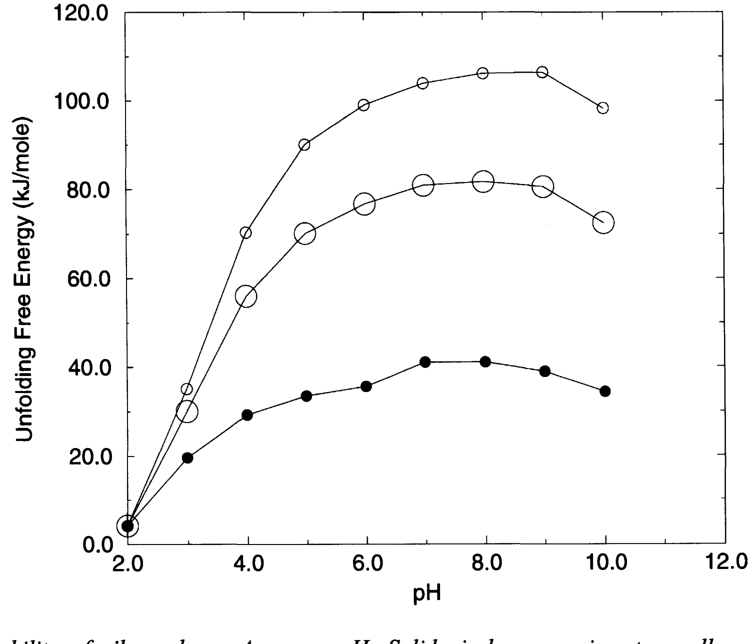


Fig. 7. Stability of ribonuclease A versus pH. Solid circles: experiment; small open circles: computed, assuming model-compound pK$_a$'s for the unfolded state; large open circles: computed, with an actual titration curve for the unfolded state.

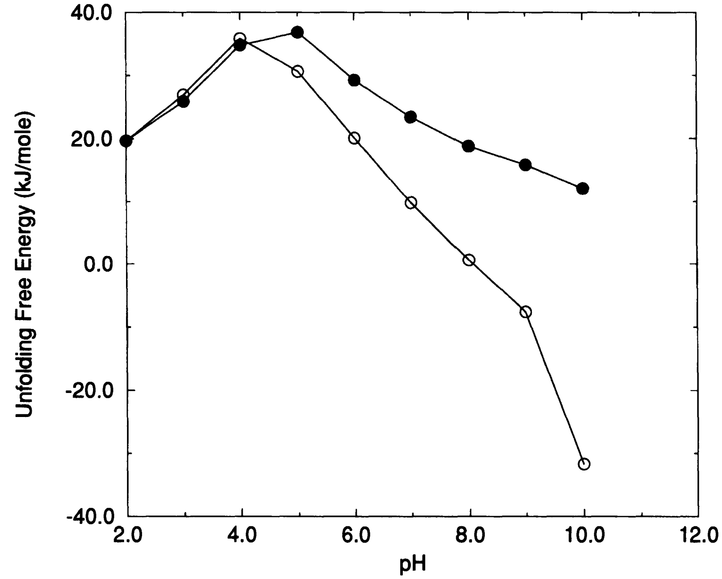


Fig. 8. Stability versus pH for ribonuclease T1. Solid circles: experiment; small open circles: computed, assuming model-compound pK_a's for the unfolded state.

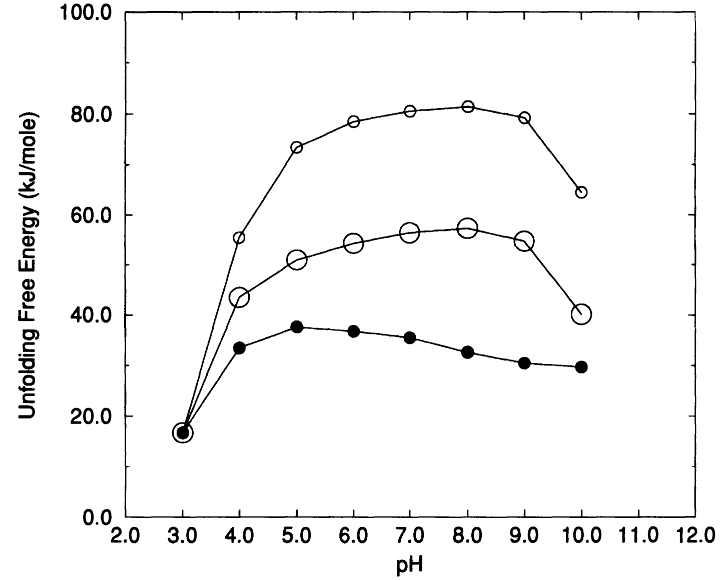


Fig. 9. Stability of barnase versus pH. Solid circles: experiment; small open circles: computed, assuming model-compound pK_a's for the unfolded state; large open circles: computed, corrected for apparent pK_a shifts of carboxylic acids in the denatured state.

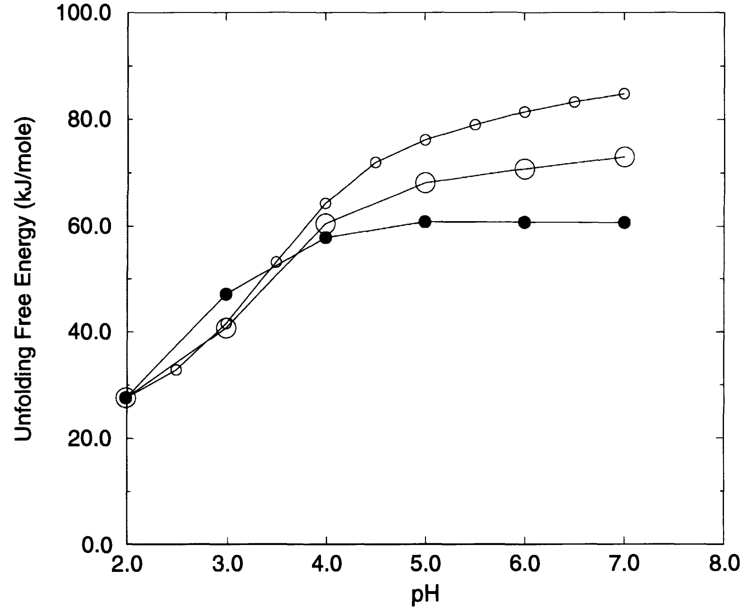


Fig. 10. Stability of hen egg-white lysozyme versus pH. Solid circles: experiment; small open circles: computed with triclinic crystal structure [85]; large open circles: computed with tetragonal crystal structure [86].

One such study finds a discrepancy between the measured stability–pH curve of barnase, and the stability–pH curve predicted from pK_a values measured in the native state and the assumption of unshifted pK_a's in the heat-denatured state. The measured stability rises more gradually between pH 3 and 6 than predicted from the titration curves of the native protein. This discrepancy can be accounted for by assuming that the pK_a's of all aspartic acid and glutamic acid residues are shifted downward by 0.4 pK_a units in the heat-denatured protein. Interestingly, including this correction for the denatured state in the theoretical stability–pH calculations brings the theoretical results closer to the measured stability–pH curves, as shown in Fig. 9. Similarly, using a measured titration curve for the unfolded state of ribonuclease A [79], in place of the assumption of unshifted pK_a's, leads to improved agreement between computed and measured stability curves (Fig. 7). The structural basis for persistent pK_a shifts in denatured states is not yet known, but the phenomenon is consistent with the notion that denatured proteins can retain a degree of structure.

In conclusion, even highly accurate calculations for native proteins may not yield excellent quantitative agreement with the measured stability–pH curves, because of persistent pK_a shifts in some denatured states. However, it should be possible to do well for other protonation-linked equilibria. For example, it should be possible to use the methods outlined here to make fairly accurate predictions of the pH-dependence of binding constants for noncovalent associations between biomolecules.

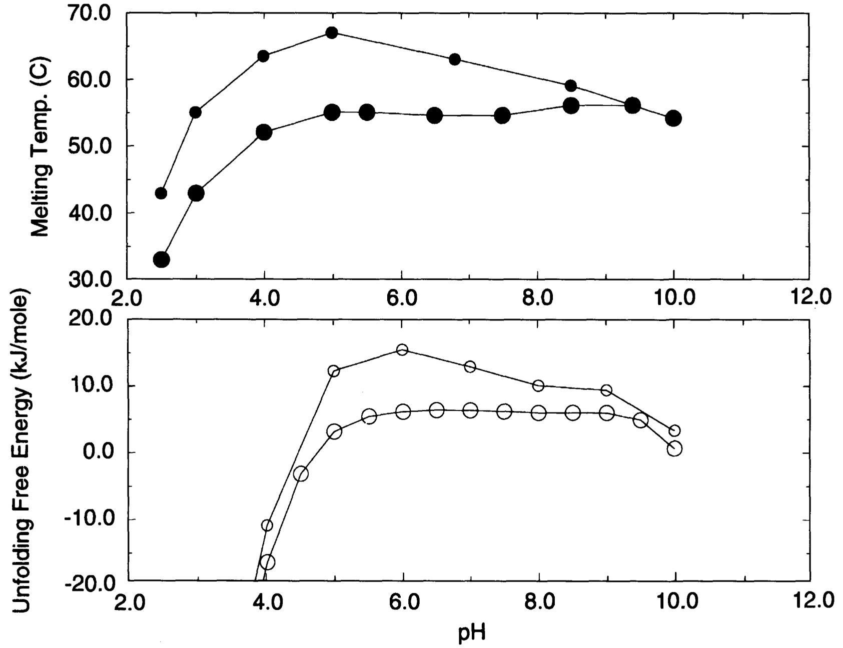


Fig. 11. Top: measured melting temperature of T4 lysozyme versus pH, computed for wild-type (small solid circles) and for mutant lacking salt-bridge between His[31] and Asp[70] (large solid circles). Bottom: stability of T4 lysozyme versus pH, computed for wild-type (small open circles) and for mutant lacking salt-bridge between His[31] and Asp[70] (large open circles).

Directions for future investigation

As outlined here, the PB model for predicting pK_a's is fairly accurate. However, further improvements in accuracy would be valuable in understanding the structure and function of biomolecules and in molecular design. Efforts to improve the accuracy will be guided by the weaknesses of existing implementations of the PB model. Explorations of the sensitivity of the results to parameters will be helpful in this effort [80]. Areas amenable to improvement include the positioning of hydrogen atoms before electrostatic calculations are performed [81,82]; the treatment of tautomer equilibria [10,83]; the treatment of the conformational flexibility of ionizable groups [10,58,75]; allowance for possible dielectric inhomogeneity of the protein; and the treatment of tightly bound solvent molecules [44]. The exploration of such issues may lead to an explanation of the unexpected observation that pK_a's computed with a high protein dielectric constant are more accurate than those computed with a seemingly more realistic low dielectric constant (see above).

It is important that enhancements be uniformly applied and broadly tested. For example, a method for treating bound solvent should clearly specify which crystallographically detected solvent molecules are to be treated explicitly. It should also be tested on a number of different proteins.

Note that the apparent accuracy of models for computing pK_a's will depend upon the data used as a test set. The characterization of models depends upon accurately measured pK_a's for groups in interesting environments. Unfortunately, most proteins possess only a small number of such groups. Furthermore, the most interesting groups are often found in relatively large proteins that are difficult to study by NMR – the usual method for making such measurements. Perhaps interesting and challenging test cases could be engineered into small, stable, well-characterized proteins. The resulting data would be very useful to theoreticians seeking to test computational methods.

Another important goal is to increase the speed of the electrostatic calculations that are used in computing the pH-dependent properties. That the PB model with a seemingly unrealistic protein dielectric constant yields rather accurate results suggests that simpler, faster, implementations might also yield accurate results. This notion seems to be borne out by recent results [84], though it is worth noting that caution is in order when one attempts to interpret data based upon highly simplified or parametrized models. Dramatic increases in speed would permit the calculation of ionization states 'on the fly' in molecular dynamics or other conformational sampling algorithms. For proteins of up to a few hundred residues, very rapid methods are already available for solving the multiple-site titration method for forces and energies. The use of such approaches might actually make it possible to improve the accuracy of pK_a's through the efficient incorporation of conformational flexibility. A full pK_a calculation could, in principle, be carried out for a large number of conformational microstates. The results could then be combined to yield predictions of experimentally observable pK_a's.

Acknowledgements

I am grateful to Drs. A.D. Robertson and E.L. Mehler for providing unpublished results, and to Drs. A. Roitberg and H.S.R. Gilson for helpful comments on the paper.

Certain commercial equipment or materials are identified in this paper in order to specify the methods adequately. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Note added in proof

Gibas and Subramaniam [Biophys. J., 71(1996)138] present a detailed study of the influence of an explicit treatment of water on computed pK_a's.

References

1. Gilson, M.K., Proteins Struct. Funct. Genet., 15(1993)266.
2. Wlodek, S.T., Antosiewicz, J., McCammon, J.A. and Gilson, M.K., In Pullman, A., Jortner, J. and Pullman, B. (Eds.) Modeling of Biomolecular Structures and Mechanisms, Kluwer, Dordrecht, 1995, pp. 25–37.
3. Antosiewicz, J., McCammon, J.A. and Gilson, M.K., J. Mol. Biol., 238(1994)415.
4. Stites, W.E., Gittis, A.G., Lattman, E.E. and Shortle, D., J. Mol. Biol., 221(1991)7.
5. Dao-Pin, S., Anderson, D.E., Baase, W.A., Dahlquist, F.W. and Matthews, B.W., Biochemistry, 30(1991)11521.
6. Varadarajan, R., Lambright, D.G. and Boxer, S.G., Biochemistry, 28(1989)3771.
7. Bender, M.L., Clement, G.E., Kezdy, F.J. and Heck, H.d'A., J. Am. Chem. Soc., 86(1964)3680.
8. Fersht, A.R. and Sperling, J., J. Mol. Biol., 74(1973)137.
9. Martell, A.E. and Smith, R.M., Critical Stability Constants, Vols. 1–IV, Plenum Press, New York, NY, 1974.
10. Antosiewicz, J., McCammon, J.A. and Gilson, M.K., Biochemistry, 35(1996)7819.
11. Wyman, J., J. Mol. Biol., 11(1965)631.
12. Schellman, J.A., Biopolymers, 14(1975)999.
13. Bashford, D. and Karplus, M., Biochemistry, 9(1990)327.
14. Mark, A.E. and van Gunsteren, W.F., J. Mol. Biol., 240(1994)167.
15. Boresch, S., Archontis, G. and Karplus, M., Proteins Struct. Funct. Genet., 20(1994)25.
16. Brady, G.P. and Sharp, K.A., J. Mol. Biol., 254(1995)77.
17. Kenneth, J. and Merz, M., J. Am. Chem. Soc., 113(1992)3572.
18. MacKerell, J.A.D., Sommer, M.S. and Karplus, M., J. Mol. Biol., 247(1995)774.
19. Lee, F.S. and Warshel, A., J. Chem. Phys., 97(1992)3100.
20. Levy, R.M., Belhadj, M. and Kitchen, D.B., J. Chem. Phys., 95(1991)3627.
21. Buono, G.S.D., Figueirido, F.E. and Levy, R.M., Proteins Struct. Funct. Genet., 20(1994)85.
22. McQuarrie, D.A., Statistical Mechanics, Harper and Row, New York, NY, 1973.
23. Urry, D.W., Peng, S.Q. and Parker, T.M., Biopolymers, 32(1992)373.
24. Urry, D.W., Peng, S.Q., Hayes, L., Jaggard, J. and Harris, R.D., Biopolymers, 30(1990)215.
25. Gilson, M.K., Curr. Opin. Struct. Biol., 5(1995)216.
26. Linderstrom-Lang, K., Compt. Rend. Trav. Lab. Carlsberg, 15(1924)7
27. Born, M., Z. Physik., 1(1920)45.
28. Tanford, C. and Kirkwood, J.G., J. Am. Chem. Soc., 79(1957)5333.
29. Tanford, C., J. Am. Chem. Soc., 79(1957)5340.
30. Shire, S.J., Hanania, G.I.H. and Gurd, F.R.N., Biochemistry, 13(1974)2967.
31. Imoto, T., Biophys. J., 44(1983)293.
32. Matthew, J.B., Gurd, F.R.N., Garie-Moreno, B.E., Flanagan, M.A., March, K.L. and Shire, S.J., Crit. Rev. Biochem., 18(1985)91.
33. Paul, C.H., J. Mol. Biol., 155(1982)53.
34. Gilson, M.K., Rashin, A.A., Fine, R. and Honig, B., J. Mol. Biol., 183(1985)503.
35. Warwicker, J. and Watson, H.C., J. Mol. Biol., 157(1982)671.
36. Klapper, I., Hagstrom, R., Fine, R., Sharp, K. and Honig, B., Proteins Struct. Funct. Genet., 1(1986)47.
37. Gilson, M.K. and Honig, B.H., Nature, 330(1987)84.
38. Sternberg, M.J., Hayes, F.R., Russell, A.J., Thomas, P.G. and Fersht, A.R., Nature, 330(1987)86.

39. Gilson, M.K. and Honig, B., Proteins Struct. Funct. Genet., 3(1988)32.
40. Gilson, M.K. and Honig, B., Proteins Struct. Funct. Genet., 4(1988)7.
41. Luty, B.A., Davis, M.E. and McCammon, J.A., J. Comput. Chem., 13(1992)768.
42. Richards, F.M., Annu. Rev. Biophys. Bioeng., 6(1977)151.
43. Bashford, D. and Gerwert, K., J. Mol. Biol., 224(1992)473.
44. Yang, A.-S., Gunner, M.R., Sampogna, R., Sharp, K. and Honig, B., Proteins Struct. Funct. Genet., 15(1993)252.
45. Antosiewicz, J., Briggs, J.M., Elcock, A.H., Gilson, M.K. and McCammon, J.A., J. Comput. Chem., 17(1996)1633.
46. Bashford, D. and Karplus, M., J. Phys. Chem., 95(1991)9556.
47. Karshikoff, A., Protein Eng., 8(1995)243.
48. Antosiewicz, J. and Porschke, D., Biochemistry, 28(1989)10072.
49. Beroza, P., Fredkin, D.R., Okamura, M.Y. and Feher, G., Proc. Natl. Acad. Sci. USA, 88(1991)5804.
50. Gilson, M.K. and Honig, B.H., Biopolymers, 25(1986)2097.
51. Simonson, T., Perahia, D. and Brunger, A.T., Biophys. J., 59(1991)670.
52. Simonson, T. and Perahia, D., Proc. Natl. Acad. Sci. USA, 92(1995)1082.
53. Nakamura, H., Sakamoto, T. and Wada, A., Protein Eng., 2(1988)177.
54. King, G., Lee, F.S. and Warshel, A., J. Chem. Phys., 95(1991)4366.
55. Smith, P.E., Brunne, R.M., Mark, A.E. and van Gunsteren, W.F., J. Phys. Chem., 97(1993)2009.
56. Lee, I. and Gilson, M.K., unpublished results.
57. Sitkoff, D., Sharp, K.A. and Honig, B., J. Phys. Chem., 98(1994)1978.
58. You, T.J. and Bashford, D., Biophys. J., 69(1995)1721.
59. Gilson, M.K., unpublished results.
60. Doctor, B.P., et al., In Rein, R. and Golombek, A. (Eds.) Computer-Assisted Modeling of Receptor–Ligand Interactions. Theoretical Aspects and Applications to Drug Design, Vol. 289, Liss, New York, NY, 1989, pp. 305–316.
61. Gentry, M.K. and Doctor, B.P., In Massoulie, J., Bacou, F., Barnard, E., Chatonnet, A., Doctor, B.P. and Quinn, D.M. (Eds.) Cholinesterases: Structures, Function, Mechanism, Genetics and Cell Biology, American Chemical Society, Washington, DC, 1991, pp. 394–398.
62. Gibney, G., Camp, S., Dionne, M., MacPhee-Quigley, K. and Taylor, P., Proc. Natl. Acad. Sci. USA, 87(1990)7546.
63. Sussman, J.L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L. and Silman, I., Science, 253(1991)872.
64. Radić, Z., Gibney, G., Kawamoto, S., MacPhee-Quigley, K., Bongiorno, C. and Taylor, P., Biochemistry, 31(1992)9760.
65. Tsukuda, H. and Blow, D.M., J. Mol. Biol., 184(1985)703.
66. Bernstein, F.C., Koetzle, T.F., Williams, T.F., Meyer Jr., G.J.B., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112(1977)535.
67. Rosenberry, T.L., Adv. Enzymol. Relat. Areas Mol. Biol., 43(1975)103.
68. Krupka, R.M., Biochemistry, 5(1966)1983.
69. Krupka, R.M., Biochemistry, 5(1966)1988.
70. Kossiakoff, A.A. and Spencer, S.A., Biochemistry, 20(1981)6462.
71. Harel, M., Schalk, I., Ehret-Sabatier, L., Bouet, F., Goeldner, M., Hirth, C., Axelsen, P., Silman, I. and Sussman, J.L., Proc. Natl. Acad. Sci. USA, 90(1993)9031.

72. Changeux, J.-P., Mol. Pharmacol., 2(1966)369.
73. Berman, H.A., Leonard, K. and Nowak, M.W., In Massoulie, J., Bacou, F., Barnard, E., Chatonnet, A., Doctor, B.P. and Quinn, D.M. (Eds.) Cholinesterases: Structures, Function, Mechanism, Genetics and Cell Biology, American Chemical Society, Washington, DC, 1991, pp. 229–234.
74. Kitz, R.J., Braswell, L.M. and Ginsberg, S., Mol. Pharmacol., 6(1969)108.
75. Yang, A.-S. and Honig, B., J. Mol. Biol., 231(1993)459.
76. Oliveberg, M., Arcus, V.L. and Fersht, A.R., Biochemistry, 34(1995)9424.
77. Tan, Y.-J., Oliveberg, M., Davis, B. and Fersht, A.R., J. Mol. Biol., 254(1995)980.
78. Robertson, A.D., unpublished results.
79. Nozaki, Y. and Tanford, C., J. Am. Chem. Soc., 89(1967)742.
80. Beroza, P. and Fredkin, D., J. Comput. Chem., 17(1996)1229.
81. Brunger, A.T. and Karplus, M., Proteins Struct. Funct. Genet., 4(1988)148.
82. Bass, M.B., Hopkins, D.F., Jaquysh, W.A.N. and Ornstein, R.L., Proteins Struct. Funct. Genet., 12(1992)266.
83. Bashford, D., Case, D.A., Dalvit, C., Tennant, L. and Wright, P.E., Biochemistry, 32(1993)8045.
84. Mehler, E.L., J. Phys. Chem., 100(1996)16006.
85. Ramanadham, M., Sieker, L.C. and Jensen, L.H., Acta Crystallogr., Sect. A, 37C(1981)33.
86. Diamond, R., J. Mol. Biol., 82(1974)371.
87. Bartik, K., Redfield, C. and Dobson, C.M., Biophys. J., 66(1994)1180.
88. Rico, M., Santoro, J., Gonzalez, C., Bruix, M. and Neira, J.L., In Cuchillo, C.M., de Llorens, R., Nogués, M.V. and Parés, X. (Eds.) Structure, Mechanism and Function of Ribonucleases, Proceedings of the 2nd International Meeting held in Sant Feliu de Guíxols, Girona, Spain, 1990, Bellaterra, Spain, 1991, pp. 9–14.
89. Antosiewicz, J., McCammon, J.A. and Gilson, M.K., Biochemistry, 35(1996)7819.
90. Brown, L.R., Marco, A.D., Wagner, G. and Wüthrich, K., Eur. J. Biochem., 62(1976)103.
91. Brown, L.R., Marco, A.D., Richarz, R., Wagner, G. and Wüthrich, K., Eur. J. Biochem., 88(1978)87.
92. Richarz, R. and Wüthrich, K., Biochemistry, 17(1978)2263.
93. Schaller, W. and Robertson, A.D., Biochemistry, 34(1995)4714.
94. Swint-Kruse, L. and Robertson, A.D., Biochemistry, 34(1995)4724.
95. Anderson, D.E., Becktel, W.J. and Dahlquist, F.W., Biochemistry, 29(1990)2403.
96. Inagaki, F., Kawano, Y., Shimada, I., Takahashi, K. and Miyazawa, T., J. Biochem., 89(1981)1185.
97. Shirley, B.A., Stanssen, P., Steyaert, J. and Pace, C.N., J. Biol. Chem., 264(1989)11621.

Semi-explicit bag model for protein solvation

Richard C. Brower[a,b] **and S. Roy Kimura**[a,c]
*[a]Center for Computational Science, [b]Electrical and Computer Engineering Department,
and [c]Biomedical Engineering Department, Boston University,
Boston, MA 02215, U.S.A.*

1. Introduction

A long-standing problem in the molecular simulation of proteins is implementing an efficient but still accurate model for the water surrounding macromolecules. It is well known that simulations (or free energy calculations) that neglect explicit water are not adequate, while the full inclusion of a large volume of water implies large, if not prohibitive, increases in computational resources [1–5].

Here we review the problem from a fundamental perspective and suggest methods to replace the 'infinite' volume of water with a small surrounding cavity (or 'bag') of water with appropriate boundary conditions [6–10]. Our goal is to find methods that are well suited to full dynamical studies of proteins in which the bounding surface should be able to move and, if necessary, with water entering and leaving the fiducial volume.

Due to limited space, we chose to outline the problem in a pedagogical style, giving citations to the vast literature at the end (see Sec. 6). We hope that this exercise in 'brain storming' will lead to some new methods or variations beyond standard practice. In our research we are experimenting with some of the suggestions made in the conclusion.

1.1. Short-range versus long-range forces

For the sake of argument, let us consider macromolecular dynamics from the strict 'molecular mechanics' viewpoint. We assume that the correct physics is reproduced by simulating the protein in an infinite (i.e., very large) box of water. The forces are found in principle from *ab initio* calculations applying the Hellman–Feynman theorem to the Born–Oppenheimer approximation of the multielectron Schrödinger equation. After the 'partial integration' of the electron coordinates for fixed location of the atoms, one is left with two classes of forces (or potential energies): short-range potentials, $V_{\text{short}}(\mathbf{x}_i)$, expressing the quantum chemistry, and long-range electrostatics, $V_{\text{long}}(\mathbf{x}_i)$, resulting from the nuclear charges and the electron charge distributions. Typical contributions to these two terms are:

short range – anharmonic bonds $(r^2 - r_0^2)^2$, repulsive core $1/r^{12}$, dispersion $1/r^6$;
long range – Coulombic $q_i q_j / r$, electronic polarization $\mathbf{p} \cdot \hat{r}/r^2$.

The long-range forces are the result of *ab initio* calculations parametrized by discrete distributions for the charge density $\rho(\mathbf{x})$ and polarizability density tensor $\gamma_{ij}(\mathbf{x})$. Assuming linear response, $P_i(\mathbf{x}) = \gamma_{ij}(\mathbf{x})E_j(\mathbf{x})$, where E_j is the jth component of the electric field. Following the standard practice, we place the $1/r^6$ dipole–dipole interaction in the short-range term, although it might in fact be useful to lump it with the other electrostatic terms. In this case the short-range problem would reduce to bonded interactions and short-range repulsions almost equivalent to those of hard spheres.

We now pose the question of how to approximate an essentially infinite volume of water by a finite fiducial volume or cavity containing explicit water very close to the macromolecule of interest (see Fig. 1). The problem is whether such an approach can in principle work and, if so, how to treat the boundary condition on the surface of the cavity. Again we should emphasize that we want to have nearby explicit water inside, so the 'cavity' is simply a useful definition of nearby (internal) and far (external) atoms at one instant of time. One may want to allow for atoms to leave and enter this volume as they do in a real system and to adjust the bounding surface to track the motion of the macromolecule. The approximation we seek involves replacing the external atoms by a mean force or reaction field. In field theory language, the problem is to find the aqueous 'vacuum' state into which the protein and its bound water form its cavity. Such a picture is used, for example, in describing quark–gluon constituents of the

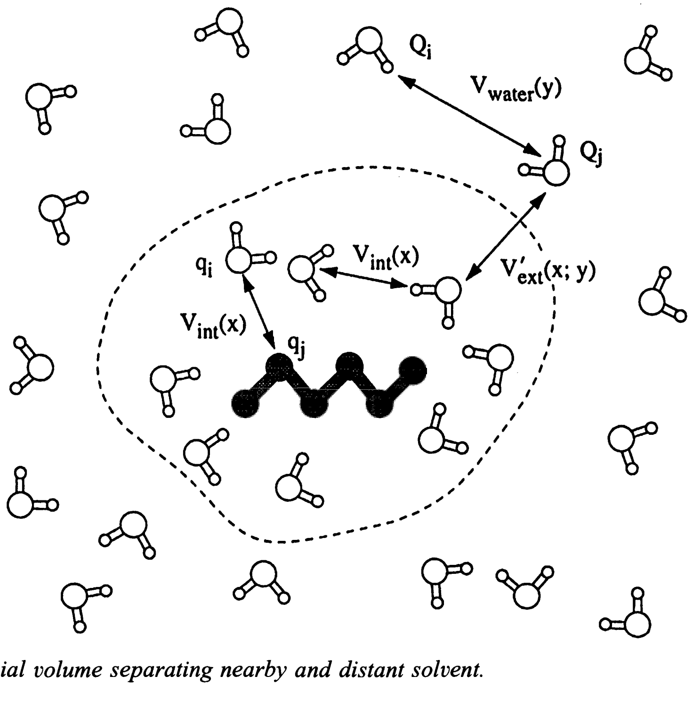


Fig. 1. Fiducial volume separating nearby and distant solvent.

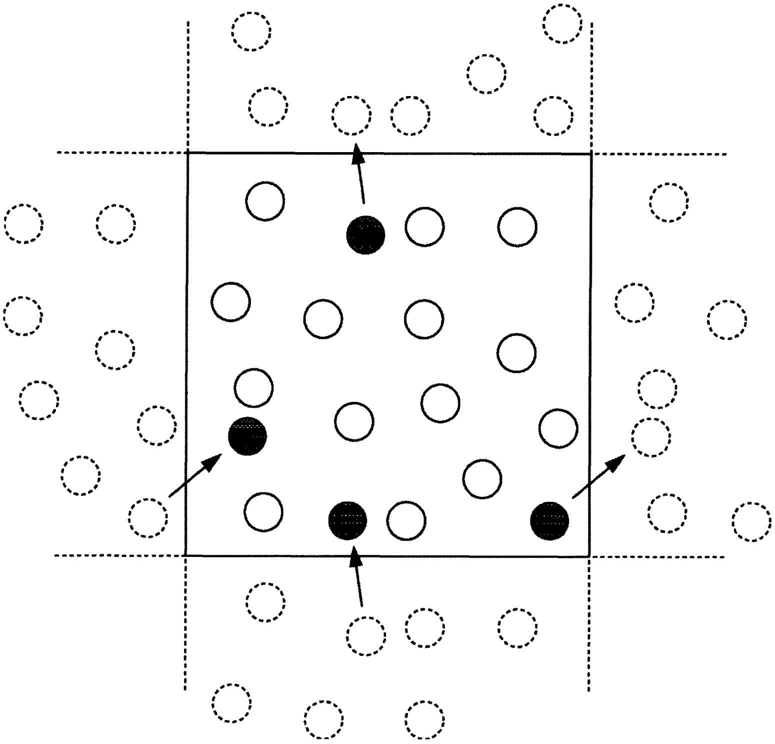


Fig. 2. Periodic box.

nucleus as a bag cut out from the surrounding confining vacuum of quantum chromodynamics (QCD). In this application, it is referred to as the MIT bag model [11].

If there were only short-range potentials (for example with nonbonded terms given by a cutoff Lennard-Jones potential), the traditional solution is to use a rectangular box and replace the surface with *periodic boundary conditions.* This allows for a 'permeable' container preserving translational invariance and momentum flux (Fig. 2). As long as the linear size of the box is long compared to the correlation length in the fluid, exponentially small finite size errors are expected.

The Coulomb potential presents a much more difficult problem. Due to the long-range 1/r potential, much bigger boxes are needed and the use of periodic boundary conditions requires great care with consideration of image charges and truncation methods (e.g., see Ref. 12) to avoid large, power-like finite volume errors. We wish to avoid these problems altogether by imagining the cavity cut out of an infinite space replacing the external forces by an effective mean potential. As long as there is a layer of explicit water near the protein, this mean description should be sufficient for a quantitatively accurate description.

2. Surface charge approach

Consider a 'snapshot' of the full thermodynamic system where all the charges are stationary. The force on an atom at x in the cavity due to all the external atoms at position y is

$$\mathbf{F}_{\mathrm{ext}}(x) = -\nabla_x V_{\mathrm{ext}}(x;\, y) \tag{1}$$

We use the compact notation (x) and (y) to refer to the entire set of coordinates internal $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ and external $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots$ to the cavity. The total potential energy (see Fig. 1)

$$V(x;\, y) = V_{\mathrm{int}}(x) + V_{\mathrm{ext}}(x;\, y) \tag{2}$$

is then split into an internal term, $V_{\mathrm{int}}(x)$, for all interactions among the atoms inside the cavity, and an external term

$$V_{\mathrm{ext}}(x;\, y) = V'_{\mathrm{ext}}(x;\, y) + V_{\mathrm{water}}(y) \tag{3}$$

which includes all the remaining interactions with the external water molecules, $V'_{\mathrm{ext}}(x;\, y)$, and among the water molecules themselves, $V_{\mathrm{water}}(y)$. Of course, since the external water–water interactions depend only on the y coordinates, it does not contribute to $\mathbf{F}_{\mathrm{ext}}(x)$. As before, we will also decompose each potential into a sum of short-range and long-range terms.

2.1. Partial integration

We now wish to convert from an instantaneous view of the external forces to a mean force. There is an exact and rigorous way to approach this within statistical mechanics. For definiteness, consider the canonical partition function (other ensembles may be used with similar results) and define G(x), the 'free energy', 'effective potential', or 'potential of mean force' – to give three of the many terms used in various fields!

$$e^{-\beta G(x)} = e^{-\beta V_{\mathrm{int}}(x)} \int d\mu(y) e^{-\beta V_{\mathrm{ext}}(x;\, y)} \tag{4}$$

The integration measure for Cartesian position coordinates is $d\mu(y) = (\mathrm{const})\, d^3y_1 d^3y_2 \cdots d^3y_N$. Of course, for water one often will want bond constraints and angular coordinates so the correct Jacobian must be included in the measure. The multiplicative constant has no effect. We may trivially separate out the effective potential due to the external water by defining $G(x) = V_{\mathrm{int}}(x) + G_{\mathrm{ext}}(x)$, and the effective external force is now given by

$$\bar{\mathbf{F}}_{\mathrm{ext}}(x) = -\nabla_x G_{\mathrm{ext}}(x) = -\langle \nabla_x V_{\mathrm{ext}}(x;\, y) \rangle \tag{5}$$

where

$$\langle(\cdots)\rangle = \frac{\int d\mu(y)e^{-\beta V_{ext}(x;y)}\ (\cdots)}{\int d\mu(y)e^{-\beta V_{ext}(x;\ y)}} \tag{6}$$

Great, we have the answer. But this is not easy to compute. The problem of doing a partial integration over the external water is as hard as the full problem. Indeed, in simulations this may dominate the entire computation.

Nonetheless, it is useful to know precisely what we are trying to model, and good models may be found for this homogeneous external vacuum state. Analytic approaches often expand the integrand in the interaction term*, $V_{ext}(x; y)$, using the link cluster expansion

$$G_{ext}(x) = \langle V_{ext}\rangle - \tfrac{1}{2}\beta(\langle V_{ext}^2\rangle - \langle V_{ext}\rangle^2) + \tfrac{1}{6}\beta^2\langle(V_{ext} - \langle V_{ext}\rangle)^3\rangle + \cdots \tag{7}$$

and try to re-sum it using various closure relations such as the hypernetted chain approximation [13], etc. We do not pursue that approach, which is a highly refined and sophisticated specialty for liquids. Our goal is a phenomenological approximation that can be tested and refined through large-scale computer simulations. For example, the first approximation might be a dielectric continuum with one adjustable parameter, the dielectric constant of water, which must be determined from experiment or computer simulations of the large system. Notice that the replacement

$$\bar{F}_{ext}(x) = -\langle\nabla_x V_{ext}(x;\ y)\rangle \rightarrow -\nabla_x\langle V_{ext}(x;\ y)\rangle \tag{8}$$

is not valid. The expression on the right-hand side is the thermodynamic internal energy (not to be confused with the internal/external decomposition of the potential). It neglects the counting of microstates or entropy in the external system, $G_{ext}(x) = \langle V(x;\ y)\rangle - T\cdot S_{ext}(x)$. Integration is, after all, just counting states. Remarkably, there are two cases where this entropy is easily accounted for: (i) pure Gaussian potentials where the entropy has *no* x dependence and can be dropped as far as forces are concerned; and (ii) linear response approximation where the free energy is proportional to the internal energy, $G = \frac{1}{2}\langle V\rangle$. Note that with linear response, the entropic effect is very large, as it represents a 100% correction to the internal energy. The short-range force is almost harmonic near equilibrium and falls into case (i), and the dielectric response falls into case (ii). These are useful guidelines, but when you mix them (as we must with $V_{ext} = V_{ext}^{short} + V_{ext}^{long}$) you are really in neither regime. However, the short-range forces only affect a small layer outside the fiducial volume, giving rise to a volume term (pressure) and surface free energy (tension), while the Coulomb force affects the entire external system. Consequently, the two terms contribute largely in different parts of the external volume and, if we do not mind

* At this point, you may wish to really separate out the external interaction term from the pure water term, $V_{ext}(x;\ y) = V'_{ext}(x;\ y) + V_{water}(y)$, and redefine $d\mu'(y) = d\mu(y)e^{-\beta V_{water}}(y)$ as a new measure or density of states for the media.

'living dangerously' for now, we will model each contribution one at a time. Thus, we introduce an *ad hoc* wall to reflect the internal atoms and turn to study the Coulomb force by itself.

2.2. Surface charges

To model the Coulomb force for all the external charges, it is useful to consider the following exact theorem:

Any arbitrary potential inside a fiducial volume, V, is given by Coulomb's law for the explicit charges plus a fictitious surface charge, σ, on the boundary, ∂V:

$$\phi(\mathbf{x}) = \sum_i \frac{q_i}{|\mathbf{x} - \mathbf{x}_i|} + \int_{\partial V} da(\xi) \frac{\sigma(\xi; y)}{|\mathbf{x} - \xi|} \tag{9}$$

Stated differently, the potential due to all the external charges,

$$\phi_{ext}(\mathbf{x}) = \sum_i \frac{Q_i}{|\mathbf{x} - \mathbf{y}_i|} \equiv \int_{\partial V} da(\xi) \frac{\sigma(\xi; y)}{|\mathbf{x} - \xi|} \tag{10}$$

can be replaced by a surface charge distribution σ(ξ; y) that depends on the charges and positions $(Q_i, \mathbf{y}_i)$ of the external atoms. In these expressions, ξ is a point on the surface, and the integral is over the boundary of the volume V.

This theorem is very similar to Green's theorem [14], except that the latter expresses the potential in terms of both a surface charge and a dipole layer. Both terms are not needed. Proving this is an elementary exercise. Imagine a single charge Q_0 outside the surface at $\mathbf{y}_0$ and find the potential for the surface as a conductor at zero potential,

$$\bar{\phi}(\mathbf{x}) = \frac{Q_0}{|\mathbf{x} - \mathbf{y}_0|} + \int_{\partial V} da(\xi) \frac{\sigma_0(\xi; \mathbf{y}_0)}{|\mathbf{x} - \xi|} \tag{11}$$

Since this solution must have zero potential inside, switching the sign of the surface charge to $-\sigma_0$ gives the equivalent surface charge needed in our theorem for a single external charge Q_0. By superposition we have our theorem. Surprisingly, this trivial observation is not easily found in the literature.

Combined with the assumption of linear response for the external media,

$$G_{ext}(x) = \tfrac{1}{2}\langle V_{ext}\rangle \tag{12}$$

which is a very reasonable approximation for the long-range electrostatic contribution, we obtain a very useful result,

$$G_{ext}^{coul}(x) = \frac{1}{2} \sum_i q_i \langle \phi_{ext}(x_i)\rangle = \frac{1}{2} \sum_i q_i \int_{\partial V} da(\xi) \frac{\langle \sigma(\xi; y)\rangle_y}{|\mathbf{x}_i - \xi|} \tag{13}$$

Note that the thermodynamic average over the external (y) coordinates has been interchanged with the integral! (see Eq. 10). Hence, the potential of mean force for any linear electrostatic system is completely described by an average surface charge on the bounding surface, $\bar{\sigma}(\xi; x) \equiv \langle \sigma(\xi; y) \rangle_y$.

It is important to notice the fundamental difference between the 'instantaneous' charge density $\sigma(\xi; y)$, which depends only on the external charge locations (y), and the average response $\bar{\sigma}(\xi; x) = \langle \sigma(\xi; y) \rangle_y$, which depends not on y but on the fixed locations, x, of the interior charges. The constant continuum dielectric model is just one example of linear response where the explicit charge density is restricted to the surface and is equal to this mean surface charge, $\bar{\sigma}(\xi; x)$. In terms of forces, we have the following interesting result. The force on a charge, q_i, due to the effective potential is

$$\bar{\mathbf{F}}_i = -\nabla_x G_{ext} = -\langle \nabla_x V_{ext} \rangle_y = -q_i \int_{\partial V} da(\xi)\, \bar{\sigma}(\xi; x) \nabla_x \left(\frac{1}{|x - \xi|} \right) \tag{14}$$

In linear response theory,

$$\begin{aligned} \bar{\mathbf{F}}_i &= -\tfrac{1}{2} \nabla_x \langle V_{ext} \rangle_y \\ &= -\frac{q_i}{2} \int_{\partial V} da(\xi) \left[\bar{\sigma}(\xi; x) \nabla_x \left(\frac{1}{|x - \xi|} \right) + \frac{1}{|x - \xi|} \nabla_x \bar{\sigma}(\xi; x) \right] \end{aligned} \tag{15}$$

In the last integral, since the first term is exactly equal to half the correct answer, the second term, which gives the linear response to the interior charges, must give exactly the same contribution.

This leads to a natural way to model the mean long-range force by determining a mean surface charge. An example of such a model is provided by the well-known replacement of the external media by a continuous uniform dielectric which we will illustrate below with a spherical cavity. Following this, we will suggest ways to modify the dielectric boundary to allow for a better treatment of short-range forces and the real flow of water through the cavity wall.

3. Boundary integral formulation of dielectric cavity

Representing the external medium by a linear dielectric continuum is an approximation to the exact mean surface charge approach. Further assuming an isotropic and homogeneous dielectric response, we are permitted to derive a self-consistent integral equation involving the surface charge. This follows from an expression of the total electrostatic potential of the system,

$$\begin{aligned} \phi_{tot}(\mathbf{x}) &= \phi_{int}(\mathbf{x}) + \phi_{ext}(\mathbf{x}) \\ &= \sum_i \frac{q_i}{|\mathbf{x} - \mathbf{x}_i|} + \int_{\partial V} da(\xi) \frac{\bar{\sigma}(\xi; x)}{|\mathbf{x} - \xi|} \end{aligned} \tag{16}$$

and the standard conditions on the normal component of the electric field near a *linear dielectric* boundary, together with a careful consideration of the surface charge contribution to the potential when $\mathbf{x} \simeq \xi$. The result is

$$\bar{\sigma}(\xi) = -\frac{\varepsilon_1 - 1}{\varepsilon_1 + 1}\sum_i \frac{q_i \hat{\mathbf{n}}(\xi)\cdot(\xi - \mathbf{x}_i)}{2\pi|\xi - \mathbf{x}_i|^3} - \frac{\varepsilon_1 - 1}{\varepsilon_1 + 1}\int_{\partial V} da(\xi')\frac{\bar{\sigma}(\xi)\hat{\mathbf{n}}(\xi)\cdot(\xi - \xi')}{2\pi|\xi - \xi'|^3} \tag{17}$$

where ε_1 is the dielectric constant of the external medium. For notational simplicity, we use $\bar{\sigma}(\xi)$ in place of $\bar{\sigma}(\xi; x)$ introduced earlier. In a more compact notation,

$$\sigma(\xi) = -\chi\,\hat{\mathbf{n}}(\xi)\cdot\mathbf{E}(\xi) \tag{18}$$

where $\chi = (\varepsilon_1 - 1)/(2\pi(\varepsilon_1 + 1))$ and $\mathbf{E} = -\nabla\phi$. This equation gives the mean induced surface charge on the boundary due to an arbitrary charge distribution inside.

Analytical solutions to the above may be obtained for a few regular boundaries such as a sphere. Of course, in such cases it is more customary to directly solve Poisson's equation and find an expression for $\phi(\mathbf{x})$. In the case of arbitrary geometries, the equation is solved numerically by dividing up the surface into finite elements; this is the well-known *boundary element method* [32,33].

In our view, the utility of this equation is its possible role in calculating long-range electrostatic forces for an arbitrarily shaped and ultimately dynamically adjusting bag containing some solvent and the solute. Of course, here we must be careful not to add too much computation to the problem in solving for the surface charge.

Below, we present calculations for the simplest example of a spherical cavity embedded in a dielectric continuum to illustrate some of the principles discussed so far. We stress that results are presented merely to illustrate the potential for the development of new methods based on this conceptual framework.

4. Results for a spherical cavity

As mentioned earlier, for a regular geometry such as a spherical cavity, we may solve Poisson's equation directly for the potential inside. The expression for the potential due to a single point charge q at $\mathbf{x}'$ inside a cavity with radius a is

$$\phi(\mathbf{x}) = \frac{q}{|\mathbf{x} - \mathbf{x}'|} + \phi_{\mathrm{ext}} \tag{19}$$

where

$$\phi_{\mathrm{ext}} = \frac{q}{a}\sum_{n=0}^{\infty}\frac{(n+1)(1-\varepsilon_1)}{\varepsilon_1(n+1)+n}\left(\frac{xx'}{a^2}\right)^n P_n(\cos\theta) \tag{20}$$

is the potential due to the reaction field of the dielectric continuum and ε_1 is the dielectric constant for water (outside). This standard example of a boundary value problem was considered by Kirkwood, who modeled the protein as a sphere embedded in a dielectric continuum to estimate pK_a shifts for ionizable groups [15,16]. In this article, we will not consider the potential due to redistribution of ions in the solvent.

Different authors, including Kirkwood himself, have sought to simplify the summation for ϕ_{ext} to render it into a more convenient form while preserving accuracy for computation, such as by expanding in $\varepsilon_0/\varepsilon_1$, which is a very small fraction (few %). We find the following exact result convenient for use in lookup tables [5]:

$$\phi_{ext} = -\frac{\varepsilon_1 - 1}{\varepsilon_1 + 1}\frac{q}{a}\frac{1}{\sqrt{1+\rho^2-2\rho z}} - \frac{\varepsilon_1 - 1}{(\varepsilon_1 + 1)^2}\frac{q}{a}\int_0^1 \frac{t^{\delta-1}\,dt}{\sqrt{1+(t\rho)^2-2t\rho z}} \tag{21}$$

where $\rho = xx'/a^2$, $z = \cos\theta$, $\delta = \varepsilon_1/(\varepsilon_1 + 1)$, and t is an integration dummy variable. Note that the first term above is the contribution from the image charge scaled appropriately by a factor involving the dielectric constants.

The force due to the dielectric response on a charge inside the cavity consists of one-body (self) and two-body (cross) contributions. Both can be computed from the gradient of the electrostatic potential energy,

$$\begin{aligned} V_{ext} &= \frac{1}{2}\sum_i\sum_j q_i q_j (G_{image_{ij}} + G_{integ_{ij}}) \\ &= \underbrace{\sum_{i<j} q_i q_j (G_{image_{ij}} + G_{integ_{ij}})}_{\text{2-body (cross) terms}} + \underbrace{\frac{1}{2}\sum_i q_i^2 (G_{image_{ij}} + G_{integ_{ij}})}_{\text{1-body (self) terms}} \end{aligned} \tag{22}$$

where the terms $G_{image_{ij}}$ and $G_{integ_{ij}}$ refer to the Green's functions for the image charge and integral contributions in Eq. 21. The force on q_i due to the solvent external to the boundary is

$$\mathbf{F}_{ext} = \mathbf{F}_{image} + \mathbf{F}_{integ} \tag{23}$$

where

$$\mathbf{F}_{image} = \frac{\varepsilon_1 - 1}{\varepsilon_1 + 1}\frac{q_i q_j}{a^3}\frac{1}{(1-\rho^2-2(w/a^2))^{3/2}}\left[\frac{x_j^2}{a^2}\mathbf{x}_i - \mathbf{x}_j\right] \tag{24}$$

$$\begin{aligned} \mathbf{F}_{integ} = \frac{\varepsilon_1 - 1}{(\varepsilon_1 + 1)^2}\frac{q_i q_j}{a^3}\Bigg[&\frac{x_j^2}{a^2}\mathbf{x}_i\int_0^1 \frac{t^{\delta+1}dt}{(1+(t\rho)^2-2t(w/a^2))^{3/2}} \\ &- \mathbf{x}_j\int_0^1 \frac{t^{\delta}dt}{(1+(t\rho)^2-2t(w/a^2)^{3/2}}\Bigg] \end{aligned} \tag{25}$$

and $w = \mathbf{x}_i \cdot \mathbf{x}_j$.

Incorporation of the above into molecular dynamics or Monte Carlo simulations is straightforward. First, we precalculate the dimensionless integrals in Eqs. 21 and 25 using Gaussian quadrature and store them in lookup tables. Then we simulate using the above potentials and forces (Eqs. 22–25) in addition to the standard molecular mechanics force field.

As an interesting side note, the theorem stated earlier involving surface charges permits one to calculate an effective dielectric constant as a function of test radius, r, in our spherical system with a charge, Q, placed in the center. This is derived by computing the net surface charge on a conducting sphere, and comparing this with the case of a dielectric sphere. The result is

$$1 - \frac{1}{\varepsilon(r)} = - \sum_{a > x_k > r} \frac{q_k}{Q} \frac{r}{|\mathbf{x}_k|} + \left(1 - \frac{1}{\varepsilon_1}\right)\frac{r}{a} \qquad (26)$$

where the sum extends over all the charges in the layer from an imaginary sphere at r to the actual dielectric at a. Note that as r → a, the expression gives the correct matching condition, $\varepsilon(r) \rightarrow \varepsilon_1$.

4.1. Estimating solvation energies of side chains

Table 1 shows example calculations of the free energy of transfer of selected amino acid side chains from vacuum to water (i.e., their solvation energies).

The electrostatic part of the solvation energy was estimated via a standard linear response assumption, i.e., by halving the average interaction energy between solvent

Table 1 *Example calculations of electrostatic solvation free energies of amino acid side-chain analogues*

Amino acid side chain		$\Delta G_{\mathrm{ele,calc}}$[a]	$\Delta G_{\mathrm{tot,exptl}}$[b]
Asn	acetamide	– 10.31	– 9.72
Asp	acetic acid	– 3.52	– 6.70
Cys	methylthiol	– 1.36	– 1.24
Gln	propionamide	– 9.33	– 9.42
Met	methylethylsulfide	– 0.09	– 1.49
Ser	methanol	– 5.41	– 5.08
Thr	ethanol	– 3.15	– 4.90
Tyr	p-cresole	– 8.59	– 6.13
Arg	N-p-guanidinium	– 58.54	– 66.07[c]
Lys	N-butylammonium	– 69.63	– 69.24

[a] Electrostatic solvation free energy from the spherical dielectric cavity model using a linear response approximation.

[b] Experimental solvation free energy from vacuum to water transfer experiments [17–19].

[c] This is a calculated value taken from Ref. 20 using finite difference Poisson–Boltzmann methods.

(both explicit and continuum) and solute (amino acid side chain),

$$\Delta G^{\mathrm{ele}}_{\mathrm{solv}} \simeq \frac{1}{2}\sum_{i} q_i \langle \phi(x_i) \rangle \tag{27}$$

where q_i and x_i are the magnitudes and positions of the partial charges on the amino acid side chain, respectively, and the angle brackets denote averaging over Monte Carlo steps for the fully charged side chain. As a rigorous verification of the use of the linear response approximation, one could carry out a thermodynamic integration over solute charge to capture the nonlinear details of the explicit solvent response. In addition, it is necessary to find an optimal radius for the explicit solvent region, and to tune the nonbonded parameters of the molecules involved (see Sec. 4.2), since, as is well known, unaltered standard molecular mechanics parameters give poor agreement between calculated and empirical solvation energies. It may also be necessary to impose geometrical constraints on the outer layer of explicit water molecules to prevent overpolarization and also to consider the transmission of thermal fluctuations from the outer bulk solvent to the explicit particles (see Sec. 6) since the dielectric continuum model is clearly a simplification that does not take into account all the degrees of freedom of the external bulk. No such attempts were made here. However, in the next section, we suggest improved treatments of the boundary to redress some of these shortcomings.

4.2. Simulation details

Side-chain analogues were placed in the center of an approximately 12 Å sphere and surrounded by ~60 water molecules. The analogues were taken in their standard initial configuration from the CHARMM [21] program. For partial charges and van der Waals radii, the CHARMM19 parameters [21] were used. The side chains were solvated using a box of pre-equilibrated TIP3P [22] water, removing all water molecules within 2.6 Å of the side chain, and those beyond 7 Å from the center. These starting coordinates were generated using the CHARMM program and were subsequently passed into an in-house implementation of the spherical model driven by dynamics and Monte Carlo algorithms. The side-chain covalent bonds were fixed by rigid constraints during simulation. Because water is generally involved in all biologically relevant simulations, we chose to develop a specific algorithm to treat explicit water rotations rather than using constrained optimization schemes such as SHAKE [23], which is a desirable alternative for treating macromolecular constraints. Averages were taken over the last 15 000 steps of 20 000 step runs. Configurations were updated using the Langevin equation [24] (equivalent to constant N,V,T Monte Carlo) at T = 298 K.

Because here our goal was to demonstrate the usefulness of the long-range electrostatic approximation, the short-range forces of external origin were modeled by a simple repulsive spherical wall potential ($1/r^{12}$) placed slightly inside the dielectric

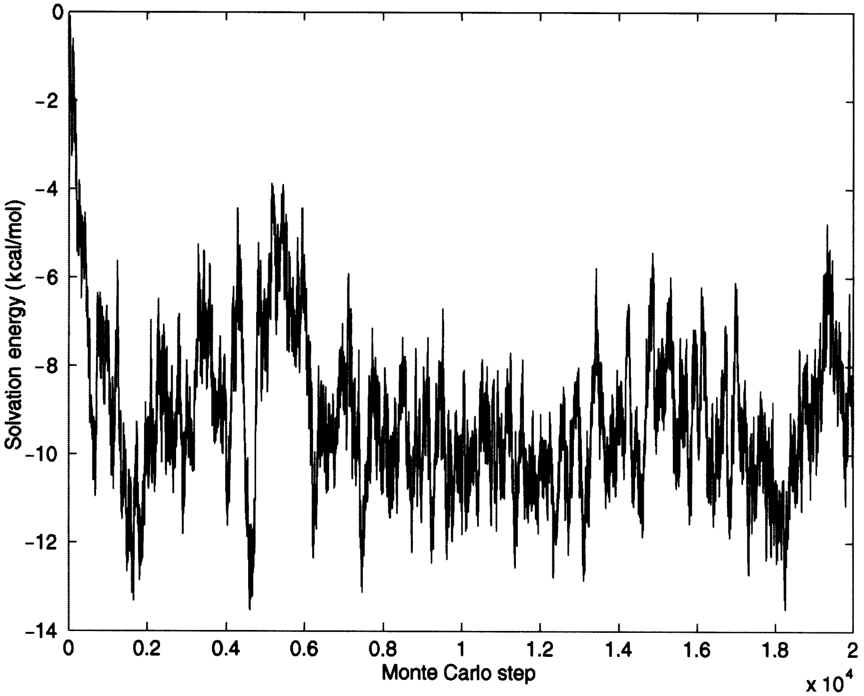


Fig. 3. Typical trajectory from side-chain solvation simulation (glutamine).

boundary. This is probably the crudest method to represent the interfacial short-range forces from the external bulk water. A reasonable radius was estimated by observing the distribution of water molecules in a test run with given cavity radius, and roughly determining the excluded volume due to the repulsive boundary. To ensure the correct pressure, the system radius was scaled at the beginning of the simulation to reflect the correct density ($\rho \simeq 0.334$ molecules/Å^3) of liquid water, with the excluded volume of the boundary force in mind. In addition to supplying pressure, the boundary is needed also to prevent evaporation.

Figure 3 shows a typical trajectory from a side-chain solvation simulation (glutamine) using the simulation protocols described above. Half the average interaction energy between solute and solvent is plotted against the Monte Carlo step. Notice that the energy falls off quickly toward equilibrium within the first few thousand steps of the simulation and is followed by equilibrium fluctuations.

Figure 4 shows the radial distribution function for an example MC simulation of 56 TIP3P water molecules in a cavity with radius 10.53 Å. A simulation protocol similar to the above was used. The tapering of the curve at larger distances reflects the finite size of our explicit water sphere.

5. Permeable boundaries and the 'polaron' model

In many respects, the spherical dielectric cavity model above is primitive and does not fully exploit the ideas presented earlier involving the division of the potential into external and internal components. In this sense, we regard this model as a 'warm-up' exercise that may help guide us toward more sophisticated methods.

There are three important issues not addressed in this model which we will briefly discuss in this section. However, all the ideas below are works in progress and we postpone the presentation of details and results to separate research articles [25].

First, we would like to minimize the distortions that an impenetrable boundary causes on the configurational evolution of the explicit water. Such distortions introduce biases in the molecular trajectories or configurational sampling which propagates into the interior. This, in turn, would require a larger explicit water layer for accurate results. The effect can be viewed essentially as a loss of entropy due to the decreased degrees of freedom of the explicit water near the boundary. Ideally, we would like to preserve the momentum flux of the molecules as in the periodic boundary case.

Second, the cavity above has a rigid, fixed shape. We would like one that can take on irregular shapes so that simulations involving only a few hydration shells closely

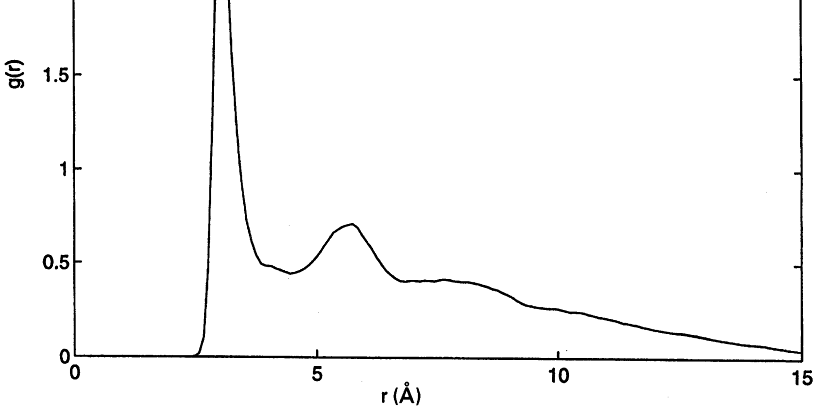


Fig. 4. Oxygen–oxygen radial distribution function for the spherical dielectric cavity model.

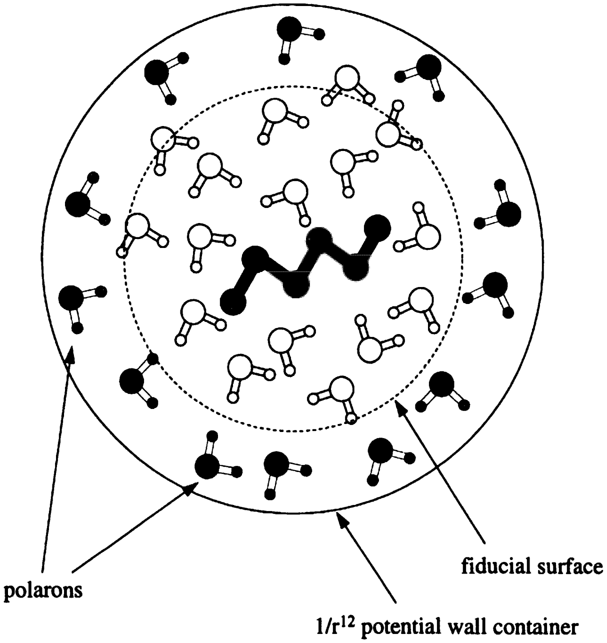


Fig. 5. Polaron approach to the dielectric sphere model.

hugging the solute of interest become possible. In this case, owing to the flexibility of the solute, we need a boundary that adjusts its shape dynamically as the simulation proceeds.

Finally, the explicit solute particles in the above are described entirely in terms of fixed partial charges and van der Waals radii. In reality, electronic polarizability plays a significant role in determining the energetics and dynamics of solvated systems [2].

We are currently pursuing four possible solutions among the many imaginable. All four rely on the determination of surface charge that reproduces bulk effects. The first two address the issue of short-range boundary distortions.

The simplest among them extends the spherical model by allowing particles to penetrate the boundary (see Fig. 5). We start with a spherical cavity and introduce an additional spherical boundary slightly inside to define a fiducial volume. We ask the question, 'what is the surface charge distribution (on the fiducial surface) that reproduces the external bulk solvent if we assumed the solvent extended to infinity?' This can be modeled quite easily. We introduce this unknown surface charge distribution during simulation by attaching appropriate charges to the water molecules in the buffer region (between the hard wall and fiducial surface). These buffer molecules, which we call 'polarons', reproduce the polarization of the bulk medium. The polaron charges are then adjusted to satisfy a discrete version of the linear integral equation

17. Each polaron is associated with a surface area patch ΔA. However, unlike the boundary element approach, we do not wish to take ΔA to zero. This is not necessary since they now represent the actual discretization of fluctuations of the underlying surface water distribution. (The polarons are intended as a mesoscopic model as discussed in Sec. 6.) In a sense, the internal boundary defines the beginning of the dielectric continuum, while the outer boundary acts as a container, supplying the appropriate short-range forces. To make the transition smooth, we may adopt a switch function that gradually turns on the 'polaron character' of a water molecule as it leaves the internal volume and enters the buffer region.

The second idea involves the use of a periodic boundary that incorporates long-range electrostatics in an unconventional fashion (see Fig. 6). In this scheme we periodically extend the short-range force and allow the molecules to enter and exit the system as in the normal periodic boundary. We then depart from this standard scenario by introducing an imaginary boundary, just inside the periodic one that

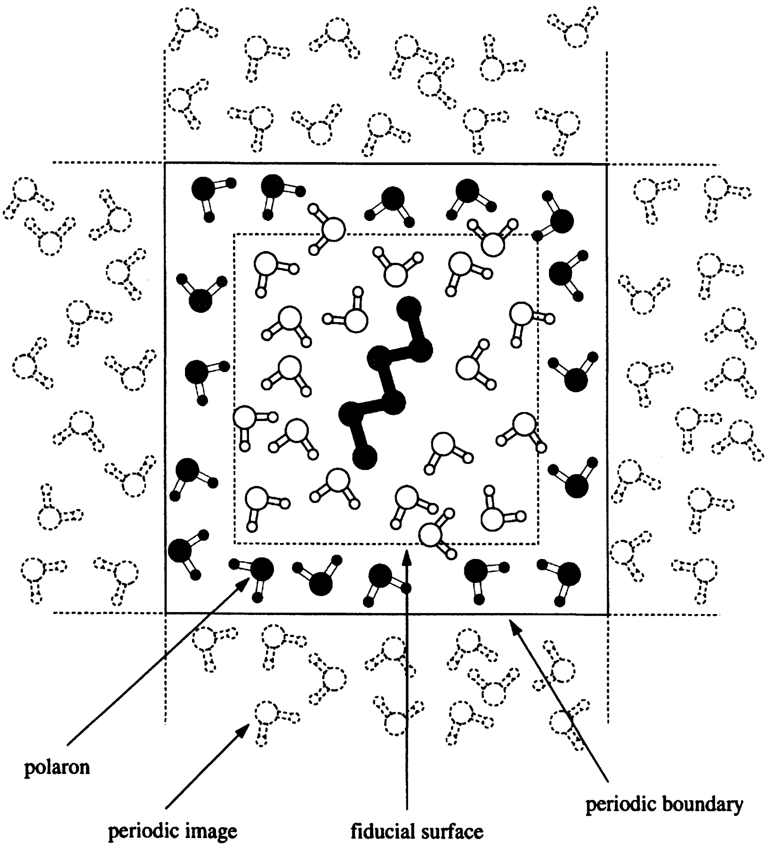


Fig. 6. Polaron approach to periodic boundary conditions.

again defines an internal fiducial volume. We solve the average surface charge problem with the appropriate polaron charges attached to the molecules in the buffer region during simulation as before. Again, we may use a switch function to smooth out the transition.

The third approach addresses the issue of flexible boundaries. Here we surround the solute of interest with one or two layers of explicit water, and allow the water droplet to define its own boundary rather than forcing it into a container. At each MD step, the boundary is defined as being the (solvent-accessible) surface of the droplet. Equation 17 is numerically solved, and the appropriate polaron charge is attached to the surface molecules. Since every MD step is a small perturbation from the previous configuration, polaron charge and surface area updates require very few additional numerical iterations. An additional force normal to the surface may be supplied to the polarons to represent the missing short-range forces that reproduce the correct pressure and prevent evaporation. This force can be used as a convenient calibration parameter to fit the model against experimental data. The challenge here is the development of efficient algorithms to approximate the surface and solve for the surface charges.

Finally, we would like to introduce a polarizable solute model by solving for the induced dipoles on each atom or charged group from the polarizability tensor (γ_{ij}) and electric field (**E**). This leads to a new set of linear equations,

$$p_n^i = \gamma_n^{ij}\, E^j(\mathbf{x}_n) \tag{28}$$

three for each dipole $\mathbf{p}_n$ at $\mathbf{x}_n$. The electric field is, of course, given by $-\nabla\phi$ where ϕ has an additional term,

$$\phi_{\mathrm{dipole}}(\mathbf{x}) = \sum_{k \neq n} \frac{\mathbf{p}_k \cdot (\mathbf{x}_k - \mathbf{x})}{|\mathbf{x}_k - \mathbf{x}|^3} \tag{29}$$

Note that the induced dipole equation (Eq. 28) has almost the same form as the surface charge equation, Eq. 18. We feel strongly that the surface charge problem can be combined with the induced polarization problem to form one efficient algorithm that simultaneously solves both. Indeed the two problems are fundamentally the same. There is an electrostatic problem in the volume outside the individual atoms and inside the bounding cavity which both react to induced polarization effects.

6. Discussion

Thus far we have reviewed a few fundamental issues related to the electrostatics of macromolecular solvation and have suggested possible computational methods that may help us understand this extremely important topic without reference to current approaches in the field. In this section, we would like to at least partially make up for our ‘sins of omission’ by giving a brief and selective mention of ideas and methods developed by other workers and comparing them with the concepts presented in this article. We will not attempt to give a comprehensive survey. For this, we refer the reader to the many review articles that address general issues in macromolecular electrostatics and related computational methods [1–5,26]. Of course, many relevant discussions can also be found in the articles of this and previous volumes of the current book series (e.g., Refs. 27–29).

From the very beginning, solvation models have been constructed with some sort of definition of a boundary that divides space into finite and infinite parts with different physical properties. The celebrated Born model [30] is an example in which an ion is modeled with a dielectric sphere and a point charge in the center embedded in another dielectric with different permittivity. This simple model was extended by Kirkwood and Tanford [15,16] to represent a protein with a dielectric sphere and arbitrary charge distribution inside.

Recent efforts in this general direction involve the modeling of macromolecules with a dielectric region whose boundary approximately delineates the van der Waals volume of the protein of interest, and which contain partial atomic charges where polar or ionizable groups exist. This is embedded in a dielectric continuum representing water, and the Poisson–Boltzmann equation is solved numerically for the electrostatic potential using finite difference methods [31]. Alternatively, Poisson's equation can be solved via the boundary element method [32,33], which has the advantage that the partial charges of the protein need not lie on grid points, and the external solvent is effectively infinite in extent.

Progress in quantum chemistry and the advent of the molecular mechanics approach, however, have made it clear that the physical forces among molecules and atoms all stem from the same source, namely the electron cloud distributions and how they interact and alter one another as a function of distance of separation. Thus, at a microscopic scale, the distinct two-phase system of water and solute has become blurred; these are nothing but two different regions in space with different mean properties. Only at macroscopic scales does this *mean* description become appropriate and the distinction clear.

Of course, if the computation were practical (or if very large amounts of computational resources were available), we could reproduce a solvated system most accurately using 'brute force' simulations that involve a very large number of explicit solvent [34–38]. If the microscopic system is carefully constructed (e.g., through consideration of electronic polarization as reviewed extensively in Ref. 2), we can in theory recover through these simulations the correct averages from very large to very small time and length scales. All that is required is for one to monitor a subset of the system (say a pair of solvent particles for pair correlations, or the particles that make up the protein) and evaluate time or Monte Carlo averages over them.

On the other hand, one rarely (if ever) requires a complete description of the entire system over *all* time and length scales. This would mean that one requires exact reproduction of the dynamics trajectories of the system as it evolves in nature. In reality, some parts of the system can be adequately described by mean properties. These can be *a priori* 'integrated out' of the probability distribution and incorporated into the potential that governs the system evolution as an effective potential. This allows the possibility for including different degrees of accuracy in the simulation by deciding how much (and which coordinates) to 'partially integrate' *a priori*, and how much to leave for the simulation to explicitly carry out. Thus, rather than the two-phase explicit versus continuum picture, it is more natural to think of the differences in length scales ranging from the macroscopic (i.e., dielectric and the Navier–Stokes equations) to the microscopic (i.e., molecular mechanics or quantum mechanics). In between, we have the *mesoscopic* picture, which does not 'exist' in nature in the traditional sense, but which we are free to construct in any fashion such that they reproduce the requi ed level of detail. The polarons in our proposed models are intended as a first attempt at such a mesoscopic description.

Mesoscopic models that incorporate both large- and small-scale properties into the particles undergoing dynamics have been an active area of research in statistical physics for many years. A few notable models that are currently receiving attention include lattice gas models [39], dissipative particle dynamics [40], and direct simulation Monte Carlo [41]. In addition, much research has gone into the development of acceleration algorithms via fast Fourier transforms or multigrid methods [42]. Application of these ideas to biological simulation may not be immediately straightforward, but one cannot preclude this possibility.

For biological macromolecules, one generally needs detailed structural information of the protein and nearby water; the rest of the bulk water affects the system only in the mean sense. Thus, from a practical simulation standpoint, it is more natural to divide the system into near and far regions rather than solute and solvent. The question then becomes one of reconciling the two descriptions at the boundary.

These notions have been exploited by many researchers in the field. One of the first studies that considered the problem of finite size simulations was that of Stace and Murrel [43], who placed a spherical shell of fixed Lennard-Jones particles at the boundary to simulate a gas-phase system.

Berkowitz and McCammon [6] proposed a model consisting of three concentric spherical regions for liquid simulations. The central sphere contains the reactive chemical system of interest including explicit solvent particles governed by molecular dynamics. This is surrounded by a buffer region (a mesoscopic transition) which contains particles that obey the Langevin equation and which thus acts as a heat bath. Outside this region is a reservoir that consists of a frozen configuration of particles taken from a prior molecular dynamics run. The three spheres are defined with respect to the central solute that lies in a much larger system of solvent particles. The spheres are thus allowed to move with respect to the 'extra' particles and coordinate axes of the larger system together with the central solute.

Brooks and Karplus [44] developed and used an effective potential (which they call a mean force field approximation or MFFA) due to the external, implicit solvent region based on pair correlation data. Their initial model was developed for nonpolar liquids. Later, Brunger et al. [7] extended the MFFA approach to simulate ST2 [45] water although the long-range effects of the electrostatic potential were not considered.

The problem of polar solvents was carefully examined by Warshel [46], who proposed the surface constrained soft sphere dipole (SCSSD) model. Here, the solute is surrounded by a collection of van der Waals spheres with point dipoles in their centers (i.e., mesoscopic water models). These are surrounded by a spherical shell of the same particles but fixed in their positions and orientations. The long-range electrostatic effects from the medium beyond the shell are calculated as a macroscopic dielectric reaction field that responds to the solute charge distribution. The shell of fixed dipoles was required to prevent *overpolarization* of the explicit solvent. This results from neglect of geometrical constraints that are normally imposed on the explicit molecules due to the (implicit) particles in the next layer out. These effects cannot be reproduced by a featureless dielectric that responds only to the solute.

Another level of detail was added by Warshel and King [47,8]. They considered the transmission of thermal fluctuations across the bulk/explicit boundary through their surface constrained all-atom solvent (SCAAS) model. Here, the solute is placed in a sphere of all-atom water molecules that are governed by regular dynamics. This is surrounded by a shell of explicit solvent which obeys a constrained potential based on a Brownian harmonic oscillator with parameters fitted against large-scale simulations. This constrained potential is meant to prevent overpolarization in a more sophisticated manner than the SCSSD model. The region beyond is given a dielectric treatment as before.

Rullmann and van Duijnen [9] constructed a model called reaction field with exclusion (RFE). In their approach, the solute and a small number of SPC [48] water molecules are placed in a spherical container. The container is defined by rejecting Monte Carlo configurations that place particles outside a given radius. This radius was made slightly smaller than that for the dielectric sphere. The solute was treated via *ab initio* quantum calculations, with an electrostatic potential governed by the dielectric reaction field of the external bulk medium in response to the solute charge distribution inside. The reaction field itself was calculated using the first term of a reformulated Kirkwood expansion (from *solute* charges only). Explicit solvent motion was computed by determining the net dipole moment of all the *solvent* molecules inside and the corresponding reaction field using Onsager's formula for a point dipole in a spherical dielectric.

Beglov and Roux [10] developed a rigorous theoretical formalism that ties the effective solvent boundary potential with the Boltzmann configurational integral, similar in spirit to the theme of this article. For electrostatics, they consider a spherical cavity embedded in a dielectric continuum and use the Kirkwood equation to compute the reaction field, except that the system radius varies according to the position of the explicit solvent molecule furthest away from the central solute.

As can be seen from the above, hybrid solvation schemes combining explicit and implicit treatment of solvent have received enormous attention over the last 20 years. As a result, much insight has been gained and useful methods have been developed. In fact, many of these methods have been applied to a variety of small biomolecular systems. These include the calculation of intrinsic pK_a's of ionizable groups in bovine pancreatic trypsin inhibitor [49], estimation of binding free energies of HIV protease inhibitors [50], and simulation of enzyme catalysis [51].

Each method described in this section has its share of strengths and weaknesses. These invariably revolve around the concepts of computational efficiency, scalability, and accuracy. It is our firm belief that the classic dilemma between cost and accuracy can be resolved by taking a fundamental perspective, being careful not to incorporate too little or too much detail. This general notion is embodied in our mesoscopic approach, in which we are allowed to include varying levels of detail into the model. We believe that the development of efficient and practical dynamic solvation schemes will remain an active area of research for many years to come.

Acknowledgements

We would like to thank Bruce Boghosian, Charles DeLisi, and Chao Zhang for helpful discussions. This work was supported in part by grants DBI-9630188 and GER-9452651 from the National Science Foundation.

References

1. Warshel, A. and Åqvist, J., Annu. Rev. Biophys. Biophys. Chem., 20(1991)267.
2. Warshel, A. and Russell, S.T., Q. Rev. Biophys., 17(1984)283.
3. Davis, M. and McCammon, J.A., Chem. Rev., 90(1990)509.
4. Harvey, S.C., Proteins Struct. Funct. Genet., 5(1989)78.
5. Zhang, C., Kimura, S.R., Weng, Z., Vajda, S., Brower, R.C. and DeLisi, C., J. Franklin Inst., submitted.
6. Berkowitz, M. and McCammon, J.A., Chem. Phys. Lett., 90(1982)215.
7. Brunger, A., Brooks III, C.L. and Karplus, M., Chem. Phys. Lett., 105(1984)495.
8. King, G. and Warshel, A., J. Chem. Phys., 91(1989)3647.
9. Rullmann, J.A. and van Duijnen, P.Th., Mol. Phys., 61(1987)293.
10. Beglov, D. and Roux, B., J. Chem. Phys., 100(1994)9050.
11. Chodos, A., Jaffe, R.L., Johnson, K., Thorn, C.B. and Wiesskopf, V.F., Phys. Rev. D, 9(1974)3471.
12. Allen, M.P. and Tildesley, D.J., Computer Simulation of Liquids, Clarendon Press, Oxford, 1987.
13. Hansen, J.P. and McDonald, I.R., Theory of Simple Liquids, Academic Press, London, 1976.
14. Jackson, J.D., Classical Electrodynamics, 2nd ed., Wiley, New York, NY, 1975.
15. Kirkwood, J.G., J. Chem. Phys., 2(1934)351.
16. Tanford, C. and Kirkwood, J.G., J. Am. Chem. Soc., 79(1957)5333.
17. Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C., Biochemistry, 20(1981)849.
18. Wolfenden, R., Biochemistry, 17(1978)199.
19. Cabani, S., Gianni, P., Mollica, V. and Lepori, L., J. Solut. Chem., 10(1981)563.
20. Sitkoff, D., Sharp, K.A. and Honig, B., J. Phys. Chem., 98(1994)1978.

21. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
22. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L., J. Chem. Phys., 79(1983)926.
23. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23(1977)327.
24. Langevin, P., Comptes Rendus, 146(1908)530.
25. Brower, R.C., Kimura, S.R., Zhang, C. and DeLisi, C., in preparation.
26. Honig, B. and Nicholls, A., Science, 268(1995)1114.
27. Smith, P.E. and van Gunsteren, W.F., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 182–212.
28. Sharp, K.A., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 147–160.
29. Berendsen, H.J.C., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 161–181.
30. Born, M., Z. Phys., 1(1920)45.
31. Sharp, K.A. and Honig, B., Annu. Rev. Biophys. Biophys. Chem., 19(1990)301.
32. Zauhar, R.J. and Morgan, R.S., J. Mol. Biol., 186(1985)815.
33. Zauhar, R.J. and Morgan, R.S., J. Comput. Chem., 9(1988)171.
34. Bash, P., Singh, U.C., Langridge, R. and Kollman, P.A., Science, 236(1987)564.
35. Jorgensen, W.L. and Briggs, J.M., J. Phys. Chem., 94(1990)1683.
36. Tirado-Rives, J. and Jorgensen, W.L., Biochemistry, 30(1991)3864.
37. Tobias, D.J. and Brooks III, C.L., Biochemistry, 30(1991)6059.
38. Van Buuren, A.R. and Berendsen, H.C., Biopolymers, 33(1993)1159.
39. Rothman, D.H. and Zaleski, S., Lattice-Gas Automata: Simple Models of Complex Hydrodynamics, Cambridge University Press, Cambridge, 1997.
40. Español, P., Phys. Rev. E, 52(1995)1734.
41. Alexander, F.J., Garcia, A.L. and Alder, B.J., Phys. Rev. Lett., 74(1995)5212.
42. Alexander, F.J., Brower, R.C., Gould, H. and Kimura, S.R., in preparation.
43. Stace, A.J. and Murrel, J.N., Mol. Phys., 33(1977)1.
44. Brooks III, C.L. and Karplus, M., J. Chem. Phys., 79(1983)6312.
45. Stillinger, F.H. and Rahman, A., J. Chem. Phys., 60(1974)1545.
46. Warshel, A., Chem. Phys. Lett., 55(1978)454.
47. Warshel, A. and King, G., Chem. Phys. Lett., 121(1985)124.
48. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F. and Hermans, J., In Pullman, B. (Ed.) Intermolecular Forces, Reidel, Dordrecht, 1981, pp. 331–342.
49. Russell, A.T. and Warshel, A., J. Mol. Biol., 185(1985)389.
50. Hansson, T. and Åqvist, J., Protein Eng., 8(1995)1137.
51. Warshel, A., Papazyan, A. and Kollman, P.A., Science, 269(1995)102.

Application of Poisson–Boltzmann solvation forces to macromolecular simulations

Adrian H. Elcock, Michael J. Potter and J. Andrew McCammon
*Department of Chemistry and Biochemistry, Department of Pharmacology,
University of California at San Diego, La Jolla, CA 92093-0365, U.S.A.*

Introduction

One of the most difficult problems encountered in the dynamical simulation of large macromolecular systems is how to deal adequately with the huge number of atomic interactions involved. For aqueous-phase simulations the computational burden associated with solvent water molecules can easily outstrip that associated with the macromolecule, even though the behavior of the solvent itself may not be of much interest. Not surprisingly therefore, considerable interest has been focused on the use of methods in which explicit solvent water molecules are replaced by an implicit dielectric continuum representation; an excellent review of such methods was given by Sharp [1] in the previous volume of this series. Perhaps the most generally accepted continuum-based approach centers on the use of the Poisson–Boltzmann (PB) equation of classical electrostatics [2], a method which owes its success to the fact that many solvation-related phenomena (with the notable exception of the hydrophobic effect) appear to be essentially electrostatic in nature. Until very recently, use of the PB approach has largely been restricted to calculations involving static representations of molecular structure, but the recent development of methods to obtain solvation *forces* from the PB equation [3] means that it can now, in principle, also be used in dynamics simulations. Applications of the former type have been comprehensively reviewed in the literature [2] and are not discussed further in this article; instead, we restrict our attention to the potential use of PB electrostatics in dynamical simulations of macromolecules.

The advantages of a PB-based approach for dynamics simulations of large systems relate not only to reductions in the size of the system, but also to the accessible timescale of the simulation. For example, the relaxation time of dissolved mobile ions that surround highly charged molecules such as DNA is typically estimated to be on the order of hundreds of picoseconds. A simulation of such a system must therefore extend for nanoseconds if absolute convergence of ion atmosphere effects is required. For such situations a PB-based approach offers special advantages since the mean distribution of mobile ions around charged molecules is readily obtained from, and is partly the basis of, the PB equation [2]. Furthermore, in cases where large-scale structural changes occur, so that the ionic atmosphere is strongly perturbed, a simple update of the solution to the PB equation is all that is required to determine the extent and effects of ion redistribution. An additional advantage is that since bulk mobile ion concentration enters explicitly into the PB equation, simulations can, in principle, be performed at varying ionic strength.

Further timescale-related advantages accrue from the omission of the molecular nature of the solvent. The large-scale motions of macromolecules, be they either the global motions of rotation and translation, or more local conformational changes, are often subject to significant damping due to solvent viscosity. In a PB-based approach the solvent, being only implicitly represented, is effectively of zero viscosity and the frictional and stochastic forces due to collisions with water molecules are neglected. Since the latter two forces are often important for redistributing energy between a molecule's vibrational modes, their absence represents something of a problem; their effects can, however, be reincorporated relatively easily, so that *correct equilibrium properties* are obtained, using the technique of stochastic dynamics (SD) [4]. Despite this slight technical difficulty, the ability to set the solvent frictional constant to a low value is expected to allow structural transitions to occur much more readily than if solvent molecules were explicitly represented.

The molecular nature of the solvent raises difficulties for conventional simulations in other ways. Ligand docking and macromolecular association, for example, are almost always accompanied by some degree of desolvation of one or both molecules. In extreme situations, such as a ligand passing down a narrow solvent-lined tunnel into the active site of an enzyme, the difficulties associated with expelling solvent molecules so that the ligand can pass are sufficiently great that the process is next to impossible to study by conventional means. Obviously, no such problem is faced when the solvent is treated simply as a dielectric continuum.

Clearly, significant advantages are to be expected from adopting a PB-based approach for simulating large macromolecular systems. What is also apparent from the above discussion, however, is that the dynamical behavior of such systems will be strongly altered by the omission of explicit solvent and ions: a combination of PB and SD methods (PB/SD) can be tuned to provide a fairly realistic description of the solute dynamics, or – by using artificially low friction coefficients – to provide more rapid configuration sampling, which still yields valid equilibrium information.

Having detailed the potential advantages of the method, it is of course necessary to demonstrate that these benefits are actually obtained in practice. An important step in this direction was recently taken in work reporting the use of a combined PB/SD method to study the conformational preferences of the small molecules dichloroethane and alanine dipeptide [5]. In the study, a conventional molecular mechanics force field was used to represent the internal bonded and nonbonded interactions (including Coulombic interactions). This 'gas-phase' force field was supplemented by electrostatic solvation forces obtained by solving the PB equation using finite difference (FD) methods. The solutes were each simulated for ~ 2 ns using stochastic dynamics to obtain probability distributions for the internal dihedral angles, which were then compared with reference distributions obtained from static energetic calculations. Probably the most important aspect of this work was the finding that very good agreement between simulated and reference conformational distributions was obtained even with relatively coarse grid spacings in the FDPB calculations.

It is obviously important to demonstrate that a PB/SD method can work well in the setting of a small-molecule model system. In most cases, however, it will be preferable to study such systems by the use of explicit solvent simulations, particularly as currently available computational resources seem more than adequate for the purpose. It remains important then to show that the method can be meaningfully applied to macromolecular systems since, as outlined above, it is in such settings that a continuum-based approach is likely to be of most use. Accordingly, as a first step towards this goal, we consider in this article some of the aspects that are more relevant to simulations of macromolecules. We begin by very briefly reviewing the expressions derived for the electrostatic forces acting in a system modeled by the PB equation. Qualitative aspects of the PB forces are then illustrated using simple models, before practical problems relating to their implementation in MD simulations are discussed. By way of an example, a relatively simple application of the method is outlined next: the simulation of a model protein–DNA interaction. Finally, future prospects and the remaining problems associated with the implementation of the PB/SD method are outlined.

Theory

In the PB model [2], the molecule of interest is modeled as a set of point charges embedded in a cavity of low dielectric (ε_m) which is itself immersed in a (generally much higher) dielectric solvent medium (ε_s). The charges and atomic radii (the latter of which are used to define the extent of the low dielectric molecular interior) are most commonly taken from conventional molecular mechanics (MM) force fields such as AMBER [6] or CHARMM [7], although parameter sets specifically designed for use in PB calculations of solvation energies have been developed [8,9]. In applications of the PB equation to macromolecular energetics, it is common to use a solute dielectric of 2–4 in a somewhat *ad hoc* attempt to include effects due to the polarizability of the macromolecule. In combination with molecular mechanics, however, it is more correct to use a solute dielectric of 1.0, since most MM force fields, and in particular those intended for use in explicit solvent simulations, are parameterized with this value of the solute dielectric in mind. This is the approach adopted here.

The total electrostatic force $\mathbf{f}_i$ on any atom of the system can be expressed as the sum of four contributions [3]:

$$\mathbf{f}_i = q_i \sum_{j \neq i}^{j=N} \frac{q_j \mathbf{r}_{ij}}{4\pi\varepsilon_m |r_{ij}|^3} + q_i \mathbf{E}_{rf} + \mathbf{f}_{dbf} + \mathbf{f}_{ibf}$$

The first of these components is the Coulombic interaction of atom i with all the other (N − 1) atoms of the system, evaluated with the solute dielectric ε_m. It should be remembered that in molecular mechanics calculations it is usual to scale 1–4 Coulombic (and van der Waals) interactions between atoms and omit such 1–2 and 1–3 interactions entirely. The same approach is adopted here for the combined PB-MM force field: the Coulombic terms are dealt with entirely by a conventional MM program.

The remaining three force components can be considered together as 'solvation' forces since they are the additional forces that result from transferring the (macro) molecule from a medium of dielectric ε_m to a dielectric of ε_s (which may also additionally contain dissolved mobile ions). The first of these components, $q_i\mathbf{E}_{rf}$, is the familiar reaction field force that expresses the interaction of the solute atom of charge q_i with the electrostatic field resulting from the polarization the charges induce in the solvent environment. The second component, $\mathbf{E}_{dbf}$, of more subtle origins, is a dielectric boundary force representing the propensity of high dielectric to move into regions of strong electrostatic fields by displacing the low dielectric (solute). The last term, $\mathbf{E}_{ibf}$, analogous to the dielectric boundary forces and present only when there is nonzero ionic strength, is a force exerted at the mobile-ion-accessible boundary expressing the tendency of such ions to move into regions of strong electrostatic fields. Expressions for the calculation of each of these components from finite difference solutions to the PB equation have been derived and discussed in detail [3]; these expressions have been implemented in the PB program UHBD [10].

As a very simple illustration of the basic qualitative behavior to be expected from solvation forces, it is instructive to examine a very simple model system. Figure 1 (upper panel) shows qualitatively the electrostatic forces acting between two oppositely charged atoms. The Coulombic forces, which are evaluated using the solute molecule's dielectric (i.e. $\varepsilon_m = \varepsilon_s = 1$), are of course attractive and directed along the line connecting the atom centers. The solvation forces, on the other hand, are repulsive and point in the opposite direction. This offsetting of the attractive Coulomb component by a repulsive solvation component yields the intuitively correct result that ion-pair formation is less favorable in a high dielectric medium (water) than in a low dielectric medium (gas phase). The origin of this effect lies in the unfavorable change in electrostatic solvation energy of the system as the ions are brought closer together. According to the simple Born [11] model of electrostatic solvation, the solvation energy of an ion is proportional to the square of the atomic charge. At infinite separation then, the solvation energy of the two-ion system is proportional to $(+1)^2 + (-1)^2$; at zero separation, however, the solvation energy is zero since the sum of the two opposite charges is zero. A similar offsetting occurs in interactions between like-charged ions (lower panel of Fig. 1). Here of course the Coulombic force is repulsive, but this is largely balanced by a favorable solvation force. Again, the net effect is intuitively reasonable: like ion-pairs are more likely to occur in water than in the gas phase. In this case, the favorable solvation force results from an increase in charge density: the solvation energy of a single ion of charge $+2$, obtained when the two ions are at zero separation, is more favorable (proportional to $(+2)^2$) than the combined solvation energy of two singly charged ions $((+1)^2 + (+1)^2)$.

This compensating effect of solvation forces is a quite general feature of PB electrostatics [2]. Figure 2 shows the electrostatic forces acting in a more complex but realistic example: that of a short DNA hexamer of sequence d(ATATAT)$_2$ in

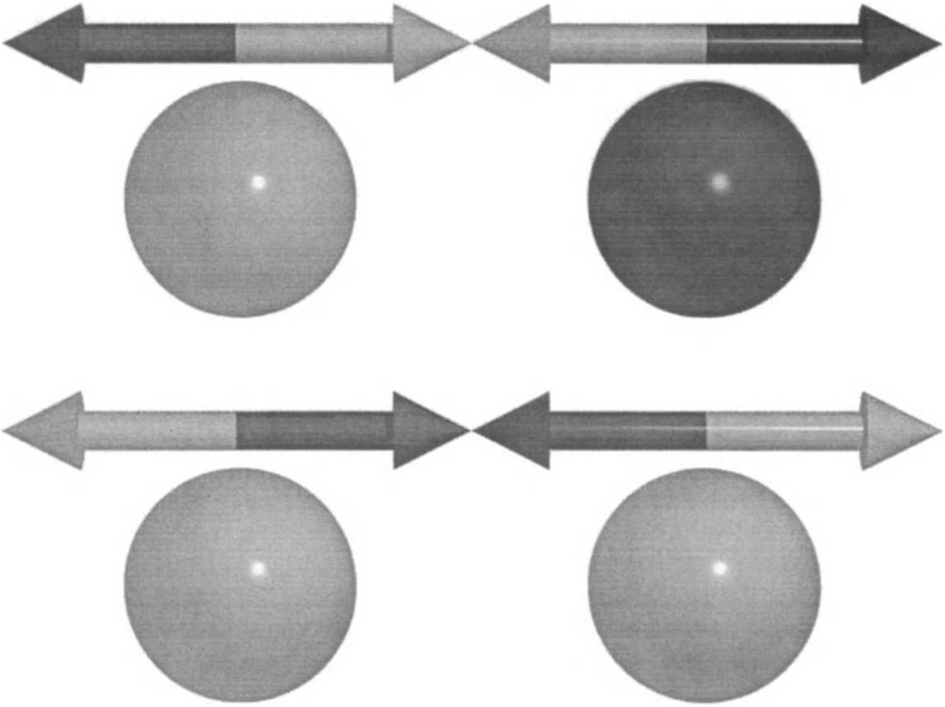


Fig. 1. Qualitative illustration of electrostatic forces acting between (top) oppositely charged and (bottom) like-charged spheres in high dielectric solvent. Coulombic forces are represented by light gray vectors and solvation forces by dark gray vectors.

a 150 mM 1–1 salt solution. In this figure, the forces acting on each of the five atoms of the phosphate group (including the O3′ and O5′ atoms) are summed together and centered on the phosphorus atom to give an indication of the overall force acting on each of the phosphate groups. Not surprisingly, the Coulombic interactions between the phosphates are repulsive and directed outwards from the minor groove where cross-strand repulsive interactions are the strongest. The solvation forces again essentially point in the opposite direction and tend to force the two strands together across the minor groove. The net forces are small, consistent with the fact that, despite the presence of large Coulombic repulsions between phosphates, the high degree of solvent screening serves to stabilize DNA in a double-helical form.

Practical aspects

Having illustrated the qualitative behavior to be expected from solvation forces, we now discuss the incorporation of these forces into dynamics simulations. In combining FDPB electrostatics with stochastic dynamics, there are two major questions

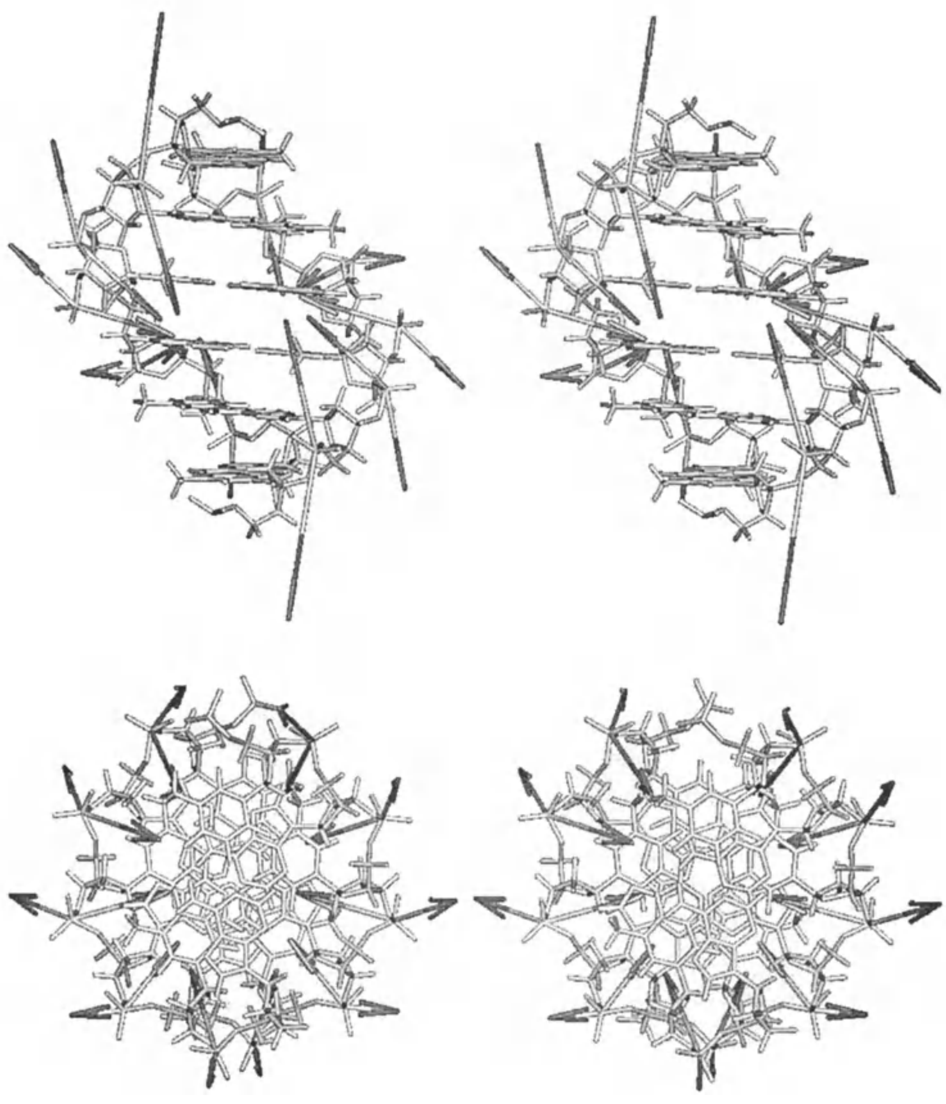


Fig. 2. Illustration of electrostatic forces acting on the phosphate groups of a DNA hexamer of sequence d(ATATAT)$_2$ (the lower view is looking down the helical axis). Arrows pointing outwards from the helical axis are Coulombic forces and those pointing inwards are solvation forces. Forces were calculated at 150 mM ionic strength.

which need to be addressed. The first is how accurately the PB forces need to be calculated to ensure a reasonable behavior of the system. Solving the PB equation using finite difference methods inevitably introduces grid dependencies to the results; such effects are a familiar aspect of PB energetics studies and similar concerns also naturally arise with force calculations. In general, a finer grid used in a PB calculation will yield more accurate (by which we mean less grid-dependent) forces and energies, but will require considerably more computer time than a similar calculation with a coarse grid. This is not usually a problem for energetics calculations since computer time is rarely a limiting factor, but for dynamics simulations, where the PB forces may have to be recalculated many thousands of times, this is likely to be a major concern. Since speed will obviously be of importance then, the question essentially boils down to one of how coarse a grid can one get away with whilst retaining forces of sufficient accuracy. In this regard it is worth repeating the most promising aspect of the work of Gilson et al. [5]: that good results could be obtained for equilibrium conformational distributions even with relatively coarse grids of around 1 Å spacing. By examining the variability of forces calculated for the DNA hexamer described above, we show below that similar results might also be obtainable with macromolecules.

The PB solvation forces acting on the atoms of the DNA hexamer were first calculated using a grid spacing of 1.0 Å, and charges and atomic radii taken from the AMBER94 force field [6]. Forces were obtained with 16 different orientations of the molecule on the grid and averaged separately for each atom; we use these average atomic forces as a reference for determining the variability of the forces on changing the grid orientation. For each of the 16 individual forces, an error was calculated as the difference between the force and the average 'correct' force acting on the same atom. Figure 3 shows the rms magnitude of this force error for each of the atoms of the DNA. A series of 12 evenly spaced peaks are visible which correspond to the atoms of the phosphate groups; also indicated is a series of six smaller double peaks corresponding to the adenine NH_2 groups. Figure 4 shows that the force errors correlate to some extent with the magnitude of the atomic charge, an effect which of course is not particularly surprising: given that the solvation forces themselves are expected to be larger in magnitude at highly charged atoms, the errors are also likely to be larger. Of the various components of the forces, the major contribution to the force errors appears to come from the dielectric boundary forces (Fig. 5); this is understandable since the discretized grid representation of the dielectric surface is expected to be particularly poor at a grid resolution of 1.0 Å. Figure 3 also shows the rms force errors obtained with a grid spacing of 0.4 Å. It is encouraging, though not particularly surprising, that the magnitudes of the force errors decrease sharply as the grid spacing is changed from 1.0 to 0.4 Å so that the largest average error for any of the atoms is reduced from around 5 kcal mol^{-1} Å^{-1} to around 1 kcal mol^{-1} Å^{-1}. The increased accuracy of the forces obtained with the smaller grid spacing is displayed graphically in Fig. 6, which plots each of the 16 individual forces calculated for a central phosphorus atom, together with their average, for the 0.4 and 1.0 Å grid spacings.

At first sight these results would appear to suggest that attempts to use a coarse grid for macromolecules with large electrostatic forces will lead to poor behavior. However, although forces obtained with the coarse grid of 1.0 Å spacing are clearly subject to a much greater degree of variability, the *average* forces obtained are in quite good agreement with the average forces calculated with the much finer grid of 0.4 Å (Fig. 7). Overall, the differences between the average forces obtained for the 0.4 and 1.0 Å grid

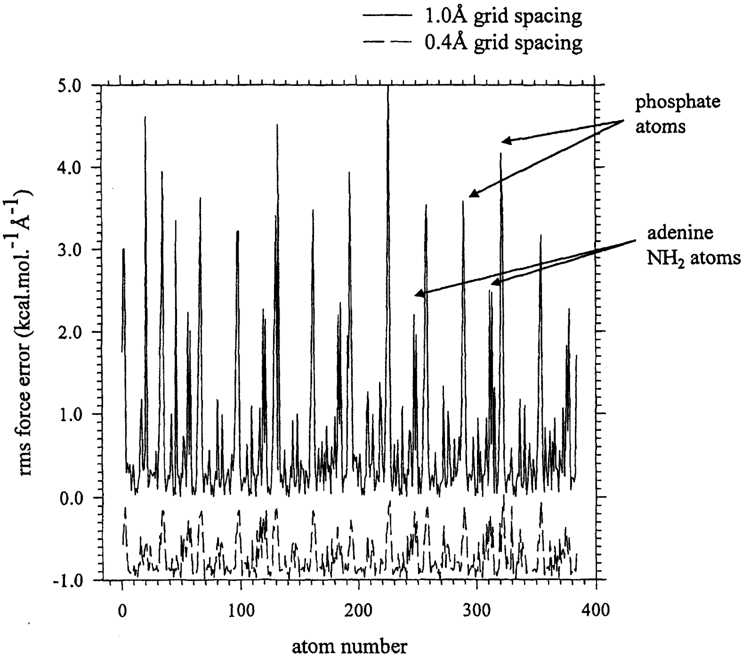


Fig. 3. Plot of rms solvation force error for each of the 384 atoms of the DNA hexamer. For clarity, values for the grid spacing of 0.4 Å are shifted down by 1 kcal mol^{-1} Å^{-1}. Atoms 1–192 are for the first DNA strand and atoms 193–382 are for the second (identical) DNA strand.

spacings are around 1 kcal mol^{-1} Å^{-1}, with errors for the phosphate atoms being around 1.4 kcal mol^{-1} Å^{-1}. The one major exception is the non-Watson–Crick paired proton of the adenine NH$_2$ groups, which clearly remains inaccurate: this is presumably a consequence of the combination of its small radius (1 Å), relatively high charge (+ 0.417e) and location at the dielectric boundary.

This overall more promising result leads directly to a discussion of the second major question to be addressed: how often to update the solvation forces. In principle of course, these forces could be recalculated at every step of an MD simulation. In practice, such an approach is, for the moment, out of the question since solving the PB equation for each new configuration of the system, particularly for macromolecular systems, would be prohibitively expensive. The most obvious way around the problem is simply to update the solvation forces only every n steps, keeping them constant for all intervening steps. However, the above results suggest that computational efficiency is not the only factor to be considered in determining the frequency of updating solvation forces. The fact that averaged forces can be relatively accurate, even though the instantaneous forces might be much more variable, indicates that the question of solvation force update frequency is also to some extent linked to the coarseness of the grid: more frequent updating might be required for coarse grids in order to provide

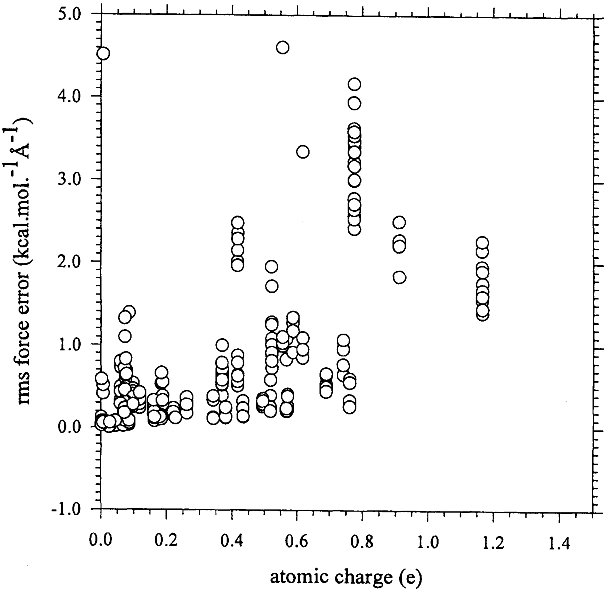


Fig. 4. Correlation of rms solvation force error with atomic charge.

a reasonable degree of averaging of the solvation forces. Obviously, this question is further connected with the relaxation properties of the system and it seems reasonable to expect that so long as the time between solvation force updates is considerably shorter than the relaxation time of the system, so that the system experiences an average of the instantaneous PB forces, reasonable results might be obtained. This will obviously vary from system to system, but for macromolecules, in which large motions of groups are subject to considerable damping, updating solvation forces every 0.1 ps should be a reasonable compromise between computational speed and force accuracy. On a related point, one could envisage more elegant schemes for updating the PB forces, such as only recalculating them when the configuration of the system has changed appreciably, for example if any atom moves more than a certain distance from its position at the previous solvation force update [5]. One could perhaps also argue that updating PB forces by simply overwriting the existing forces is undesirable since the forces change discontinuously. Indeed, Niedermeier and Schulten [12], in the first application of PB solvation forces to macromolecular dynamics (albeit using a model neglecting dielectric boundary forces), used an exponential scaling method for smoothly updating solvation forces. It should be remembered, however, that there is perhaps little point in adopting a method for smoothly updating the solvation forces when the forces themselves are subject to considerable numerical inaccuracy.

Questions of PB force accuracy and update frequencies are not the only concerns that must be faced in implementing a combined PB/SD method. An equally important

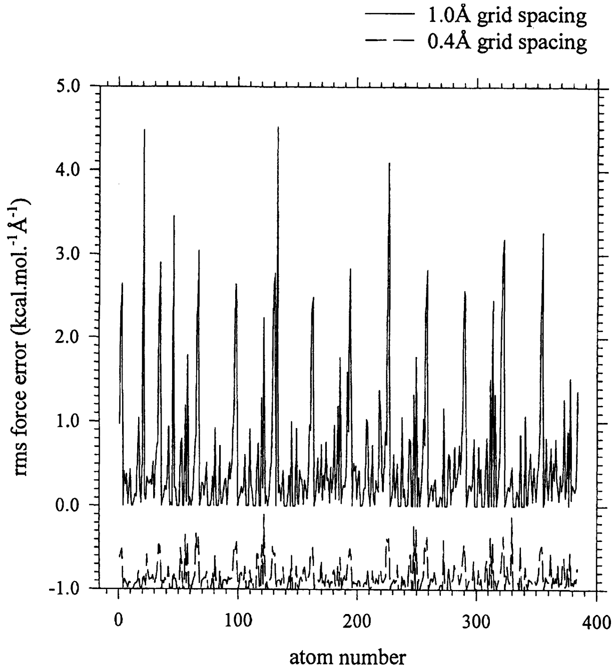


Fig. 5. Plot of rms dielectric boundary force error for each of the 384 atoms of the DNA hexamer. For clarity, values for the grid spacing of 0.4 Å are shifted down by 1 kcal mol^{-1} Å^{-1}.

problem, and one that is much less easily tested, concerns the overall balance of forces. As seen earlier for DNA, but also true for most systems with strong electrostatic forces, is the fact that Coulombic and solvation forces largely compensate one another so that the *net* electrostatic force is modest, even though the individual components may be large in magnitude. This requirement that large force components cancel out to leave a small net force would appear to represent a considerable challenge for proper parameterization of the method. Whilst it is, in some cases (e.g. amino acid side-chains), possible to adjust the atomic parameters (charges and radii) so as to reproduce experimental hydration free energies [8], this is not always the case owing to a lack of adequate experimental data (e.g. nucleic acid bases [13]). In any case, since the overall binding affinity of two molecules results from a balance of Coulombic and solvation contributions, there is no reason to believe that parameters developed simply to fit solvation energies will necessarily provide a good description of intermolecular interactions. As it stands, it appears that perhaps the only truly unambiguous method of verifying a suitable overall balance of forces is to run a long dynamics simulation with a view to assessing the long-time stability and behavior of the structure. In keeping with the considerable amount of care put into developing conventional explicit solvent force fields [6,7], it is therefore likely that the

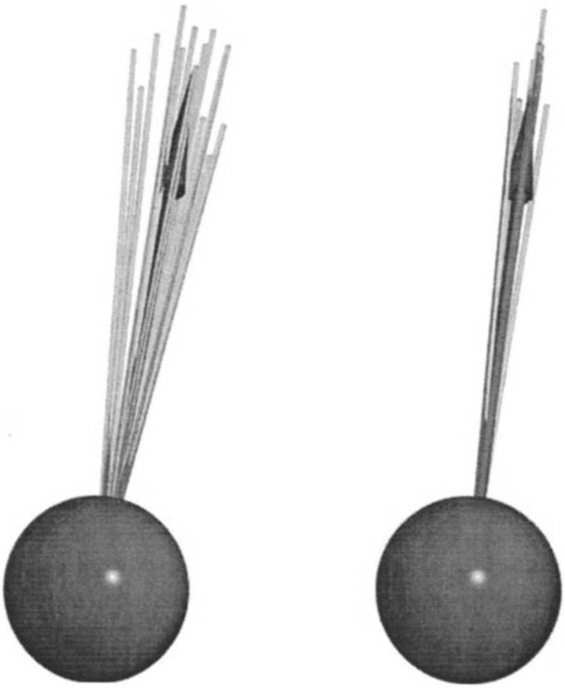


Fig. 6. Illustration of the variation of solvation forces acting on a central phosphorus atom of the DNA hexamer for 16 orientations using grid spacings of (left) 1.0 Å and (right) 0.4 Å.

development of a properly balanced force field for use in PB/SD simulations will be a laborious undertaking.

At this stage, it is perhaps worth mentioning one final effect likely to be encountered specifically in continuum solvent simulations. The atomic fluctuations that naturally occur in the course of dynamics often result in the formation of small cavities; in conventional explicit simulations, solvent will only enter such cavities if they persist for some time. This is due both to the time requirements of the purely diffusive solvent exchange process and to the presence of often significant energetic barriers. However, PB/SD simulations allow high dielectric solvent to enter such cavities instantaneously since the dielectric map is recalculated each time the solvation forces are updated. Instantaneous solvent relaxation is normally viewed as a major advantage of a continuum-based approach; in this case, it serves to increase the frequency of significant fluctuations occurring relative to explicit simulations, since the solvent, being fully relaxed, effectively lowers the kinetic barrier to any given conformational change. Whether this effect is actually an advantage or not will depend to a large extent on one's viewpoint: in terms of allowing fast exploration of configuration space as might be required in free energy calculations, the effect is clearly advantageous; for simulations aimed at studying a particular (hopefully stable) macromolecular conformation

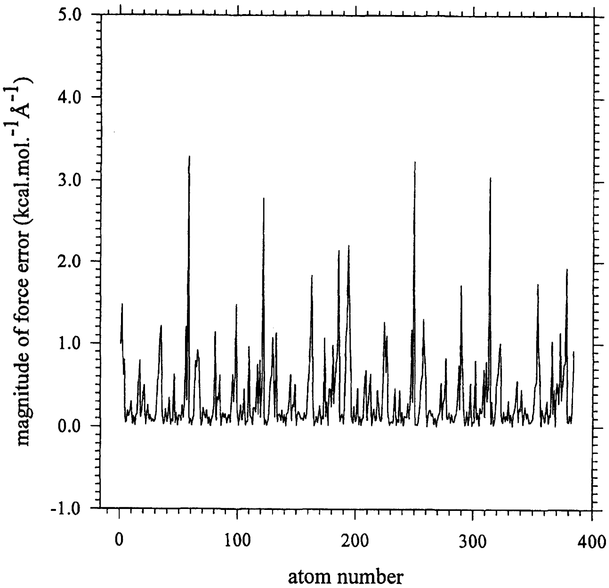


Fig. 7. Plot of the magnitude of the difference between average solvation forces calculated at 0.4 and 1.0 Å grid spacings for each of the 384 atoms of the DNA hexamer.

on the other hand, the effect is potentially a significant drawback, allowing a much greater degree of conformational freedom than might otherwise be desirable. The effect can be alleviated, but not eliminated, by the use of a probe-accessible surface definition [14] which requires cavities to attain a reasonable size before becoming solvent-accessible.

Example application to a model protein–DNA association

The above section has outlined some of the more important questions to be addressed in implementing a combined PB/SD method, and it will be necessary to deal with each of these questions properly in turn before applying the method to systems of real interest. As an indication of the potential utility of the method however, in this section we describe one relatively straightforward application to a model protein–DNA association reaction.

Proteins often induce large and surprising structural changes in DNA upon association, the most famous example so far being provided by the crystal structure of the TATA-box binding protein (TBP) complexed with DNA [15,16]. The protein binds in the minor groove of the DNA, bending the DNA away from the protein surface through an angle of nearly 90° and causing a dramatic increase in the minor groove width. These structural changes, which were unprecedented at the time the crystal structures were first announced, have since been observed in other protein–DNA complexes such as the sex-determining SRY protein [17]. It had previously been suggested [18] that a significant contribution to the groove-opening effect might come from the loss of solvent screening of phosphate charges which accompanies protein binding: the expulsion of high dielectric solvent by the approach of a low dielectric protein ought to increase cross-strand phosphate repulsions.

We recently investigated this idea by using the combined PB/SD method [19]. A model protein consisting of seven atoms (each of radius 10 Å), designed to fit snugly in the DNA minor groove, was constructed and placed at a distance of 30 Å from a DNA 16-mer (Fig. 8). The model protein was *uncharged* and did not *directly* interact with the DNA in any way (i.e. there were no Coulombic or van der Waals interactions); the protein therefore simply acts to define a region of low dielectric immersed in the otherwise high dielectric solvent.

The response of the DNA to the approach of this model protein into the minor groove was then investigated by hybrid PB/SD simulations carried out using a combination of the CHARMM [20] and UHBD [10] molecular simulation programs, with information transfer being controlled by simple command scripting. All dynamics calculations were performed within a version of CHARMM modified to read in PB solvation forces calculated periodically by UHBD; once read in, the solvation forces

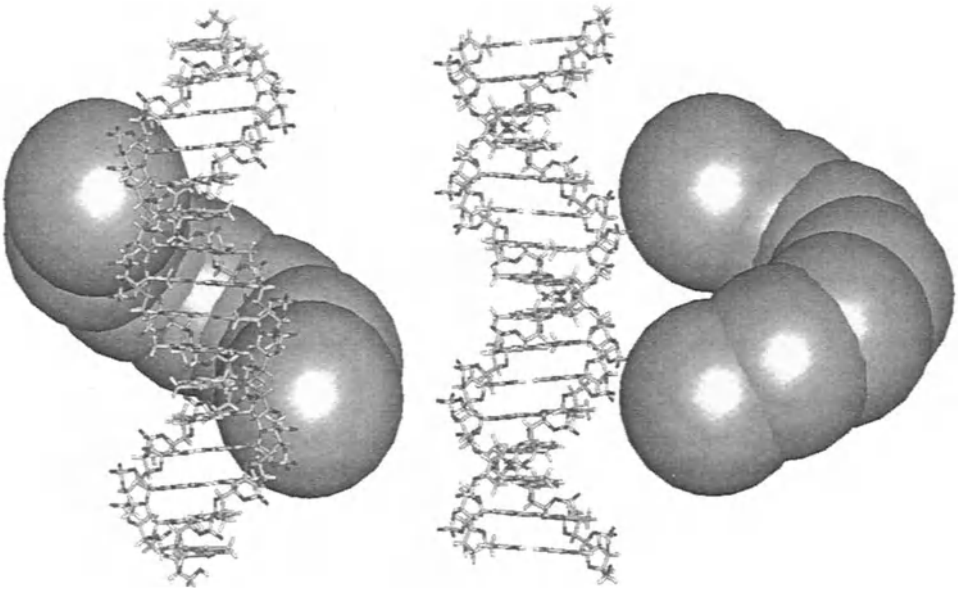


Fig. 8. Views of the model protein–DNA system at a separation distance of 30 Å. The protein, represented here by gray spheres, consists of seven atoms, each of zero charge and radius 10 Å, arranged to match the minor groove of DNA at a separation distance of 7.5 Å. The DNA sequence used in the simulations is d(ATATATATATATATAT)$_2$.

were reapplied at each timestep of the simulation. Interactions between DNA atoms were represented using the CHARMM22 force field [7] with the addition of weak harmonic restraints (10 kcal mol^{-1} Å^{-1}) to maintain base-pairing during the simulations. Solvation forces modeling the effects of transferring the system from the gas phase to a 150 mM 1–1 salt solution were obtained by solving the PB equation with UHBD on a 75^3 grid of spacing 1 Å. To be consistent with the use of an MM force field for the bonded and nonbonded interactions, a solute dielectric of 1.0 was used in the PB calculations. The response of the DNA to the approach of the protein was modeled with the technique of stochastic dynamics at 300 K using a low friction coefficient of 6.5 ps^{-1} [5] to allow a fast structural response of the system. The protein was moved closer to the DNA in 2.5 Å steps (to a final separation of 7.5 Å), with 1 ps of dynamics being performed at each protein–DNA distance using a timestep of 1 fs. PB solvation forces were updated every 0.1 ps: in the light of the earlier discussion concerning the interplay between force accuracy and update frequency, this should be sufficient to ensure a reasonable averaging of the solvation forces.

The overall protein–DNA association reaction is therefore completed in 10 ps. Of course, this is several orders of magnitude faster than the real process, so the simulations are not intended to provide a realistic model for the dynamic process of protein–DNA association. Instead, the purpose of the present simulations is to investigate only the likely structural consequences of protein approach. In other words, we stress here the use of the combined PB/SD method as a means of simulating average structural responses rather than as a method for the simulation of true macromolecular dynamics. To ensure that the structural changes that occur during the simulation actually result from the approach of the protein, and not from some other artifact of the system such as a gross imbalance in the potential functions or the use of a coarse grid spacing, a control simulation of the DNA alone, from which the protein was omitted, was performed in an exactly analogous fashion.

Perhaps not surprisingly, the approach of the low dielectric protein does indeed cause significant structural changes in the DNA. This is best illustrated by comparing the protein–DNA complex obtained at the end of the simulations with what would have been obtained had the DNA structure been held fixed (Fig. 9). It can be seen that with the DNA held fixed, a very considerable degree of structural overlap occurs at the position of closest approach of the protein to the DNA. In particular, it will be noticed that a number of the DNA phosphate groups are completely hidden. In contrast, the structure obtained at the end of the PB/SD simulation shows the DNA phosphates no longer embedded within the protein but, instead, fully exposed to the high dielectric solvent. This structural change is effected (or accompanied) by a dramatic opening-up of the minor groove, an effect which is not observed in the control simulation of the DNA alone: the central interstrand phosphate–phosphate distances in the protein–DNA system are on average 5.1 Å longer than those in the control simulation.

An energetic interpretation can be placed on these structural changes by examining the solvation energy of the system as a function of the protein–DNA distance (Fig. 10). When the DNA is held fixed, the approach of the protein, and the accompanying

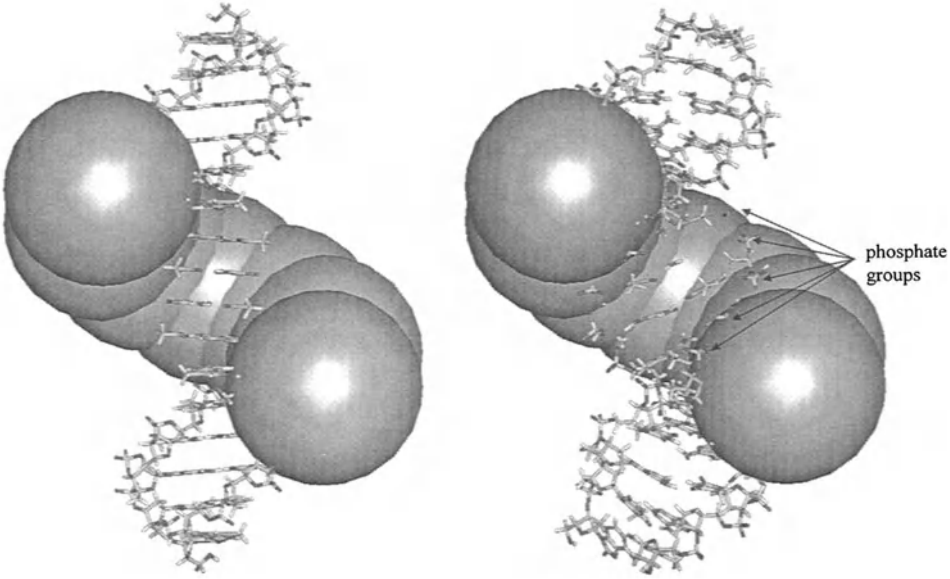


Fig. 9. Structure of the protein–DNA system at the end of the simulation (right) compared with what would have been obtained if the DNA had been held rigid (left).

exclusion of high dielectric solvent, causes a large unfavorable change in the solvation of the DNA. When the DNA is allowed to move in the PB/SD simulations, its structural response is such as to limit this loss of solvation, primarily by pushing the phosphate groups out towards the regions of nearby high dielectric. It is important to emphasize again that, in these simulations, no direct forces acted between the protein and the DNA; groove opening does not result, for example, from steric clashes with the protein since no van der Waals interactions operate between the protein and the DNA. The structural changes observed therefore result *only* from changes in the solvent and ionic environment of the DNA and, in particular, the decreased solvent shielding of phosphate repulsions.

The simulations suggest, therefore, that large structural changes in DNA can be induced without introducing new charge–charge interactions between the protein and the DNA, by modifying the charge–charge interactions already present within the DNA. Real protein–DNA systems are of course much more complicated than the simple model system investigated here, and other effects, such as hydrophobic and charge–charge interactions between the protein and the DNA, will undoubtedly be important in determining the overall protein–DNA structure [18,21]. As such, it will probably be difficult to provide unambiguous experimental support for the presence or contribution of dielectric effects. Despite this, the probable importance of such effects suggested by the present results is underlined by the fact that the DNA-binding

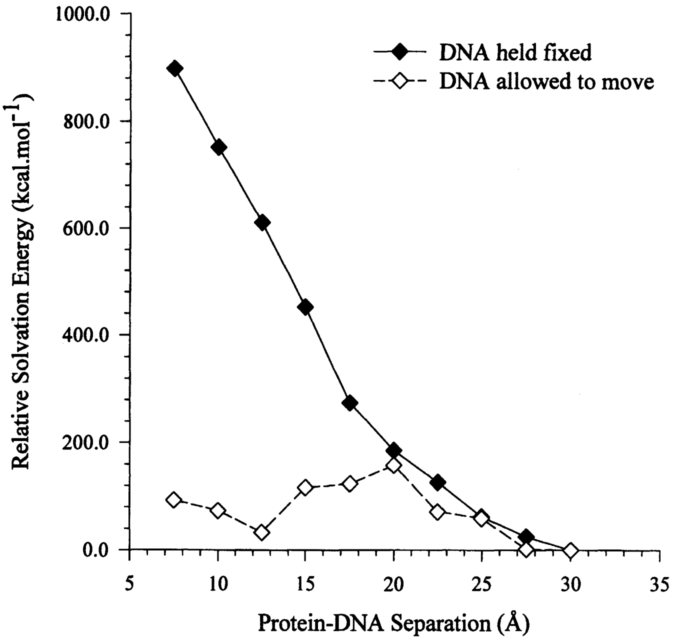


Fig. 10. Relative electrostatic solvation energy of the protein–DNA system as a function of the protein–DNA separation.

domains of the proteins inducing the most dramatic changes in DNA structure are extremely hydrophobic [18].

Remaining problems and future prospects

The above results show that the PB/SD method can be usefully applied to approach interesting problems which would otherwise be difficult or impossible to study. The next important step will be to demonstrate that the method is capable of producing stable simulations of macromolecules over much longer periods of time. As discussed earlier, such simulations should provide the most unambiguous test of the method's utility, an aspect which may ultimately rely on the development of a suitably balanced and parameterized force field. In terms of systems for study, DNA undoubtedly represents one of the most challenging owing to the magnitude of the Coulombic and solvation forces present, and as such there is good reason to believe that a method capable of allowing stable simulations of DNA will also be of more general use. It should be pointed out that just this situation has recently been reached for explicit solvent simulations of DNA: simulations extending to 2 ns have been reported, indicating a high level of stability in the structures [22,23]. Clearly, to have full confidence in the method, it will be necessary for PB/SD simulations to demonstrate a similar degree of stability, although it should be remembered that the timescale of conformational changes in a PB/SD simulation can be made much shorter than that obtained in conventional MD simulations. Because of this latter aspect, it is worth pointing out that comparisons of the relative computational efficiency of continuum and explicit methods based solely on CPU time per picosecond of simulation are not likely to be useful.

In addition to parameterization issues, other aspects of the method will also require further investigation. The optimization of factors such as solvation force update frequency, solvent friction constants, etc. will be important, though undoubtedly, to an extent, system dependent. Such aspects are currently under study for small-molecule systems [24]. In identifying areas for future study, it is worth emphasizing that our discussion has focused exclusively on electrostatic aspects, omitting any mention of other effects such as hydrophobic forces which are clearly of major importance for macromolecular stability. Hydrophobic energies are of course commonly expressed as being proportional to the solvent-accessible surface area, and the recent development of computationally efficient methods for the calculation of surface-area-dependent forces [25] should facilitate their inclusion in more physically complete continuum-based dynamics simulations. Ultimately, for a combined PB/SD method to be of general use for macromolecular simulations, it will be necessary to address all of the above aspects. With further developmental work, however, PB/SD should become a valuable simulation method, providing information complementary to that obtainable using explicit solvent techniques.

Acknowledgements

This work was supported in part by grants from the National Institutes of Health, the National Science Foundation and the National Science Foundation MetaCenter Program. A.H.E. is a Wellcome Trust International Prize Traveling Research Fellow supported under Grant 041157/Z/94/Z. M.J.P. is a Howard Hughes Medical Institute Predoctoral Fellow.

References

1. Sharp, K.A., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 147–160.
2. Honig, B. and Nicholls, A., Science, 268(1995)1144.
3. Gilson, M.K., Davis, M.E., Luty, B.A. and McCammon, J.A., J. Phys. Chem., 97(1993)3591.
4. Brooks III, C.L., Karplus, M. and Pettitt, B.M., Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics, Advances in Chemical Physics, Vol. LXXI, Wiley, New York, NY, 1988.
5. Gilson, M.K., McCammon, J.A. and Madura, J.D., J. Comput. Chem., 16(1995)1081.

6. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A., J. Am. Chem. Soc., 117(1995)5179.
7. Mackerell Jr., A.D., Wiórkiewicz-Kuczera, J. and Karplus, M., J. Am. Chem. Soc., 117(1995)11946.
8. Sitkoff, D., Sharp, K.A. and Honig, B., J. Phys. Chem., 98(1994)1978.
9. Sitkoff, D., Ben-Tal, N. and Honig, B., J. Phys. Chem., 100(1996)2744.
10. Madura, J.D., Briggs, J.M., Wade, R.C., Davis, M.E., Luty, B.A., Ilin, A., Antosiewicz, J., Gilson, M.K., Bagheri, B., Scott, L.R. and McCammon, J.A., Comput. Phys. Commun., 91(1995)57.
11. Born, M., Z. Phys., 1(1920)45.
12. Niedermeier, C. and Schulten, K., Mol. Sim., 8(1992)361.
13. Cullis, P.M. and Wolfenden, R., Biochemistry, 20(1981)3024.
14. Gilson, M.K., Sharp, K.A. and Honig, B., J. Comput. Chem., 9(1988)327.
15. Kim, Y., Geiger, J.H., Hahn, S. and Sigler, P.B., Nature, 365(1993)512.
16. Kim, J.L., Nikolov, D.B. and Burley, S.K., Nature, 365(1993)520.
17. Werner, M.H., Huth, J.R., Gronenborn, A.M. and Clore, G.M., Cell, 81(1995)705.
18. Travers, A.A., Nature Struct. Biol., 2(1995)615.
19. Elcock, A.H. and McCammon, J.A., J. Am. Chem. Soc., 118(1996)3787.
20. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
21. Werner, M.H., Gronenborn, A.M. and Clore, G.M., Science, 271(1996)778.
22. York, D.M., Yang, W.T., Lee, H., Darden, T. and Pedersen, L.G., J. Am. Chem. Soc., 117(1995)5001.
23. Cheatham, T.E., Miller, J.L., Fox, T., Darden, T.A. and Kollman, P.A., J. Am. Chem. Soc., 117(1995)4193.
24. Smart, J.L., Marrone, T.J., Gilson, M.K. and McCammon, J.A., in preparation.
25. Sridharan, S., Nicholls, A. and Sharp, K.A., J. Comput. Chem., 16(1995)1038.

Part III
Structure refinement

Time-averaging crystallographic refinement

Celia A. Schiffer
Genentech Inc., 460 Point San Bruno Boulevard, South San Francisco,
CA 94080, U.S.A.

Introduction

New techniques in crystallography, the use of synchrotron radiation and freezing techniques, have led to a quickly expanding number of protein structures determined at high resolution. High-resolution data provide for a more detailed description of the protein, allowing analysis of the relative flexibility of different regions of the molecule and of the details of water structure.

The conformations of a protein allowed within the crystal are conventionally modeled by isotropic temperature factors and sometimes by multiple conformations of a particular side chain. A better representation of the atomic mobility is obtained by application of the method of time-averaging refinement. This method fits the data with an ensemble of structures, generated by molecular dynamics, rather than a single structure [1–8]. This method was initially developed for the refinement of NMR structures to properly account for conflicting data that cannot simultaneously be satisfied by one structure [9,10]. The method of time-averaging refinement was first applied to protein crystallography in the refinement of phospholipase A_2 [1]. Given sufficient data, this method gives a better representation of the conformational variability of a biological molecule than can be obtained by the use of either isotropic or anisotropic temperature factors [3]. The force field used in the simulation restricts the structures of the ensemble to be of low energy and physically reasonable; thus, conformational variability is represented by multiple sterically reasonable configurations.

Refinement procedure

The term restraining the refinement to the experimental structure factors in time averaging is V^{sf}_{restr}, which is added to the physical potential energy function, V_{phys}. This potential function restrains the system to the X-ray data (Eq. 1) just as in traditional molecular dynamics or simulated annealing refinement.

$$V = V_{phys} + V^{sf}_{restr} \tag{1}$$

The difference from traditional refinement is that V^{sf}_{restr} is a function of an ensemble of structures rather than a single structure (Eq. 2).

$$V^{sf}_{restr} = \tfrac{1}{2}k^{sf} \textstyle\sum_s [\,|F_{obs}(\vec{s})| - k_{sc}\,|<F_{calc}(\vec{s})>|\,]^2 \tag{2}$$

and

$$< F_{calc}(\vec{s}) >_{\tau_x,t} = [\tau_x(1 - e^{-t/\tau_x})]^{-1} \int_0^t e^{-(t-t')/\tau_x} F_{calc}(\vec{s}; \vec{r}(t'))\,dt' \tag{3}$$

There are two force constants in the experimental restraining potential, V^{sf}_{restr}. k_{sc} scales the calculated to the observed structure factors and k^{sf} scales the restraining potential to the physical potential. The force constant k^{sf} must be chosen carefully since the time-averaging potential is dependent on previous configurations of the molecule and does not conserve energy. The optimal force constant is one that still refines the ensemble to the X-ray data, but does not heat the system more than 15% over the bath temperature. Empirical tests of such a force constant scale the X-ray term to be approximately 5% of the magnitude of the rest of the force [4].

τ_x is the relaxation time over which an ensemble of structures is accumulated and averaged. The relaxation time is related to temperature factors, since both parameters affect the spatial distribution of the electron density due to a particular atom. The temperature factor gives an instantaneous contribution, whereas the averaging contributes as many atom positions as are taken into account within time τ_x. Temperature factors should not be independently refined during a time-averaging refinement simulation as this would be redundant; rather a small constant temperature factor (such as 2 Å^2) should be assigned to all atoms. However, choosing a reasonable relaxation period is critical to the success of the refinement. τ_x needs to be sufficiently long to allow mobile parts of the molecule to cover configurational space adequately. Too short a τ_x will result in the ensemble of structures sampling a region of space that is between the observed conformations and a poorer fit to the data. In practice, the constraints of computer time mean that fast relaxing disorder can be properly accounted for, but slow relaxing disorder will not be sampled. Thus in setting up a refinement simulation, the longest practical relaxation time should be chosen.

The exponential decay function in Eq. 3 is built into the standard formula for averaging the structure factor over a time period t, during the course of the refinement. The simulation time should be at least 10-fold longer than the relaxation time to reduce model bias toward the starting configuration. The force calculated from Eq. 2 depends on the rate of change of the time-averaged structure factor which is dependent on both the length of the simulation, t, and the relaxation time, τ_x. This allows for the calculation of a running R-value:

$$R(\text{running}) = \frac{\sum |F_{obs} - < F_{calc} >_{\tau_x,t}|}{\sum |F_{obs}|} \tag{4}$$

which can be monitored over the course of the simulation.

A limitation of time-averaging refinement is that the ratio of observations to unknowns must be sufficiently high to describe multiple conformations. Protein molecules have restricted geometries, which are determined by the peptide bonds and the sterically allowed conformations of the side chains. Thus, individual atoms are not free to move independently of each other. This restricted geometry reduces the number of independent parameters that must be fit by the data in determining a structure. In time-averaging refinement, the structures of the ensemble refined to the data are also not independent of each other; rather they are related both by the restrictions of chemistry and by their relationship to each other in time. To signal overfitting of the data, a free R-value [3,11] can be used during the refinement simulation. The refinement is proceeding well if the free R-value stays coupled with the refined R-value during the simulation. However, if the free R-value diverges from the refined R-value and starts to ascend, this is a clear indication that the refinement simulation is not set up appropriately.

Expanding the asymmetric unit to a full unit cell provides additional searching capabilities in time averaging [4,7]. This can enhance the sampling statistics over conformational space, since each asymmetric unit can fit the data slightly differently. Refining the full unit cell can also provide independent verification of how frequently a particular event is likely to occur.

A time-averaging refinement simulation provides a unique ability to examine water structure and the relative accessibility of water sites around a protein structure. To accomplish this, the simulation box, either the asymmetric unit or the unit cell must be filled with water. This full hydration of the unit cell also allows the use of a force field with full charges, and thus should better mimic the electrostatic environment that exists within a crystal. However, use of the correct protonation state to mimic the pH within the crystals is essential, since incorrect electrostatics may disrupt the crystal lattice.

Analysis of results

Analyzing and usefully interpreting the results of a time-average refinement is not trivial. Structure factors from time-averaging crystallographic refinement are calculated from the ensemble of structures in the simulation. At the end of the simulations, the structure factors are averaged over the analysis period to provide an average calculated set of structure factors for the refinement defined as

$$< F_{calc}(\vec{s}) >_t = t^{-1} \int_0^t F_{calc}(\vec{s}, \vec{r}(t'))\, dt' \qquad (5)$$

and this leads to the calculation of a final R-value:

$$R = \frac{\sum | F_{obs} - < F_{calc} >_{\infty,t} |}{\sum | F_{obs} |} \qquad (6)$$

All the molecular motions that occurred in the refinement simulation over the analysis period are averaged into these structure factors. Electron density maps generated from these calculated structure factors and average calculated phases provide a view of the predominant motions sampled in the calculation. Difference maps assess the fit to the observed data. For example, if the protonation state of the molecule is incorrect, the electrostatics for the simulation box will be incorrect, forcing many side chains into the wrong conformation and producing many difference peaks in the electron density maps.

Atomic coordinates are saved frequently during the simulation which allows for a detailed analysis of the ensemble of structures. From the atomic trajectories, a temperature factor for the protein atoms can be calculated,

$$B = 8/3\,\pi^2 \, < d^2 > \tag{7}$$

where $< d^2 >$ is the mean-square positional fluctuation. This is a useful quantity to assess the types of motion that occur and allows for comparisons with the temperature factors belonging to the conventionally (single-structure) refined crystal structure. This is especially pertinent for α-carbon atoms whose positions tend to be less variable. Dihedral angles for the side chains as a function of time can also be extracted from the trajectories.

The water structure has to be assessed somewhat differently since, during the course of the refinement simulation, the water molecules are free to move throughout the simulation box. The relative accessibility of water sites around the protein molecule is more interesting than the trajectory of any particular water molecule. These sites can be mapped by recording the positions of the water molecules at every timestep on a fine three-dimensional grid [12]. The contribution of each water molecule can be spread over grid points by a Gaussian and normalized so that the sum over all grid points of contributions of one water molecule equals one. Over the analysis period, grid points that are frequently occupied by water molecules can be defined as a water site. These water sites can then be compared for correspondence with peaks in the electron density map.

Once these water sites are determined, they can be characterized in terms of the number of water molecules visiting a site, the percentage of time a site was occupied, and the shape of the visiting distribution. A water site temperature factor can be calculated using positions, $\vec{r}_i$, of visiting water molecules around the site position, $\vec{r}_{site}$:

$$B \;=\; 8/3^2\,\pi^2 < (\vec{r}_i - \vec{r}_{site})^2 > \;= 8/3\,\pi^2 < \Delta\vec{r}^2 > \tag{8}$$

The anisotropy of a water site can be defined as

$$A \;=\; 2 < \Delta r_i^2 > /(< \Delta r_j^2 > \;+\; < \Delta r_k^2 >) - 1.0 \tag{9}$$

where $< \Delta r_i^2 >$, $< \Delta r_j^2 >$, $< \Delta r_k^2 >$ are the mean-squared displacements from the water site in three orthogonal directions, and where $< \Delta r_i^2 >$ is the largest in magnitude of the three displacements. Finally, if an entire unit cell was refined in the simulation, the characteristics of symmetry-related water sites from the trajectory analysis can be compared for reliability.

Conclusions

Time averaging provides a useful method for the detailed refinement of high-resolution crystal structures. Analysis of the conformations of the protein's side chains and loops characterizes the attainable flexibility of the molecule within the limits of the experimental data. This description of the protein is more accurate than either isotropic or anisotropic temperature factors. In addition, the ensemble of protein and water configurations can be analyzed to determine the relative accessibility of ordered water around a protein molecule, a determination which is not possible by any other experimental analysis. As computers become faster and high-resolution data become available for more proteins, time-averaging refinement will become a practical method for further characterization of a protein's structure and flexibility.

Acknowledgements

The author would like to thank A.A. Kossiakoff and W.F. van Gunsteren for their support.

References

1. Gros, P., van Gunsteren, W.F. and Hol, W.G.J., Science, 249(1990)1149.
2. Gros, P. and van Gunsteren, W.F., Mol. Sim., 10(1993)377.
3. Schiffer, C.A., Gros, P. and van Gunsteren, W.F., Acta Crystallogr., Sect. D, 51(1995)85.
4. Schiffer, C.A., Kossiakoff, A.A. and van Gunsteren, W.F., in preparation.
5. Clarage, J.B. and Phillips, G.N., Acta Crystallogr., Sect. D, 50(1994)24.
6. Clarage, J.B., Romo, T., Andrews, B.K. and Pettitt, B.M., Proc. Natl. Acad. Sci. USA, 92(1995)3288.
7. Burling, F.T. and Brünger, A.T., Isr. J. Chem., 34(1994)165.
8. Burling, F.T., Weis, W.I., Flaherty, K.M. and Brünger, A.T., Science, 271(1996)72.
9. Torda, A.E., Scheek, R.M. and van Gunsteren, W.F., Chem. Phys. Lett., 157(1989)289.
10. Torda, A.E., Scheek, R.M. and van Gunsteren, W.F., J. Mol. Biol., 214(1990)223.
11. Brünger, A.T., Nature, 355(1992)472.
12. Lounnas, V. and Pettitt, B.M., Proteins, 18(1994)133.

Incorporation of solvation energy contributions for energy refinement and folding of proteins

Werner Braun
Sealy Center for Structural Biology, University of Texas Medical Branch at Galveston, Galveston, TX 77555-1157, U.S.A.

Introduction

Protein structures determined from X-ray or NMR data are generally refined by including empirical energy terms for intramolecular interactions [1–3]. A good stereochemical quality of an experimentally determined protein structure is a necessary requirement for a high-resolution structure [4–7]. Several force fields for intramolecular interactions in proteins are nowadays in widespread use [8–14]. However, a low intramolecular energy of an experimentally determined structure does not prove that this structure is correct, as the analysis of incorrectly determined experimental and deliberately misfolded protein structures shows [15,16]. It is necessary to include the protein–solvent interaction in the refinement process.

Various models for treating the protein–solvent interaction have been suggested and their strengths and limitations were critically evaluated in several recent reviews [17–19]. Protein–solvent interaction can be computed with either explicit moving solvent particles [20–24], or in continuum models based on the Poisson– Boltzmann equation [25,26] or on the solvent accessible surface area [27–33]. Treating protein–solvent interactions in a continuum approximation is an order-of-magnitude faster than computations with explicit molecules. This efficiency makes this approach attractive for refinement calculations and studies of protein docking and folding.

The first methods calculated the accessible surface area of an atom by numerical integration of the accessible arc lengths of cross sections. These numerical methods, which have been optimized over the years, are robust and easy to implement. However, analytical expressions of the areas and their gradients with respect to the atomic coordinates are needed for use in energy refinement programs. An analytical approximation of the surface area was proposed for that reason [34]. Connolly [35–37] first presented an exact analytical solution of the integration of the accessible surface area. Richmond [38] suggested a different mathematical representation of intersecting circles.

Several improvements of the continuum approach were necessary to allow its application to the refinement of protein structures [39–42]. Recently, we showed that these calculations can be further simplified by new and computationally more efficient equations, which calculate the derivative exactly [43,44]. Several computational aspects of this approach are still being developed, with regard to handling singularities or introducing approximations for higher efficiency [45–49].

An important aspect of the surface area approach is the choice of the atomic solvation parameter sets. Different atomic solvation parameters have been tested for their ability to discriminate between native and alternative folds and their merits in the energy refinement of protein structures derived from NMR data [50–56].

The next sections describe basic mathematical features for calculating the accessible surface areas and their derivatives analytically. The strengths and limitations of each refinement method are evaluated and compared to the use of explicit water molecules in unconstrained and constrained molecular dynamics calculation.

Calculation of the solvent accessible surface area

Chothia [29] correlated the gain in free energy for the transfer of residues from the surface of a protein to the interior with the loss in accessible surface area, and estimated a free-energy gain of 24 cal/mol Å^2 from the observed linear correlation. Considering the different polarity of atoms or atom groups on the surface of the protein and using more precisely measured transfer energies of amino acid residue analogues [33], Eisenberg and McLachlan [32] calculated the free energy of interaction of the protein with water, E_{hyd}, by

$$E_{hyd} = \sum_{i=1,n} \sigma_i A_i \tag{1}$$

where A_i is the solvent accessible surface area of atom i and σ_i is a 'solvation parameter' depending on the atom type. The atoms are treated as spheres where the radii of the spheres are the van der Waals radii enlarged by 1.4 Å, to represent the water molecule rolling over the van der Waals surface of the protein [27]. An analytical expression for the A_i and their derivatives is needed to include the protein–solvent interaction E_{hyd} in an empirical energy function for energy minimization or molecular dynamics calculations.

Following the general procedure of Connolly [35,36] and Richmond [38], we have further developed the basic equations for the analytical calculation of the accessible surface area and its derivatives [43,44]. Figure 1 shows a typical arrangement of spheres occluding parts of sphere i. The solvent accessible surface A_i of sphere i is enclosed by p intersecting accessible arcs. The global Gauss–Bonnet theorem for the case of intersecting spheres leads to an analytical expression for A_i which can be calculated from the arc lengths and tangential vectors at the end points of the accessible parts of the intersecting circles:

$$A_i = r_i^2 \left[2\pi + \sum_{\lambda=1,p} \Omega_{\lambda,\lambda+1} + \sum_{\lambda=1,p} \cos\Theta\,\Phi \right] \tag{2}$$

The polar angle Θ is the opening angle of the circle of intersection at the center of sphere i (Fig. 2). The distance from the center of sphere i to the center of the

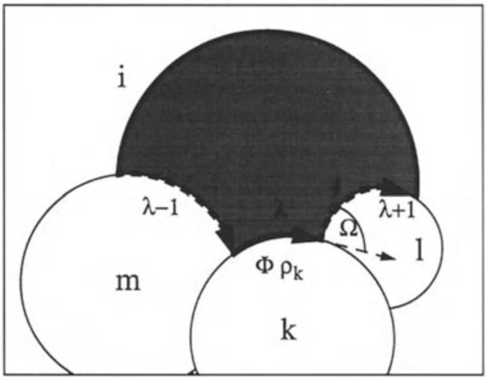


Fig. 1. Accessible surface area of a sphere i (dark area), intersecting with three other spheres k, l and m. Two spheres i and k form a circle of intersection, which is in general partially solvent exposed. The solvent-exposed part is the arc λ (bold line). The end points of λ are defined by cutting two other circles of intersections λ − 1 and λ + 1 (dashed lines) with spheres m and l. The length of the arc λ is given by Φ ρ_k where Φ is the angle at the center of the circle corresponding to the solvent accessible part and ρ_k is the radius of the circle of intersection. The angle between the tangential vectors of successive arcs is given by Ω.

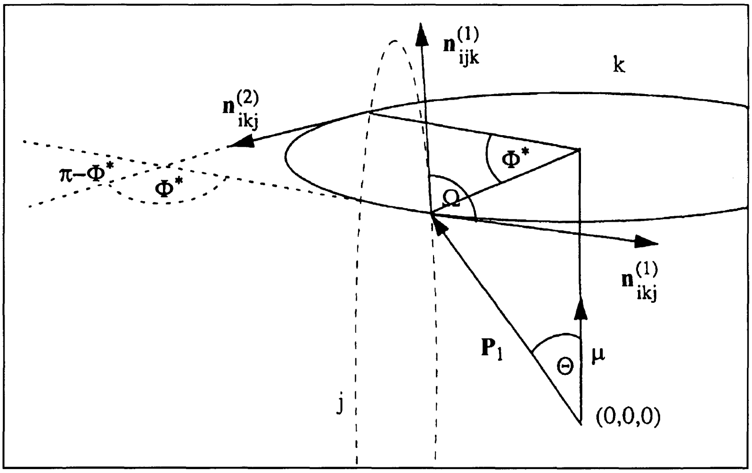


Fig. 2. Definition of the polar angle Θ, the angle Φ corresponding to the accessible arc, and the angle Ω between tangential vectors $\boldsymbol{n}^{(p)}_{ijk}$. The complementary angle Φ = 2π − Φ corresponding to the buried arc is shown. The three indices ijk label the three spheres which intersect in the considered point. The first two indices, i and j, give the intersection circle to which the vector is tangential. The number of the intersection point p (1 or 2) is written in superscript. The intersection points can be classified as 'entry' or 'exit' points. These are the points where one enters or leaves the buried arc when moving on the oriented intersection circle. For example, $\boldsymbol{P}_l$ is an exit point of the intersection circle k.*

intersection circle k is α. The polar angle Θ can be calculated from α and the radius of the sphere r_i by

$$\cos\Theta = \frac{\alpha}{r_i} \tag{3}$$

The angles Φ and Ω can be calculated from the tangential vectors $\mathbf{n}_{ijk}^{(p)}$. As the angle of the accessible part of the intersection circle Φ can be larger than π, special consideration has to be taken for the two cases [43]. For $\Phi < \pi$ the angle is given by

$$\Phi = \arccos\left(\mathbf{n}_{ikj}^{(1)} \cdot \mathbf{n}_{ikj}^{(2)}\right) \tag{4}$$

For the angle Ω no special case needs to be treated separately, as Ω cannot be larger than π:

$$\Omega = \arccos\left(\mathbf{n}_{ikj}^{(1)} \cdot \mathbf{n}_{ijk}^{(1)}\right) \tag{5}$$

The tangential vectors themselves can be calculated from the coordinates of the intersection points $\mathbf{P}_1$ and $\mathbf{P}_2$. The equations for the three intersecting spheres lead to

$$(\mathbf{x}_j - \mathbf{P})^2 = r_j^2 \tag{6}$$

$$(\mathbf{x}_k - \mathbf{P})^2 = r_k^2 \tag{7}$$

$$\mathbf{P}^2 = r_i^2 \tag{8}$$

A convenient decomposition for the intersection points $\mathbf{P}_1$ and $\mathbf{P}_2$ is in three orthonormal vectors $\boldsymbol{\mu}$, $\mathbf{v}$ and $\boldsymbol{\omega}$ (Fig. 3):

$$\mathbf{P}_{1,2} = \alpha\boldsymbol{\mu} + \beta\mathbf{v} + \gamma_{1,2}\boldsymbol{\omega} \tag{9}$$

All individual components of $\mathbf{P}_{1,2}$ can be calculated from Eqs. 6–8 (see Ref. 43):

$$\alpha = \frac{d_k^2 + r_i^2 - r_k^2}{2d_k} \tag{10}$$

$$\beta = \frac{1}{\sin\varphi}(g_j - g_k \cos\varphi) \tag{11}$$

$$\gamma_{1,2} = \pm\sqrt{r_i^2 - \alpha^2 - \beta^2} \tag{12}$$

$$\boldsymbol{\mu} = \frac{\mathbf{x}_k}{d_k} \tag{13}$$

$$\mathbf{v} = \frac{1}{\sin\varphi}\left(\frac{\mathbf{x}_j}{d_j} - \cos\varphi\boldsymbol{\mu}\right) \tag{14}$$

$$\boldsymbol{\omega} = \boldsymbol{\mu} \wedge \mathbf{v} \tag{15}$$

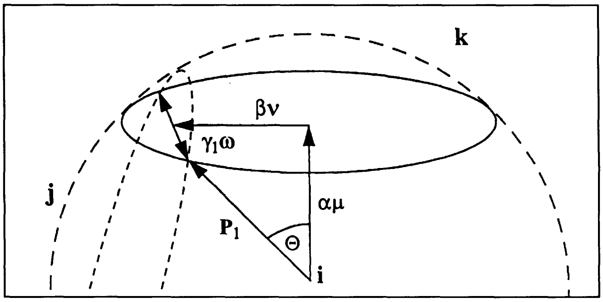


Fig. 3. The sphere i is supposed to be in the origin of the coordinate system. The center of any other sphere k is then determined by the vector $\mathbf{x}_k = (x^1, x^2, x^3)$ with $|\mathbf{x}_k| = d_k$. Three spheres i, k and j intersect in the two points $\mathbf{P}_1$ and $\mathbf{P}_2$. The intersecting points can be decomposed in three orthonormal vectors, $\boldsymbol{\mu}$, $\mathbf{v}$ and $\boldsymbol{\omega}$. The unit vector $\boldsymbol{\mu}$ points from the center of sphere i to the center of sphere k, and $\mathbf{v}$ is orthogonal to $\boldsymbol{\mu}$, pointing towards the center of j.

Calculations of the derivatives

The derivatives of the scalars α, β, γ and the unit vectors **μ**, **ν**, **ω** can be directly calculated from Eqs. 10–15. With the equations for the intersection points and their derivatives, we can now calculate the values cos Θ, Φ and Ω in the Gauss–Bonnet formula. For the calculation of the gradient of the solvation energy, Eq. 1, we need the derivative of every solvent accessible surface A_i with respect to all atom coordinates. The matrix $\partial A_i/\partial x_k$ is, however, sparse, as only those derivatives are different from zero where the sphere k intersects sphere i. From the Gauss–Bonnet theorem we have

$$\frac{\partial A_i}{\partial x_k} = r^2\left[\sum_{\lambda=1,p}\frac{\partial\Omega_{\lambda,\lambda+1}}{\partial x_k} + \sum_{\lambda=1,p}\frac{\partial\cos\Theta}{\partial x_k}\Phi + \sum_{\lambda=1,p}\cos\Theta\frac{\partial\Phi}{\partial x_k}\right] \tag{16}$$

All dependencies of the angles Ω, Θ and Φ on the atomic coordinates of sphere k have to be considered in Eq. 16. Using the notation

$$\partial = \frac{\partial}{\partial x_k^m} \tag{17}$$

the derivatives of the individual terms in Eq. 16 can be calculated as

$$\partial\Omega = \frac{1}{\sin\Omega}\left(\partial\mathbf{n}_{ikj}^{(1)}\cdot\mathbf{n}_{ijk}^{(1)} + \mathbf{n}_{ikj}^{(1)}\cdot\partial\mathbf{n}_{ijk}^{(1)}\right) \tag{18}$$

$$\partial\cos\Theta = \frac{\partial\alpha}{r_i} \tag{19}$$

$$\partial\Phi = \partial\arccos\left(\mathbf{n}_{ikj}^{(1)}\cdot\mathbf{n}_{ikj}^{(2)}\right) = \frac{1}{\sin\Phi}\left[\partial\mathbf{n}_{ikj}^{(1)}\cdot\mathbf{n}_{ikj}^{(2)} + \mathbf{n}_{ikj}^{(1)}\cdot\partial\mathbf{n}_{ikj}^{(2)}\right] \tag{20}$$

Practical implementations

The program MSEED [39] calculates surface areas and derivatives by first locating only accessible vertices P connected by edges. Vertices are common points of three intersecting spheres, connected by intersecting circles (edges). The time-saving procedure in MSEED is based on a fast method for searching for vertices only through following accessible edges. The CPU time required to calculate the solvation energy E_{hyd} is not the time-limiting factor, as the method scales with a factor between $n^{2/3}$ and n. However, buried surfaces are not found by this method, which might be a limitation in certain applications.

In the program FANTOM [42,43] two internal lists are used to speed up the calculation: a near-neighbor list for intersecting atoms and a list of atoms on the surface. The number of atoms in this intersection list is only dependent on the average packing density of the protein and not on the overall size of the protein. Thus, the major part of the CPU time needed for the calculations increases linearly with the size of the protein. The intersection points of the atoms i, k and j are calculated from Eqs. 9–15. A further aspect for practical implementation is the handling of singularities, where for example two spheres are contacting each other, or three atoms are collinear. These critical cases can be handled through careful implementation [42,43,45].

A great variety of techniques have been described to improve on the efficiency of surface calculations. These methods use new hardware architecture such as parallel computers, implement more sophisticated search methods or simplify the earlier approaches to reduce CPU time while assuring that the calculated values of the surface area or derivatives are still good approximations of the values.

Implementation on parallel computers

The method of Lee and Richards [27] has been implemented on an Intel parallel computer with 64 processors [57]. The accessible surface area of each sphere is calculated by sectioning the space occupied by the protein in a set of parallel planes (z-sections) with a certain distance ΔZ between neighboring planes. The lengths of accessible arcs, L_a, are determined for each sphere i on a given plane. The accessible surface area is then computed by numerical integration along the z-axis:

$$A_i = \sum_a \frac{r_i}{\sqrt{r_i - z_a}} DL_a \qquad (21)$$

The calculation of the surface area of each atom can be performed independently from the calculation for other atoms. Partitioning and mapping of the atoms to the available processor has been arranged to balance the work load. The overall efficiency of the parallel computation was reported to be 67% for 64 processors and almost 90% for 16 processors.

The analytical calculation of the surface areas and their gradients was also ported to a parallel machine, an Intel Paragon [43]. The computation of the individual areas A_i by Eq. 2 and their gradients by Eq. 16 can be independently processed if the coordinates of all atoms and their radii have been broadcast to all processors. One master processor broadcasts the data, synchronizes the parallel computation, and collects the result. The remaining $(n - 1)$ slave processors calculate the solvation energy and its gradient for approximately $M/(n - 1)$ atoms for a protein of M atoms. An almost optimal load balance was achieved by assigning every $(n - 1)$th atom to the same slave processor. In that particular study, only the solvation energy was calculated in parallel as it is the most time-consuming part. The fraction of time to calculate the protein–solvent interaction dropped to about 10% of the total time, using 20 processors in the parallel version of FANTOM.

Improvement of the Shrake and Rupley method

The Shrake and Rupley method (SR) [28] generates a certain number N_{tot} of test points which are uniformly distributed on the surface of the sphere. Then it counts how many of these points are not buried by other atoms (N_{acc}). The accessible surface area of the sphere can then be approximated from the ratio N_{acc}/N_{tot}:

$$A_i = 4\pi r_i^2 \frac{N_{acc}}{N_{tot}} \tag{22}$$

The method can be improved by replacing the test points on the surface of the sphere with spherical triangles covering the whole sphere. In the GEPOL procedure [58], 60 spherical triangles cover the spheres, and the tesselation can be improved further by a finer granularity at the boundaries. Triangles are checked to determine if they are buried by other spheres, and the accessible surface area is estimated by summing the surface of all exposed triangles. Numerical tests proved that the convergence of the surface areas towards the correct values is faster with this method than when using uniformly distributed test points.

Another variant of the SR method places the molecule in a cubic grid [59]. All grid points within the molecule are assigned a value of 1, while those outside have zero value. A test point is not occluded by neighboring atoms if its grid point has value zero. This test can be performed without calculating distances to neighboring atoms. The time saving is similar to the time saving using a near-neighbor list, but because most of the operations can be done with logical bit operations, the method is reported to be a factor of 8 faster than the SR method [59].

Another method [60] restricts the test for the occlusion of test points by neighboring atoms only to certain test points. This is achieved by appropriately distributing and numbering the test points on the atoms (Z-ordering). A substantial reduction of the number of checks and a factor of 3 in speed-up were reported.

More sophisticated lattice methods reduce the number of checks by using precalculated libraries [61], or combine fast methods for creating a list of neighboring atoms by cubic grids, as used routinely in distance geometry methods [62], with grid search methods for the surface points (double cubic lattice method) [47]. However, most of these recent methods have not yet been included in refinement procedures.

Probabilistic methods

Wodak and Janin [34] greatly simplified the accessible surface area problem by applying a statistical approach. The accessible surface area A_i of a sphere with radius r_i in the presence of a second intersecting sphere with radius r_k is a simple analytical function of the interatomic distance d between the centers of the two spheres. Analytical integration or applying Eq. 2 leads to

$$A_i = 2\pi r_i \left[r_i + \frac{d^2 + r_i^2 - r_k^2}{2d} \right] \tag{23}$$

The fraction of A_i to the total area of this sphere can be regarded as the probability for a point on the surface of sphere i being outside of an intersecting sphere k. Assuming that all intersecting spheres are randomly and independently distributed around the sphere i, the probability for a point on the surface of sphere i being outside of all intersecting spheres can be calculated from the product of the individual probabilities. It was shown that the total accessible surface areas of globular proteins can be approximated within 2–4% error by this method. The error for the accessibility of individual residues was about 20%. These low error rates might be low enough such that protein–solvent interactions can be reliably included in simplified energy potentials derived from data sets of known protein structures [63].

In this probabilistic approach, the first and second derivatives of the hydration energy can be calculated. A modification of this approach by calibrating the surface area in 270 small molecules empirically [64] has been included in the energy minimization program MacroModel [65].

Approximate methods for the derivatives

An approximation of the derivative of the accessible surface areas has been recently given by a geometrical argument [48]. The algorithm combines some aspects of the SR method to determine the fractional accessibility of circles of intersections. The method was shown to be comparable in speed to the MSEED approach, but as in all numerical methods the speed of the method depends on the desired level of accuracy. For hen egg-white lysozyme, an overall root-mean-square error of the gradient of 1–2% was achieved, compared to the gradient calculated by ANAREA. It remains to be seen how useful these approximated derivatives are in actual refinement calculations. As the absolute values of the gradients are decreased in optimization, small errors in gradients can prohibit or significantly slow down convergence.

Methods for calculating excluded volumes and their derivatives

Volume-dependent energy terms have also been proposed to calculate solvation energy in a continuum method [66]. A number of algorithms have been derived to calculate the excluded volume of a protein. Fast numerical methods [65,67,68] classify cubes of a grid lattice containing the protein to be exterior, surface or interior (white, gray or black). Critical cubes, mainly surface cubes, are reclassified on a grid of half mesh size. The approximate volume is given by the sum of the volumes of the interior and half of the volumes of the surface cubes. This procedure eliminates most of the interior atoms early in the calculation. By several iterations of dividing the mesh size in half, any required accuracy of the volume can be achieved.

Derivatives of volumes are needed for refinement methods. A fast analytical approach for calculating the first and second derivatives of the excluded volume as a function of atomic coordinates has been described [69]. The molecule is cut by parallel planes (z-sections), which are typically 0.1–0.2 Å apart, and the contribution for the first and second derivatives of each z-section is calculated and summed up to give the total contribution. For each accessible arc on a given z-section, the force acting on the 'central atom', i.e. the atom which is currently considered in the loop over all atoms of the proteins, is calculated by integration of

$$dF = P \cdot n\,dS \tag{24}$$

The analytical energy-minimized structures of hen egg-white lysozyme have been compared with the pressure-induced changes as determined by a crystal structure of 1000 atm. Volume changes are less than 1%, so structural changes have to be analyzed carefully. The first and second derivatives of excluded volumes as a function of atomic coordinates have been worked out, implemented in a truncated Newton optimization method. Pressure-induced changes were calculated and were found to be qualitatively in agreement with the experimentally observed values.

Comparison of different atomic solvation parameter sets

Several new studies derived atomic solvation parameters σ_i from transfer free energies of amino acid analogues [40,50–52], along similar lines as Eisenberg and McLachlan [32] derived it from transfer energies between n-octanol to water [33], but with enlarged data sets. Another approach uses the preference of atoms or atom groups to be on the surface in three-dimensional protein structures [63,70–72]. Schiffer et al. [41,55] derived a parameter set in direct relation with their molecular mechanics force field, AMBER [12]. They calculated the effect of surrounding water molecules with a distance-dependent dielectric constant in a simulation of bovine pancreatic trypsin inhibitor (BPTI) and determined the solvation parameters that most closely fit this simulation.

Empirical energy functions are used to test whether the native protein structure has the lowest energy within a set of alternative three-dimensional structures. A second criterion for the validity is the stability of the native structure in energy minimization, Monte Carlo or molecular dynamics calculations. A third criterion tests the ability to drive perturbed structures towards the native structure.

BPTI, a classic protein for testing new computational procedures, was used to evaluate atomic solvation parameter sets for their ability to differentiate native from near-native conformations [52]. A set of 39 conformations, generated with Monte Carlo methods in the vicinity of the X-ray structure with rms deviations of 0.68–1.33 Å, were used as test conformations [73] for several 'atomic solvation parameter sets. The test data sets included the classic parameter sets from Eisenberg and McLachlan [32] and Ooi et al. [50]. As expected, minimization of the conformational energy without including solvent contribution did not favor the native conformation. If solvent contribution was included, the conformation with the lowest rms deviation from the X-ray structure had the lowest energy for the E&M parameters, but not for the OONS parameters.

A related question addressed in the same study [52] was: Is there a monotonic relation between energy and the rms deviation from the native structure? This would be very useful for the predictive power of an energy function, although it is not clear to what extent a realistic model for the protein–solvent interactions should show this monotony. The authors used the Kendall coefficient for a quantitative comparison of the data sets. They found the highest concordance for a data set derived from the NMR coupling constants of peptides in water. This study clearly showed that there are several atomic solvation parameter sets whose energy terms used during refinement should improve the 3D structure. However, the study does not allow a final judgement of the relative merits of the parameter sets, as the number of sampled conformations is relatively small. Further, the maximal deviation to the native structure was, with 1.3 Å for C^α atoms, only slightly greater than the deviation of the solution to the crystal structure, as estimated by comparing NMR and X-ray structures [5].

Refinement of proteins with protein–solvent terms

All 39 conformations of BPTI were also subjected to energy minimization with three different solvation parameters [53]. The concordance coefficient for two of the data sets increased, but calculations with the data set which had the highest concordance in the previous test produced unrealistically large perturbations. For the two other data sets, the rank of the conformation, according to the rms deviation from the X-ray structure PTI4 with lowest energy, significantly decreased from 22 to 4 and from 20 to 1. The studies were complemented with Monte Carlo plus minimization (MCM) [74] starting from the X-ray structure PTI5. The structure stayed nearer to the initial starting structure (within 1.8 Å) than minimization without solvent, but actually converged more towards the X-ray structure PTI4. This compares fairly well with a recent molecular dynamics study of BPTI using explicit water molecules, where the final backbone deviation from the initial NMR structures after 1–1.4 ns is about 1.5 Å, and the deviation from the crystal structure is about 2 Å [75].

The effect of protein–solvent interactions on refining protein structures with and without NMR constraints was investigated for the α-amylase inhibitor, tendamistat [42]. Four parameter sets from Richards [31], Ooi et al. [50], Vila et al. [52], and Wesson and Eisenberg [40] were tested. The refined structures changed only slightly by 0.5 Å backbone rms deviation from the initial unrefined structures with all four parameter sets. Without constraints the best parameter set (Richards) produced a final structure with a significantly smaller deviation from the NMR structure (backbone root-mean-square deviation (rmsd) of 1.1 Å) compared to the minimization *in vacuo* (1.7 Å).

A more systematic study of the quality of atomic solvation parameters in the refinement procedure was carried out with 25 BPTI and tendamistat conformations [54]. These conformations were perturbed from the NMR solution structures in a range of 0.4–2.7 Å backbone rms deviation. Two parameter sets from Ooi et al. [50] and Wesson and Eisenberg [40] were compared to the restraining force of the total accessible surface or the nonpolar part of the surface. The *in vacuo* energy function did not improve the perturbed structures as measured by the backbone rms deviation before and after energy refinement. The surprising result was that simple parameter sets were as efficient as the more sophisticated parameters in driving the perturbed structures back towards the solution structures.

Comparison with molecular dynamics simulation using explicit water molecules

Molecular dynamics calculations with explicit water molecules yield valuable information on the time scales and the amount of fluctuations of the backbone and side-chain motions in water, of the diffusion of the surrounding water molecules and of residence times for surface water molecules. Teeter [17] and Daggett and Levitt [24] have critically reviewed a large number of these calculations. One particular aspect is the accuracy with which these simulations can reproduce known NMR or X-ray structures. The rms deviations of the α-carbons from the crystal or NMR solution structures are similar to those found by refining protein structures with surface area potentials. It remains to be seen to what extent molecular dynamics calculations with explicit water molecules can reproduce the X–ray or NMR structures, if the trajectories are started from perturbed solution structures.

Application to folding and prediction studies

The usefulness of surface area potentials on folding or prediction of the 3D structures of proteins have been examined in a few studies. As part of a more general procedure for predicting the 3D structure of proteins, solvation models were included in calculating the three-dimensional structure of the avian pancreatic polypeptide [76] and rat galanin [77]. In both cases the protein–solvent contribution was an important selection criterion.

The three-helix bundle protein E*r*-10 was folded from an unfolded state containing three preformed helices by energy minimizations and Monte Carlo simulations with highly weighted protein–solvent interactions [43]. The unfolded structures that were minimized with the intramolecular interactions alone did not change their conformations towards the native structure, even though their energy values dropped considerably. In contrast, all structures obtained by minimizing the intramolecular and the protein–solvent interaction had significantly lower rmsd values compared to the NMR reference structure. The Monte Carlo simulations with the adaptive temperature schedule [78] produced structures which resemble the native E*r*-10 structure. The three structures with the lowest energies had rmsd values of 4.7, 3.8 and 3.0 Å compared to the correct structure.

The solvation contribution to the free energy of folding, ΔG_S, was also examined for several parameter sets [79,80]. A linear correlation of ΔG_S with the number of residues was obtained for all parameter sets. However, the absolute value and even the sign of ΔG_S varies between the data sets. Some of the parameter sets resulted in negative ΔG_S values, meaning that the folded state is favored over the unfolded state, as expected, but other data sets lead to positive values. Several of these data sets were derived with different intramolecular force fields and the analysis is based on the assumption that the intramolecular contribution is independent of the protein–solvent interaction. Currently there is no parameter set available which is best for all purposes.

Conclusions

Protein structure refinement including the protein–solvent interaction in a continuum approximation is a computationally efficient and practical approach. The quality of the resulting protein structures is similar to structures obtained with molecular dynamics calculation using explicit water molecules. Realistic atomic solvation parameters must still be optimized and tested for a variety of different protein folds. A good parameter set should: (i) favor the native state over all alternative structures; (ii) drive perturbed structures towards the native state; and (iii) give realistic estimates of the protein–solvent interaction.

Acknowledgements

I thank Dr. C.H. Schein for a critical reading of the manuscript. Support from the NSF program (BIR-9632326) is gratefully acknowledged.

References

1. Hendrickson, W.A. and Wüthrich, K. (Eds.) Macromolecular Structures, Current Biology, London, 1995.
2. Jack, A. and Levitt, M., Acta Crystallogr., Sect. A, 34(1978)931.

3. Brünger, A.T., J. Mol. Biol., 203(1988)803.
4. Némethy, G. and Scheraga, H.A., FASEB J., 4(1990)3189.
5. Clore, G.M., Robien, M.A. and Gronenborn, A.M., J. Mol. Biol., 231(1993)82.
6. Morris, A.L., MacArthur, W., Hutchinson, E.G. and Thornton, J.M., Proteins, 12(1992)345.
7. Wüthrich, K., Acta Crystallogr., Sect. D, 51(1995)249.
8. Hagler, A.T., Huler, E. and Lifson, S., J. Am. Chem. Soc., 96(1974)5319.
9. Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A., J. Phys. Chem., 79(1975)2361.
10. Némethy, G., Pottle, M.S. and Scheraga, H.A., J. Phys. Chem., 87(1983)1883.
11. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
12. Weiner, P.K., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7(1983)230.
13. Van Gunsteren, W.F. and Berendsen, H.J.C., J. Mol. Biol., 176(1984)559.
14. Brünger, A.T., XPLOR Version 3.1 User Manual, Yale University Press, New Haven, CT, 1992.
15. Bränden, C.I. and Jones, T.A., Nature, 343(1990)687.
16. Novotny, J., Bruccoleri, R. and Karplus, M., J. Mol. Biol., 177(1984)787.
17. Teeter, M.M., Annu. Rev. Biophys. Biophys. Chem., 20(1991)577.
18. Smith, P.E. and Pettitt, B.M., J. Phys. Chem., 98(1994)9700.
19. Honig, B. and Nicholls, A., Science, 268(1995)1144.
20. Van Gunsteren, W.F. and Karplus, M., Biochemistry, 21(1982)2259.
21. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L., J. Chem. Phys., 79(1983)926.
22. Bash, P.A., Singh, U.C., Langridge, R. and Kollman, P.A., Science, 236(1987)564.
23. Levitt, M. and Sharon, R., Proc. Natl. Acad. Sci. USA, 85(1988)7557.
24. Daggett, V. and Levitt, M.A., Proc. Natl. Acad. Sci. USA, 89(1992)5142.
25. Warwicker, J. and Watson, H.C., J. Mol. Biol., 157(1982)671.
26. Klapper, I., Hagstrom, R., Fine, R., Sharp, K. and Honig, B., Proteins, 1(1986)47.
27. Lee, B. and Richards, F.M., J. Mol. Biol., 55(1971)379.
28. Shrake, A. and Rupley, J.A., J. Mol. Biol., 79(1973)351.
29. Chothia, C., Nature, 248(1974)338.
30. Nozaki, Y. and Tanford, C., J. Biol. Chem., 246(1971)2211.
31. Richards, F.M., Annu. Rev. Biophys. Bioeng., 6(1977)151.
32. Eisenberg, D. and McLachlan, A.D., Nature, 316(1986)199.
33. Fauchère, J.L. and Pliska, V., Eur. J. Med. Chem.-Chim. Ther., 18(1983)369.
34. Wodak, S.J. and Janin, J., Proc. Natl. Acad. Sci. USA, 77(1980)1736.
35. Connolly, M.L., J. Appl. Crystallogr., 16(1983)548.
36. Connolly, M.L., J. Am. Chem. Soc., 107(1985)1118.
37. Connolly, M.L., J. Mol. Graph., 11(1993)139.
38. Richmond, T.J., J. Mol. Biol., 178(1984)63.
39. Perrot, G., Cheng, B., Gibson, K.D., Vila, J., Palmer, K.A., Nayeem, A., Maigret, B. and Scheraga, H.A., J. Comput. Chem., 13(1992)1.
40. Wesson, L. and Eisenberg, D., Protein Sci., 1(1992)227.
41. Schiffer, C.A., Caldwell, J.W., Stroud, R.M. and Kollman, P.A., Protein Sci., 1(1992)396.
42. von Freyberg, B. and Braun, W., J. Comput. Chem., 14(1993)510.
43. Mumenthaler, Ch. and Braun, W., J. Mol. Mod., 1(1995)1.

44. The full paper is available on-line on www: http://science.springer.de/jmm/jmm.htm.
45. Eisenhaber, F. and Argos, P., J. Comput. Chem., 14(1993)1272.
46. Petitjean, M., J. Comput. Chem., 15(1994)507.
47. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. and Scharf, M., J. Comput. Chem., 16(1995)273.
48. Sridharan, S., Nicholls, A. and Sharp, K.Z., J. Comput. Chem., 16(1995)1038.
49. Cossi, M., Mennucci, B. and Cammi, R., J. Comput. Chem., 17(1996)57.
50. Ooi, T., Oobatake, M., Némethy, G. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 84(1987)3084.
51. Kang, Y.K., Gibson, K.D., Némethy, G. and Scheraga, H.A., J. Phys. Chem., 92(1988)4739.
52. Vila, J., Williams, R.L., Vasquez, M. and Scheraga, H.A., Proteins, 10(1991)199.
53. Williams, R.L., Vila, J., Perrot, G. and Scheraga, H.A., Proteins, 14(1992)110.
54. von Freyberg, B., Richmond, T.J. and Braun, W., J. Mol. Biol., 233(1993)275.
55. Schiffer, C.A., Caldwell, J.W., Kollman, P.A. and Stroud, R.M., Mol. Sim., 10(1993)121.
56. Stouten, P.F.W., Frömmel, C., Nakamura, H. and Sander, C., Mol. Sim., 10(1993)97.
57. Martino, R.L., Johnson, C.A., Suh, E.B., Trus, B.L. and Yap, T.K., Science, 265(1994)902.
58. Silla, E., Villar, F., Nilsson, O., Pascual-Ahuir, J.L. and Tapia, O., J. Mol. Graph., 8(1990)168.
59. Wang, H. and Levinthal, C., J. Comput. Chem., 12(1991)868.
60. Abagyan, R., Totrov, M. and Kuznetsov, D., J. Comput. Chem., 15(1994)488.
61. LeGrand, S.M. and Merz Jr., K.M.M., J. Comput. Chem., 14(1993)349.
62. Braun, W. and Gō, N., J. Mol. Biol., 186(1985)611.
63. Kocher, J.-P.A., Rooman, M.J. and Wodak, S.J., J. Mol. Biol., 235(1994)1598.
64. Hasel, W., Hendrickson, T.F. and Still, W.C., Tetrahedron Comput. Methodol., 1(1988)103.
65. Mohamadi, F., Richards, N.G.J., Guida, W.C., Liskamp, R., Lipton, M., Caufield, C., Chang, G., Hendrickson, Th. and Still, W.C., J. Comput. Chem., 11(1990)440.
66. Kang, Y.K., Gibson, K.D., Némethy, G. and Scheraga, H.A., J. Phys. Chem., 92(1988)4735.
67. Higo, J. and Gō, N., J. Comput. Chem., 10(1989)376.
68. Karfunkel, H.R. and Eyraud, V., J. Comput. Chem., 10(1989)628.
69. Kundrot, C.E., Ponder, J.W. and Richards, F.M., J. Comput. Chem., 12(1991)402.
70. Delarue, M. and Koehl, P., J. Mol. Biol., 249(1995)675.
71. Wang, Y., Zhang, H., Li, W. and Scott, R.A., Proc. Natl. Acad. Sci. USA, 92(1995)709.
72. Wang, Y., Zhang, H. and Scott, R.A., Protein Sci., 4(1995)1402.
73. Ripoll, D.R., Piela, L., Vasquez, M. and Scheraga, H.A., Proteins, 10(1991)188.
74. Li, Z. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 84(1987)6611.
75. Brunne, R.M., Berndt, K.D., Güntert, P., Wüthrich, K. and van Gunsteren, W.F., Proteins, 23(1995)49.
76. Liwo, A., Pincus, M.R., Wawak, R.J., Rackovsky, S. and Scheraga, H.A., Protein Sci., 2(1993)1715.
77. Liwo, A., Oldziej, S., Ciarkowski, J., Kupryszewski, G., Pincus, M.R., Wawak, R.J., Rackovsky, S. and Scheraga, H.A., J. Protein Chem., 13(1994)375.
78. von Freyberg, B. and Braun, W., J. Comput. Chem., 12(1991)1065.
79. Chichie, L.M., Gregoret, L.M., Cohen, F.E. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 87(1990)3240.
80. Juffer, A.H., Eisenhaber, F., Hubbard, S.J., Walther, D. and Argos, P., Protein Sci., 4(1995)2499.

Normal mode analysis of biomolecular dynamics

David A. Case
Department of Molecular Biology, The Scripps Research Institute,
La Jolla, CA 92037, U.S.A.

Introduction

The past decade has seen an impressive advance in the application of molecular simulation methods to problems in chemistry and biochemistry. As computer hardware has become faster and software environments more sophisticated, the amount of detailed information available and its expected level of accuracy has grown steadily [1]. But it has become increasingly clear that the 'easy' part of a simulation project is setting up and carrying out the calculations, and the hard part generally lies in extracting useful data from among the very many things that can be calculated from a trajectory or Monte Carlo simulation. Normal mode analysis provides an approximate but analytical description of the dynamics, and has long been recognized as an important limiting case for molecular dynamics in condensed phases [2]. A principal limitation arises from the fact that normal modes are defined by an expansion about a particular point on the potential energy surface, and hence have difficulty describing transitions from one local minimum to another. The quasiharmonic and 'instantaneous' mode theories discussed below attempt to ameliorate some of this neglect of the 'rugged' nature of protein energy landscapes. Yet there remains a 'paradoxical aspect' [3] of biomolecular dynamics that is still the subject of considerable study: even though the energy surface contains many local minima, proteins behave in some ways as though the energy surface were harmonic, and normal mode analyses are often more correct than one might expect. This article reviews some recent experience on the application of normal mode ideas to biomolecules, looking at how this technique describes short-timescale motion as well as longer timescale, collective motions. The use of harmonic ideas in the analysis of crystallographic and NMR data will also be outlined.

Fundamentals of normal mode analysis

Basic ideas

The basic idea of normal mode analysis is to expand the potential function V(**x**) in a Taylor series expansion about some point x_0:

$$V(\mathbf{x}) = V(\mathbf{x}_0) + \mathbf{g}^T(\mathbf{x} - \mathbf{x}_0) + \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T\mathbf{F}(\mathbf{x} - \mathbf{x}_0) + \cdots \qquad (1)$$

If the gradient **g** of the potential vanishes at this point and third- and higher order derivatives are ignored, it is straightforward [4,5] to show that the dynamics of the system can be described in terms of the normal mode directions and frequencies $\mathbf{Q}_i$, ω_i, which satisfy

$$\begin{aligned} \mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{-1/2}\mathbf{Q}_i &= \omega_i^2\mathbf{Q}_i \\ \mathbf{Q}_i \cdot \mathbf{Q}_j &= \delta_{ij} \end{aligned} \tag{2}$$

In Cartesian coordinates, the matrix **M** contains atomic masses on its diagonal, and the Hessian matrix **F** contains the second derivatives of the potential energy evaluated at $\mathbf{x}_0$. The time evolution of the system is then

$$x_i(t) = x_i(0) + 2^{1/2} \sum_k Q_{ik} m_i^{-1/2} \sigma_k \cos(\omega_k t + \delta_k) \tag{3}$$

where σ_k is an amplitude, $\omega_k$the angular frequency and δ_k the phase of the kth normal mode of motion. The phases and amplitudes depend upon the positions and velocities at time $t = 0$. It is conventional in molecular problems to divide the frequencies ω_i by the speed of light to report the results in cm^{-1} units.

A straightforward computation of normal modes in Cartesian coordinates thus involves a numerical diagonalization of a matrix of size 3N × 3N, for a molecule with N atoms. It is now about 10 years since computers have become powerful enough to allow normal mode calculations to be carried out on proteins and nucleic acids [5–9], following earlier and influential studies on peptides. With present-day computers, it is not difficult to study proteins up to about 150 amino acids with an all-atom model, or to about 200 amino acids using a united atom description in which hydrogens bonded to carbon are not explicitly represented. A common approximation for larger systems assumes that bond lengths and angles are fixed. This can reduce the size of the matrix involved by about an order of magnitude. Calculations can be carried out by direct construction of the potential and kinetic energy matrices in (curvilinear) internal coordinates [10,11], or through matrix partitioning techniques that start from Cartesian derivatives [12–14]. In general, reductions of the dimensionality of the expansion space have noticeable but not overwhelming effects on the resulting normal mode description of the dynamics. The directions of the lower frequency modes are largely preserved, but frequencies in general are higher in the lower dimensional space [15,16], suggesting that small fluctuations in bond lengths and bond angles have the effect of allowing the dihedral angles to become more flexible. Many practical aspects of computing modes for large molecules can be found in recent articles by Brooks et al. [14] and Wako et al. [11].

In the normal mode coordinate frame, each mode is independent of the rest, and average quantities can generally be written as sums of contributions from each mode:

$$< f(\mathbf{x}) > \approx f(\mathbf{x}_0) + \tfrac{1}{2} \sum_i^{3N} \sum_{k,l} \frac{\partial^2 f(x_0)}{\partial x_k \partial x_l} (m_k m_l)^{-1/2} Q_{ik} Q_{il} \sigma_i^2 + \cdots \tag{4}$$

The thermal averages of the second moments σ_i^2 of the amplitude distributions can be calculated for both classical and quantum statistics [17]:

$$\sigma_{i,\text{class}}^2 = kT/\omega_i^2, \qquad \sigma_{i,\text{qm}}^2 = \frac{h}{4\pi\omega_i} \coth \frac{h\omega_i}{4\pi kT} \tag{5}$$

where h and k are the Planck and Boltzmann constants. The two statistics coincide in the limits of low frequency or high temperature. For biomolecules, the most important difference is generally that higher frequency modes have little amplitude in classical statistics but have nonnegligible zero-point motion in quantum statistics. Harmonic models thus provide one of the few practical ways for including quantum effects in biomolecular simulations.

Time-dependent averages can also be determined for normal mode dynamics. The most common case is a correlation function, where f is the product of two time-dependent functions, $f = A(\mathbf{x}, 0) \cdot B(\mathbf{x}, \tau)$. In the special case where A and B are linear functions of $\mathbf{x}$, Eq. 4 becomes

$$< A(0) \cdot B(\tau) > \; = \sum_i^{3N} \sum_{k,l} \frac{\partial A}{\partial x_k} \frac{\partial B}{\partial x_l} (m_k m_l)^{-1/2} Q_{ik} Q_{il} \sigma_i^2 \cos(\omega_i \tau) + <A> \; <B> \tag{6}$$

For small (infinitesimal) displacements from equilibrium, the calculation of averages such as those in Eq. 4 is independent of the underlying coordinate system used to describe the molecule. This is not true for larger fluctuations, which can become significant for low-frequency motions of biomolecules: a finite displacement along a normal mode direction in a curvilinear coordinate system (such as one defined through bond lengths and angles) will generally have a different character than one expressed in Cartesian coordinates. Each normal mode represents a concerted motion of the molecule that is linear in the space in which the analysis is carried out; if two coordinate systems are related by a nonlinear transformation, the predicted fluctuations (and even the average structure) can depend upon which coordinate system is used. As a simple example, consider the torsional mode about the symmetry axis of a methyl group. In dihedral angle space, the protons move in a circle, maintaining constant C-H bond lengths, whereas in Cartesian coordinates, the hydrogen atoms move off on a tangent, distorting internal bond lengths and angles. Similar behavior is seen for many other types of local motions. Sunada and Gō [10] have analyzed such effects in BPTI by computing the transformation coefficients between Cartesian and dihedral angle space through second order. They find good agreement between the two coordinate systems only if second-order terms are included. It is also possible to obtain 'exact' results even for finite fluctuations by carrying out Monte Carlo simulations on the harmonic potential surface, where the transformation between curvilinear (angle) space and Cartesian coordinates is performed exactly at each instantaneous conformation, and not merely for the minimum-energy conformation. Because of the smoothness and simplicity of the potential, these simulations converge relatively quickly. For most properties examined so far, it appears that a second-order expansion about the average position gives results nearly identical to those from the Monte Carlo simulations [10].

Internal coordinate frames related to dihedral angle variables, however, have their own problems when applied to large molecules, especially when closed rings (or 'quasirings' closed by hydrogen bonds) are present. This leads to a consideration of the question of finding an 'optimal' coordinate frame, in which the harmonic approximation would be most correct even for finite displacements from equilibrium. A somewhat similar problem is faced in the development of 'natural internal coordinates' for geometry optimization [18–20], where again the goal is to minimize anharmonic couplings between potential displacement directions. Application of these ideas to macromolecules might lead to interesting results.

Langevin modes

It is also possible to solve for normal mode dynamics in the presence of viscous damping by a continuum 'solvent' [21]. In this approach, Newton's equations are replaced by Langevin equations that include terms describing viscous damping and random (white) noise.

$$m\ddot{\mathbf{x}} = -V'(\mathbf{x}) - \zeta\mathbf{v} + \mathbf{r}(t) \tag{7}$$

Here $\mathbf{v}$ is the velocity vector, ζ is the friction matrix and $\mathbf{r}(t)$ is a vector of random numbers. The random numbers follow Gaussian distribution with the following properties:

$$\begin{aligned} &< r_i(t) > \; = 0 \\ &< r_i(t)\, r_j(t') > \; = 2\zeta_{ij}\,\delta(t - t')/k_BT \end{aligned} \tag{8}$$

Equation 8 is a fluctuation–dissipation relation that ensures that the long-timescale behavior of the system converges to an equilibrium one characterized by the temperature T. Expanding the potential to quadratic terms as in Eq. 1, and defining for convenience $\alpha \equiv \mathbf{M}^{1/2}(\mathbf{x} - \bar{\mathbf{x}})$ and $\nu \equiv \dot{\alpha}$, yields a matrix version of the Langevin dynamics:

$$\begin{aligned} \begin{pmatrix} \dot{\alpha} \\ \dot{\nu} \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ -\mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{-1/2} & -\mathbf{M}^{-1/2}\zeta\mathbf{M}^{-1/2} \end{pmatrix} \begin{pmatrix} \alpha \\ \nu \end{pmatrix} + \begin{pmatrix} 0 \\ R(t) \end{pmatrix} \\ &\equiv \mathbf{A}\begin{pmatrix} \alpha \\ \nu \end{pmatrix} + \begin{pmatrix} 0 \\ R(t) \end{pmatrix} \end{aligned} \tag{9}$$

The random numbers $R_i(t)$ now satisfy

$$\begin{aligned} &< R_i(t) > \; = 0 \\ &< R_i(t) R_j(t') > \; = 2m_i^{-1/2}\zeta_{ij}m_j^{-1/2}\,\delta(t - t')/k_BT \end{aligned} \tag{10}$$

If the macromolecule has N atoms, then $\mathbf{A}$ is a 6N x 6N matrix and is nonsymmetric. It is useful to construct the distributions arising from these stochastic differential equations from solutions to the homogeneous (ordinary) equations that are obtained when R(t) vanishes [21]. These solutions involve the propagator exp(At), which can be expressed in terms of an eigenanalysis of the matrix **A**. Dynamical averages can then be computed using an analogue of Eq. 6, where the frequencies are now complex, describing in general damped oscillatory motions.

The coupling between solvent and solute is often represented by a 'bead' model in which each atom is a source of friction, with some corrections to represent the effects of burial or of hydrodynamic interactions between atoms [21]; alternatively, effective friction couplings can be extracted from molecular dynamics simulations [22,23]. Computational details can be found in some of the early publications [21,24]. Calculations on small proteins and nucleic acids using a bead model for fractional coupling to solvent indicated that most modes with vacuum frequencies below about 75 cm^{-1} would become overdamped, and that frequency shifts could be significant [24]. These frictional models, however, may overestimate solvent damping, especially for lower frequency motions [25]. An analysis of molecular dynamics simulations of BPTI in water suggested a model in which the frictional coupling was nearly the same in all modes, with a value near 47 cm^{-1}, so that modes with effective frequencies below about 23 cm^{-1} would become overdamped [23]. Analysis of inelastic neutron scattering data [26] led to a model in which the effective friction is a Gaussian function of the frequency, so that the lower frequency collective modes experience a greater frictional damping than do more localized, higher frequency modes (cf. Fig. 1). In this model, modes below 15 cm^{-1} are overdamped, and some frictional damping effects are noticeable up to about 75 cm^{-1}.

A classic study of the effects of solvent damping on vibrational motions involves the 'hinge-bending' motion between two domains of lysozyme, which was originally analyzed in terms of the energy profile along an assumed bending coordinate, and found to be overdamped [27]. More recent normal mode investigations of this system provide a detailed description of the nature of the hinge-bending coordinate; projections of Monte Carlo or molecular dynamics trajectories onto the normal mode coordinates support the basic features of the normal mode analysis and allow the dynamics to be analyzed in terms of harmonic and anharmonic contributions [3,28].

Quasiharmonic analysis

Another important extension of normal modes is to 'quasiharmonic' behavior, in which effective modes are computed such that the second moments of the amplitude distribution match those found in a Monte Carlo or molecular dynamics simulation using the complete, anharmonic force field [29,30]. The basic idea is to compute the fluctuation matrix from a dynamics or Monte Carlo simulation:

$$\sigma_{ij} = <(x_i - \bar{x}_i)(x_j - \bar{x}_j)> \qquad (11)$$

and to assume that the complete conformational probability distribution is approximately a multivariate Gaussian:

$$P(\mathbf{x}) = (2\pi)^{-n/2}|\det \sigma|^{-1/2} \exp[-\tfrac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T \sigma^{-1}(\mathbf{x} - \bar{\mathbf{x}})] \qquad (12)$$

The probability distribution can also be related to the potential energy:

$$P(\mathbf{x}) \approx \exp[-V(\mathbf{x})/kT] \tag{13}$$

In the quasiharmonic model, V is a quadratic function of position:

$$V(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^{T}\mathbf{F}^{\text{quasi}}(\mathbf{x}-\bar{\mathbf{x}}) \tag{14}$$

so that the effective force constant matrix becomes the inverse of the fluctuation matrix found in the simulation:

$$\mathbf{F}^{\text{quasi}} = kT[\sigma]^{-1} \tag{15}$$

Since $\mathbf{F}^{\text{quasi}}$ and σ have common eigenvectors, the quasiharmonic modes can be determined from the mass-weighted fluctuation matrix, and it is not necessary to explicitly construct $\mathbf{F}^{\text{quasi}}$.

It is important to recognize that this sort of quasiharmonic analysis is based on a static analysis of the fluctuation matrix and not on any time-series analysis of actual motions. Many features of the correlation matrix, including aspects of the 'low-frequency' behavior, are present even in fairly short molecular dynamics simulations [30,31]. The convergence characteristics of these modes (and, indeed, of molecular simulations in general) are still the subject of study, and I will return to this subject below.

The quasiharmonic assumption that the distribution of configurations is a multivariate Gaussian provides an analytical form that permits the calculation of quantities such as the vibrational entropy, which are otherwise hard to estimate [29,32,33]. As with true normal mode analysis, a contribution to thermodynamic quantities can be associated with each mode; the overall entropy can be expressed in terms of the logarithm of the determinant of σ, which is less expensive to determine than the full eigenvalue analysis. This approach has been used recently, for example, to estimate energetic consequences of cross-links to protein stability [34], or entropic effects associated with protein oligomerization [35].

The quasiharmonic approach includes some effects of anharmonic terms in the potential, at least to the extent that they influence the mean-square displacements, but it still assumes distributions that are unimodal in character. Tests of this assumption using MD simulations give mixed results. The atomic displacements in an α-helix appear to be approximately Gaussian over a wide temperature range [36], and MD simulations on lysozyme suggest that most atoms have fluctuations that are highly anisotropic but only slightly anharmonic [37]. Other studies, including those on myoglobin [38], crambin [39], mellitin [40], and lysozyme [41], have found distribution functions with more than one maximum, especially along low-frequency directions. It is possible to account approximately for such results in estimates of the entropy [33], but any significant deviations from unimodal behavior are very difficult to accommodate into a quasiharmonic description.

Alternative connections between MD simulation results and harmonic models can also be drawn. Bialek and Goldstein [42] discuss a model in which the effective

Hessian is derived from the mean curvature rather than by fitting the elements of σ. Roitberg et al. [43] have described an interesting approach that uses a vibrational self-consistent field approach, expanding the potential to quartic terms in normal mode coordinates. This leads to the construction of a mean-field potential for each mode, and solution of the quantum vibrational problem as a series of one-dimensional Schrödinger equations. Application to BPTI showed significant anharmonic contributions to the effective frequencies and to computed mean-square displacements. Although this approach is not 'quasiharmonic' in the sense used here, it does allow the interpretation of anharmonic results within the framework of normal mode analysis, and may lead to a variety of interesting calculations on quantum effects on protein dynamics, particularly at low temperature.

Harmonic descriptions of biomolecular dynamics

When considering the applications of normal mode analyses to large molecules, it is convenient to divide the discussion into considerations of short-timescale dynamics and long-timescale, collective motions. Most early applications to biomolecules used simplified force fields and concentrated on the low-frequency behaviors that characterize overall molecular flexibility and which were thought to be reasonably independent of force field details. Recent studies have emphasized the wealth of information that can be gleaned from the analysis of biomolecular dynamics even on very short timescales. The following sections consider these two areas.

There is a close connection between mode frequency and the collective character of the atomic motions. Figure 1 plots a 'collectivity index' κ_i against frequency for crambin [44]. This index is proportional to the exponential of the 'information entropy' of the eigenvector:

$$\kappa_i = N^{-1} \exp\left[-\sum_n u_{i,n}^2 \log u_{i,n}^2 \right] \tag{16}$$

where $u_{i,n}$ measures the extent of motion of atom n in mode i, normalized such that the sum of its squares is unity [44]. κ_i can vary from 1/N to 1, where a value of 1 indicates maximal collectivity where all the eigenvector amplitudes are the same (as for the modes describing global translations). There is a clear drop-off in collectivity as the frequency increases, with large collectivity only found in modes below 200 cm^{-1}.

Short-timescale dynamics

At short timescales, transitions from one local minimum region to another are minimal, and an analysis that uses normal mode ideas makes good sense. Beyond estimating the intrinsic vibrational frequencies, interest in the dynamical behavior in this region often centers on the anharmonic transfer of energy between modes, and the accompanying loss of phase coherence for individual modes. These are closely related

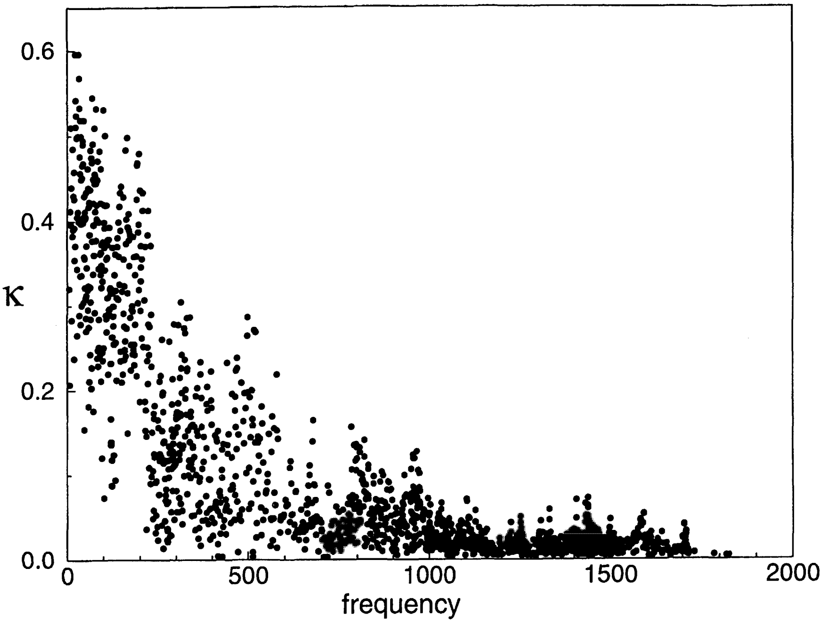


Fig. 1. Mode collectivity κ_i (Eq. 16) as a function of mode frequency (in cm^{-1}) for normal modes of crambin. Adapted from Ref. 44.

to important questions about the nature of the potential energy surface and the efficiency of vibrational energy transfer. Some typical features are illustrated in Fig. 2, which compares solvated molecular dynamics, normal mode and Langevin mode predictions for a time correlation function related to NMR relaxation in a zinc-finger peptide [45]. Here the short-time oscillations predicted by the gas-phase normal modes reproduce the behavior of the solvated MD simulation for about 0.5 ps, but beyond that the dephasing behavior arising from collisions with solvent molecules and from anharmonic interactions within the protein itself leads to divergent behavior. The simple Langevin treatment shown here (using a 'bead' model for frictional interactions) forces the oscillations in the correlation function to decay (as they will never do in the pure normal mode treatment), but this damping takes place on much too short a timescale.

Temperature and velocity echoes. Some recent and novel simulation techniques point the way to a more general exploration of this sort of short-timescale behavior [46–49]. In the simplest experiment [46], kinetic energy is 'quenched' (all velocities are instantaneously set to zero) at two different times separated by a variable evolution time t_1. A spontaneous echo (drop in temperature) can then appear at time t_1 following the second quench. The intensity of the echo is related to vibrational frequencies of the system and to the strength of anharmonic coupling between modes. It is typical to find periodic modes in proteins with phase coherence times on the order

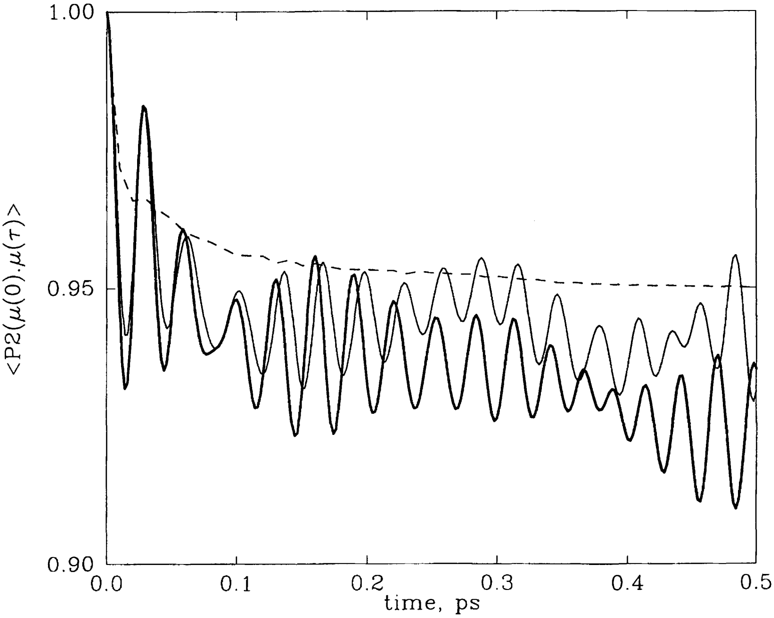


Fig. 2. Orientational correlation function (Eq. 18) of the C^{α}-H^{α} bond of Ala^{15} in a zinc-finger peptide [45]. Heavy solid curve: solvated molecular dynamics simulation; light solid curve: gas-phase normal mode calculation; dashed curve: Langevin mode calculation.

of 1 ps. A Fourier transform into the frequency domain gives information about the density of states that appears to be in good agreement with the results from inelastic neutron scattering. These results suggest that a large amount of information is available from very short time (subpicosecond) dynamical behavior. Extensions to other types of velocity reassignment protocols [48,49] suggest that a significant amount of interesting dynamical behavior may be obtained by this general approach.

Instantaneous normal modes. It has been recognized for some time that one of the salient features of protein dynamics is the existence of many local conformational minima, and that transitions from (the vicinity of) one local minimum to another represent an important feature in macromolecular dynamics [50,51]. The normal mode model, which expands the potential about a single local minimum, does not directly include contributions from such transitions on a rough or corrugated surface. Recent studies on peptide and protein systems show that between 30% and 70% of the total atomic fluctuation arises from transitions between minima, with normal mode theories working better for the more tightly constrained systems [31,52]. An interesting approach to understanding the dynamics of fluid systems that do not oscillate about a single or small number of conformational minima involves the calculations of modes about the instantaneous configurations sampled in a simulation. Since these are in general not local minima, the frequency spectrum contains both real and imaginary components, and the nature and distribution of these 'unstable' modes can be related to dynamical quantities [53–57]. Straub and Thirumalai [58,59] have applied such ideas to the ribonuclease S-peptide, computing the instantaneous normal mode spectrum between 40 and 500 K. The number and character of the unstable modes can be used to characterize the distribution of barriers between 'conformational substates'. At room temperature, about 4% of the modes are unstable, and this value is predicted to increase to about 10% in the high-temperature limit. A distribution of barrier heights between conformers that fits the frequency data has the following form:

$$g(E_B) = a\theta(E_{low} - E_B) + bE_B e^{-E_B/E_0} \qquad (17)$$

Here there is a constant density of low-energy barriers for $E_B < E_{low}$ ($\theta(E)$ is the Heaviside function), and a Poisson distribution of higher energy barriers with a maximum at E_0. For the S-peptide, $a = 0.325\,(\text{kcal/mol})^{-1}$, $E_{low} = 0.2$ kcal/mol, $b = 0.13\,(\text{kcal/mol})^{-2}$, and $E_0 = 1$ kcal/mol. This fit yields a broad distribution of barrier heights that includes many very small barriers. It should be of considerable interest to see what distributions are obtained in other biomolecular systems.

Long-timescale motions

As indicated above, considerable attention has been given to using normal mode or quasiharmonic analysis to probe low-frequency, longer timescale motions in biopolymers. Since the amplitude of mode fluctuations is inversely proportional to frequency (cf. Eq. 5), normal modes have the attractive feature of describing motions that contribute most to atomic fluctuations in terms of a relatively small number of mode directions and frequencies: a rough rule of thumb is that 1% of the modes contribute up to 90% of the overall root-mean-square atomic fluctuations. This has inspired an interest in the characterization of low-frequency modes along with hopes that interesting domain movements [60] might appear as identifiable modes, or as a combination of a small number of modes. A further hope is that the 'essential dynamics' [41] of proteins might involve motions confined to a subspace of low-frequency normal modes.

A fairly large amount of literature now exists in which large-scale collective motions of proteins have been studied with normal mode calculations [61]. In addition to the analyses of the lysozyme hinge-bending modes mentioned above, interdomain motions of G-actin (with ADP and calcium bound) [62] and a 'mitten mode' of an epidermal growth factor [63] have recently been characterized through normal mode analyses. Molecular dynamics simulations of myoglobin have been analyzed in terms of rigid-helix and side-chain dynamics, with low-frequency rigid-helix vibrations being characterized through Fourier transforms of velocity autocorrelation functions [64].

An obvious problem in evaluating harmonic models for long-timescale motions is the lack of a secure standard for comparison. It is generally not possible to run molecular dynamics simulations for long enough periods of time to obtain statistically meaningful information about nanosecond-scale motions. For short peptides of four to six amino acids, simulations from 10 to 100 ns sometimes appear to be approaching an equilibration among various conformational states [65], and studies using simplified solvent models (such as Brownian dynamics or mean-field simulations) can study much longer timescales [66]. But it is clear that most picosecond-to-nanosecond simulations of proteins and nucleic acids in water are not well equilibrated, at least for some aspects of interatomic correlations [67], and that most of the simulations on this timescale contain 'rare events' that complicate straightforward analyses based on the assumption of an equilibrated sample [68–70]. The nature of the protein energy landscape is such that there are likely to be conformational transitions on nearly all timescales [51] so that any individual time segment of a simulation will probably not be in equilibrium with respect to some types of motion. Slow motions that involve relatively large or correlated conformational changes may dominate the lowest frequencies in a quasiharmonic analysis. While individual low-frequency quasiharmonic modes are sensitive to the details of the trajectory that produced them, it is less clear how to interpret the subspace spanned collectively by the large-amplitude eigenvectors. Balsera et al. [69] have concluded from a scaling analysis and computational studies on G-actin that the character of the large-amplitude quasiharmonic subspace is so dependent upon the details of the underlying simulation that the 'essential dynamics' of the system cannot be ascertained from it. The quasiharmonic analysis tends to mix ordinary normal modes together, so that the density of states is a smoother function of frequency than for a true normal mode distribution, and much of the distinction between local and global character (cf. Fig. 1) is lost [31]. Further study will be required to ascertain the extent to which quasiharmonic directions are useful in describing or rationalizing the global behavior of biomolecular motions.

Normal modes in crystallographic and NMR refinement

Thermal parameters in crystallography. One nice feature of the normal mode description is that it provides a compact description of dynamical behavior in which the major contributions to atomic fluctuations are dominated by a relatively small number of low-frequency modes. The quasiharmonic picture can thus be viewed as a description of molecular motion parametrized by a fairly small number of adjustable parameters. Two groups have used this idea to refine the temperature factors in proteins, using either the frequencies [71] or frequencies plus mode-mixing parameters as adjustable parameters [72–74]. The results appear to give a good picture of fluctuations that contribute to the thermal parameters, although the caveats discussed above concerning distribution functions with multiple peaks should be kept in mind.

There are a variety of potential advantages to using a normal-mode-based model for thermal parameters compared to more conventional B-factor refinements that appear to be borne out in practice. First, this model provides a clear distinction between true internal motions and 'external' contributions to thermal parameters arising from lattice vibrations or crystalline disorder. Second, the model includes important aspects of anisotropic and correlated atomic motions without the introduction of an unmanageably high number of adjustable parameters. Finally, the mode adjustment procedure appears to be robust, yielding behavior in a 'free R-factor' analysis [75] that is better than that of more conventional isotropic B-factor refinement [74].

Diffuse scattering. In addition to Bragg diffraction intensity that appears at reciprocal lattice points, crystals with internal thermal fluctuations also exhibit diffuse scattering that arises from correlated fluctuations in the average electron density [76]. Studies on diffuse scattering in some protein crystals suggest that correlated atomic displacements are complex and liquid-like, with correlations that decay over a relaxation distance of about 6 Å [67,77,78]. It is difficult to capture such effects in molecular dynamics simulations, and some evidence suggests that a normal mode model gives a good account of most diffuse features [78]. As with ordinary B-factors, normal mode analysis can be used to create a motional model in which the directions of low-frequency eigenvectors are kept fixed, but the amplitudes of motions (or, equivalently, the effective frequencies) are treated as adjustable parameters to be modified to fit experimental data. An extension of this idea allows the mixing of some low-frequency modes, creating a more flexible model with a larger number of adjustable parameters. Preliminary applications of these ideas to lysozyme have been reported [79], and more quantitative studies should help establish the extent to which the effective mode model captures the fundamental aspects of atomic motions that are reflected in diffuse scattering. General questions about the frequency distribution and damping of low-frequency vibrations can also be addressed by the analysis of inelastic neutron scattering, discussed elsewhere [26,80,81].

Normal modes and NMR relaxation. In biomolecular NMR, a great deal of potential important structural information can be obtained from the analysis of dipolar spin relaxation. Longitudinal spin relaxation rates in biomolecules are determined by time correlation functions of the general form [82]

$$C(\tau) \approx < P_2[\cos\chi(\tau)]/[r^3(0)r^3(\tau)] > \tag{18}$$

where r is the distance between spins and χ is the angle between the interspin vector at time 0 and at time τ. These correlation functions can be readily computed from formulations like that in Eq. 6 [82,83], and 'order parameters' that reflect the effects of internal motion on relaxation rates can be compared with experimental measurements. NMR relaxation experiments are commonly interpreted in terms of an intermediate-level 'model-free' analysis which assumes that the spectral density function for dipolar relaxation can be written as the sum of two Lorentzians [82,84]:

$$J(\omega) = \frac{1}{2\pi} < r^{-6} > \left(\frac{S^2\tau_c}{1+\omega^2\tau_c^2} + \frac{(1-S^2)\tau_{tot}}{1+\omega^2\tau_{tot}^2}\right) \tag{19}$$

where

$$\tau_{tot}^{-1} \equiv \tau_c^{-1} + \tau_e^{-1} \tag{20}$$

and τ_e is an effective internal correlation time. S^2 is then related to a plateau value (if it exists) of correlation functions like that in Eq. 18 and τ_e to the rate at which the plateau value is reached [84].

Vibrations and other rapid motions scale NOESY and ROESY cross-relaxation rates by the same quantity, which has been called the dynamic scaling factor γ [85] or correction factor Q [86]:

$$\gamma \equiv Q = \frac{\Gamma^{dyn}}{\Gamma^{stat}} = <r>^6 \ <r^{-6}> \ S^2 \equiv RS^2 \tag{21}$$

where Γ^{dyn} corresponds to the dynamically averaged relaxation rate and Γ^{stat} to a static reference structure. A value of $\gamma < 1$ reflects a dominance of angular over radial motions, with the reverse for $\gamma > 1$. Henry and Szabo [83] have considered vibrational contributions to these quantities, performing the average in internal rather than Cartesian coordinates, and provide explicit formulas for dipole–dipole interactions that separate radial and angular behavior.

$$\gamma \approx \left\{1 + \frac{1}{2r_{eq}^2}\left[15 \sum_{\alpha\beta}^{3} \zeta_\alpha \zeta_\beta < \Delta_\alpha \Delta_\beta > \ - 3\sum_{\gamma}^{3} < \Delta_\gamma^2 > \right]\right\}^2 \tag{22}$$

The dependence of NMR order parameters on the coordinate system used to describe finite displacements can be significant, and Cartesian coordinates are often poorly suited to the description of local motions that may often be dominated by floppy torsions. More correct results are probably obtained by the use of Eq. 22, or by higher-than-first-order expansions of both the operator involved in correlation decay [24] and of the fluctuations about equilibrium in an internal coordinate system [10]. These corrections can become significant for order parameters far from unity.

Recently, normal mode analyses of NMR order parameters have been reported for BPTI [87], for a zinc-finger peptide [45] and for crambin [44]. Figure 3 shows some results for BPTI for heteronuclear relaxation of the backbone C^α-H^α spin pairs. More flexible parts of the backbone (lower order parameters) are found at the N- and C-termini of the chain and in the loop region of the antiparallel β-sheet, near residue 27. Results assuming quantum statistics exhibit lower order parameters, since the presence of zero-point oscillations leads to a wider spatial probability distribution. The systematic difference between classical and quantum results is around 5%, and this effect should be corrected for in comparisons of experiment with classical MD simulations.

For homonuclear proton–proton cross-relaxation as observed by NOESY or ROESY experiments, quantum effects are found to be much smaller than in the heteronuclear case. This is a consequence of the fact that proton pairs are separated by larger distances, and hence the motion of their internuclear vector is less susceptible to high-frequency local vibrations. Typical quantum corrections are less than 1% for

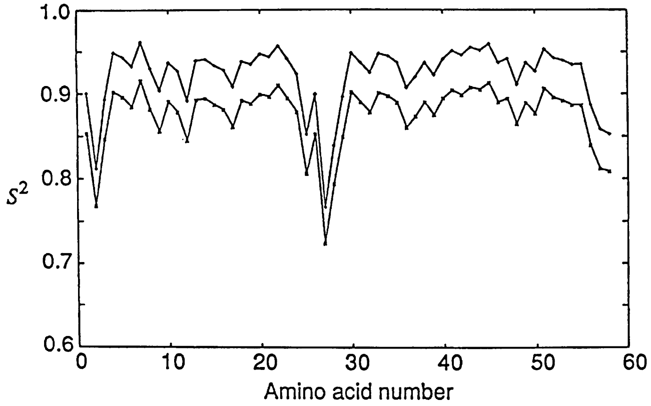


Fig. 3. Theoretical heteronuclear angular order parameters S^2 of the C^{α}-H^{α} atom pairs of the BPTI backbone calculated from a normal mode analysis at a temperature of 309 K [87]. The upper curve corresponds to classical statistics and the lower curve to quantum statistics.

interresidue pairs [45], and below 3% for geminal proton pairs [87]. Molecular dynamics simulations predict order parameters with similar trends but with larger deviations from unity, indicative of motions affecting spin relaxation that are not included in the normal mode picture [45,82].

NMR refinement. As in crystallographic refinement, the quasiharmonic normal mode description can be viewed as a model for molecular motions containing a relatively small number of adjustable parameters (the effective frequencies of the low-lying modes) that can be fit to experimental data. In particular, this approach offers a way to analyze both heteronuclear and homonuclear spin relaxation parameters in a common framework. Figure 4 shows some sample calculations of N-H and

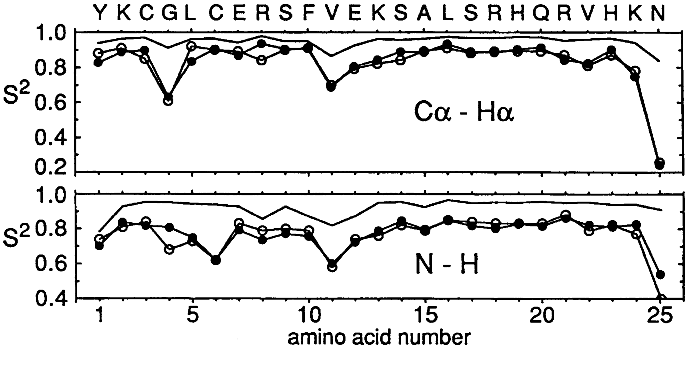


Fig. 4. Order parameters for a zinc-finger peptide. Solid line: results from normal mode analysis; open circles: results from solvated molecular dynamics simulation; filled circles: normal mode values with adjusted frequencies. Adapted from Ref. 88.

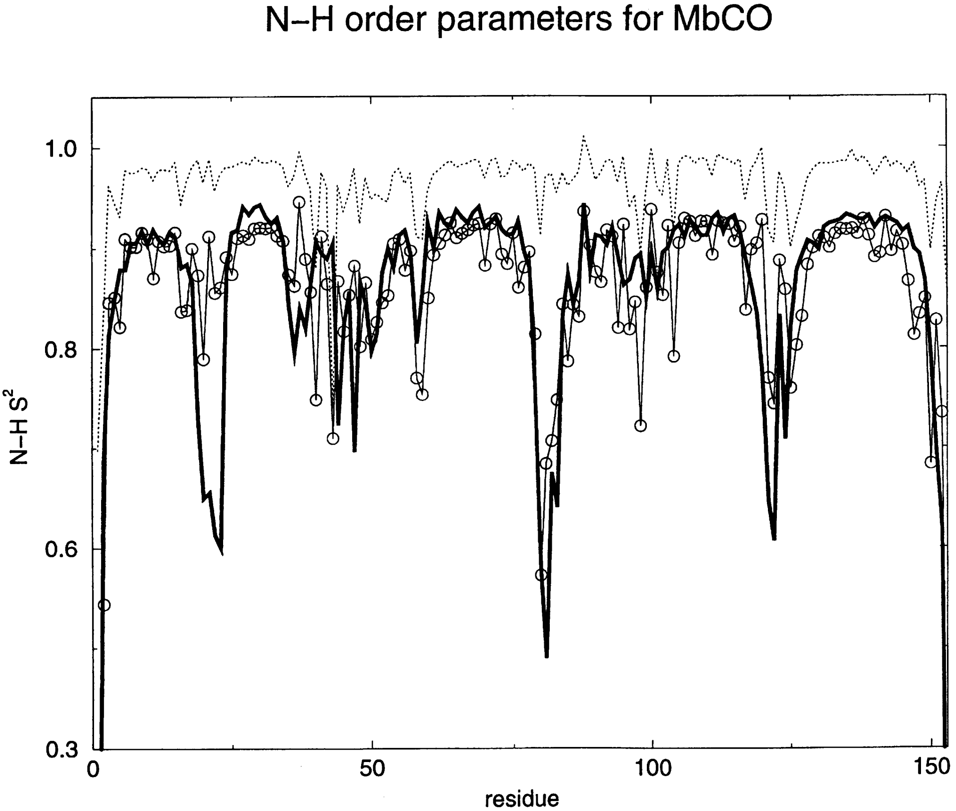


Fig. 5. Order parameters for myoglobin-CO. Dotted line: results from normal mode analysis; solid line: results from 1.5 ns solvated molecular dynamics simulation; open circles: normal mode values with adjusted frequencies as described in the text.

C^{α}-H order parameters for a zinc-finger peptide [88], where a solvated molecular dynamics simulation [45] was used to generate the 'experimental' order parameters. The upper curve shows order parameters computed from a normal mode analysis; as is typical, the fluctuations are underestimated in this approximation compared to those seen in the solvated dynamics simulation. The open circles show the MD target values, and the filled circles show the fitted results with the adjustment of 250 frequencies. The fit is essentially perfect, except for the N-H order parameter for residue 4, indicating that the low-frequency space provides a useful expansion space for describing motions involved in ^{15}N and ^{13}C relaxation in proteins.

A similar, but more challenging, example is shown in Fig. 5, which looks at N-H order parameters in myoglobin. Again, the 'experimental' results are from a solvated molecular dynamics simulation of 1.5 ns duration [89]. The upper curve shows the order parameters computed from a normal mode calculation, and the open circles show the results when the amplitudes and mixing parameters of the 26 lowest modes are adjusted, along with a 'global' scaling of frequencies under 1000 cm^{-1} (i.e. a single additional scale factor multiplied these normal mode frequencies). This is probably a somewhat more realistic test of this method, since the target order parameters vary over a wider range (which is probably more representative of larger proteins) and because fewer modes (and hence fewer adjustable parameters) were used in the fitting procedure. Overall, the character of the fitted motion fits the target values fairly well, except for a dip near residue 24 (between the A and B helices) where even a scaling of the low-frequency amplitude parameters cannot yield order parameters much below 0.8.

The adjusted frequency normal mode models illustrated in Figs. 4 and 5 provide a description of the dynamics that can also be used to predict motional contributions to proton–proton nuclear Overhauser peaks that are involved in structure determination. In principle, this provides an approach by which protein structure and dynamics can be refined against both homonuclear (proton–proton) and heteronuclear (^{15}N and ^{13}C) NMR relaxation data. Preliminary studies suggest that many features of the dynamical scaling factors for proton pairs are reproduced by dynamical models fit to heteronuclear (^{15}N and ^{13}C) relaxation data, but that large correction factors are often underestimated [88]. It seems likely that some combination of normal mode analysis with models that allow larger conformational transitions (such as jumps between alternate side-chain conformers) will be required to carry out realistic refinements of this sort.

Conclusions

Normal mode analyses continue to occupy an important niche in the dynamical analyses of biomolecules by providing a compact and analytical representation of an important limiting case. The directions of the low-frequency modes often provide useful quantities for the description of correlated motion even in the presence of significant anharmonicity. Extensions to disordered or significantly anharmonic systems provide interesting insights into protein dynamics, and suggest new approaches to the analysis of experiments and molecular dynamics simulations.

Acknowledgements

This work was supported by NIH Grant GM-38794. I thank Rafael Brüschweiler and Jon Baker for helpful discussions.

References

1. Van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993.
2. McCammon, J.A., Wolynes, P.G. and Karplus, M., Biochemistry, 18(1979)928.

3. Horiuchi, T. and Gō, N., Proteins Struct. Funct. Genet., 10(1991)106.
4. Goldstein, H., Classical Mechanics, Addison-Wesley, Reading, MA, 1980.
5. Levitt, M., Sander, C. and Stern, P.S., J. Mol. Biol., 181(1985)423.
6. Gō, N., Noguti, T. and Nishikawa, T., Proc. Natl. Acad. Sci. USA, 80(1983)3696.
7. Brooks, B. and Karplus, M., Proc. Natl. Acad. Sci. USA, 80(1983)6571.
8. Tidor, B., Irikura, K.K., Brooks, B.R. and Karplus, M., J. Biomol. Struct. Dyn., 1(1983)231.
9. Rudolph, B.R. and Case, D.A., Biopolymers, 28(1989)851.
10. Sunada, S. and Gō, N., J. Comput. Chem., 16(1995)328.
11. Wako, H., Endo, S., Nagayama, K. and Gō, N., Comput. Phys. Commun., 91(1995)233.
12. Hao, M.-H. and Harvey, S.C., Biopolymers, 32(1992)1393.
13. Hao, M.-H. and Scheraga, H.A., Biopolymers, 34(1994)321.
14. Brooks, B.R., Janezic, D. and Karplus, M., J. Comput. Chem., 16(1995)1522.
15. Kitao, A., Hayward, S. and Gō, N., Biophys. Chem., 52(1994)107.
16. Janezic, D. and Brooks, B.R., J. Comput. Chem., 16(1995)1543.
17. McQuarrie, D.A., Statistical Mechanics, Harper & Row, New York, NY, 1976.
18. Forgarasi, G., Zhou, X., Taylor, P.W. and Pulay, P., J. Am. Chem. Soc., 114(1992)8191.
19. Pulay, P. and Fogarasi, G., J. Chem. Phys., 96(1992)2856.
20. Baker, J., J. Comput. Chem., 9(1993)1085.
21. Lamm, G. and Szabo, A., J. Chem. Phys., 85(1986)7334.
22. Kitao, A., Hirata, F. and Gō, N., Chem. Phys., 158(1991)447.
23. Hayward, S., Kitao, A., Hirata, F. and Gō, N., J. Mol. Biol., 234(1993)1207.
24. Kottalam, J. and Case, D.A., Biopolymers, 29(1990)1409.
25. Liu, F., Horton, J., Mayne, C.L., Xiang, T. and Grant, D.M., J. Am. Chem. Soc., 114(1992)5281.
26. Smith, J.C., Q. Rev. Biophys., 24(1991)227.
27. McCammon, J.A., Gelin, B., Karplus, M. and Wolynes, P.G., Nature, 262(1976)325.
28. Gibrat, J.-F. and Gō, N., Proteins Struct. Funct. Genet., 8(1990)258.
29. Levy, R.M., Karplus, M., Kushick, J. and Perahia, D., Macromolecules, 17(1984)1370.
30. Teeter, M.M. and Case, D.A., J. Phys. Chem., 94(1990)8091.
31. Janezic, D., Venable, R.M. and Brooks, B.R., J. Comput. Chem., 16(1995)1554.
32. Karplus, M. and Kushick, J.N., Macromolecules, 14(1981)325.
33. DiNola, A., Berendsen, H.J.C. and Edholm, O., Macromolecules, 17(1984)2044.
34. Tidor, B. and Karplus, M., Proteins Struct. Funct. Genet., 15(1993)71.
35. Tidor, B. and Karplus, M., J. Mol. Biol., 238(1994)405.
36. Perahia, D., Levy, R.M. and Karplus, M., Biopolymers, 29(1990)645.
37. Ichiye, T. and Karplus, M., Proteins Struct. Funct. Genet., 2(1987)236.
38. Kuriyan, J., Petsko, G.A., Levy, R.M. and Karplus, M., J. Mol. Biol., 190(1986)227.
39. García, A.E., Phys. Rev. Lett., 68(1992)2696.
40. Kitao, A., Hirata, F. and Gō, N., J. Phys. Chem., 97(1993)10231.
41. Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C., Proteins, 17(1993)412.
42. Bialek, W. and Goldstein, R.F., Biophys. J., 48(1985)1027.
43. Roitberg, A., Gerber, R.B., Elber, R. and Ratner, M.A., Science, 268(1995)1319.
44. Brüschweiler, R., J. Chem. Phys., 102(1995)3396.
45. Palmer, A.G. and Case, D.A., J. Am. Chem. Soc., 114(1992)9059.
46. Becker, O.M. and Karplus, M., Phys. Rev. Lett., 70(1993)3514.
47. Xu, D., Schulten, K., Becker, O.M. and Karplus, M., J. Chem. Phys., 103(1995)3112.
48. Xu, D. and Schulten, K., J. Chem. Phys., 103(1995)3124.

49. Schulten, K., Lu, H. and Bai, L., In Flyvbjerg, H., Hertz, J., Jensen, M.H., Mouritsen, O.G. and Sneppen, K. (Eds.) Physics of Biological Systems – From Molecules to Species, Springer, New York, NY, 1997, pp. 117–152.
50. Stillinger, F.H. and Weber, T.A., Science, 225(1984)983.
51. Frauenfelder, H., Sligar, S.G. and Wolynes, P.G., Science, 254(1991)1598.
52. Thomas, A., Roux, B. and Smith, J.C., Biopolymers, 33(1993)1249.
53. LaViolette, R.A. and Stillinger, F.H., J. Chem. Phys., 83(1985)4079.
54. Seeley, G. and Keyes, T., J. Chem. Phys., 91(1989)5581.
55. Buchner, M., Ladanyi, B.M. and Stratt, R.M., J. Chem. Phys., 97(1992)8522.
56. Madan, B. and Keyes, T., J. Chem. Phys., 98(1993)3342.
57. Keyes, T., J. Chem. Phys., 103(1995)9810.
58. Straub, J.E. and Thirumalai, D., Proc. Natl. Acad. Sci. USA, 90(1993)809.
59. Straub, J.E., Rashkin, A.B. and Thirumalai, D., J. Am. Chem. Soc., 116(1994)2049.
60. Gerstein, M., Lesk, A.M. and Chothia, C., Biochemistry, 33(1994)6739.
61. Case, D.A., Curr. Opin. Struct. Biol., 4(1994)285.
62. Tirion, M.M. and Ben-Avraham, D., J. Mol. Biol., 230(1993)186.
63. Ikura, T. and Gō, N., Proteins, 16(1993)423.
64. Furois-Corbin, S., Smith, J.C. and Kneller, G.R., Proteins, 16(1993)141.
65. Brooks III, C.L. and Case, D.A., Chem. Rev., 93(1993)2487.
66. De Loof, H., Harvey, S.C., Segrest, J.P. and Pastor, R.W., Biochemistry, 30(1991)2099.
67. Clarage, J.B., Romo, T., Andrews, B.K., Pettitt, B.M. and Phillips Jr., G.N., Proc. Natl. Acad. Sci. USA, 92(1995)3288.
68. Chandrasekhar, I., Clore, G.M., Szabo, A., Gronenborn, A.M. and Brooks, B.R., J. Mol. Biol., 226(1992)239.
69. Balsera, M.A., Wriggers, W., Oono, Y. and Schulten, K., J. Phys. Chem., 100(1996)2567.
70. Hünenberger, P.H., Mark, A.W. and van Gunsteren, W.F., J. Mol. Biol., 252(1995)492.
71. Diamond, R., Acta Crystallogr., Sect. A, 46(1990)425.
72. Kidera, A. and Gō, N., J. Mol. Biol., 225(1992)457.
73. Kidera, A., Inaka, K., Matsushima, M. and Gō, N., J. Mol. Biol., 225(1992)477.
74. Kidera, A., Inaka, K., Matsushima, M. and Gō, N., Protein Sci., 3(1994)92.
75. Brünger, A.T., Nature, 355(1991)472.
76. Benoit, J.-P. and Doucet, J., Q. Rev. Biophys., 28(1995)131.
77. Clarage, J.B., Clarage, M.S., Phillips, W.C., Sweet, R.M. and Caspar, D.L.D., Proteins Struct. Funct. Genet., 12(1992)145.
78. Paure, F., Micu, A., Pérahia, D., Doucet, J., Smith, J.C. and Benoit, J.-P., Nature Struct. Biol., 1(1994)124.
79. Mizuguchi, K., Kidera, A. and Gō, N., Proteins, 18(1994)34.
80. Smith, J.C. and Kneller, G.R., Mol. Sim., 10(1993)363.
81. Kneller, G.R., Doster, W., Settles, M., Cusack, S. and Smith, J.C., J. Chem. Phys., 97(1992)8864.
82. Brüschweiler, R. and Case, D.A., Prog. NMR Spectrosc., 26(1994)27.
83. Henry, E.R. and Szabo, A., J. Chem. Phys., 82(1985)4753.
84. Lipari, G. and Szabo, A., J. Am. Chem. Soc., 104(1982)4546.
85. Brüschweiler, R., Roux, B., Blackledge, M., Griesinger, C., Karplus, M. and Ernst, R.R., J. Am. Chem. Soc., 114(1992)2289.
86. Post, C.B., J. Mol. Biol., 224(1992)1087.
87. Brüschweiler, R., J. Am. Chem. Soc., 114(1992)5341.
88. Brüschweiler, R. and Case, D.A., Phys. Rev. Lett., 72(1994)940.
89. Hirst, J.D. and Brooks III, C.L., J. Mol. Biol., 243(1994)173.

Part IV
Simulation of large systems

Dynamics of biomolecules: Simulation versus X-ray, neutron and infrared experiment

Jeremy C. Smith
Molecular Simulation Group, SBPM/DBCM, Commissariat à l'Energie Atomique,
CEA-Saclay, F-91191 Gif-sur-Yvette Cedex, France

1. Introduction

In the present chapter we examine the use of computer simulation in the interpretation of experiments on the dynamics in condensed phases of small molecules, polymers and biological macromolecules. Computer simulation provides a stepping stone between experiment and simplified descriptions of the physical behaviour of complex systems. There are two stages involved in this. The first is the comparison of physical quantities that are measurable or derivable from measurements with the same quantities derived from simulation. The second is the interpretation of the experimental results using the detailed information present in the simulation.

The comparison with experiment can be made at several levels. The first, and most common, is in the comparison of 'derived' quantities that are not directly measurable: for example, a set of average crystal coordinates or a diffusion constant. A comparison at this level is convenient in that the quantities involved describe directly the structure and dynamics of the system. However, the obtention of these quantities, from experiment and/or simulation, may involve approximation and model-dependent data analysis. For example, to obtain experimentally a set of average crystallographic coordinates, the imposition of a physical model to interpret an electron density map is required. To reduce these problems a comparison can be made at the level of the measured quantities themselves, such as diffraction intensities, dynamic structure factors or absorption coefficients. A comparison at this level still involves some approximation. For example, background corrections have to be made in the experimental data reduction. However, fewer approximations are necessary as to the structure and dynamics of the sample itself and comparison with experiment is normally more direct. This approach requires a little more work on the part of the computer simulation team, as methods for calculating experimental intensities from simulation configurations must be developed.

Having made the comparison with experiment, one may then make an assessment as to whether the simulation agrees sufficiently well to be useful in interpreting the experiment in detail. In cases where the agreement is not good, the determination of the cause of the discrepancy is often instructive. The errors may arise from the simulation model or from the assumptions used in the experimental data reduction, or both. In the case where the quantities examined agree, the simulation can be decomposed so as to isolate the principal components responsible for the observed intensities. Sometimes the dynamics involved can be described by an analytical model.

The spectroscopic techniques that have been most frequently used to investigate biomolecular dynamics are those that are commonly available in laboratories, e.g. nuclear magnetic resonance (NMR), fluorescence and Mössbauer spectroscopies. However, these methods involve motions on timescales that are not well sampled by molecular dynamics simulation using present computer power. Moreover, the establishment of relations linking NMR and fluorescence with atomic motion is fraught with theoretical difficulties. The experimental techniques examined here were chosen for their suitability for the examination of motions presently accessible to atomic-detail computer simulation. Far-infrared and neutron spectroscopy probe dynamics on subnanosecond timescales that can be well sampled with present-day molecular dynamics simulations. Moreover, underlying relations between dynamics and measurement are relatively easy to express formally for these techniques. Neutron scattering gives information on self- and cross-correlations in atomic motions. Far-infrared spectroscopy provides a description of charge fluctuations and is thus a promising tool for examining the pervasive problem of modelling electrostatics in biomolecular simulation. X-ray crystallography is also examined here, as a method that does not as yet give temporal dynamical information but which, when combined with computer simulation, is a potentially powerful probe of atomic fluctuations. All three of these techniques share the property of using expensive sources not commonly available in the laboratory. Neutrons are produced by a nuclear reactor or spallation source. The X-ray and far-infrared experiments discussed here were performed using intense synchrotron radiation, although in favourable cases laboratory sources may also prove to be useful.

This chapter is divided into three sections. In the first, the basic principles of X-ray and neutron scattering and far-infrared absorption and their relations with computer simulation are examined. In the last two sections, examples of the interpretation of the experiments are given using harmonic analyses and molecular dynamics simulation performed with the CHARMM program and potential function [1]. In the second section, a variety of atomic motions is examined in a number of condensed-phase systems: small-molecule liquids and crystals, and polymer crystals. The molecular crystal dynamics examined involves vibrational and diffusive motions. The combined simulation–experimental procedure is particularly useful for the characterization of the dynamics of biological macromolecules, where the direct interpretation of experiments using simplified models is made difficult by the diversity of the vibrational and diffusive motions present. In the final section, therefore, we derive a picture of global, picosecond (ps)-timescale protein dynamics by combining X-ray and neutron scattering experiments with harmonic and molecular dynamics calculations.

2. X-ray, neutron and far-infrared experiments and their relation to simulation

2.1. Dynamic structure factor

We first examine the relation between particle dynamics and the scattering of radiation in the case where both the energy and momentum transferred between the sample and the incident radiation are measured. Linear response theory allows dynamic structure factors to be written in terms of equilibrium fluctuations of the sample. For neutron scattering from a system of identical particles, this is as follows [2–4]:

$$S_{coh}(\vec{Q}, \omega) = \frac{1}{2\pi} \int\int dt\, d^3r\, e^{i(\vec{Q}\cdot\vec{r} - \omega t)} G(\vec{r}, t) \tag{1}$$

$$S_{inc}(\vec{Q}, \omega) = \frac{1}{2\pi} \int\int dt\, d^3r\, e^{i(\vec{Q}\cdot\vec{r} - \omega t)} G_s(\vec{r}, t) \tag{2}$$

where $\vec{Q}$ is the scattering wavevector, ω is the energy transfer, and the subscripts coh and inc refer to coherent and incoherent scattering, discussed later. $G_s(\vec{r}, t)$ and $G(\vec{r}, t)$ are van Hove correlation functions which, for a system of N particles undergoing classical dynamics, are defined as follows:

$$G(\vec{r}, t) = \frac{1}{N} \sum_{i,j} \langle \delta(\vec{r} - \vec{R}_i(t) + \vec{R}_j(0)) \rangle \tag{3}$$

$$G_s(\vec{r}, t) = \frac{1}{N} \sum_{i} \langle \delta(\vec{r} - \vec{R}_i(t) + \vec{R}_i(0)) \rangle \tag{4}$$

where $\vec{R}_i(t)$ is the position vector of the ith scattering nucleus and $\langle \cdots \rangle$ indicates an ensemble average.

$G(\vec{r}, t)$ is the probability that, given a particle at the origin at time $t = 0$, any particle (including the original particle) is at $\vec{r}$ at time t. $G_s(\vec{r}, t)$ is the probability that, given a particle at the origin at time $t = 0$, the same particle is at $\vec{r}$ at time t.

Equation 1 has an equivalent form in X-ray scattering, where the scattered intensity is given as follows [5]:

$$|F(\vec{Q}, \omega)|^2 = \frac{1}{2\pi} \int\int d^3r\, dt\, P(\vec{r}, t)\, e^{i(\vec{Q}\cdot\vec{r} - \omega t)} \tag{5}$$

where $P(\vec{r}, t)$ is the spatiotemporal Patterson function given by

$$P(\vec{r}, t) = \int\int d\vec{R}\, dt\, \rho(\vec{R}, t) \rho(\vec{r} + \vec{R}, t + \tau) \tag{6}$$

and $\rho(\vec{r}, t)$ is the time-dependent electron density. Unfortunately, X-ray photons with wavelengths corresponding to atomic distances have energies much higher than those associated with thermal fluctuations. For example, an X-ray photon of 1.8 Å wavelength has an energy of 6.9 keV corresponding to a temperature of 8×10^7 K. X-ray detectors have not been sufficiently sensitive to measure the minute fractional energy changes associated with molecular fluctuations, and so the practical exploitation of Eq. 5 has been difficult. Therefore, in subsequent discussions of X-ray scattering we examine only the cases where inelastic and elastic scattering are indistinguishable experimentally. In contrast to X-rays, the mass of the neutron is such that the energy exchanged in exciting or de-exciting ps-timescale thermal motions is a large fraction of the incident energy and can be measured precisely. A thermal neutron of 1.8 Å wavelength has an energy of 25 meV corresponding to k_BT at 300 K. To further examine the neutron scattering case, we perform space Fourier transformation of the van Hove correlation functions:

$$S_{coh}(\vec{Q}, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt\, e^{-i\omega t} I_{coh}(\vec{Q}, t) \tag{7}$$

$$I_{coh}(\vec{Q}, t) = \frac{1}{N} \sum_{i,j} b^{*}_{i,coh} b_{j,coh} \langle e^{-i\vec{Q}\cdot\vec{R}_i(0)} e^{i\vec{Q}\cdot\vec{R}_j(t)} \rangle \tag{8}$$

$$S_{inc}(\vec{Q}, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt\, e^{-i\omega t} I_{inc}(\vec{Q}, t) \tag{9}$$

$$I_{inc}(\vec{Q}, t) = \frac{1}{N} \sum_{i} b^{2}_{i,inc} \langle e^{-i\vec{Q}\cdot\vec{R}_i(0)} e^{i\vec{Q}\cdot\vec{R}_i(t)} \rangle \tag{10}$$

Neutrons are scattered by the nuclei of the sample. Due to the random distribution of nuclear spins in the sample, the scattered intensity will contain a *coherent* part arising from the average neutron–nucleus potential and an *incoherent* part arising from fluctuations from the average. The coherent scattering arises from self- and cross-correlations of atomic motions and the incoherent scattering arises from single atom motions. Each isotope has a coherent scattering length $b_{i,coh}$ and an incoherent scattering length $b_{i,inc}$ which defines the strength of the interaction between the nucleus of the atom and the neutron. We see from Eqs. 7 and 9 that the coherent and incoherent dynamic structure factors are time Fourier transforms of the coherent and incoherent *intermediate scattering functions* $I_{coh}(\vec{Q}, t)$ and $I_{inc}(\vec{Q}, t)$. $S_{inc}(\vec{Q}, \omega)$ and $S_{coh}(\vec{Q}, \omega)$ may contain elastic ($\omega = 0$) and inelastic ($\omega \neq 0$) parts. The elastic scattering process probes correlations of atomic positions at long times, whereas the inelastic scattering process probes position correlations as a function of time.

2.2. Incoherent neutron scattering

Neutron scattering from organic molecules is dominated by incoherent scattering from the hydrogen atoms. This is largely because the incoherent scattering cross-section ($4\pi b^2_{inc}$) of hydrogen is $\simeq 15$ times greater than the total scattering cross-sections of carbon, nitrogen or oxygen. In the systems examined here, incoherent scattering thus essentially gives information on the self-correlations of hydrogen atom motions.

The intermediate scattering functions are quantum-mechanical time-correlation functions that are replaced by classical time-correlation functions if they are calculated from molecular dynamics simulations. This leads to a problem with the detailed balance condition, which relates the intensities of neutron energy loss and gain processes as follows:

$$S_{inc}(\vec{Q}, \omega) = e^{\beta\hbar\omega} S_{inc}(-\vec{Q}, -\omega) \tag{11}$$

where $\beta = 1/k_BT$.

The detailed balance condition does not hold in the classical limit ħ → 0. To correct for this, one can apply a semiclassical formula (given here for isotropic systems, such as for polycrystalline, powder or solution samples [4]):

$$S_{inc}(Q, \omega) \approx \frac{\beta\hbar\omega}{1 - e^{-\beta\hbar\omega}} S_{inc,cl}(Q, \omega) \tag{12}$$

The semiclassical correction (Eq. 12) is an approximation valid only in the linear response regime $\hbar\omega < k_BT$. A program for calculating neutron scattering properties from molecular dynamics simulations has recently been published [6].

In practice, the measured incoherent scattering energy spectrum is divided into elastic, quasielastic and inelastic scattering. Inelastic scattering arises from vibrations. Quasielastic scattering is typically Lorentzian or a sum of Lorentzians centred on $\omega = 0$, and arises from diffusive motions in the sample. Elastic scattering gives information on the self-probability distributions of the hydrogen atoms in the sample. We now examine these forms of scattering in more detail.

2.2.1. Quasielastic incoherent scattering

It is useful for subsequent analysis to review here the procedure commonly used to extract dynamical data directly from experimental incoherent quasielastic neutron scattering profiles [7]. It is assumed that the atomic position vectors can be decomposed into two contributions, one due to diffusive motion, $\vec{r}_{i,d}(t)$, and the other from vibrations, $\vec{u}_{i,v}(t)$, i.e.

$$\vec{R}_i(t) = \vec{r}_{i,d}(t) + \vec{u}_{i,v}(t) \tag{13}$$

Combining Eq. 13 with Eq. 10 and assuming that $\vec{r}_{i,d}(t)$ and $\vec{u}_{i,v}(t)$ are uncorrelated, one obtains

$$I_{inc}(\vec{Q}, t) = I_d(\vec{Q}, t)\, I_v(\vec{Q}, t) \tag{14}$$

where $I_d(\vec{Q}, t)$ and $I_v(\vec{Q}, t)$ are obtained by substituting $\vec{R}_i(t)$ in Eq. 10 with $\vec{r}_{i,d}(t)$ and $\vec{u}_{i,v}(t)$, respectively.

The Fourier transform of Eq. 14 gives

$$S(\vec{Q}, \omega) = S_d(\vec{Q}, \omega) \otimes S_v(\vec{Q}, \omega) \tag{15}$$

where $S_d(\vec{Q}, \omega)$ and $S_v(\vec{Q}, \omega)$ are obtained by the Fourier transformation of $I_d(\vec{Q}, t)$ and $I_v(\vec{Q}, t)$ and the symbol ⊗ denotes the convolution product.

The vibrational intermediate scattering function is given by [4]:

$$I_v(\vec{Q}, t) = \sum_i b_i^2 e^{-\langle(\vec{Q}\cdot\vec{u}_{i,v})^2\rangle} e^{\langle[\vec{Q}\cdot\vec{u}_{i,v}(0)][\vec{Q}\cdot\vec{u}_{i,v}(t)]\rangle} \tag{16}$$

To derive an analytically tractable form of $S_{inc}(\vec{Q}, \omega)$ in the quasielastic energy window (typically $-15\,\mathrm{cm}^{-1} < \hbar\omega < 15\,\mathrm{cm}^{-1}$), we align $\vec{Q}$ with the Cartesian axis x in the laboratory frame. Assuming that (i) $\langle u_{v,x}^2 \rangle$, the x-axis vibrational mean-square displacement, is the same for all the hydrogens and (ii) $Q^2\langle u_{v,x}^2 \rangle \ll 1$, $S_v(\vec{Q}, \omega)$ can be expressed as follows:

$$S_v(\vec{Q}, \omega) = e^{-Q^2\langle u_{v,x}^2 \rangle}[\delta(\omega) + S_v^{inel}(\vec{Q}, \omega)] \tag{17}$$

where $e^{-Q^2\langle u_{v,x}^2 \rangle} S_v^{inel}(\vec{Q}, \omega)$ is the vibrational inelastic dynamic structure factor. Combining Eqs. 15 and 17 one obtains

$$S_{inc}(\vec{Q}, \omega) = e^{-Q^2\langle u_{v,x}^2 \rangle}[S_d(\vec{Q}, \omega) + S_d(\vec{Q}, \omega) \otimes S_v^{inel}(\vec{Q}, \omega)] \tag{18}$$

$S_v^{inel}(\vec{Q}, \omega)$ will contain high-frequency inelastic peaks due to intramolecular vibrations that fall outside the quasielastic energy window and may also contain intensity within the energy window. The contribution within the energy window is assumed to be due to the lattice phonon background. The density of states, $g(\omega)$, corresponding to the latter contribution is assumed to be given by the Debye model, i.e. $g(\omega) \propto \omega^2$. Given that $Q^2\langle u_{v,x}^2 \rangle \ll 1$, then $S_v^{inel}(\vec{Q}, \omega) \sim Q^2\langle u_{v,x}^2 \rangle g(\omega)$. If $\langle u_{v,x}^2 \rangle \propto \omega^{-2}$ in the quasielastic region, then $S_v^{inel}(\vec{Q}, \omega) \propto Q^2$. We therefore represent $S_d(\vec{Q}, \omega) \otimes S_v^{inel}(\vec{Q}, \omega)$ as an energy-independent background, $B(\vec{Q})$, leading to the equation

$$S_{inc}(\vec{Q}, \omega) = e^{-Q^2\langle u_{v,x}^2 \rangle} S_d(\vec{Q}, \omega) + B(\vec{Q}) \tag{19}$$

In directionally averaged versions of Eq. 19, the above mean-square displacement is replaced by the corresponding two- or three-dimensional quantity divided by a factor of 2 in the two-dimensional case and six for a spherically averaged dynamic structure factor.

2.2.2. Elastic incoherent structure factor

$I_d(\vec{Q}, t)$ can be separated into time-dependent and time-independent parts as follows:

$$I_d(\vec{Q}, t) = A_0(\vec{Q}) + I'_d(\vec{Q}, t) \tag{20}$$

The elastic incoherent structure factor (EISF), $A_0(\vec{Q})$, is defined as follows [7]:

$$A_0(\vec{Q}) = \lim_{t \to \infty} I_d(\vec{Q}, t) = \int d^3r\, e^{i\vec{q}\cdot\vec{r}} \lim_{t \to \infty} G_d(\vec{r}, t) \tag{21}$$

where $G_d(\vec{r}, t)$ is the contribution to the van Hove self-correlation function due to diffusive motion. $A_0(\vec{Q})$ is thus determined by the diffusive contribution to the space probability distribution of the hydrogen nuclei.

Taking the Fourier transform of Eq. 20 and combining it with Eq. 19 yields

$$S(\vec{Q}, \omega) = e^{-Q^2\langle u_v^2 \rangle}[A_0(\vec{Q})\delta(\omega) + S'_d(\vec{Q}, \omega)] + B(\vec{Q}) \tag{22}$$

This equation contains three terms, representing the elastic $[e^{-Q^2\langle u_v^2 \rangle} A_0(\vec{Q})\delta(\omega)]$, quasielastic $[e^{-Q^2\langle u_v^2 \rangle} S'_d(\vec{Q}, \omega)]$ and inelastic $[B(\vec{Q})]$ scattering.

Experimentally, the scattering spectra will have a finite energy resolution, given by a resolution function, $R(\omega)$. Incorporating this effect in Eq. 22, the dynamic structure factor becomes

$$S(\vec{Q}, \omega) = e^{-Q^2\langle u_v^2\rangle}[A_0(\vec{Q})\,R(\omega) + S_d'(\vec{Q}, \omega) \otimes R(\omega)] + B(\vec{Q}) \tag{23}$$

$A_0(\vec{Q})$ and $S'(\vec{Q}, \omega)$ may be extracted from experiment by fitting Eq. 23 to the measured scattering profiles. For this, it is necessary to assume *a priori* parametric forms for $A_0(\vec{Q})$ and $S'(\vec{Q}, \omega)$; these depend on the dynamical model that one wishes to fit. Several such models, such as continuous diffusion on a circle or sphere or jumps between sites, are described in Ref. 7. It turns out that $A_0(\vec{Q})$ and $S'(\vec{Q}, \omega)$ obtained from experiment may also depend on the instrumental resolution function. If slow motions occur in the system, the dynamic structure factor may contain quasielastic contributions with widths much narrower than that of $R(\omega)$. These contributions will then be experimentally indistinguishable from the elastic scattering and the extracted experimental EISF will be different from the EISF in the long-time limit.

Extraction of $A_0(\vec{Q})$ from a molecular dynamics simulation. This assumes that we are able to determine the diffusive contribution to the atomic trajectories. In this case the EISF can be obtained in two ways: from the long-time limit of $I_d(\vec{Q}, t)$ using Eq. 21 or, assuming that the position vector of any given atom is uncorrelated with itself at infinite time, the EISF can be written as follows (cf. Eq. 10):

$$A_0(\vec{Q}) = \sum_i b_i^2 |\langle e^{i\vec{Q}\cdot\vec{r}_{i,d}}\rangle|^2 \tag{24}$$

If the full molecular dynamics trajectories are used, without separation into diffusive and nonvibrational components, a different EISF, which we call $A_{0,\mathrm{tot}}(\vec{Q})$, that includes contributions from all types of motions can be calculated:

$$A_{0,\mathrm{tot}}(\vec{Q}) = \sum_i b_i^2 |\langle e^{i\vec{Q}\cdot\vec{R}_i}\rangle|^2 \tag{25}$$

Given the assumptions used in deriving Eq. 22, we can write

$$A_{0,\mathrm{tot}}(\vec{Q}) = e^{-Q^2\langle u_v^2\rangle} A_0(\vec{Q}) \tag{26}$$

2.2.3. *Inelastic incoherent scattering*

Scattering intensity. For a system executing harmonic dynamics, the transform in Eq. 2 can be performed analytically and the result expanded in a power series over the normal modes in the sample. The following expression is obtained:

$$S_{\mathrm{inc}}(\vec{Q}, \omega) = \sum_i b_{\mathrm{inc}}^2 \exp[-2W_i(\vec{Q})] \prod_\lambda \left[\sum_{n_\lambda} \exp(n_\lambda \hbar\omega_\lambda \beta/2) \times I_{n_\lambda}\left(\frac{\hbar(\vec{Q}\cdot\vec{e}_{\lambda,i})^2}{2M\omega_\lambda \sinh(\hbar\omega_\lambda\beta/2)}\right)\right] \delta\left(\omega - \sum_\lambda n_\lambda \omega_\lambda\right) \tag{27}$$

In Eq. 27, M is the hydrogen mass, λ labels the mode, $\vec{e}_{\lambda,i}$ is the atomic eigenvector for hydrogen i in mode λ, and ω_λ is the mode angular frequency. n_λ is the number of quanta of energy $\hbar\omega_\lambda$ exchanged between the neutron and mode λ.

$W_i(\vec{Q})$ is the exponent of the Debye–Waller factor, $\exp[-2W_i(\vec{Q})]$, for hydrogen atom i and is given as follows:

$$2W_i(\vec{Q}) = \frac{1}{2NM}\sum_\lambda \frac{\hbar(\vec{Q}\cdot\vec{e}_{\lambda,i})^2}{\omega_\lambda}[2n(\omega_\lambda)+1] = Q^2 < u^2_{Q,i} > \tag{28}$$

In Eq. 28, N is the number of modes, $n(\omega_\lambda)$ is the Bose occupancy and $< u^2_{Q,i} >$ is the mean-square displacement for atom i in the direction of $\vec{Q}$.

Equation 27 is an exact quantum-mechanical expression for the scattered intensity. A detailed interpretation of this equation is given in Ref. 8. Inserting the calculated eigenvectors and eigenvalues in the equation allows the calculation of the incoherent scattering in the harmonic approximation for processes involving any desired number of quanta exchanged between the neutrons and the sample, e.g. one-phonon scattering involving the exchange of one quantum of energy $\hbar\omega_\lambda$, two-phonon scattering, and so on.

The label λ in Eq. 27 runs over all the modes of the sample. In the cases examined here, normal-mode analyses have been performed for isolated molecules (proteins) and molecular crystals. In the case of an isolated molecule, λ runs over the $3N-6$ normal modes of the molecule, where N is the number of atoms. In the case of a crystal, λ runs over the phonon modes in the asymmetric unit of the first Brillouin zone.

Vibrational density of states. The vibrational density of states, $G(\omega)$, is related to the classical dynamical structure factor by

$$\begin{aligned} G(\omega) &= \lim_{Q\to 0}\frac{\omega^2}{Q^2}S_{cl}(Q,\omega) \\ &= \frac{1}{N}\sum_i \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, e^{-i\omega t}\langle \vec{v}_i(0)\vec{v}_i(t)\rangle \end{aligned} \tag{29}$$

where $\langle \vec{v}_i(0)\vec{v}_i(t)\rangle$ is the autocorrelation function of the velocity, $\vec{v}_i(t)$ of atom i. $g(\omega)$ is the kinetic energy of the hydrogen atoms in the system as a function of frequency. Equation 29 holds formally also in the quantum case. Experimentally, $G(\omega)$ can be obtained in principle by performing the extrapolation of $S(\vec{Q},\omega)$ to $Q=0$. From molecular dynamics simulations, $G(\omega)$ can be calculated as the Fourier transform of the velocity autocorrelation function. For harmonic analysis, $G(\omega)$ is simply the cross-section-weighted frequency distribution, i.e.

$$G(\omega) = \sum_{i,\lambda}\frac{b^2_{inc}|\vec{e}_{\lambda,i}|^2}{m_i}\delta(\omega-\omega_2) \tag{30}$$

2.3. Coherent scattering of X-rays and neutrons

Coherent scattering allows cross-correlations in atomic positions to be probed. Here we examine the following types of coherent scattering: (i) solution scattering of neutrons (orientationally averaged without energy analysis); (ii) scattering of X-rays and neutrons by crystals without energy analysis (diffraction and diffuse scattering); and (iii) coherent inelastic neutron scattering by crystals.

2.3.1. Solution scattering

Solution scattering from biological macromolecules has been employed particularly at small scattering angles to provide low-resolution information on molecular structure. Small-angle neutron scattering has the advantage over its X-ray counterpart that, due to the difference in sign of the coherent scattering lengths of deuterium and hydrogen, experimental conditions can be optimized so as to obtain a good contrast of the macromolecule over the solution. Time integration of Eq. 7 and orientational averaging then lead to the following equation for the scattered intensity:

$$S_{coh}(Q) = K^2 < |F(\vec{Q})|^2 > \tag{31}$$

where $< |F(\vec{Q})|^2 >$ is the form factor of the protein molecule and K is its average contrast given by

$$K = \frac{1}{V}\int_V (\rho(\vec{r}) - \rho^0)\, d\vec{r} \tag{32}$$

where V is the volume giving rise to the contrast, $\rho(\vec{r})$ is the coherent neutron scattering length density in the protein at point $\vec{r}$, and ρ^0 is the solvent scattering length density.

Neutron small-angle scattering has been applied to derive the configurational distribution of phosphoglycerate kinase (PGK) strongly denatured in 4 M guanidinium hydrochloride solution [9]. The denaturing of the protein produces a clear change in the scattering profile and a large increase of the radius of gyration, R_g, from 24 Å in the native form to 78 Å in the denatured form. To interpret the data, a model was derived in which the excess scattering density associated with the protein is pictured as a freely jointed chain of N spheres of radius L linked by rigid bonds of length 2L. The form factor for this is as follows:

$$< |F(\vec{Q})|^2 > = \sum_{l=1}^{N} \sum_{m=1}^{N} |F(\vec{Q})|^2 \left[\frac{\sin(2QL)}{2QL}\right]^{|l-m|} \tag{33}$$

The space Fourier transform of $S_{coh}(Q)$ gives the radial density distribution function P(r). Fits of the freely jointed spheres model to the experimental P(r) for denatured PGK are shown in Fig. 1. A good fit was found to have ~100 spheres with radius ~8.5 Å. The freely jointed spheres description was used to define bounds for the generation of atomic-detail molecular models for individual configurations of the denatured chain, one of which is depicted in Fig. 2. Further analysis of the

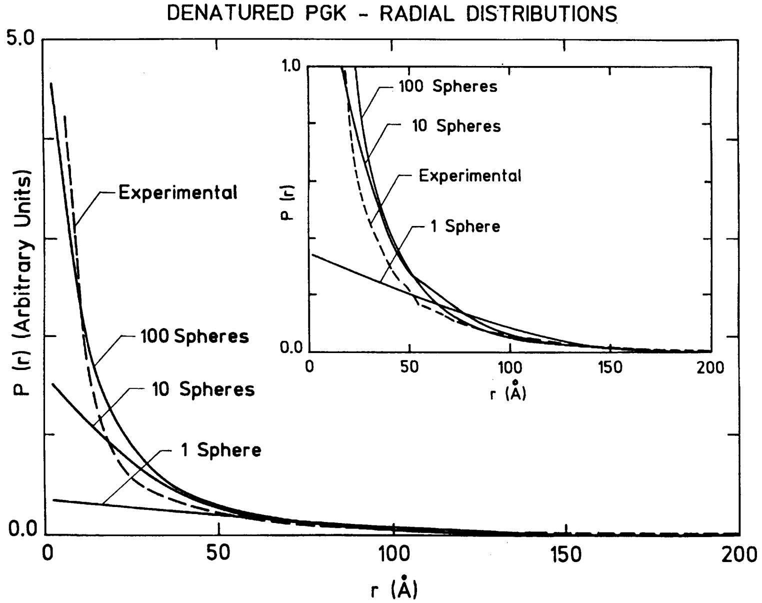


Fig. 1. Radial distribution functions for phosphoglycerate kinase denatured in 4 M guanidine hydrochloride solution. (- - -): experimental distribution function obtained from small-angle neutron scattering data; (—): distribution function obtained from the freely jointed chain of spheres with 1, 10 and 100 spheres. From Ref. 9.

experimental data using random polymer theory showed that over the range $3R_g^{-1} < Q < 0.20\ Å^{-1}$, the polypeptide behaves as an excluded volume chain [10].

2.3.2. Diffraction by crystals

In an X-ray crystallography experiment, the instantaneous scattered intensity is given by [5]:

$$I_{hkl} = |F_{hkl}|^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i f_j^* \exp[i\vec{Q} \cdot (\vec{R}_i - \vec{R}_j)] \tag{34}$$

where F_{hkl} is the structure factor, $\vec{R}_i$ is the position vector of atom i in the crystal and f_i is the X-ray atomic form factor. For neutron diffraction f_i is replaced by the coherent scattering length.

It is not feasible to insert into Eq. 34 the atomic positions for all the atoms in the crystal for every instant in the time of the experiment. Rather, the intensity must be evaluated in terms of statistical relationships between the positions. One approach is to consider a real crystal as a superposition of an ideal periodic structure with slight perturbations. When exposed to X-rays, the real crystal gives rise to two scattering components: the set of Bragg reflections arising from the periodic structure, and

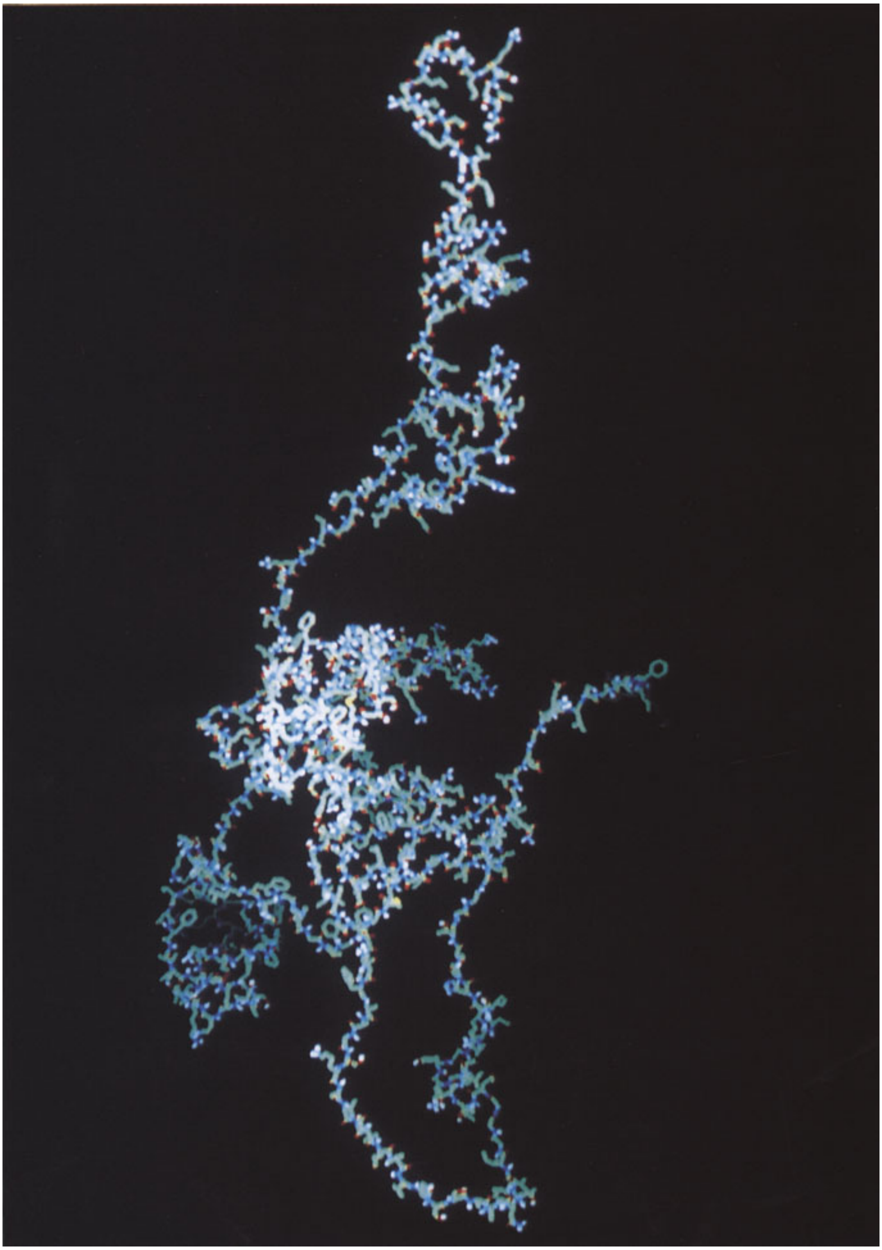


Fig. 2. Full view of a configuration of denatured phosphoglycerate kinase generated by combining the low-resolution small-angle scattering data and the freely jointed chain of spheres model in Fig. 1 with molecular modelling. Details are given in Ref. 9.

scattering outside the Bragg spots (diffuse scattering) that arises from the structural perturbations:

$$I_{hkl} = I^{B}_{hkl} + I^{D}_{hkl} \tag{35}$$

where I^{B}_{hkl} is the Bragg intensity, found at integer values of h, k and l, and I^{D}_{hkl} is the diffuse scattering, not confined to integer values of h, k and l.

In terms of structure factors, the various intensities are given by [5]:

$$I_{hkl} = |F_{hkl}|^2 \tag{36}$$

$$I^{B}_{hkl} = | < F_{hkl} > |^2 \tag{37}$$

$$I^{D}_{hkl} = |\Delta F_{hkl}|^2 \tag{38}$$

where ΔF_{hkl} is the Fourier transform of the electron density perturbation.

Bragg diffraction. The Bragg peak intensity reduction due to atomic displacements is described by the well-known 'temperature' factors. Assuming that the position $\vec{R}_i$ can be decomposed into an average position, $< \vec{R}_i >$, and an infinitesimal displacement, $\vec{u}_i = \delta\vec{R}_i = \vec{R}_i - < \vec{R}_i >$, then the X-ray structure factors can be expressed as follows:

$$F_{hkl} = \sum_{i=1}^{N} f_i(\vec{Q}) \exp(i\vec{Q} \cdot < \vec{R}_i >) \exp(W_i(Q)) \tag{39}$$

where $W_i(Q) = -\frac{1}{3} < u^2_{i,Q} > Q^2$ and $< u^2_{i,Q} >$ is the mean-square displacement in the direction of Q. $W_i(Q)$ is the Debye–Waller factor and is equivalent to that given in Eq. 28 for neutron scattering.

Temperature factors are of interest to structural biologists mainly as a means of deriving qualitative information on the fluctuations of segments of a macromolecule. However, X-ray temperature factor analysis has drawbacks. One of the most serious is the possible presence of a static disorder contribution to the atomic fluctuations. This

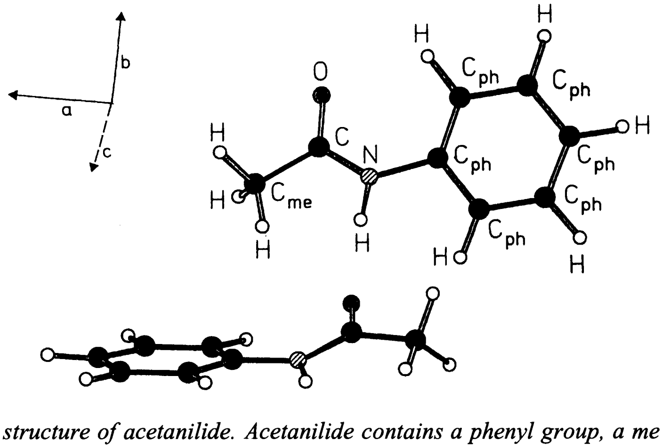


Fig. 3. Crystal structure of acetanilide. Acetanilide contains a phenyl group, a methyl group and a peptide group that links the molecules of the crystal together via hydrogen bonds into parallel chains. From Ref. 11.

Table 1 Mean-square displacements of hydrogen atoms in crystalline acetanilide at 15 K

	I	II	III		I	II	III
Methyl H	a	0.0359	0.0352	Phenyl H_{para}	a	0.0126	0.0118
	b	0.0366	0.0438		b	0.0273	0.0278
	c	0.0253	0.0258		c	0.0279	0.0258
Isotropic		0.0326	0.0349	Isotropic		0.0226	0.0218
Methyl H	a	0.0150	0.0160	Phenyl H_{meta}	a	0.0215	0.0220
	b	0.0482	0.0510		b	0.0189	0.0194
	c	0.0328	0.0336		c	0.0265	0.0246
Isotropic		0.0320	0.0335	Isotropic		0.0223	0.0220
Methyl H	a	0.0301	0.0286	Phenyl H_{meta}	a	0.0154	0.0162
	b	0.0162	0.0172		b	0.0218	0.0210
	c	0.0483	0.0606		c	0.0296	0.0256
Isotropic		0.0315	0.0354	Isotropic		0.0222	0.0209
Amide H	a	0.0178	0.0192	Phenyl H_{ortho}	a	0.0159	0.0160
	b	0.0110	0.0120		b	0.0189	0.0194
	c	0.0258	0.0300		c	0.0286	0.0250
Isotropic		0.0182	0.0204	Isotropic		0.0211	0.0201
				Phenyl H_{ortho}	a	0.0203	0.0200
					b	0.0170	0.0164
					c	0.0264	0.0254
				Isotropic		0.0212	0.0206

Column I: hydrogen atoms of acetanilide. Phenyl hydrogens are named according to their position relative to the N-substitution site. a, b, c refer to the crystallographic directions. Column II: anisotropic (a, b, c crystallographic directions) and isotropic mean-square displacements (Å^2), from neutron diffraction data [12]. Column III: anisotropic (a, b, c crystallographic directions) and isotropic mean-square displacements (Å^2), from harmonic analysis. From Ref. 11.

cannot be distinguished from the dynamic disorder due to the absence of energy analysis of the scattered X-ray photons. For quantitative work with X-rays, one approach is to choose a system in which there is negligible static disorder and in which the harmonic approximation is valid. An example of such a system is acetanilide, (C_6H_5-CONH-CH_3), at 15 K. Acetanilide is shown in Fig. 3. In recent work [11] the molecular mechanics force field was parametrized for this crystal and normal-mode analyses were performed in the full configurational space of the crystal, i.e. including all intramolecular and intermolecular degrees of freedom. As a quantitative test of the accuracy of the force field, anisotropic quantum-mechanical mean-square displacements of the hydrogen atoms were calculated in each Cartesian direction as a sum over the phonon normal modes using Eq. 28 and compared with experimental neutron diffraction temperature factors [12]. The experimental and theoretical temperature factors are presented in Table 1. The values of the mean-square displacements are in excellent agreement. As we shall see later, the forms and frequencies of the individual vibrational modes that sum to give $< u^2_{Q,i} >$ are themselves also in good agreement with experiment.

2.3.3. X-ray diffuse scattering

Any perturbation from ideal space-group symmetry in a crystal will give rise to diffuse scattering. The X-ray diffuse scattering intensity, I^D_{hkl}, at some point (hkl) in reciprocal space can be written as

$$I^D_{hkl} = N\sum_m \langle (F_n - \langle F\rangle)(F_{n+m} - \langle F\rangle)^*\rangle \exp(-\vec{Q}\cdot\vec{R}_m) \tag{40}$$

where F_n is the structure factor of the nth unit cell, and the sum $\sum_m$ runs over the relative position vectors $\vec{R}_m$ between the unit cells. The function $\langle (F_n - \langle F\rangle)(F_{n+m} - \langle F\rangle)^*\rangle$ is determined by correlations between atomic displacements.

If the diffuse scattering of dynamical origin contributes significantly to the measured scattering, it may provide information on the nature of correlated motions in biological macromolecules that may themselves be of functional significance. To examine this possibility it is necessary to construct dynamical models of the crystal, to calculate their diffuse scattering and to compare with experiment. The advent of high-intensity synchrotron sources and image plate detectors has allowed good-quality X-ray diffuse scattering images to be obtained from macromolecular crystals.

A program exists, named SERENA (Scattering of Ex-Rays Elucidated by Numerical Analysis), for calculating X-ray diffuse scattering intensities from configurations of atoms in molecular crystals [13]. The configurations are conveniently derived from molecular dynamics simulations, although in principle any collection of configurations can be used. SERENA calculates structure factors from the individual configurations and performs the required averages in Eqs. 36–38.

Displacements correlated within unit cells but not between them lead to *very diffuse scattering* that is not associated with the Bragg peaks. This can be conveniently explored using present-day simulations of biological macromolecules. However, motions correlated over distances larger than the size of the simulation model will clearly not be included. Due to computational requirements this has excluded the use of atomic-detail molecular dynamics in the examination of the diffuse scattering resulting from correlated displacements of biological macromolecules in different unit cells. The use of periodic boundary conditions in the simulation suppresses motions of wavelength longer than the box edge. Displacements correlated between different unit cells lead to characteristic haloes around or streaks between the Bragg spots. Diffuse streaks in lysozyme diffraction patterns have been described using rigid-body displacements of the molecules in adjacent unit cells [14]. The haloes around the Bragg peaks (thermal diffuse scattering) may be due to lattice vibrations. They occur in protein crystals [15], but have not yet been examined using molecular simulation and remain something of a mystery. Thermal diffuse scattering in small-molecule crystals has been examined experimentally and using molecular simulation [16]. Indeed, in favourable circumstances, it is possible to energy-analyse neutron thermal diffuse scattering to obtain both the frequencies and wavevectors of the lattice modes concerned, as discussed below.

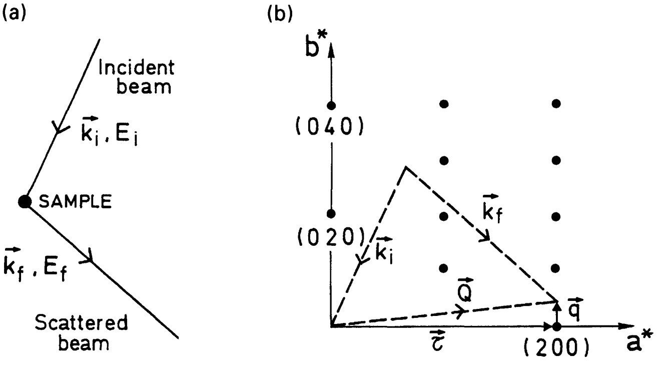


Fig. 4. Schematic vector diagrams illustrating the use of coherent inelastic neutron scattering to determine phonon dispersion relationships: (a) scattering in real space; (b) scattering triangles illustrating the momentum transfer, $\vec{Q}$, of the neutrons in relation to the reciprocal lattice vector of the sample, $\vec{\tau}$, and the phonon wavevector, $\vec{q}$. Dots represent Bragg reflections.

2.3.4. Coherent inelastic neutron scattering

The use of coherent neutron scattering with simultaneous energy and momentum resolution provides a probe of time-dependent pair correlations in atomic motions. Coherent inelastic neutron scattering is therefore particularly useful for examining lattice dynamics in molecular crystals and holds promise for the characterization of correlated motions in biological macromolecules [17]. A property of lattice modes is that for particular wavevectors there are well-defined frequencies; the relations between these two quantities are the phonon dispersion relations. Neutron scattering is the only effective technique for determining phonon dispersion curves. The scattering geometry used is illustrated in Fig. 4. The following momentum conservation law is obeyed:

$$\vec{k}_i - \vec{k}_f = \vec{Q} = \vec{\tau} + \vec{q} \tag{41}$$

where $\vec{k}_i$ and $\vec{k}_f$ are the initial and final neutron wavevectors. The vibrational excitations have wavevector $\vec{q}$ which is measured from a Brillouin zone centre (Bragg peak) located at $\vec{\tau}$, a reciprocal lattice vector.

If the displacements of the atoms are given in terms of the harmonic normal modes of vibration for the crystal, the coherent one-phonon inelastic neutron scattering cross-section is given by

$$S_{coh}(\vec{Q}, \omega) = \sum_j S_j(\vec{Q}, \omega) \tag{42}$$

where the summation is over all vibrational modes of the crystal. For one mode one has

$$S_j(\vec{Q}, \omega) = \frac{\langle n(\omega_j(\vec{q})) + 1 \rangle}{\omega_j(\vec{q})} |F_j(\vec{Q}, \vec{q})|^2 \delta(\omega \pm \omega_j(\vec{q}))\delta(\vec{Q} \pm \vec{q} - \tau) \tag{43}$$

In this expression $\omega_j(\vec{q})$ is the frequency of the phonon with wavevector $\vec{q}$ belonging to phonon dispersion branch j. $\langle n(\omega_j(\vec{q})) + 1\rangle$ is the Bose factor. $|F_j(\vec{Q}, \vec{q})|^2$ is given by

$$|F_j(\vec{Q}, \vec{q})|^2 = \sum_k m_k^{1/2} b_k(\vec{Q} \cdot \vec{e}_k^j(\vec{q})) \exp(i\vec{Q} \cdot \vec{R}_k) \exp(-2W_k(\vec{Q})) \tag{44}$$

where m_k is the mass of atom k, b_k is its coherent scattering length and $\vec{R}_k$ its position. $W_k(Q)$ in Eq. 44 is the exponent of the Debye–Waller factor and is given by Eq. 28. $\vec{e}_k^j(\vec{q})$ is the eigenvector of the kth atom in the jth mode and describes the pattern of the displacements in one unit cell. $\vec{e}_k^j(\vec{q})$ has 3s components, where s is the number of atoms in the unit cell. For any direction of $\vec{q}$ in the Brillouin zone there are 3s dispersion curves.

2.4. Far-infrared absorption spectroscopy

Infrared absorption spectroscopy is a standard technique in structural biology. However, its use has been primarily limited to the examination of high-frequency ($\hbar\omega \gg k_BT$ at 300 K) local vibrations in macromolecules and their relation with structure. For several reasons, far-infrared spectroscopy, corresponding to ps-time-scale vibrations, has been much less commonly applied to biological problems. Firstly, the efficiency of laboratory sources is reduced in the far-infrared region. This, together with the high absorption of water in the far-infrared, renders difficult the obtention of good-quality spectra from biological macromolecules. However, the advent of high-intensity synchrotron far-infrared sources may overcome these difficulties [18]. Secondly, the vibrations in the far-infrared region are difficult to assign unambiguously. However, these vibrations are of particular interest in molecular simulation as they involve collective modes and hydrogen-bond vibrations, i.e. vibrations influenced by electrostatic and van der Waals interactions. Moreover, as infrared absorption arises from charge fluctuations, the combination of simulation with far-infrared experiment should provide information on the charge fluctuations associated with displacements along soft degrees of freedom in biological systems.

Atomic charges have two distinguishable effects on far-infrared absorption spectra. The first is indirect: Coulombic interactions between charges play a role in determining atomic dynamics. In molecular dynamics simulation this enters into the potential energy function used for calculating the forces between the atoms. The second effect is that the atomic charge fluctuations associated with the nuclear position fluctuations directly determine the absorption. Although, in reality, the charge phenomena involved in both the above effects are the same, for practical and interpretational purposes they are commonly treated separately. In particular, whereas the explicit inclusion of polarization terms is not a feature of some successful interaction potentials for molecular dynamics simulation, it is a requirement for the spectroscopic activation of certain observed far-infrared features. Therefore, for the calculation of infrared intensities, a charge fluctuation model including polarization can be applied *a posteriori* to the atomic trajectories generated by molecular dynamics simulation.

2.4.1. Absorption coefficient

The infrared absorption coefficient, I(ω), of a system is given by [19]:

$$I(\omega) = \frac{4\pi^2}{3cn(\omega)V}\frac{\omega}{\hbar}(1 - e^{-\beta\hbar\omega})C(\omega) \tag{45}$$

$$C(\omega) = \frac{1}{2\pi}\int dt\, e^{-i\omega t} < \vec{M}(0)\cdot\vec{M}(t) > \tag{46}$$

where c is the velocity of light, n(ω) is the refractive index of the medium, $\vec{M}$ is the dipole moment of the system and V its volume. The quantum-mechanical C(ω) in Eq. 46 fulfils the detailed balance condition that we have already seen in neutron scattering in Sec. 2.2:

$$C(-\omega) = e^{-\beta\hbar\omega}C(\omega) \tag{47}$$

Molecular dynamics simulations allow the computation of the classical limit, $C_{cl}(\omega)$, of the dipole moment autocorrelation function. As mentioned for neutron scattering, Eq. 47 does not hold in the classical limit ($\hbar \rightarrow 0$) where $C_{cl}(-\omega) = C_{cl}(\omega)$. To correct the classical correlation function and to re-establish Eq. 47, the classical time-correlation function can be identified as the real part of its quantum-mechanical counterpart. This leads to the following transformation:

$$C_{cl}(\omega) \rightarrow \frac{2}{1 + e^{-\beta\hbar\omega}}C_{cl}(\omega) \tag{48}$$

The absorption then becomes

$$I(\omega) = \frac{4\pi^2}{3cn(\omega)V}\frac{\omega}{\hbar}\tanh(\beta\hbar\omega/2)C_{cl}(\omega) \tag{49}$$

2.4.2. Calculation of the system dipole moment

The time dependence of the total dipole moment of the system, $\vec{M}(t)$, can be calculated *a posteriori* from the nuclear trajectories generated by a molecular dynamics simulation. The dipole moment of a given molecule is expressed as the sum of a permanent part, $\vec{p}$, that is independent of the environment of the molecule and an induced part, $\vec{d}$, that depends on the local electric field. $\vec{d}$ has the following form:

$$\vec{d} = \sum_i \vec{\mu}_i \tag{50}$$

where the $\vec{\mu}_i$ are the induced point dipoles on the individual atoms. One method for calculating the $\vec{\mu}_i$ is based on a procedure, originally introduced by Applequist [20] and modified by Thole [21], for calculating the polarizability tensor of small molecules, given the atomic coordinates and the isotropic atomic polarizabilities. This method has recently been applied to examine far-infrared absorption from water [22] and holds particular promise for macromolecular calculations. We now briefly review this method.

Consider a system of N atoms in an external electric field, $\vec{E}_{ext}$. Each atom, i, possesses a polarizability, α_i, and is polarized by $\vec{E}_{ext}$, giving rise to an induced dipole moment $\vec{\mu}_i$ which itself contributes to the total electric field $\vec{E}_{tot}$. This can be written in the following way:

$$\vec{\mu}_i = \alpha_i \vec{E}_{tot} \tag{51}$$

where

$$\vec{E}_{tot} = \vec{E}_{ext} + \sum_{j \neq i} \mathbf{T}_{ij} \vec{\mu}_j \tag{52}$$

The total electric field contains the external field, $\vec{E}_{ext}$, and the field due to the induced dipoles, $\sum_{j \neq i} \mathbf{T}_{ij} \vec{\mu}_j$, where $\mathbf{T}_{ij}$ is the induced dipole tensor. Equation 52 can be written in the following matrix form:

$$\begin{bmatrix} \alpha_1^{-1} & -\mathbf{T}_{12} & \cdots & -\mathbf{T}_{1N} \\ -\mathbf{T}_{21} & \alpha_2^{-1} & \cdots & -\mathbf{T}_{2N} \\ \vdots & \vdots & & \vdots \\ -\mathbf{T}_{N1} & -\mathbf{T}_{N2} & \cdots & \alpha_N^{-1} \end{bmatrix} \begin{bmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \\ \vdots \\ \vec{\mu}_N \end{bmatrix} = \begin{bmatrix} \vec{E}_{ext} \\ \vec{E}_{ext} \\ \vdots \\ \vec{E}_{ext} \end{bmatrix} \tag{53}$$

Multiplying Eq. 53 by the inverse of the matrix and summing over the atoms reduces it to the simple form

$$\vec{\mu} = \mathbf{A} \vec{E}_{ext} \tag{54}$$

where $\vec{\mu}$ is the total induced dipole moment of the system. It is proportional to the external field. **A** is the polarizability tensor including induced atomic dipole interactions. Applequist chose a point dipole model for $\mathbf{T}_{ij}$:

$$\mathbf{T}_{ij} = \nabla_{\vec{r}_i} \nabla_{\vec{r}_j} \frac{1}{r_{ij}} \tag{55}$$

$$= \frac{3}{r_{ij}^5} \begin{bmatrix} x^2 & xy & xz \\ yx & y^2 & yz \\ zx & zy & z^2 \end{bmatrix} - \frac{1}{r_{ij}^3} \mathbf{I} \tag{56}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the positions of atoms i and j, x, y, z are the Cartesian coordinates of $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$, and **I** is the identity matrix. In this model, atom i is subjected to the dipolar part of the field generated by an infinitely small charge distribution placed at the centre of atom j. However, when the interatomic distance r_{ij} approaches $s_{ij} = (4\alpha_i \alpha_j)^{1/6}$ the model can lead to unphysically large values for the induced dipole moments. Thole noticed this problem and introduced a damping effect [21]. If $r_{ij} \leq s_{ij} = a(\alpha_i \alpha_j)^{1/6}$, where a is a constant, the potential $1/r_{ij}$ which generates the dipolar field (cf. Eq. 55) is replaced by that generated by a delocalized isotropic charge distribution, $\rho(\vec{r})$. To reproduce experimental polarizability tensors of small molecules, using Eq. 54, with the polarizabilities α_i as fitted parameters, Thole tried several kinds of radial dependences for $\rho(\vec{r})$. The best distribution found gave $a = 1.662$ and the following form for $\mathbf{T}_{ij}$:

$$\mathbf{T}_{ij} = \frac{3v_{ij}^4}{r_{ij}^5}\begin{bmatrix} x^2 & xy & xz \\ yx & y^2 & yz \\ zx & zy & z^2 \end{bmatrix} - \frac{(4v_{ij}^3 - 3v_{ij}^4)}{r_{ij}^3}\mathbf{I} \tag{57}$$

where v_{ij} has the form:

$$v_{ij} = \begin{cases} \dfrac{r_{ij}}{s_{ij}} & \text{if } r_{ij} < s_{ij} \\ 1 & \text{otherwise} \end{cases} \tag{58}$$

To calculate the induced part of the dipole moment of a system of molecules, the above procedure must be adapted. The external field $\vec{E}_{ext}$ is replaced by $\vec{E}_i$, the field acting on the ith atom due to the permanent charges q_j. For self-consistency within the Thole representation, the permanent charges are also represented by isotropic delocalized charge distributions. Consequently, the classical formula for $\vec{E}_i$ is modified as follows [21]:

$$\vec{E}_i = \sum_j (4v_{ij}^3 - 3v_{ij}^4)\, q_j \vec{r}_{ij}/r_{ij}^3 \tag{59}$$

where v_{ij} has the same definition as in Eq. 58.

Equation 51 then becomes

$$\vec{\mu}_i = \alpha_i\left(\vec{E}_i + \sum_{j \neq i} \mathbf{T}_{ij}\vec{\mu}_j\right) \tag{60}$$

To be consistent with the molecular mechanics potential, intramolecular permanent charge interactions involving 1–2 and 1–3 interactions can be excluded from the induced dipole calculation. All other dipole–dipole interactions are included as in Ref. 21.

To calculate the induced dipoles on the molecules, Eq. 60 can be solved for each molecular dynamics trajectory frame generated. This can be done using an iterative procedure in which the $\vec{\mu}_i$ vectors calculated in the nth step are used as $\vec{\mu}_j$ in the $(n + 1)$st step. The iteration is repeated until convergence of $\vec{\mu}_i$. The iteration of Eq. 60 leads to a self-consistent representation of the local field and the induced dipoles.

In Eq. 60, two terms require summations over atom pairs: $\vec{E}_i$ and $\sum_{j \neq i}\mathbf{T}_{ij}\vec{\mu}_j$. These summations are expensive to compute for large numbers of atoms. For this and for interpretational reasons, it is therefore useful to examine if the form of the calculated spectrum is sensitive to the presence of approximations in their evaluation. The method used to approximate $\vec{E}_i$ need not, *per se*, be the same as that used to approximate $\sum_{j \neq i} \mathbf{T}_{ij}\vec{\mu}_j$. Indeed, the fact that the dipole interaction is short-range (it varies as $1/r^3$) suggests that an approximation to the long-range interactions in the calculation of $\sum_{j \neq i} \mathbf{T}_{ij}\vec{\mu}_j$ might be appropriate. However, the field $\vec{E}_i$, due to the permanent point charges, contains a significant long-range component. Therefore, the use of a spherical truncation in the evaluation of $\vec{E}_i$, although rapid, is likely to lead to errors. An Ewald sum would be more accurate. Another possibility is a method intermediate between the spherical cutoff and a full all-atom pair calculation, involving adding a reaction field to the cutoff sphere. The reaction field simulates an infinite dielectric medium outside the sphere. The induced dipoles are then given by [23]:

$$\vec{\mu}_i = \alpha_i \left[\vec{E}_i + \sum_{\substack{j \neq i \\ r_{ij} < R_c}} \left(\mathbf{T}_{ij}\vec{\mu}_j + \frac{2(\varepsilon_{rf} - 1)}{2\varepsilon_{rf} + 1} \frac{\vec{\mu}_j}{R_c^3} \right) \right] \tag{61}$$

where ε_{rf} is the relative dielectric constant of the medium outside the sphere and R_c is the cutoff. Equation 61 is simple to implement.

3. Dynamics in small-molecule condensed phases

In this section examples are given of the combination of molecular simulation with experiment for the determination of ps-timescale dynamics of crystals and liquids of small molecules. The examples are chosen to cover a wide range of dynamical phenomena: local and collective vibrations, lattice modes, and diffusive motions of molecules and parts of molecules.

3.1. Vibrational dynamics

3.1.1. Anharmonic local vibrations in acetanilide

The crystalline state provides structurally well-characterized systems enabling detailed studies of environmental effects on molecular motions. Using a molecular mechanics force field, it is possible in principle to obtain a complete description of the ground-state nuclear dynamics of a molecular crystal, by working in the full, 3N-dimensional configurational space (where N is the number of atoms in the crystal) using computer simulation methods. In this way an attempt can be made to describe the structural and dynamical features of the crystal in a unified fashion.

An optimized molecular mechanics potential function has been obtained for the acetanilide crystal (shown in Fig. 3) by performing energy minimizations and harmonic analyses of the crystal and adjusting the parameters of the function so as to reproduce low-temperature structural and spectroscopic data [11]. The resulting normal modes provide a description of the low-temperature intramolecular and lattice vibrations. With one exception, all the fundamental frequencies of the intramolecular modes in the refined force field were within 3% of their values obtained at the centre of the Brillouin zone by optical spectroscopy. Most of the mode assignments were in agreement with previous assignment schemes. Moreover, the calculated crystal field splitting of the vibrational bands (into eight distinct components for acetanilide) was also found to be in quantitative agreement with experiment.

Using the results of the harmonic analysis, incoherent inelastic neutron scattering intensities were calculated using Eq. 27 assuming the presence of one-, two- and three-phonon scattering processes. The results at 25 K are compared with experiment in Fig. 5. Because of the large hydrogen displacement in methyl torsion, this peak is by far the strongest feature of the experimental spectrum. The peak is narrow and well resolved at 145 cm^{-1}. The average intensity of the lattice mode peaks ($< 100\ cm^{-1}$) is ~20% of the methyl torsional peak at 145 cm^{-1}.

Having refined the force field in the harmonic approximation, it was possible to use the full, anharmonic potential function in molecular dynamics simulations. Simulations of the acetanilide crystal were performed using periodic boundary conditions, at 80, 140 and 300 K, and the temperature dependence of the hydrogen-weighted vibrational density of states was calculated using Eq. 29.

Methyl libration. Figure 6 presents the density of states, G(ω), for the methyl hydrogens calculated from the harmonic phonon analysis and from the simulations.

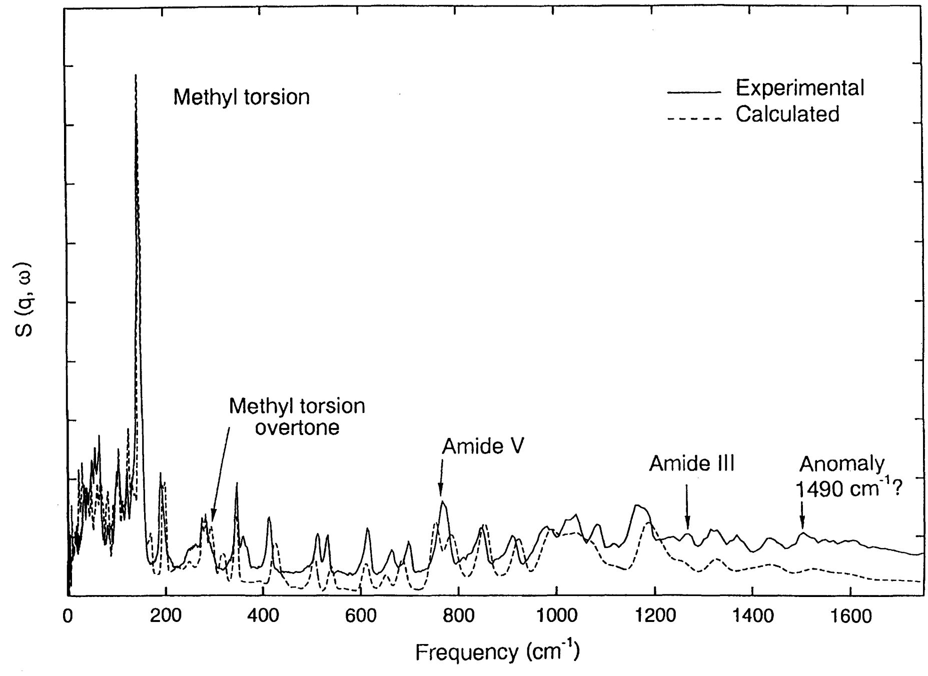


Fig. 5. Incoherent neutron scattering dynamic structure factor S_{inc} ($\vec{Q}$, ω), measured at 25 K using the spectrometer TFXA at the spallation source at the Rutherford–Appleton Laboratory, Oxford, and calculated using the results of a normal-mode analysis of the crystal and assuming the presence of one-, two- and three-phonon scattering, using Eq. 27. From Ref. 11.

The methyl torsional mode is at 145 cm^{-1} at 80 K, and shifts downward and broadens with increasing temperature above 80 K, to reach ~115 cm^{-1} at 300 K. This temperature dependence is in good agreement with that observed experimentally [24,25]. The experiments indicate the onset, above 100 K, of a downward shift from 142 cm^{-1} at 100 K to 125 cm^{-1} at 300 K. The downward shift and broadening are due to the anharmonic nonbonded environment of the methyl group and the presence of frictional damping. The peak shift is also associated with the onset of torsional transitions of the methyl group on the ps timescale, which occur in the simulations at 140 and 300 K.

Peptide hydrogen-bond vibration. A significant temperature dependence of the NH out-of-plane bands was observed in the acetanilide simulations. Figure 7 displays G(ω) for the amide hydrogen from the phonon calculations and from the molecular dynamics simulations. This region contains three separate bands, all involving NH out-of-plane motion. The peak maxima at 750 and 785 cm^{-1} shift downward by 14 cm^{-1} at 300 K and the bands change in form. As the temperature rises the bands broaden, and the highest frequency band near 785 cm^{-1} shifts downward more than that centred at 778 cm^{-1}, eventually merging with it above 140 K. These observations are again in quantitative agreement with experiment [12,24] and indicate that

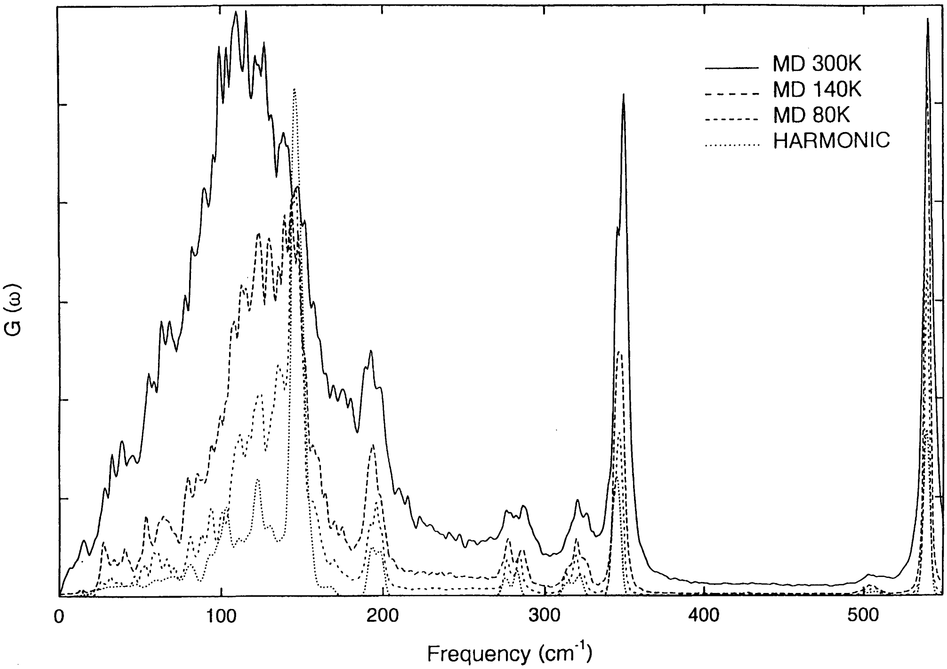


Fig. 6. G(ω) for the methyl hydrogens in the acetanilide crystal calculated from a harmonic analysis and from molecular dynamics simulations at 80, 140 and 300 K. From Ref. 11.

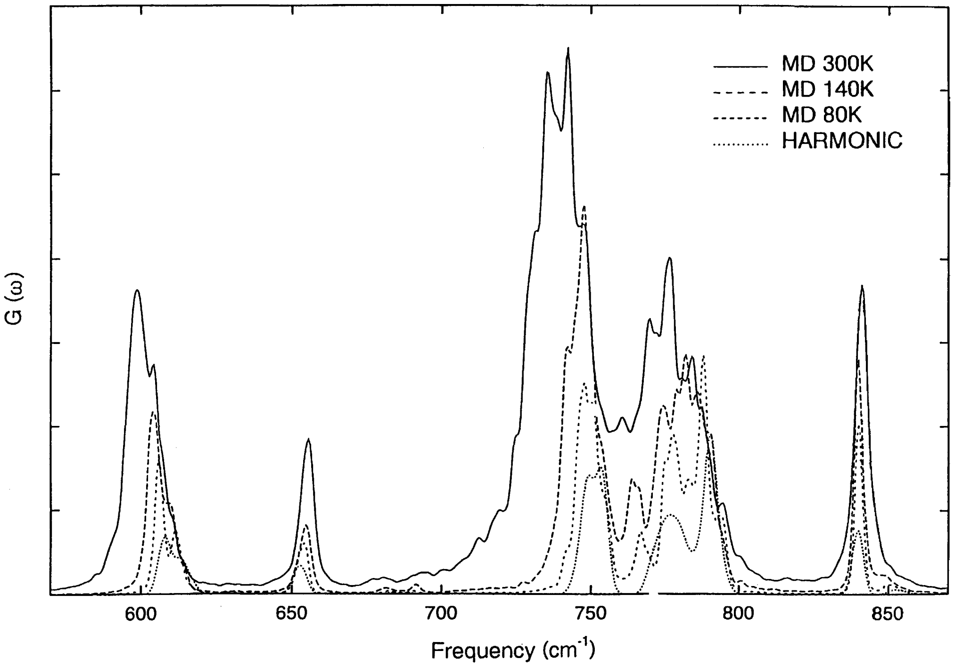


Fig. 7. G(ω) for the amide hydrogens in the acetanilide crystal calculated from a harmonic analysis and from molecular dynamics simulations at 80, 140 and 300 K. From Ref. 11.

sensitive details of the temperature-dependent anharmonic hydrogen-bond dynamics are reproduced by the molecular mechanics force field.

3.1.2. Hydrogen-bond dynamics in water

The interpretation of the far-infrared spectrum of water presents some interesting questions concerning the nature of the charge fluctuations leading to the absorption profile. The experimental 300 K far-infrared spectrum of water contains a wide absorption band at $\sim 600\,\mathrm{cm}^{-1}$ due to librations (rotations) of water molecules in their local hydrogen-bond networks, and a band at $200\,\mathrm{cm}^{-1}$ due to hydrogen-bond stretching [26]. Attempts have been made to reproduce these features using molecular dynamics simulation [27,28], but quantitative agreement with the experimental spectrum is lacking. In recent work [22] with the TIP3P potential [29], an improved agreement with experiment was obtained by using the self-consistent polarization method described in Sec. 2.4. Other questions that were addressed concerned the role of long-range electrostatic interactions in determining the induced dipoles and the dynamics associated with the far-infrared absorption.

Simulations were performed using two different methods for representing the electrostatic interactions in the potential function (spherical cutoff and Ewald sum) and the system dipole moment, $\vec{M}$, was calculated using two different methods for including the induced dipole contribution, i.e. the terms on the right-hand side of Eq. 60.

. *Induced dipole method EWRF.* The polarizability tensor of water is contained in the analytical form of the dipolar field tensor, thus leading to a self-consistent evaluation of the induced dipoles, i.e. Eq. 60 is solved iteratively to convergence. The field due to the permanent charges is calculated by an Ewald sum.

Induced dipole method NOINIT. This is a noniterative method in which the induced dipole on each water molecule is obtained by multiplying the experimental water molecule polarizability tensor by the electric field created by the permanent charges of the other atoms. This field is evaluated with an Ewald sum. The permanent charges used are those from the TIP3P model, i.e. they are increased from their gas-phase values. This method is identical to that used previously with the SPC potential [28].

Figure 8 compares the experimental far-infrared spectrum of water at 300 K with spectra calculated from the molecular dynamics simulations using the above methods. Decomposition of the calculated spectra indicates that the 600 cm^{-1} band is due to fluctuations of the positions of the permanent charges in the simulation, whereas the 200 cm^{-1} band is due to induced dipole fluctuations. The comparison of the curves in Fig. 8 shows that when the self-consistent, iterative method is used to calculate the induced dipoles, the ~200 cm^{-1} translational band produces a strong shoulder in the spectrum, whereas this is not the case for the spectrum calculated with the noniterative method.

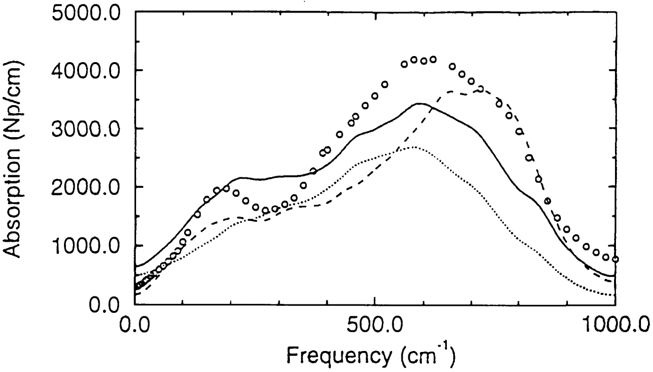


Fig. 8. Far-infrared spectra of water at 300 K. (○): experimental spectrum from Ref. 26. (—): spectrum calculated from a molecular dynamics simulation performed with Ewald summation. Induced dipole contribution calculated with method EWRF. (- - -): spectrum calculated from a molecular dynamics simulation performed with spherical truncation. Induced dipole contribution calculated with method EWRF. (· · ·): spectrum calculated from a molecular dynamics simulation performed with Ewald summation. Induced dipole contribution calculated with method NOINIT. From Ref. 22.

The frequency of the ~600 cm^{-1} rotational band is significantly influenced by long-range electrostatic interactions in the potential function. This is visible in Fig. 8 as a shift to high frequencies when the electrostatic interactions in the potential function are spherically truncated compared to the simulation in which an Ewald summation was used. Further calculations showed that the mean-field vibrational behaviour of individual molecules is similar in both the spherical truncation and Ewald simulations and that the frequency shift arises from differences in the dynamics of the relative orientations of different molecules. Cross-correlations in the static orientations of water molecules can be quantified with the Kirkwood G-factor, G_k, defined as

$$G_k = \sum_{i,j} \frac{< \vec{p}_i \cdot \vec{p}_j >}{Np^2} \tag{62}$$

where the $\vec{p}_i$ are the permanent dipoles and N is the number of atoms. G_k was found to be 0.15 in the spherical truncation simulation and 5.06 in the EWALD simulation, i.e. the EWALD simulation contains considerably more pair orientational ordering. The strong dependence of dipole–dipole correlations and associated dielectric quantities on electrostatic truncation had been observed previously in simulations of dipolar liquids [30–32]. The dependence calls into question the accuracy of calculations of the static dielectric constant of biological macromolecules using simulations with spherical truncation.

In contrast to the rotational band, the induced dipole fluctuations were found to be relatively insensitive to the long-range effects; the effective isotropy of the water molecular polarization tensor leads to a decoupling of the induced dipole fluctuations from the dynamical intermolecular orientational correlations of the permanent dipoles.

3.1.3. Collective vibrations in polyacetylene

Collective vibrations in molecular crystals are of particular interest in molecular simulation as a probe of nonbonded interactions. Here we consider two crystals exhibiting low-frequency collective vibrations: one, L-alanine, in which well-defined phonon dispersion relations exist and another, polyacetylene, in which low-frequency vibrations also exist but are subject to significant anharmonic effects.

Polyacetylene, $(CH)_x$, is the simplest example of a conjugated polymer. Interest in the physical properties of this molecule has been intensified with the finding that it can be chemically doped to 'metallic' levels of conductivity [33]. Polyacetylene can be crystallized in forms in which the C-C single bond dihedral angles are either mostly cis or mostly trans. Inelastic neutron scattering experiments from stretch-oriented cis-rich and trans-rich polyacetylene have enabled the vibrational density of states, G(ω), of the system to be determined in directions parallel ($G_{\|}$) and perpendicular ($G_{\perp}$) to the average chain axes depicted in Fig. 9 [34,35]. The experimental G(ω)'s were found to be highly anisotropic and to exhibit considerable differences between the cis and trans conformers. Also, a marked change in the experimental G(ω) was found on

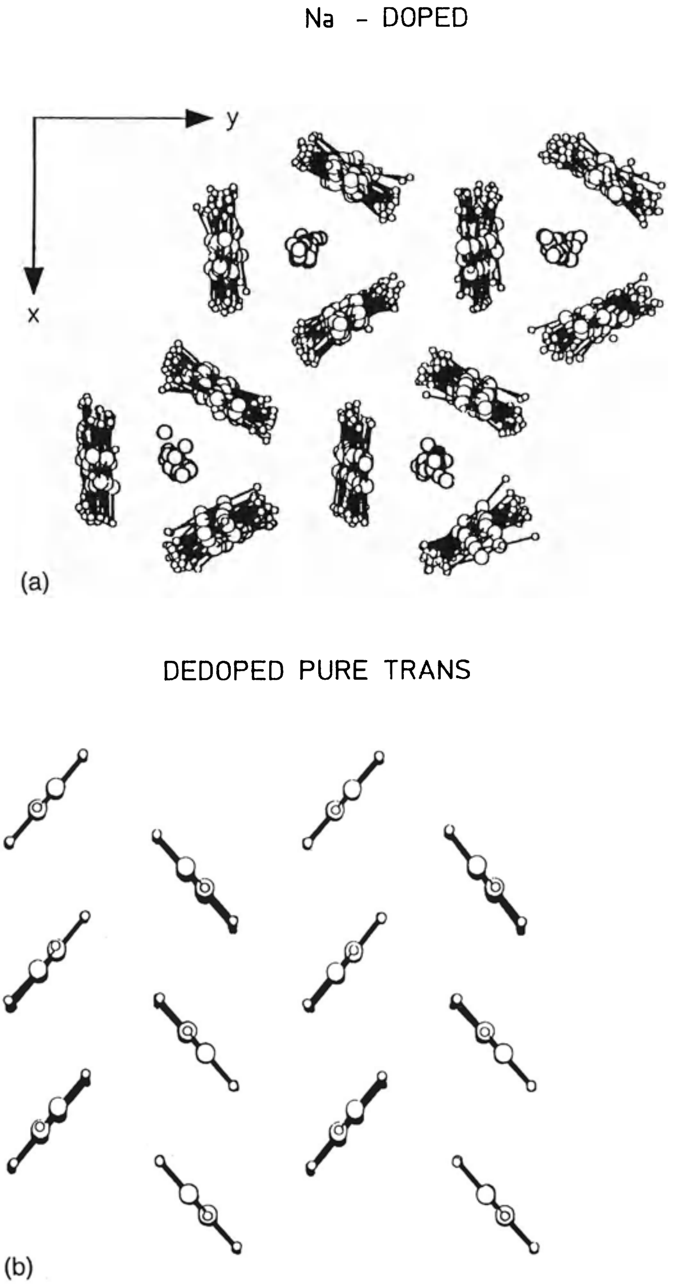


Fig. 9. (a) Snapshot of the simulation primary box of sodium-doped polyacetylene. This view is approximately perpendicular to the channel axis. (b) Energy-minimized structure of pure polyacetylene. From Ref. 37.

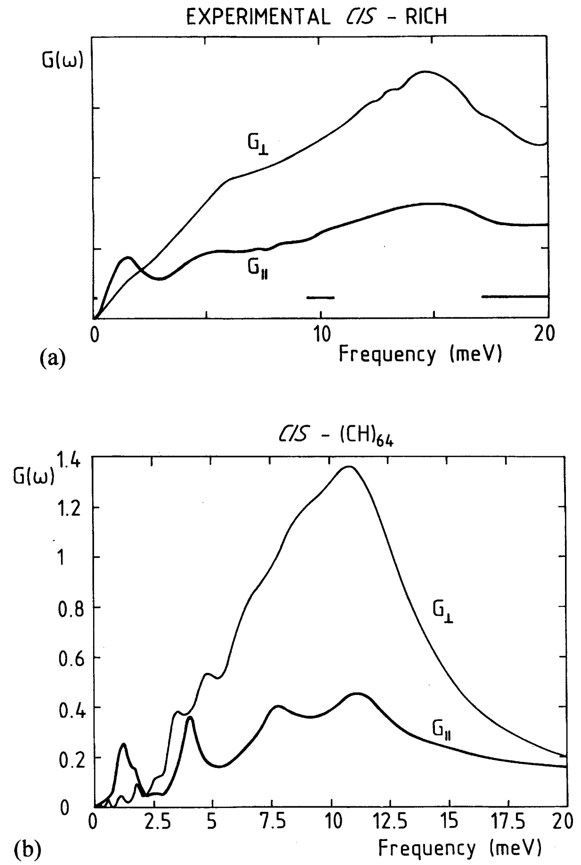


Fig. 10. (a) Experimental $G_{\parallel}$ and $G_{\perp}$ for cis-rich polyacetylene. (b) $G_{\parallel}$ and $G_{\perp}$ derived from cis-(CH)$_{64}$ simulation. The units of the simulation-derived $G(\omega)$ are Å^2 ps. The simulation-derived spectra have been convoluted with the experimental instrumental resolution function. From Ref. 36.

doping with sodium. These results threw down a challenge to molecular simulation. Would the dependence of the density of states on conformation, geometry and doping be reproduced? To examine this, molecular dynamics simulations were performed on pure and doped polyacetylene with potential function parameters transferred from small-molecule work [36,37].

Figure 10 presents the experimental densities of states for cis-(CH)$_x$ over the 0–20 meV range (0–160 cm^{-1}), parallel and perpendicular to the chain axes*. Also shown in Fig. 10 are the corresponding quantities derived from a molecular dynamics

*In neutron scattering a variety of energy and frequency units are used: 1 meV = 8.07 cm^{-1} = 0.24 THz = 11.61 K.

simulation of the crystal. In the experimental $G_{\parallel}$, there is a distinct peak at 1.5 meV and a broad peak with a maximum at 15 meV. In the simulation, both the 1.5 meV peak and the broad peak are present although the maximum of the latter is at 11 meV. In $G_{\perp}$, the 1.5 meV peak is absent in both simulation and experiment whereas the broad peak persists.

Figure 11 presents the experimental and simulation-derived data for trans-$(CH)_x$. The spectra are markedly different from those in the cis case. The changes seen experimentally are also present in the simulations. Figure 12 shows the simulation-derived and experimental $G_{\perp}$ spectra of sodium-doped trans-$(CH)_{64}$. There is a considerable difference in $G_{\perp}$ between the doped and undoped species in the simulations. In the pure system the intensity increases steeply from 5 meV to a plateau

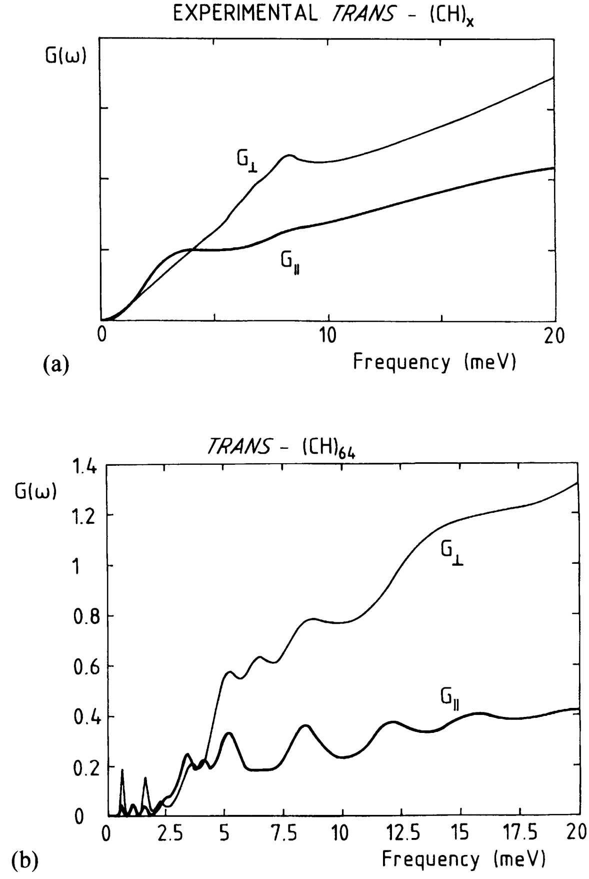


Fig. 11. (a) Experimental $G_{\parallel}$ and $G_{\perp}$ for trans-rich $(CH)_x$. (b) $G_{\parallel}$ and $G_{\perp}$ from trans-$(CH)_{64}$ simulation. The simulation-derived spectra have been convoluted with the experimental instrumental resolution function. From Ref. 36.

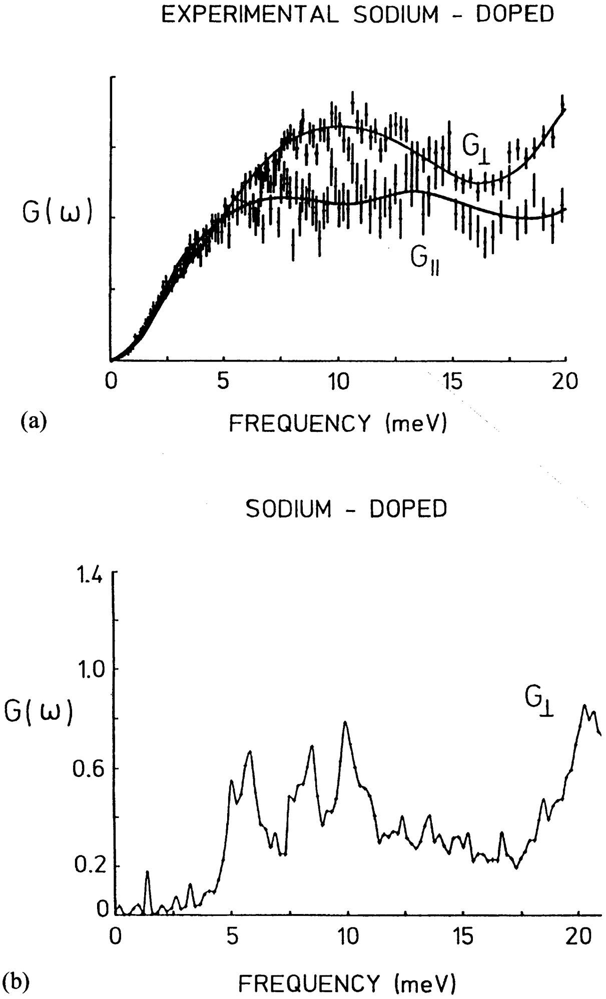


Fig. 12. (a) Experimental and (b) simulation-derived $G_\perp$ for sodium-doped polyacetylene. From Ref. 37.

at ~15–20 meV, whereas in doped $(CH)_{64}$ a broad minimum is present at ~12–18 meV. The experimental data were collected on an instrument with a relatively low neutron flux, worsening the counting statistics. Nevertheless, the broad minimum in the $G_\perp$ spectrum at 12–18 meV is clearly present in the experimental data, in accord with the simulation results.

The features of the experimental densities of states are sufficiently well reproduced by simulation that a detailed examination of the simulation-derived spectra was considered useful. This analysis showed that the $G_\parallel$ feature in the pure species at

~1.5 meV in cis and 4 meV in trans is a rigid-molecule vibration that is independent of chain length. Higher frequency vibrations (5 meV < 20 meV) are coupled intramolecular C-C torsions and do show a considerable variation with chain length. It is these C-C torsional vibrations that are modified on doping with sodium.

3.1.4. Lattice vibrations in L-alanine

Zwitterionic L-alanine ($+H_3N$-$C(CH_3)$-CO_2-) is a dipolar molecule that forms large, well-ordered crystals in which the molecules form hydrogen-bonded columns. The strong interactions lead to the presence of well-defined intramolecular and intermolecular vibrations which can usefully be described using harmonic theory.

Coherent inelastic neutron scattering experiments have been combined with normal-mode analyses to examine the collective vibrations in L-alanine [16].

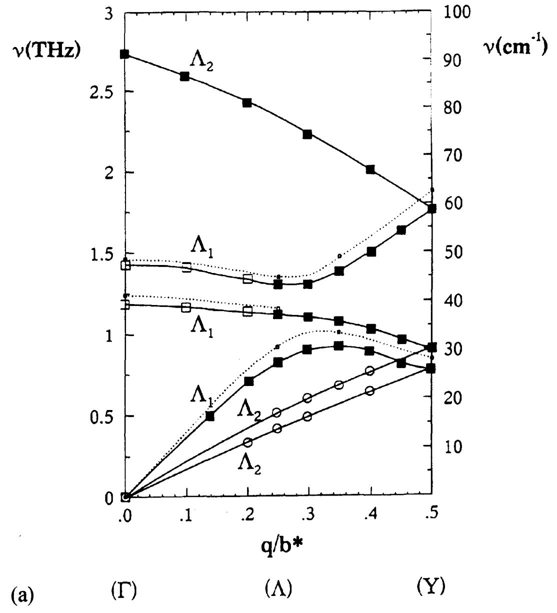


Fig. 13. (a) Dispersion curves for crystalline zwitterionic L-alanine at room temperature along the b crystallographic direction determined by coherent inelastic neutron scattering. The full circle and full square symbols are associated with phonon modes observed in predominantly transverse and purely longitudinal configurations, respectively, i.e., for vectors $\vec{Q}$ and $\vec{q}$ perpendicular and parallel to one another, respectively. They correspond to measurements performed around the strong Bragg reflections (200), (040) and (002). The empty square symbols are neutron data points obtained around the (330), (103) and (202) reciprocal lattice point in a mixed configuration. Solid lines indicate the most probable connectivity of the dispersion curves and dashed lines correspond to measurements performed at low temperature, $T = 100$ K. (b) Theoretical dispersion curves for L-alanine determined from normal-mode analysis. From Ref. 16.*

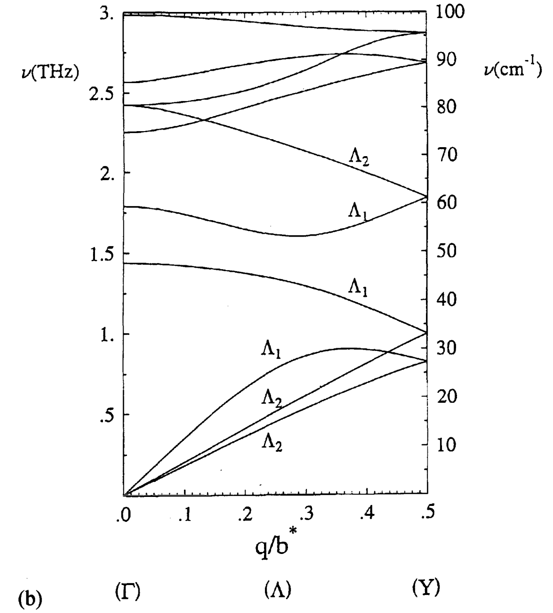


Fig. 13. (continued).

Ab initio quantum-chemical calculations were performed to determine the ammonium: carboxylate hydrogen-bonding interaction energy curve and bond rotational potentials. The molecular mechanics potential function was parametrized to fit the *ab initio* results. Using the potential function, normal-mode calculations were performed in the full configurational space of the crystal. Experiments were performed to obtain coherent inelastic neutron scattering intensities, $S_{coh}(\vec{Q}, \omega)$, and to trace the phonon dispersion relations along the three principal axes of the first Brillouin zone.

Figure 13 shows the experimental phonon frequencies $\nu_i(\vec{q})$ ($\nu = \omega/2\pi$) for several modes propagating along the crystallographic direction $\vec{b}^*$. The solid lines represent the most probable paths for the dispersion curves $\nu_i(\vec{q})$. The theoretical dispersion curves are also given. The forms of the calculated branches are similar to those determined experimentally. At the border of the zone the frequencies are in good agreement, in the range 46–62 cm^{-1} compared with 50–62 cm^{-1} experimentally. At the zone centre the frequencies of the optical modes are ~10 cm^{-1} higher than the experimental values and the gap between the frequencies of the second and the third optical modes is smaller in the calculations than in the experiment.

A complete representation of the experimental and calculated neutron data involves a three-dimensional plot of the intensities versus $\vec{q}$ and ω. This is shown for the $\vec{b}^*$ direction in Fig. 14. The comparison shows that the positions and relative intensities of the theoretical and experimental peaks corresponding to the acoustic and the first two optic modes are in agreement.

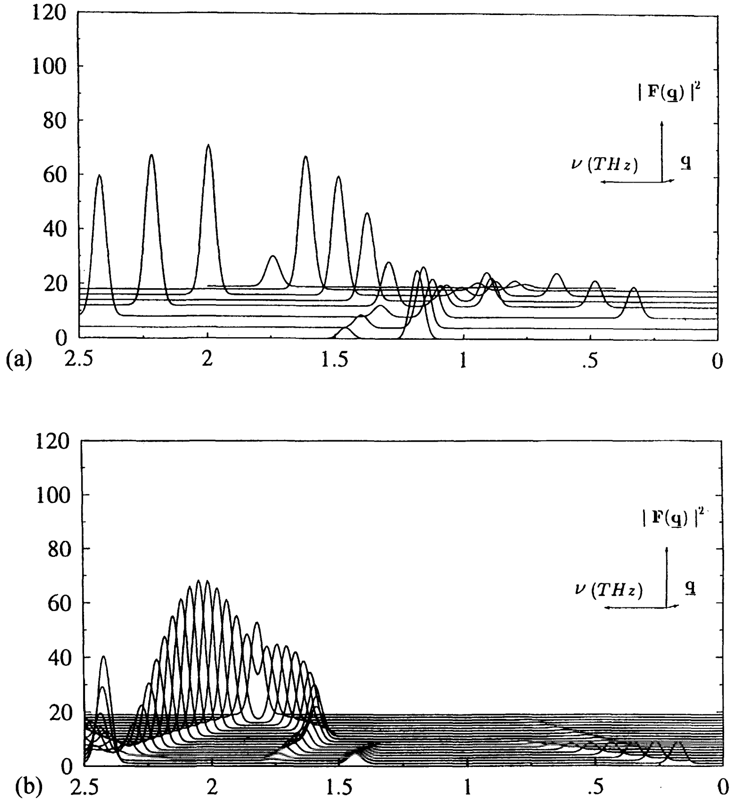


Fig. 14. (a) Experimental and (b) calculated coherent inelastic neutron intensities for crystalline zwitterionic L-alanine in the b direction. From Ref. 16.*

The calculated sound velocities, mean-square displacements, dispersion curves and coherent inelastic neutron scattering intensities were found to be mostly in quantitative agreement with experiment. An exception is the low-wavevector portion of the longitudinal acoustic branch in the $\vec{c}^*$ direction, for which the associated experimentally determined sound velocity is a remarkably high 6.6 km/s, twice the theoretical value. The $\vec{c}^*$ direction is roughly parallel to the aligned zwitterionic dipoles. It is possible that polarization effects and/or long-range dipolar correlations, such as were discussed for water in the previous section, play a role in determining the high sound velocity and were not completely represented in the potential function.

3.2. Diffusive motions in molecular crystals

Molecules or parts of molecules in crystals can undergo diffusive motion on timescales accessible to molecular dynamics simulation. These motions can also be probed using incoherent quasielastic and elastic neutron scattering.

3.2.1. Methyl group rotation in the alanine dipeptide

The three methyl groups of the alanine dipeptide, (CH_3-CONH-$C^{\alpha}H(C^{\beta}H_3)$-CONH-$CH_3$), provide a system where various vibrational and diffusive motions combine to produce measured spectra and where molecular dynamics can be usefully applied to unravel their contributions [38]. One of the methyl groups is the side chain, the intrinsic (gas-phase) rotational barrier for which is ~3 kcal/mol, and the other two methyl groups are adjacent to double bonds and have intrinsic barriers of ~0 kcal/mol. In what follows we refer loosely to the side-chain and terminal groups as 'hindered' and 'free' methyls, respectively. This terminology refers to the intrinsic barrier; in the crystal an effective rotational barrier exists for all methyls due to the influence of nonbonded interactions.

Figure 15 shows the Q-dependence of the experimental elastic scattering, $A_{tot}(Q)$. Also shown is $A_{tot}(Q)$ derived, using Eq. 25, from all the hydrogens in molecular dynamics simulations of the dipeptide crystal. The dashed lines represent $A_{tot}(Q)$ derived using a correction formula described below. Manipulation of Eq. 27 indicates that an EISF that is Gaussian in Q (linear in Fig. 15) is consistent with the presence of harmonic motion. At 300 K in both experiment and simulation the EISFs are nonlinear. At 50 K both curves are almost Gaussian in Q.

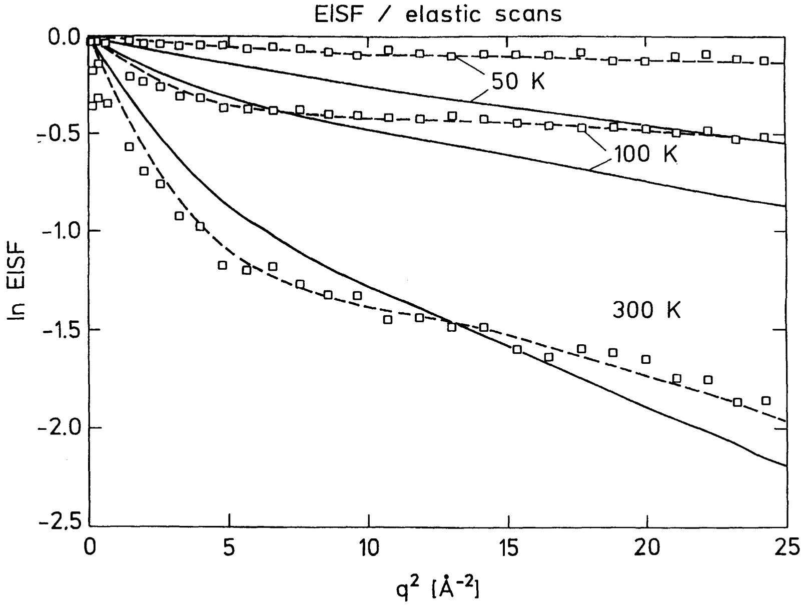


Fig. 15. Log of the elastic intensity versus q^2 for the crystalline alanine dipeptide. (□): experimental; (—): from molecular dynamics simulation; (- - -): using a correction formula (Eq. 63). The values of p, the population factor, are 0 for the simulation at 50 K, 0.15 at 100 K and 0.35 at 300 K. The corresponding values for the subtracted vibrational rms displacements, u_{Δ}^2, are 0.130 Å^{2}at 50 K, 0.133 at 100 K and 0.140 at 300 K. From Ref. 38.

We now examine decompositions of the simulation-derived $A_{tot}(Q)$. Figure 16a shows the contribution to $A_{tot}(Q)$ from the translational rigid-body motion of the free and hindered methyl groups. The hindered and free groups present almost identical translational elastic scattering, linear in Q^2, i.e. vibrational. Figure 16b presents the equivalent rotational contribution. For the free groups, nonlinearity is clear at all temperatures. The form of the curve at 300 K is close to that of the spherical Bessel function found for the EISF calculated from an analytical model of continuous diffusion of the methyl hydrogens on a circle [7]. The hindered groups have a comparatively weak Q-dependence with some nonlinearity. The hindered methyl rotational motion is not as fully developed in the simulations as it is experimentally.

The elastic scattering at high Q in the simulations is increased with respect to the experiment at all temperatures. This is mainly due to the fact that, in the experiment, the detector efficiencies were normalized with respect to the 22 K scattering, leading to an experimental underestimation of the vibrational mean-square fluctuations at higher temperatures.

One can summarize the above considerations formally in the following 'correction formula' for the EISF:

$$A_{tot,corr}(Q) = A_{tot,sim}(Q)\cdot\frac{A_{+}(Q)}{A_{-}(Q)} \tag{63}$$

$A_{+}(Q)$ is the EISF for rotational motion to be added to the simulation and $A_{-}(Q)$ is the EISF for vibrational motion to be subtracted from the simulation:

$$A_{+}(Q) = 1 - p(1 - A_{rot}(Q)) \tag{64}$$

$$A_{-}(Q) = e^{-Q^2\langle u_{\Delta}^2\rangle} \tag{65}$$

In Eq. 64, p is the fraction of methyl groups not showing rotational transitions in the simulation that do undergo full rotation in the timescale resolvable by the experiment. $A_{rot}(Q)$ is the EISF of the additional rotational motion. For a rotational jump model this is typically a linear combination of zeroth-order spherical Bessel functions. In Eq. 65, $\langle u_{\Delta}^2\rangle$ denotes the mean-square vibrational fluctuation to be subtracted. In Fig. 15 the dashed lines show the corrected EISF using the simple model described above. The parameters p and $\langle u_{\Delta}^2\rangle$ are given in the figure caption.

The picture of the methyl group dynamics that arises from the simulation–experimental analysis is illustrated qualitatively in Fig. 17, which shows simulation-derived time series for methyl dihedral angles calculated from a free and a hindered group. Only librations and no transitions for both types are seen at 50 and 100 K. At 300 K the free methyl group undergoes several rotational transitions involving 2–3 rad displacements in the 10 ps time period. The time series shows diffusive characteristics. For the hindered methyl, one jump-like transition is seen at 300 K, associated with a forcing and subsequent damping of the librational oscillation.

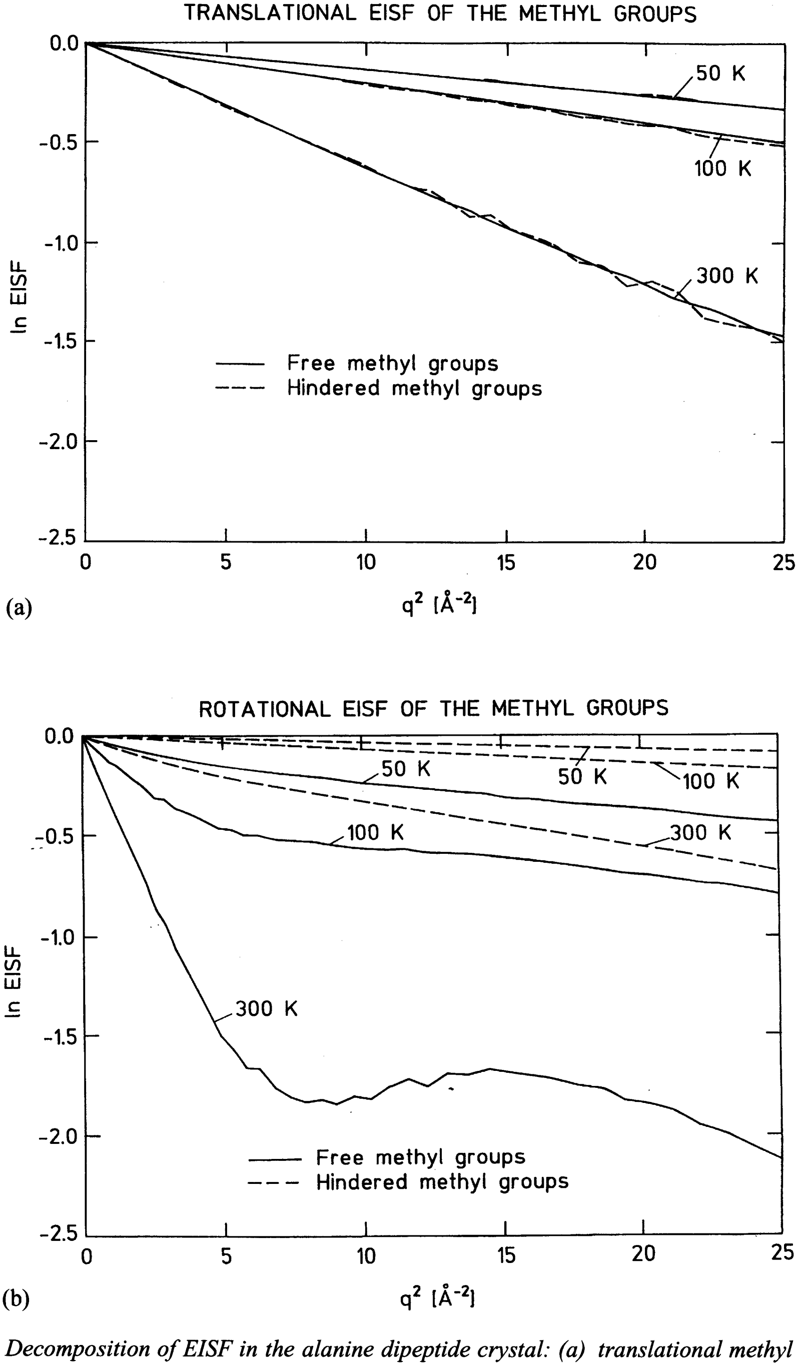


Fig. 16. Decomposition of EISF in the alanine dipeptide crystal: (a) translational methyl components; (b) rotational methyl components. From Ref. 38.

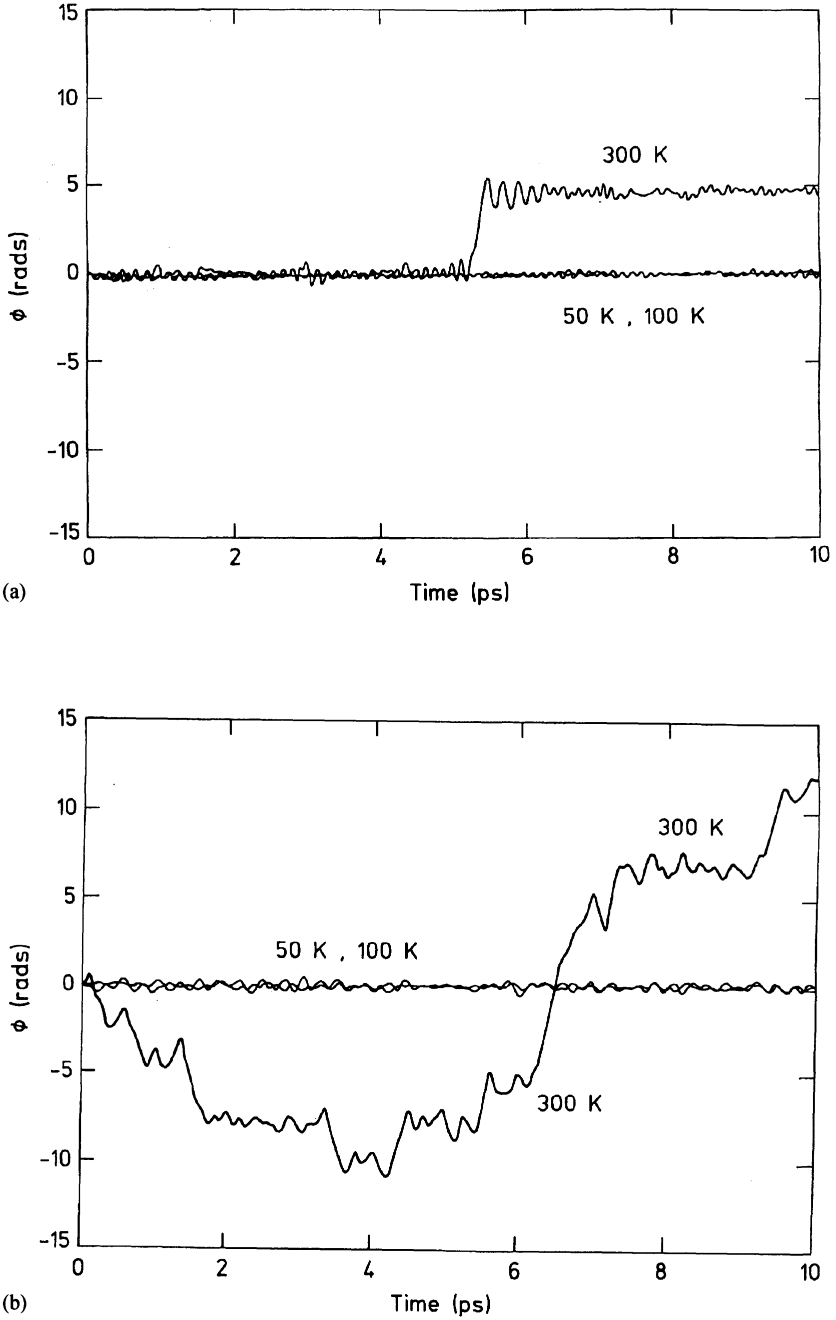


Fig. 17. Time series for methyl dihedral angles (C_3-axis) from molecular dynamics simulations of the alanine dipeptide crystal: (a) a hindered methyl; (b) a free methyl. From Ref. 38.

3.2.2. Alkane diffusion in urea inclusion compounds

Urea inclusion compounds are attractive systems for the characterization of the dynamics of n-alkane chains in a confined environment. In these systems the urea 'host' molecules form a hydrogen-bonded network containing parallel channels into which linear 'guest' molecules pack (see Fig. 18). Incoherent quasielastic neutron scattering experiments have been performed on n-nonadecane–urea at 180 K in which $S_{inc}(\vec{Q}, \omega)$ was determined with $\vec{Q}$ oriented parallel and perpendicular to the channel axes [41].

Varying the simulation model. The effect of varying the molecular dynamics simulation model system on the calculated quasielastic neutron scattering profiles was examined [39,40]. Simulations were performed with differing numbers of n-nonadecane molecules per channel and by varying the packing distance between the molecules. The effect of varying the alkane repeat distance along the channel axis on the calculated quasielastic scattering is shown in Fig. 19a, b in the $Q_{\|}$ and $Q_{\perp}$ geometries together with the corresponding experimental data. Clearly, the calculated quasielastic profiles depend strongly on the intrachannel alkane–alkane interactions. In simulations MU1 and MU3 the alkane molecule centres of mass are separated by ∞ and ~29 Å, respectively. In simulations MU5 and MU10 the repeat distance is 26.44 Å, the experimental value. In MU10 there are 10 molecules per channel in the primary box, whereas there are 5 in MU5. Figure 19 shows that simulations MU1 and MU3, in which the guest molecules are further apart than observed experimentally,

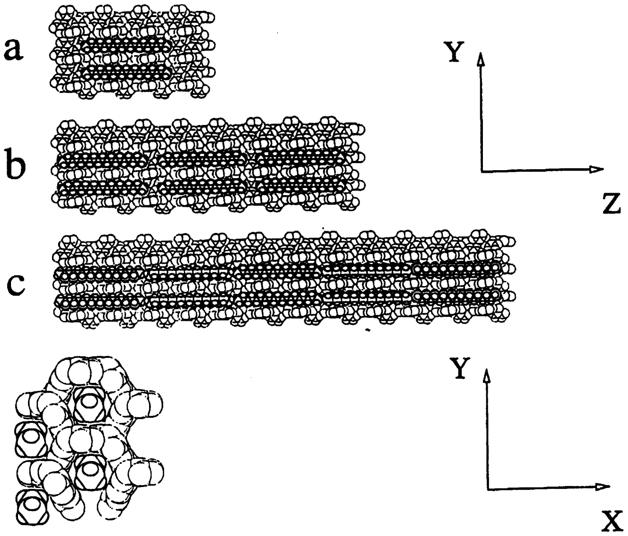


Fig. 18. Simulation primary boxes for the urea–n-nonadecane inclusion compound. Upper: in the YZ plane, (a) model MU1, (b) model MU3, (c) model MU5. The primary box of MU10 is that of MU5 but doubled in the Z direction. Lower: in the XY plane. From Ref. 40.

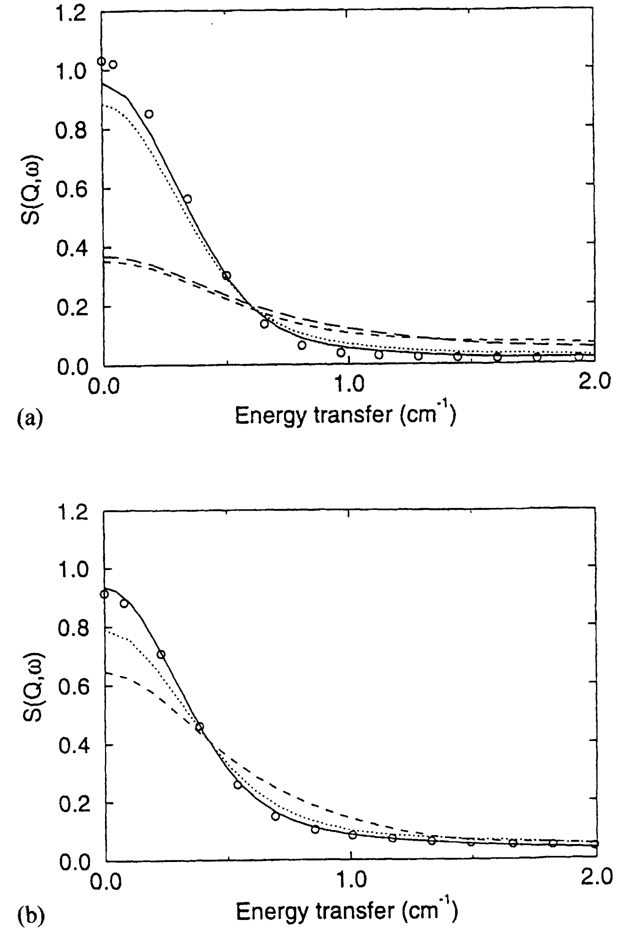


Fig. 19. (a) Dynamic structure factor for the urea–alkane inclusion compound at 180 K in the $Q_{\|}$ geometry. $Q = 1.0\,\text{Å}^{-1}$. (b) Dynamic structure factor in the $Q_{\perp}$ geometry. $Q = 1.0\,\text{Å}^{-1}$. Experiment (○), simulation MU10 (—), simulation MU5 (· · ·), simulation MU3 (- - -), simulation MU1 (– –). From Ref. 40.

are not in agreement with experiment. Their quasielastic spectra in the $Q_{\|}$ direction are too broad, indicating the presence of too fast diffusive motion, and the chains are not sufficiently confined. Figure 19b indicates that the guest–guest interactions also noticeably affect the dynamics in the orthogonal $Q_{\perp}$ directions. In both the $Q_{\|}$ and $Q_{\perp}$ geometries, the best agreement with experiment is obtained with the experimentally determined n-nonadecane repeat distance. Other calculations indicated that fixing the urea molecules also leads to quantitative disagreement with the experimental quasielastic spectra.

Effect of instrumental energy resolution. The effect of finite instrumental energy resolution on the scattering functions in the alkane–urea system has been investigated. Figure 20 shows the intermediate scattering functions perpendicular to the channel axis, calculated with and without taking into account the energy resolution of the experimental spectra in Fig. 19. When the I(Q, t) derived from the resolution-broadened dynamic structure factor has reached its long-time limit, the I(Q, t) calculated without resolution broadening continues to decrease. Therefore, the effect of the instrumental resolution is to hide slower relaxation processes in the simulation, leading to an overestimation of the long-time limit of I(Q, t): the EISF. Thus, slow motions exist in the simulation that were not detected by the experiment.

To further investigate the slow rotational motions, probability distributions were calculated from the simulations [40]. The rotational probability distributions for two of the molecules are shown in Fig. 21. As the chains are, in principle, equivalent, we see that the rotational dynamics has not converged over the 330 ps timescale of the simulation. However, averaging over the 20 chains in the simulation produces an approximately sixfold symmetric probability distribution (not shown), as expected from the hexagonal symmetry of the host structure. From this distribution, $\langle p(\Phi) \rangle$, a potential of mean force was calculated, and is shown in Fig. 22. The potential of mean force presents a barrier to rotational transition of $\sim k_BT$ at 180 K.

The rotational probability distributions determine the EISF. An experimental EISF has been determined with $\vec{Q}$ oriented perpendicular to the channel axis [41]. It is therefore of interest to compare this with EISFs calculated from the simulation probability distributions.

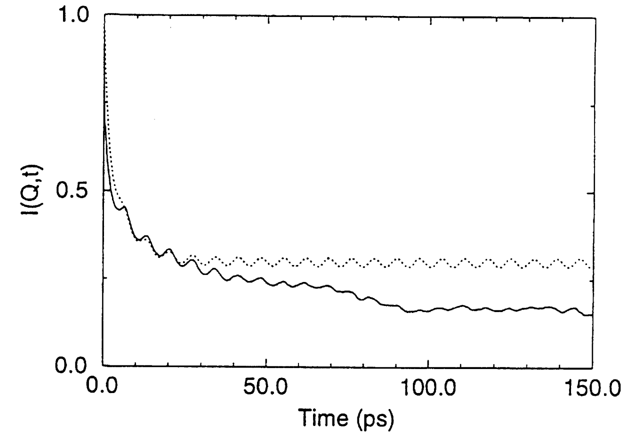


Fig. 20. Intermediate scattering function, I(Q, t) for the urea–alkane system in the $Q_\perp$ geometry calculated from molecular dynamics simulation. Q = 1.72 Å^{-1}. (—): without instrumental energy resolution effect; (· · ·): with instrumental energy resolution effect. From Ref. 40.

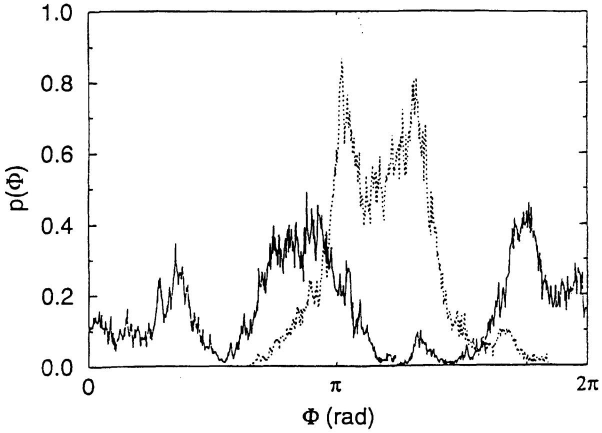


Fig. 21. Distribution of the rotational angle, ϕ, for two alkane chains of simulation MU5. From Ref. 40.

Figure 23 presents EISF curves calculated using five different methods:

(i) $A_{0,\mathrm{exp}}(Q)$ is the EISF calculated by fitting Eq. 23 to the experimental $S(\vec{Q}, \omega)$ using an analytical model in which the molecules perform restricted rotational diffusion [40]. The model was also fitted to the simulation-derived intermediate scattering function, $I(\vec{Q}, t)$, which was damped by the experimental instrumental energy resolution function. An EISF identical to experiment was obtained, and the simulation and experiment are thus in accord.

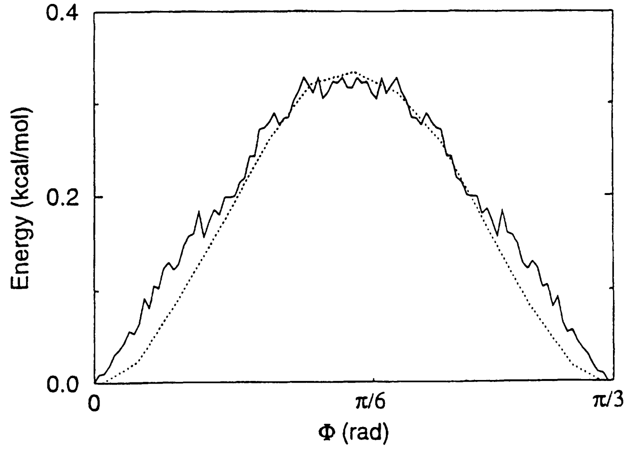


Fig. 22. (—): potential of mean force from symmetrized rotational distribution of alkane chains in the urea–alkane inclusion compound calculated from molecular dynamics simulation at 180 K; (· · ·): corresponding potential energy curve. From Ref. 40.

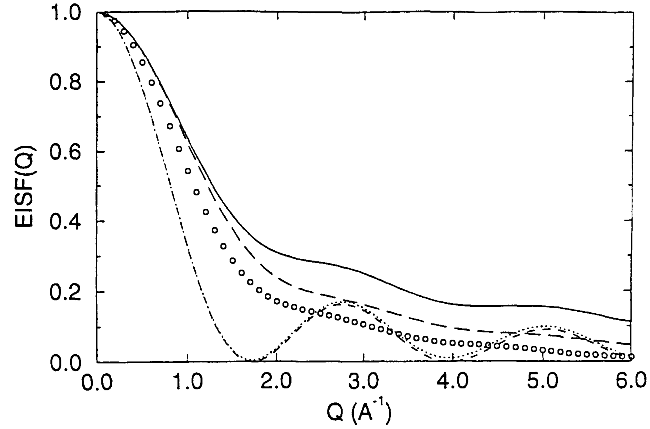


Fig. 23. EISF calculated from simulation MU5. $\langle A_{0,\Phi}(Q)\rangle$ (—); $A_{0,\langle\Phi\rangle}(Q)$ (· · ·); $A_{0,unif}(Q)$ (- - -); $A_{0,exp}(Q)$ (— — —); $A_{0,tot}(Q)$ (O). *From Ref. 40.*

(ii) $A_{0,tot}(Q)$ is the EISF derived from simulation MU5 as a sum over the atoms using Eq. 25 and is calculated from the full molecular dynamics trajectories including all the degrees of freedom of the alkane molecules.

(iii) $\langle A_{0,\Phi}(Q)\rangle$ is calculated from simulation MU5 as follows. Assuming no intramolecular motion, i.e. $u_{i,v} = 0$ and $I_v(\vec{Q}, t) = 1$ in Eq. 13, and that each molecule i performs only rotations in the plane perpendicular to the chain axis (the XY plane), the EISF can be expressed as

$$A_{0,i}(\vec{Q}) = \left| \int_0^{2\pi} p_i(\Phi)\, e^{-i\vec{Q}\cdot\vec{r}(\Phi)}\, d\Phi \right|^2 \tag{66}$$

where $\vec{r}(\Phi)$ is the vector

$$\begin{cases} r_0 \cos(\Phi) \\ r_0 \sin(\Phi) \end{cases} \tag{67}$$

and r_0 is the radius of gyration. Analysis of the alkane–urea simulations showed that the quasielastic scattering does indeed arise from the alkane molecules acting as rigid bodies. Averaging over all the orientations of $\vec{Q}$ in the XY plane, we obtain

$$A_{0,i}(Q) = \frac{1}{2\pi} \int_0^{2\pi} d\theta \left| \int_0^{2\pi} p_i(\Phi)\, e^{-i\vec{Q}(\theta)\cdot\vec{r}(\Phi)}\, d\Phi \right|^2 \tag{68}$$

where $\vec{Q}(\theta)$ is the vector

$$\begin{cases} Q \cos(\theta) \\ Q \sin(\theta) \end{cases} \tag{69}$$

The mean value, $\langle A_{0,\Phi}(Q)\rangle$, is calculated from simulation MU5 as the average of the $A_{0,i}(Q)$ of the 20 individual chains in the simulation, i.e.

$$\langle A_{0,\Phi}(Q)\rangle = \frac{1}{20}\sum_{i=1}^{20} A_{0,i}(Q) \tag{70}$$

(iv) $A_{0,\langle\Phi\rangle}(Q)$ is the EISF calculated from simulation MU5 using the distribution $\langle p(\Phi)\rangle$, averaged over the molecules and the sixfold symmetry:

$$\langle p(\Phi)\rangle = \frac{1}{N}\sum_{i} p_i(\Phi) \tag{71}$$

i.e., by analogy with Eq. 68,

$$A_{0,\langle\Phi\rangle}(Q) = \frac{1}{2\pi}\int_0^{2\pi} d\theta \left| \int_0^{2\pi} \langle p(\Phi)\rangle e^{-i\vec{Q}(\theta)\cdot\vec{r}(\Phi)} d\Phi \right|^2 \tag{72}$$

This approach resembles that employed in a recent analysis of the dynamics of a pure alkane rotator phase [42].

(v) $A_{0,\mathrm{unif}}(Q)$ is calculated using Eq. 68 and a flat distribution for $p_i(\Phi)$, i.e. $p_i(\Phi) = 1/2\pi$.

Figure 23 contains essentially two forms of curve, one in which the EISF has converged and is zero at $Q = 1.75$ Å^{-1} and $Q = 3.95$ Å^{-1}, and the other in which it has not. Averaging over $p_i(\Phi)$ (as in $A_{0,\langle\Phi\rangle}(Q)$) rather than over $A_{0,i}(Q)$ (as in $\langle A_{0,\Phi}(Q)\rangle$) has a dramatic effect on the calculated EISF and leads to an effectively converged structure factor. Lengthening further the simulation would be expected to lead to small changes in $\langle p(\Phi)\rangle$. The accompanying changes in the EISF would be expected to be negligible. That the EISF is relatively insensitive to details of $\langle p(\Phi)\rangle$ is illustrated by the fact that the EISF calculated using a flat distribution for $p(\Phi)$ and that using $\langle p(\Phi)\rangle$ are hardly distinguishable in Fig. 23. Indeed, it turns out that $\langle p(\Phi)\rangle$ possesses a symmetry that gives rise to an EISF identical to that obtained with a flat distribution. Other forms of $\langle p(\Phi)\rangle$ lead to markedly different EISFs [7].

According to the ergodic principle, the Φ distribution averaged over the 20 alkane chains of simulation MU5 is equivalent to that of a single chain sampled over a length 20 times longer than the simulation MU5, i.e. 6.56 ns. This gives an estimate of the upper limit of the time required to reach the long-time limit of the time-correlation function, $I_d(\vec{Q}, t)$.

Three of the curves shown in Fig. 23 are not close to convergence: $\langle A_{0,\Phi}(Q)\rangle$, $A_{0,\mathrm{exp}}(Q)$ and $A_{0,\mathrm{tot}}(Q)$. Although these curves are broadly similar, the differences between them are significant and merit consideration. $\langle A_{0,\Phi}(Q)\rangle$ and $A_{0,\mathrm{tot}}(Q)$ were both derived from simulation MU5 by averaging over the EISFs from the individual molecules. However, $\langle A_{0,\Phi}(Q)\rangle$ was derived using a rigid-molecule fit to the alkane atom trajectories, whereas $A_{0,\mathrm{tot}}(Q)$ includes all atomic motions in the XY plane. Therefore, internal and off-axis motions of the alkane chains have been eliminated from $\langle A_{0,\Phi}(Q)\rangle$ but are present in $A_{0,\mathrm{tot}}(Q)$. As a result, $A_{0,\mathrm{tot}}(Q)$ is lower than $\langle A_{0,\Phi}(Q)\rangle$. With a sufficiently long simulation, the individual $p_i(\Phi)$ would be expected to reach $\langle p(\Phi)\rangle$ such that $\langle A_{0,\Phi}(Q)\rangle$ would reach $A_{0,\langle\Phi\rangle}(Q)$. $A_{0,tot}(Q)$ would reach a form similar to $A_{0,\langle\Phi\rangle}(Q)$, but not exactly the same due to the additional motions contributing to $A_{0,tot}(Q)$.

In summary, therefore, the work on alkanes in urea provides an example where, due to the instrumental energy resolution, the experimentally measured quantity has not converged to its long-time limit. In contrast, with appropriate averaging techniques, molecular simulation can be used to derive the long-time behaviour. This happy situation is the reverse of that most commonly encountered in simulation studies.

4. Protein dynamics

Picosecond-timescale and Å-lengthscale dynamics in native proteins are of particular interest as they are accessible to molecular dynamics simulation. The ps timescale is also interesting physically as all the different types of motion discussed in the previous section on small molecules occur in proteins on this timescale at physiological temperatures, i.e. underdamped vibrations, overdamped vibrations, continuous and jump diffusion. Thus, ps-timescale protein dynamics possesses considerable complexity and the combination of experiment and simulation is necessary to unravel the components of the atomic motions present.

The combination of simulation and neutron scattering in the analysis of internal motions in globular proteins was reviewed in 1991 [43]. Here we briefly recall these results and complement them with some newer findings involving the comparison of simulation with neutron and X-ray diffuse scattering.

4.1. Vibrations

Vibrations in proteins can be conveniently examined using normal-mode analysis of isolated molecules. The results of such analyses indicate the presence of a variety of vibrations, with frequencies upward of a few cm^{-1}. In most cases the very lowest frequency modes dominate the calculated mean-square displacements. For example, in a normal-mode analysis of lysozyme, 80% of the mass-weighted mean-square fluctuation was found to originate from a small number of modes with frequencies $<30\,\text{cm}^{-1}$ [44]. The very low frequency modes are large-amplitude, delocalized, correlated vibrations.

4.1.1. Incoherent inelastic neutron scattering – vibrational amplitudes and damping properties

Incoherent inelastic neutron scattering, combined with normal-mode analysis, is well suited to examine low-frequency vibrations in proteins. This is primarily due to the fact that large-amplitude displacements scatter neutrons strongly. Experiments on bovine pancreatic neutron inhibitor (BPTI), combined with normal-mode analysis of the isolated protein, demonstrated that low-frequency underdamped vibrations do exist in the protein [45]. A comparison of absolute scattering cross-sections indicated that the average vibrational amplitudes were in quantitative agreement. An improved agreement with experiment was obtained by introducing a friction coefficient for each mode in a damped Langevin oscillator description [46]. The distribution of friction coefficients p(ω), obtained by fitting to the experiment, follows a Gaussian form, i.e.

$$p(\omega) = A\exp\left[\frac{\omega^2}{2\sigma^2}\right] \tag{73}$$

with $A = 30\,\mathrm{cm}^{-1}$ and $\sigma^2 = 225\,\mathrm{cm}^{-2}$. Therefore, modes with frequencies $<16\,\mathrm{cm}^{-1}$ are overdamped, that is, they do not oscillate. Thus, the very lowest frequency modes predicted by harmonic models, for example, the lysozyme hinge bending mode, do not vibrate at the calculated frequencies. They are either absent or overdamped. If overdamped, however, they can still be a source of correlated motions in proteins.

Femtosecond spectroscopic experiments have provided evidence implying low-frequency vibrations in primary electronic transitions in functional photosynthetic reaction centres [47]. The lowest frequency vibration identified had a frequency of $15\,\mathrm{cm}^{-1}$ and was underdamped, close to being critically damped. Although the form of this vibration and how it influences the electronic transitions are not yet clear, it is interesting to note that $15\,\mathrm{cm}^{-1}$ corresponds to frequencies of the lowest frequency, underdamped collective vibrations detected in small proteins using neutrons. Moreover, the damping characteristics of the vibration inferred from the femtosecond spectroscopic studies are similar to what would be expected from the damping scheme introduced phenomenologically in Eq. 73, and also with the damping properties of a $15\,\mathrm{cm}^{-1}$ rigid-helix vibration characterized in a molecular dynamics simulation of myoglobin [48].

Molecular dynamics simulation has shown that the very low frequency vibrations of myoglobin can be described in terms of rigid-helix motions [48]. However, rigid-helix motions contribute only about 30% of the mean-square displacements of helix atoms in this protein. A simplified description of the large-amplitude internal helix motions in polyalanine and myoglobin, using the P-curve algorithm [49], has recently been given [50].

Experimental incoherent neutron scattering data have been collected on tRNA and a comparison has been made with normal-mode calculations [51]. At low temperatures, a broad peak is seen in the dynamic structure factor due to the low-frequency modes. This peak is centred at $\sim 40\,\mathrm{cm}^{-1}$, somewhat higher in frequency than that observed so far in small, globular proteins. The vibrational frequency distribution calculated from the normal-mode analysis rises to a broad maximum at $\sim 50\,\mathrm{cm}^{-1}$, in general accord with experiment. However, the lowest frequency vibrations in the harmonic model ($<40\,\mathrm{cm}^{-1}$) were not present in the experimental sample. This may be due to the strong equilibrium solvation effects expected on the tRNA atoms and not included in the harmonic analysis.

4.1.2. Vibrational density of states

The low-frequency portion of the vibrational density of states, G(ω), for BPTI has been determined experimentally using Eq. 29. Subsequently, attempts were made to reproduce this frequency distribution using molecular simulation. In an initial study, normal-mode analyses were performed with different electrostatic truncation schemes [46]. The resulting G(ω)'s are compared with experiment in Fig. 24. G(ω) obtained using electrostatic truncation smoothed by a cubic switching function was found to be in better agreement with experiment than that obtained using no electrostatic truncation. One possible explanation for this is that the effect of the switch function mimics the effect of the environment in the experimental powder sample. By the analysis of two 100 ps simulations of BPTI, one in water and one in vacuum, a model of frictional damping was developed that describes the effect of water on G(ω) [52]. Of the two simulations, G(ω) calculated for the protein in water resembled more closely the form of the experimental function. It was shown that treating each vacuum principal mode as an independent damped Langevin oscillator, with the natural

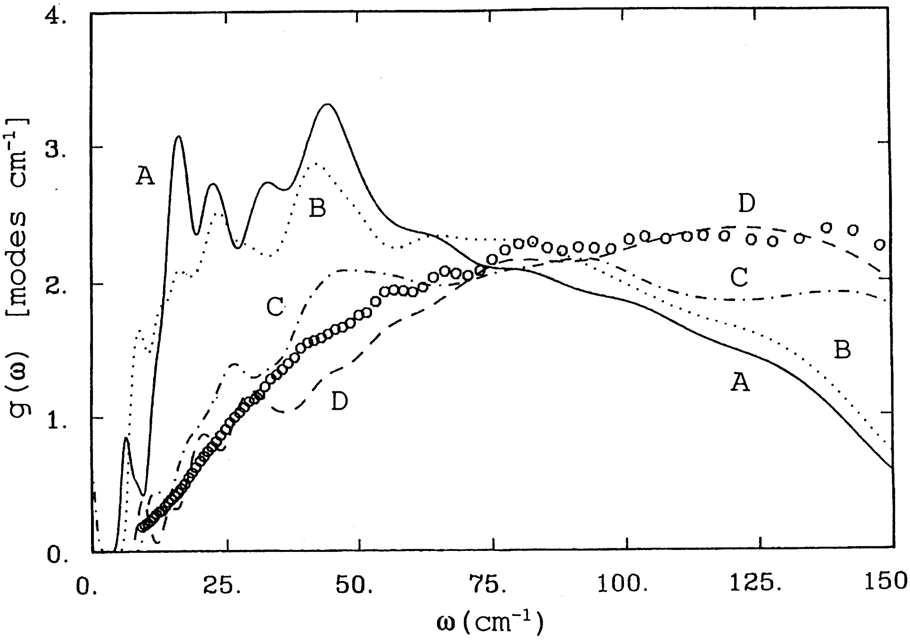


Fig. 24. Density of states, G(ω), for BPTI from experiment (circles) and from four normal-mode analyses. The analysis corresponding to curve A was performed in the extended atom approximation with no electrostatic truncation. Curve B used the extended atom approximation and shift electrostatic truncation at 7.5 Å. Curve C used the extended atom method and switch electrostatic truncation (from 6.5 to 7.5 Å). Curve D included all the atoms explicitly and used switch truncation. From Ref. 46.

frequency of each mode given by its vacuum effective frequency, and assigning all modes a friction coefficient of 47 cm^{-1}, gives a G(ω) closely similar to that obtained in the solution simulation and in the experiment. This damping scheme 'is different from that proposed in Eq. 73, but the frequency of critical damping (23.5 cm^{-1}) is similar.

The existence of temperature echoes in a molecular dynamics simulation of BPTI has been demonstrated [53]. Temperature echo involves the application of a sequence of two cooling pulses: the first creates a coherent vibrational state and the second selects the mode(s) that will echo. The frequency dependence of the depth of the echo was shown to have the same form as the experimental G(ω) for BPTI. More recent work has demonstrated that, although the echo depth is related to the density of states, vibrational dephasing due to anharmonicity of the protein also plays an important role [54]. G(ω) has also been calculated from normal-mode analyses in which the effect of the environment on the protein vibrations has been approximated [55]. The experimental G(ω) for BPTI has been employed in calculations of the low-temperature heat capacity of the protein [56].

4.1.3. Far-infrared spectroscopy

Using the National Synchrotron Light Source at Brookhaven, far-infrared absorption in the frequency range 15–45 cm^{-1} was detected in samples of lysozyme at different hydrations [18]. The form of the absorption profile was found to be temperature independent but varied significantly with the hydration of the protein. At higher hydrations the profile closely resembles that of water in the region 20–45 cm^{-1}. At low hydration marked differences were seen, with, in particular, the appearance of an absorption maximum at 19 cm^{-1}. A parallel theoretical investigation has been undertaken [57]. Preliminary results suggest that far-infrared absorption from lysozyme contains a significant component from induced dipole absorption.

4.2. Diffusive motion

4.2.1. Incoherent quasielastic neutron scattering

Above ~200 K there is a nonvibrational component to protein atom dynamics that has been detected using several experimental techniques including neutron scattering [43,58]. The dynamical transition is also present in molecular dynamics simulations [59]. There is evidence that the nonvibrational dynamics is of particular importance for the functioning of some proteins, e.g. in ligand binding [60] or proton transfer reactions [61]. Inelastic neutron scattering measurements on bacteriorhodopsin have shown that the ability of the protein to functionally relax and complete its photocycle is strongly correlated with the onset of anharmonic dynamics in the membrane [61]. Neutron experiment indicates that a dynamical transition also occurs in tRNA [51]. The transition starts at a slightly lower temperature, ~180 K, and is somewhat sharper than in proteins. This may be related to the relative simplicity of the tRNA structure.

Various models for the nonvibrational atomic motions in proteins at 300 K have been proposed. Most of them are based on the idea of transitions between conformational substates and assume individual or collective stochastic jump dynamics of the atoms between minima on the potential energy surface of the folded protein [58,62,63]. However, the observed neutron scattering profiles could originate instead from continuous diffusive motion and/or from overdamped harmonic motion. To determine the nature of the nonvibrational component requires an *ab initio* test of a given model hypothesis. This test can be made using molecular dynamics simulations, by determining to what extent the hypothetical atomic motion contributes to the simulated atomic trajectories. The contribution to the simulation-derived intensity from the simplified model dynamics can then be calculated and compared with experiment.

The dynamical transition of proteins is often discussed within the framework of the liquid-glass transition [64]. In this context one may ask whether a granularity of the high-temperature 'liquid' phase exists, i.e. whether it is possible to define subunits of proteins that can be treated analogously to molecules in a liquid. In a recent analysis the individual side chains attached to the protein backbone were considered as rigid subunits and their contribution to the neutron scattering profiles of myoglobin at physiological temperatures was calculated [65,66].

To determine the contribution of the 'side-chain liquid' to the dynamics, rigid reference structures of each side chain were fitted to the corresponding structures in each time frame of a molecular dynamics trajectory of myogobin. In this way a trajectory of the fitted atomic positions was built up and could be analysed in the same manner as the full trajectory, enabling a quantitative calculation of the rigid side-chain contribution to the neutron scattering. For the fitting procedure the C^α atoms on the protein backbone were included in the side-chain reference structures and constrained to coincide with the C^α positions in each time frame of the full molecular dynamics trajectory. In other words, the fitted rigid side chains were pinioned to the C^α atoms.

Figure 25 shows the quasielastic neutron spectra obtained from experiment and simulation. Here we are primarily concerned with the nonvibrational contribution that dominates the scattering for frequencies <1 meV. The experimental and simulation data match well in the accessible frequency range (10^{-2} to 1 meV). Clearly the rigid side-chain motions account completely for the full dynamics.

That the diffusive motion leading to the quasielastic scattering arises from rigid side-chain motions may seem surprising as most side chains contain rotatable dihedral degrees of freedom. Indeed, torsional jump models have been used to describe the quasielastic scattering from proteins. However, it is clear from Fig. 25 that conformational transitions of the side chains, although present in the simulation, do not contribute significantly to the quasielastic scattering.

The quasielastic scattering and average mean-square displacement contain a dominant component from rigid-body diffusive motions of the side chains. The displacements are caused by collisions between atoms in different side chains [48].

These 'kicks' are transmitted through the side chains via stiff covalent forces. The result is rigid-body displacement of all the atoms in a side chain. Side-chain collisions are very frequent since the atomic packing density in a protein is comparable to that of a solid. After some time the form of a side chain may change due to a torsional transition, giving rise to an error in the fitted atomic positions. However, the continuous diffusive motion of the side chains is still well described since the requirement for this is that consecutive side-chain conformations be similar. Although the side chains are flexible, they behave as rigid bodies with respect to the diffusive, liquid-like motion detected in the neutron scattering experiments.

The EISF(Q) was calculated from the simulation using Eq. 25. The experimental and simulated EISFs are plotted in Fig. 26 and match well. Comparing the rigid side-chain contribution with the result from the full trajectory shows, again, almost perfect agreement. This means that the average volume accessible to the hydrogen atoms is well sampled by the rigid side-chain motions. This conclusion is supported by calculations of the time-dependent mean-square displacements, shown in Fig. 27. In contrast, the rigid-helix displacements make a minor contribution to the helix-atom mean-square displacements.

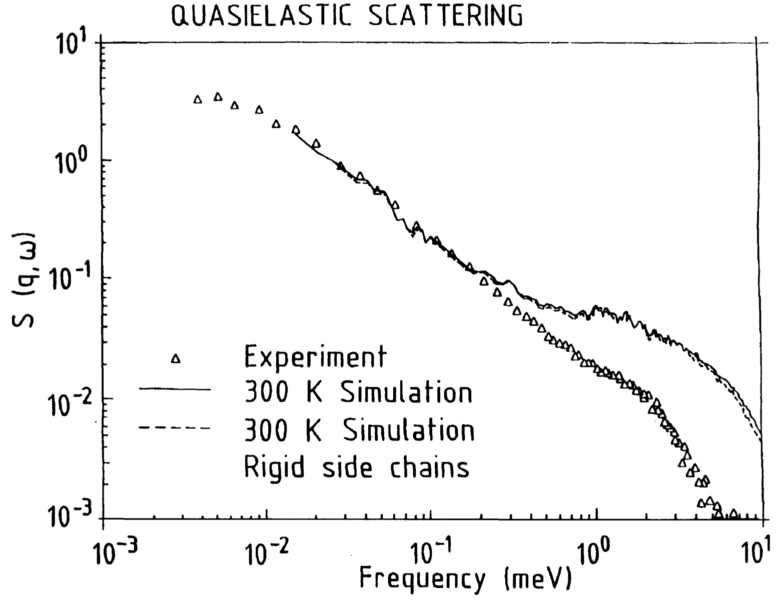


Fig. 25. Log/log plot of $S(\vec{q}, \omega)$ versus ω for myoglobin at 300 K obtained from experiment, Ref. 58 (triangles), the full molecular dynamics simulation trajectory, including internal side-chain motions (solid lines), and the fitted rigid side-chain trajectory (dashed lines). To reduce statistical errors, the experimental data were obtained by averaging over roughly Q-independent scattering profiles in a q-range of 2 Å^{-1} [58]. The simulation data represent $S(\vec{Q}, \omega)$ for $Q = 1$ Å^{-1}. From Ref. 66.

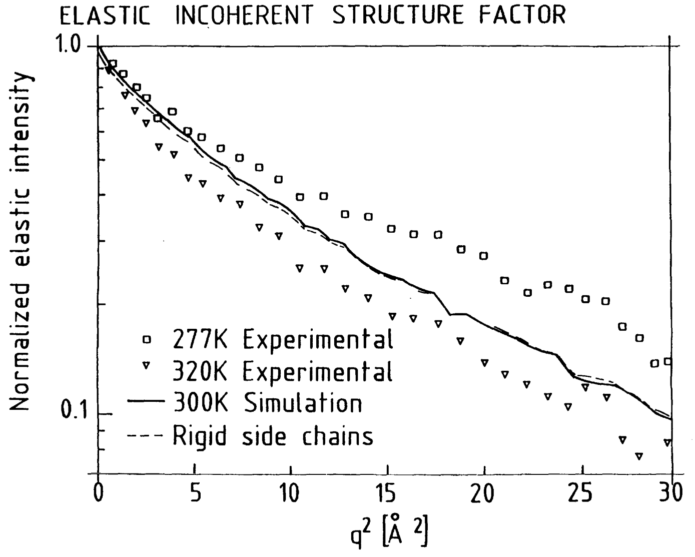


Fig. 26. Elastic incoherent structure factor for myoglobin at 300 K from the experiment, full simulation and rigid side-chain model, represented as in Fig. 25. From Ref. 66.

4.2.2. X-Ray diffuse scattering and correlated motions in lysozyme

The very diffuse scattering found in crystals of lysozyme and insulin has also been described in terms of 'liquid-like' motions [67,68], although in a different fashion to that described above; the diffuse scattering was interpreted using a phenomenological model of random atomic displacements correlated over distances <6.0 Å. This description excludes contributions to the scattering arising from correlations over longer distances. Ligand binding and cooperativity often require conformational change involving correlated displacements of atoms [69]. A simple model for the long-distance transmission of information across a protein involves the activation and amplification of correlated motions that are present in the unperturbed protein. Although long-range correlated fluctuations are required for functional, dynamic information transfer, it is not clear to what extent they contribute to equilibrium thermal fluctuations in proteins. It is therefore important to know whether equilibrium motions in proteins can indeed be correlated over long distances or whether anharmonic and damping effects destroy such correlations.

To further examine the dynamical origins of X-ray diffuse scattering by proteins, experimental scattering was measured from orthorhombic lysozyme crystals and compared with patterns calculated using molecular simulation [70]. The low intensity of the very diffuse scattering necessitated the use of synchrotron radiation with image plate detection. An experimental scattering pattern is shown in Fig. 28 together with

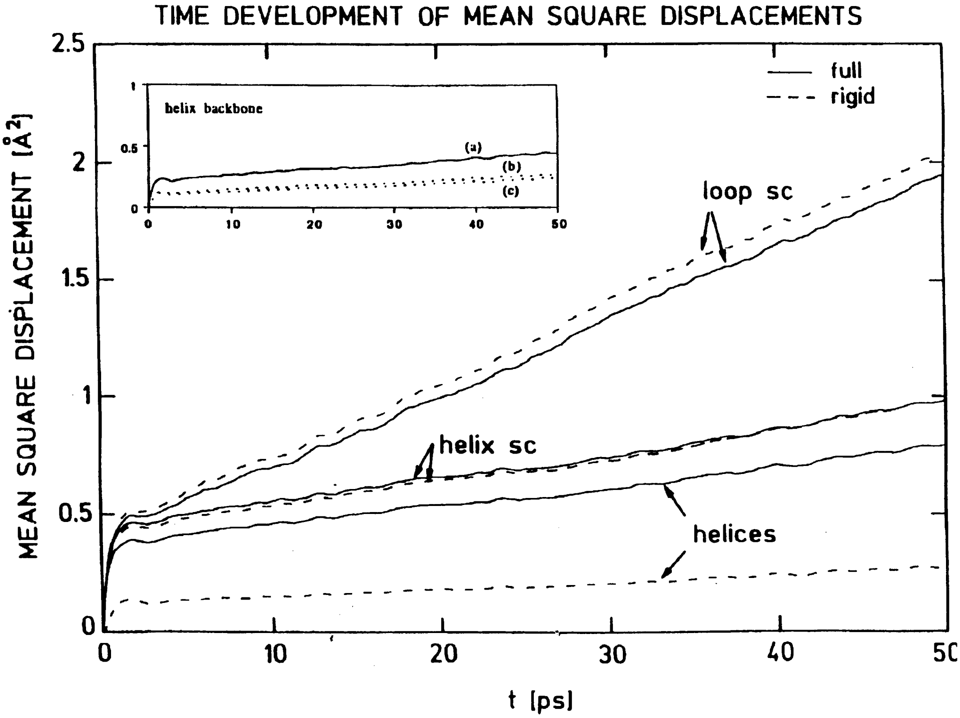


Fig. 27. Time development of the average mean-square displacement (normalized per atom), $\langle \Delta \vec{x}^2 \rangle(t)$, from a molecular dynamics simulation of myoglobin at 300 K and from a fitted rigid side-chain trajectory (sc = side chain). From Ref. 48. Inset: mean-square displacements from the full trajectory (curve a) and rigid-helix main-chain atom trajectories (curves b and c). Curve b: rigid-body fit performed using only the N, C^α and C atoms; curve c: rigid-body fit performed using all the helix atoms (main chain and side chain).

patterns calculated from a normal-mode analysis and from a molecular dynamics simulation of the isolated lysozyme molecule. Only the 15 very lowest frequency modes from the harmonic analysis were required to produce a converged pattern – the addition of further modes did not modify significantly the calculated pattern. This is partly because the lowest frequency modes dominate the mean-square displacements in the harmonic approximation and partly because they are correlated over many atoms, the diffuse scattering intensity being proportional to the number of atoms involved. The average position of the diffuse ring is reproduced by both the normal modes and the molecular dynamics. However, a closer examination reveals that the fine details are better reproduced by the normal modes.

That the scattering pattern obtained from a harmonic description of the lysozyme dynamics is in reasonable accord with the observed data is consistent with the idea

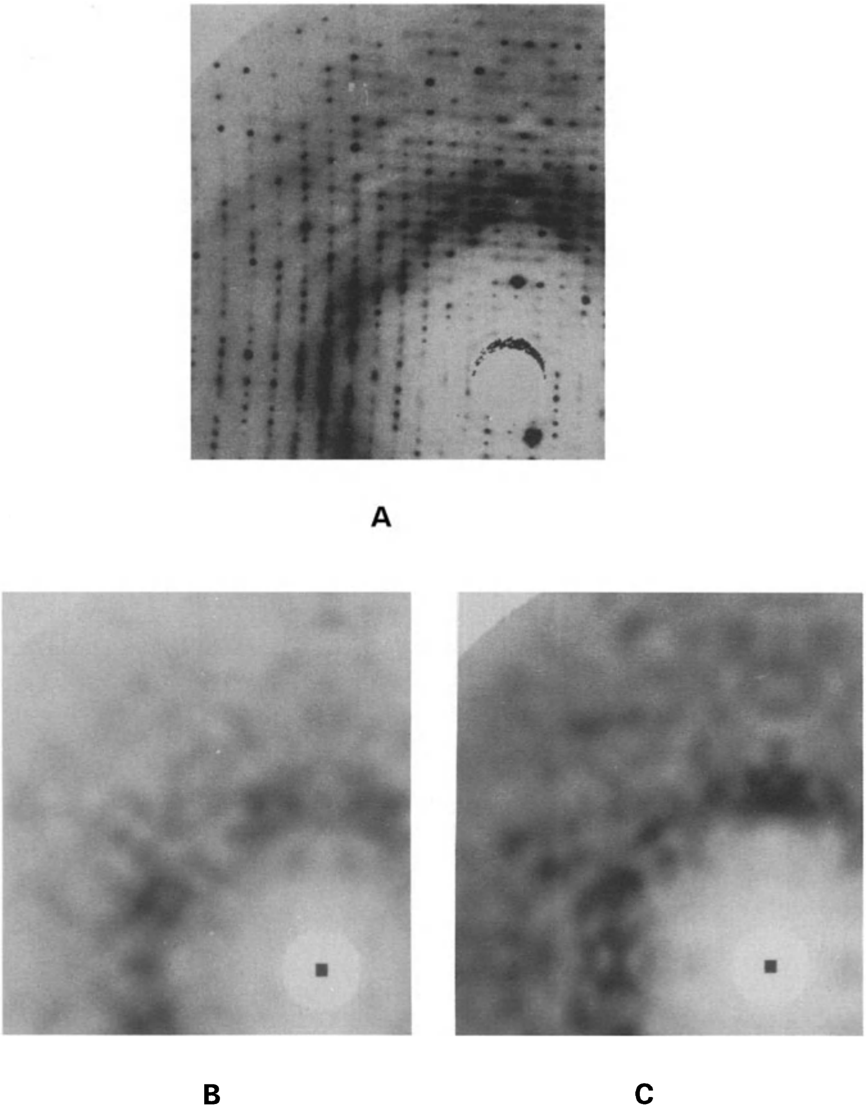


Fig. 28. Diffuse X-ray scattering patterns from orthorhombic hen egg-white lysozyme: (A) experimental; (B) from a normal-mode analysis of the lysozyme molecule; (C) from a molecular dynamics simulation. See Ref. 70 for details.

that intramolecular displacements correlated over long distances can exist, in contrast to the conclusions of the previous analyses of lysozyme and insulin [67,68]. However, as discussed in the quasielastic neutron scattering analysis, a large fraction of the atomic displacements at 300 K do originate from 'liquid-like' motions, meaning nonvibrational, diffusive dynamics [66]. In myoglobin, the major contribution to the atomic mean-square displacements can be described as diffusive motions of the side chains acting as rigid bodies, like molecules in a liquid [48].

That nonvibrational, diffusive motions exist in proteins does not contradict the above-mentioned simulation/diffuse scattering results for lysozyme. The correlated motions visible in the X-ray pattern can be diffusive or vibrational. Frictional damping of the modes, as would be incorporated in a damped Langevin oscillator description, does not affect their amplitudes and directions. Therefore, frictional damping would not affect the calculated diffuse scattering, as the diffuse scattering does not depend on the time evolution of the atomic displacements, that is, whether they vibrate or not. Thus, it is conceivable that the modes contributing to the diffuse scattering pattern are a combination of underdamped and overdamped vibrations, the latter containing a diffusive element.

The question arises as to why the diffuse scattering function calculated using a harmonic approximation to the potential function is in closer accord than that calculated using the full potential function with molecular dynamics. One reason for this is that the molecular dynamics permits the average structure to drift from the average crystallographic structure more than the harmonic analysis. Incomplete representation of the environment of the protein will exacerbate this problem. Furthermore, the diffuse scattering calculated from molecular dynamics simulations of proteins is found to converge very slowly, with significant variations over timescales of hundreds of picoseconds [71,72]. The calculation of converged diffuse scattering from molecular dynamics simulation represents a notable challenge for the future.

5. Conclusions

The combination of simulation with scattering and absorption experiments allows a wide range of dynamical phenomena to be characterized in condensed-phase molecular systems. The work described in the present article testifies to the versatility of empirical potential energy functions of the molecular mechanics type in describing motions on a range of timescales from 10^{-15} to 10^{-10} s, i.e. from fs, localized vibrations to ~100 ps activated processes.

As simulation with a molecular mechanics potential function provides, in principle, a complete description of the structure and dynamics of a crystal, many associated experimental properties can be computed. The vibrations in crystals that can be explored range from lattice phonons to localized intramolecular vibrations and the effect of the crystal environment on the intramolecular band centres and their splitting. Rather subtle anharmonic effects on soft vibrations can also be accurately represented, as evidenced by the temperature dependence of the hydrogen-bonded NH out-of-plane bending mode of the peptide group acetanilide, and its methyl torsion. The experimental determination and theoretical description of lattice vibrations in molecular crystals have hitherto mainly been confined to systems containing only a few atoms per unit cell. The work on L-alanine described here demonstrates how coherent inelastic neutron scattering experiments can be combined with harmonic analyses to characterize the low-frequency, collective vibrations of a crystal containing 52 atoms per unit cell. Coherent inelastic neutron scattering experiments on the dynamics of crystalline adenine have also been reported [73]. In a similar vein, overdamped and under-damped acoustic phonons have been observed in wet-spun DNA fibres [74]. The extension to biological macromolecular crystals awaits.

The analysis of water leads to a picture of the charge fluctuations associated with far-infrared absorption in which long-range interactions and polarization play important roles. The method described for calculating polarization effects on infrared spectra can, in principle, be extended to the calculation of the dynamical trajectories themselves. In this way a unified charge model that reproduces both the dynamics and the far-infrared absorption would be obtained. The atomic polarizabilities derived using the method described are transferable to a range of molecular systems [75]. This opens up the possibility of applying the method to many systems, including large, flexible macromolecules for which a molecular polarizability matrix method would not be useful. Calculations of the far-infrared spectra of the hydrated protein, lysozyme, using the present method, are underway in our laboratory and may afford a means of interpreting hydration-dependent experimental spectra obtained recently using synchrotron radiation [18].

Neutrons are not easy to get hold of. To produce them in controlled conditions requires a nuclear reactor or a spallation source, the cost of which does not fall into the budget of an average structural biology laboratory! Moreover, even from these sources neutron fluxes are very low, typically $\sim 10^7$ neutrons/(cm^2 s) compared, for instance, to X-ray fluxes at a synchrotron ($\sim 10^{12}$ photons/(cm^2 s)). Thus, neutron experiments suffer from a counting statistical problem that has limited the range of applications to those with large samples ($\sim 10^{-1}$ g) and long counting times ($\sim$days). However, there is some hope that a future European neutron source will be built with $\sim$ 50 times the flux of the world's most powerful present facility [76]. This would open up a new range of inelastic experiments on biological macromolecules involving specific H/D labelling, and spin echo measurements of coherent scattering that would provide information on ns-timescale correlations. However, by the time such a source is built (may be around 15 years from now), concurrent progress in computer power and simulation methodology is likely to have pushed the timescale of events accessible to atomic-detail computer simulation well beyond the ns regime. Nevertheless, the general strategy outlined here for combining simulation with experiment will still be applicable and simulations will be required for the reliable interpretation of experiments probing the long-time dynamics of complex biological systems. The detailed information on the forces present in interatomic potential energy functions will thus be incorporated into the analysis of the experimental data, and an unequivocal description of the behaviour will be obtained by decomposition of simulations.

Acknowledgements

The following members and ex-members of the Molecular Simulation group at Saclay contributed to the work described in the present chapter: Michel Ferrand, Sylvie Furois-Corbin, Larry Hayward, Stephanie Héry, Gerald Kneller, Alex Micu, Frederico Nardi, Benoit Roux and Marc Souaille. The author would like to thank the above people and the numerous external collaborators for their valuable contributions.

References

1. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
2. Van Hove, L., Phys. Rev., 95(1954)249.
3. Van Hove, L., Physica, 24(1958)404.
4. Lovesey, S., Theory of Thermal Neutron Scattering from Condensed Matter, International Series of Monographs on Physics, Vol. 72, Oxford Science, Oxford, 1984.
5. Cowley, J.M., Diffraction Physics, North-Holland, Amsterdam, 1975.
6. Kneller, G., Keiner, V., Kneller, M. and Schiller, M., Comput. Phys. Commun., 91(1995)191.
7. Bee, M., Quasielastic Neutron Scattering: Principles and Applications in Solid State Chemistry, Biology and Materials Science, Adam Hilger, Bristol, 1988.
8. Smith, J.C., Cusack, S., Brooks, B., Pezzeca, U. and Karplus, M., J. Chem. Phys., 85(1986)3636.
9. Calmettes, P., Roux, B., Durand, D., Desmadril, M. and Smith, J.C., J. Mol. Biol., 231(1993)840.
10. Calmettes, P., Durand, D., Desmadril, M., Minard, P., Receveur, V. and Smith, J.C., Biophys. Chem., 53(1994)105.
11. Hayward, R.L., Middendorf, H.D., Wanderlingh, U. and Smith, J.C., J. Chem. Phys., 102(1995)5525.
12. Barthes, M., Kellouai, H., Page, G., Moret, J., Johnson, S.W. and Eckert, J., Physica D, 68(1993)45.
13. Micu, A. and Smith, J.C., Comput. Phys. Commun., 91(1995)331.
14. Doucet, J. and Benoît, J.-P., Nature, 325(1987)643.
15. Glover, I.D., Harris, G.W., Helliwell, J.R. and Moss, D.S., Acta Crystallogr., B47(1991)960.
16. Micu, A., Durand, D., Quilichini, M., Field, M.J. and Smith, J.C., J. Phys. Chem., 99(1995)5645.
17. Bellissent-Funel, M.C., Teixera, J., Chen, S.H., Dorner, B., Middendorf, H.D. and Crespi, H.L., Biophys. J., 56(1989)713.
18. Moeller, K.D., Williams, G.P., Steinhauser, S., Hirschmugl, C. and Smith, J.C., Biophys. J., 61(1992)276.
19. Gordon, R.G., Advances in Magnetic Resonance, Vol. 3, Academic Press, New York, NY, 1968.
20. Applequist, J., Carl, J.R. and Fung, K.K., J. Am. Chem. Soc., 94(1972)2952.
21. Thole, B.T., Chem. Phys., 59(1981)341.
22. Souaille, M. and Smith, J.C., Mol. Phys., 87(1996)1333.

23. Neumann, M. and Steinhauser, O., Mol. Phys., 39(1980)437.
24. Barthes, M., Eckert, J., Johnson, S.W., Moret, J., Swanson, B.I. and Unkefer, C.J., J. Phys. I France, 2(1992)1929.
25. Johnston, C.T., Agnew, S.F., Eckert, J., Jones, L.H., Swanson, B.I. and Unkefer, C.J., J. Phys. Chem., 95(1991)5281.
26. Hasted, J.B., Husain, S.K., Frescura, F.A.M. and Birch, J.R., Chem. Phys. Lett., 118(1985)622.
27. Madden, P.A. and Impey, R.W., Chem. Phys. Lett., 123(1986)502.
28. Guillot, B., J. Chem. Phys., 95(1991)1543.
29. Jorgensen, W.L., Chandrasekhar, J.D., Madura, R.W., Impey, R.W. and Klein, M.L., J. Chem. Phys., 79(1983)926.
30. Andrea, T.A., Swope, W.C. and Andersen, H.C., J. Chem. Phys., 79(1983)4576.
31. Neumann, M., Steinhauser, O. and Pawley, G.S., Mol. Phys., 52(1984)97.
32. Neumann, M., J. Chem. Phys., 85(1986)1567.
33. Proceedings International Congress on Synthetic Metals 1990 [Synth. Met., 49 (1991)].
34. Sauvajol, J.L., Djurado, D., Dianoux, A.J., Theophilou, N. and Fischer, J.E., Phys. Rev. B, 43(1991)14305.
35. Sauvajol, J.L., Djurado, D., Dianoux, A.J. and Fischer, J.E., J. Chim. Phys., 89(1992)969.
36. Dianoux, A.J., Kneller, G.R., Sauvajol, J.L. and Smith, J.C., J. Chem. Phys., 99(1993)5586.
37. Dianoux, A.J., Kneller, G.R., Sauvajol, J.L. and Smith, J.C., J. Chem. Phys., 101(1994)634.
38. Kneller, G.R., Doster, W., Settles, M., Cusack, S. and Smith, J.C., J. Chem. Phys., 97(1992)8864.
39. Souaille, M., Smith, J.C., Dianoux, A.J. and Guillaume, F., In Baus, M., Rull, L.F. and Ryckaert, J.-P. (Eds.) Observation, Prediction and Simulation of Phase Transitions in Complex Fluids, NATO ASI Series C, Vol. 460, Kluwer, Dordrecht, 1995, p. 609.
40. Souaille, M., Guillaume, F. and Smith, J.C., J. Chem. Phys., 105(1996)1.
41. Guillaume, F., Sourisseau, C. and Dianoux, A.J., J. Chim. Phys., 88(1991)1721.
42. Ryckaert, J.-P., Klein, M.L. and MacDonald, I.R., Mol. Phys., 83(1994)439.
43. Smith, J.C., Q. Rev. Biophys., 24(1991)227.
44. Go, N., Biophys. Chem., 35(1990)105.
45. Cusack, S., Smith, J.C., Finney, J.L., Tidor, B. and Karplus, M., J. Mol. Biol., 202(1988)903.
46. Smith, J.C., Cusack, S., Tidor, B. and Karplus, M., J. Chem. Phys., 93(1990)2974.
47. Vos, M.H., Rappaport, F., Lambry, J.-C., Breton, J. and Martin, J.-L., Nature, 363(1993)320.
48. Furois-Corbin, S., Smith, J.C. and Kneller, G.R., Proteins Struct. Funct. Genet., 16(1993)141.
49. Sklenar, H., Etchebest, C. and Lavery, R., Proteins Struct. Funct. Genet., 6(1989)46.
50. Furois-Corbin, S., Smith, J.C. and Lavery, R., Biopolymers, 35(1995)555.
51. Nardi, F., Doster, W., Tidor, B., Karplus, M., Cusack, S. and Smith, J.C., Isr. J. Chem., 34(1994)233.
52. Hayward, S., Kitao, A., Hirata, F. and Go, N., J. Mol. Biol., 234(1993)1207.
53. Becker, O.M. and Karplus, M., Phys. Rev. Lett., 70(1993)3514.
54. Xu, D., Schulten, K., Becker, O.M. and Karplus, M., J. Chem. Phys., 103(1995)3112.
55. Yoshioki, S., J. Comput. Chem., 15(1994)684.
56. Edelman, J., Biopolymers, 32(1992)209.
57. Souaille, M. and Smith, J.C., in preparation.
58. Doster, W., Cusack, S. and Petry, W., Nature, 337(1989)754.

59. Smith, J.C., Kuczera, K. and Karplus, M., Proc. Natl. Acad. Sci. USA, 87(1990)1601.
60. Rasmussen, B.F., Stock, A.M., Ringe, D. and Petsko, G.A., Nature, 357(1992)423.
61. Ferrand, M., Dianoux, A.J., Petry, W. and Zaccai, G., Proc. Natl. Acad. Sci. USA, 90(1993)9668.
62. Loncharich, R.J. and Brooks, B.R., J. Mol. Biol., 213(1990)351.
63. Elber, R. and Karplus, M., Science, 235(1987)318.
64. Angell, C.A., Proc. Natl. Acad. Sci. USA, 92(1995)6675.
65. Smith, J.C. and Kneller, G.R., Mol. Sim., 10(1993)363.
66. Kneller, G.R. and Smith, J.C., J. Mol. Biol., 242(1994)181.
67. Caspar, D.L.D., Clarage, J., Salunke, D.M. and Clarage, M., Nature, 332(1988)659.
68. Clarage, J., Clarage, M., Phillips, W., Sweet, R. and Caspar, D., Proteins Struct. Funct. Genet., 12(1992)145.
69. Gerstein, M., Lesk, A.M. and Chothia, C., Biochemistry, 33(1994)6739.
70. Faure, P., Micu, A., Perahia, D., Doucet, J., Smith, J.C. and Benoît, J.-P., Nature Struct. Biol., 2(1994)124.
71. S. Héry, Rapport de Thèse de Diplôme d'Etudes Approfondies, Université Paris VI, 1994.
72. Clarage, J.B., Romo, T., Andrews, B.K., Pettitt, B.M. and Phillips, G.N., Proc. Natl. Acad. Sci. USA, 92(1995)3288.
73. Martel, P., Frank, V. and Hennion, M., Phys. Rev. A, 41(1990)7006.
74. Grimm, H., Stiller, H., Majkrzak, C.F., Rupprecht, A. and Dahlborg, U., Phys. Rev. Lett., 59(1987)1780.
75. Voisin, C. and Cartier, A., J. Mol. Struct., 286(1993)35.
76. Report, European Science Foundation Workshop on Past, Present and Future Uses of Neutron Scattering, Autrans, European Science Foundation, 1996.

Part V
Protein folding

Protein structure prediction by global energy optimization

Ruben A. Abagyan
The Skirball Institute of Biomolecular Medicine, Biochemistry Department,
NYU Medical Center, New York University, 540 First Avenue, New York, NY 10016, U.S.A.

Introduction

The theoretical prediction of biomolecular structure from first principles and without the crutches of experimental restraints remains a dream. Most theoreticians agree that the answer is the global minimum of the free energy function [1,2], but disagree about strategies to find the minimum. Several schools of thought have formed over the years: dynamicists [3–11], minimalists [12–32], and synthesists [2,33–44]. Dynamicists believe that sufficiently long simulations of a quasi-continuous trajectory of molecular dynamics of atomic models *in vacuo* or in water will solve the problem using new generations of computers, code parallelization [45,46], and optimized simulation techniques. Minimalists, unwilling to play power games and too impatient to wait until new generations of processors cover the next mile of a hundred-mile road, simplify the system by using a reduced atomic representation or a lattice, inventing a potential and then enjoying the luxury of always finding the global minimum of their energies as well as most of the other possible states for a chain of up to a hundred simplified residues [27,47]. The third school shares the impatience of minimalists, yet resists the temptation to use simple models since it appears that accuracy is a pivotal issue. Synthesists focus on the development of algorithms to replace molecular dynamics as a generator of conformational changes [42,43,48,35] and the design of methods of energy calculation which combine accuracy and speed.

Let us list some of the ideas and assumptions of the synthesists, including the author, which the following review is based upon.

1. Oscillations of bond lengths and bond angles are not essential for protein structure prediction and some of these degrees of freedom are not even excited at room temperature. Therefore, using torsion angle space instead of Cartesian coordinate space is highly preferable since it reduces the dimensionality of the problem by a factor of 7, eliminates fast oscillations, and smoothens the energy landscape.

2. A continuous molecular dynamics trajectory is not really necessary for structure prediction. The optimal structure can be found by a global optimization algorithm making much larger steps.

3. Explicit consideration of water molecules can also be sacrificed in simulations of folding for the following reasons: (i) too many additional degrees of freedom; (ii) a really large box is necessary because of the long-range nature of the electrostatic

interactions; and (iii) the relaxation time of water molecules after a large conformational change is prohibitively long. Concurrently, the solvation effect can be evaluated by continuous approximations more efficiently and, potentially, more accurately.

4. The correct conformation and an enormous number of alternative conformations of a polypeptide chain may have very close energies. A high accuracy of energy calculations is absolutely essential to recognize the correct answer.

This review justifies and describes an emerging general method of biased probability global optimization of a detailed energy function in the internal coordinate space of arbitrarily constrained molecular models, and demonstrates that the same method can be applied uniformly to a wide variety of modeling problems such as peptide folding, homology modeling, protein design, macromolecular docking, and domain rearrangements.

A better model for structure prediction

If polypeptide models can be slightly simplified by constraining bond lengths and bond angles to their ideal values, without a critical loss of accuracy, then this strategy is strongly preferable. Not only because in an unconstrained Cartesian representation the number of degrees of freedom is 7 times larger, or because soft and hard modes are mixed in a Cartesian coordinate, but also because the local energy surface becomes smoother and minimization with respect to the torsion angles has a much larger radius of convergence [40].

The radius of convergence can be evaluated by generating a series of randomly distorted conformations and testing the ability of a minimization procedure to restore the initial energy-minimized conformation. Such an experiment was performed with a small globular protein, 29-residue trypsin inhibitor, determined by NMR spectroscopy [49]. To compare the radii of convergence in torsion and Cartesian spaces, we performed two series of energy minimizations starting from 600 randomly distorted conformations with average torsion angle deviations up to 43° and Cartesian rms deviations up to 5 Å. The initial undistorted conformation was obtained separately for each of the two series, by regularization and minimization of the ECEPP/2 energy [34] with respect to torsion angles, and by molecular dynamics and minimization of the CHARMM energy [4] in Cartesian coordinate space, respectively.

The results revealed a qualitative difference between the two representations and the corresponding energy landscapes. While the torsion minimization restored with high accuracy most of the conformations distorted up to 2 Å of coordinate rms deviation or 12° of torsion rms deviation, and some conformations even at amplitudes of 33°, minimization with soft bonds and angles in Cartesian space could not fully restore the optimal conformation even after relatively small distortions; conformational changes resulting from minimization were small, and the starting conformation was never reached even with 0.25 Å accuracy if deviation exceeded 5° or about 1 Å (Fig. 1). Usage of molecular dynamics in combination with local minimization did not change the fundamental picture. Based on this test, we may conclude that

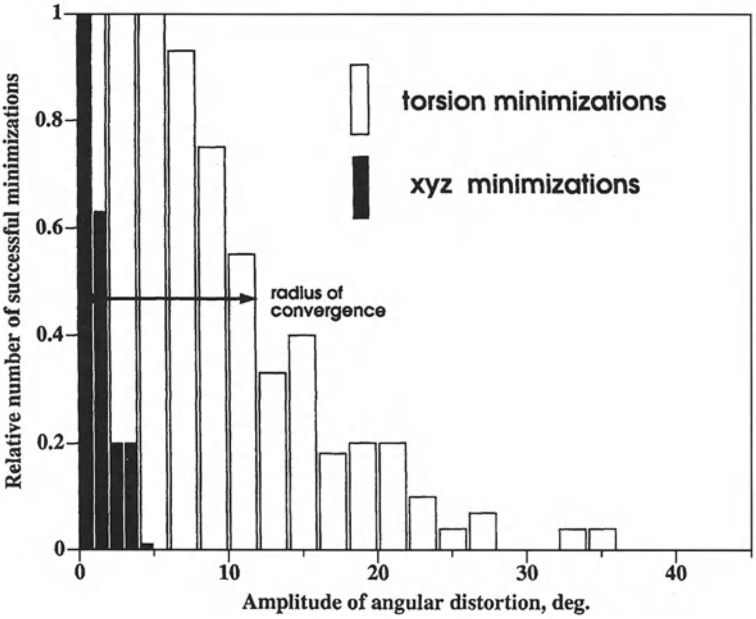


Fig. 1. Relative number of randomly distorted conformations which restored the initial low-energy conformation with 0.25 Å accuracy after energy minimization in Cartesian space (solid) and torsion space (open) [40].

unconstrained models in Cartesian space appear to be 'glassy', i.e. they always find a local minimum close to the start.

Conversely, protein models with fixed covalent geometry are much more 'elastic', i.e. they tend to travel far to find the minimum. An important consequence of the observed difference between radii of convergence is that the number of energy minima in torsion space is an exponentially small fraction of the number of minima in Cartesian space. The histogram in Fig. 1 suggests that, for the considered example, the average well size for each dimension is about 5 times smaller. This number taken to a power of the number of degrees of freedom makes an astronomically large number, suggesting that torsion space is a better choice for problems of global minimization and refinement.

It is tempting to declare conclusively that only fixed covalent geometry should be used for large-scale protein structure predictions, but we still do not know if the accuracy of the idealized geometry approximation is sufficient (see the discussion of the accuracy of energy calculations below). Deformations of bond angles even by a few degrees may be essential for tight, buried, or proline-containing turns. Nonetheless, there are several reasons, although not decisive, to believe that the accuracy of the torsion model may still be sufficient. First, in the ECEPP force field [33,34,36] special measures are taken to reduce barriers of rotation due to the fixed geometry (the repulsion between two atoms connected via three bonds is reduced by half). Second, a distortion of the bond angle by several degrees may usually be compensated by small changes in the surrounding torsion angles, although particularly difficult geometries do exist. Third, most of the protein structures determined by X-ray crystallography at high resolution can be represented by relaxed standard covalent geometry models with 0.2–0.4 Å rms deviation from experimental coordinates [50].

Internal coordinate mechanics (ICM)

If we do want to impose covalent geometry constraints, at least sometimes, or even form rigid bodies with frozen internal structure, it is important to choose a representation of molecular geometry which makes these operations simple. It is certainly not easy in Cartesian representation, where individual atoms are not geometrically connected (i.e. change of one coordinate of an atom does not affect positions of other atoms). On the other hand, pure torsion space is incomplete and cannot become a universal alternative to the Cartesian representation. Fortunately, a few extensions make the torsion representation of molecular geometry much more versatile and applicable to a broader range of modeling tasks [51,52,40].

The ICM tree was an attempt to design a more general representation of several arbitrarily constrained molecules potentially containing flexible bond angles and bond lengths. The topological tree of an ICM model of an arbitrarily constrained multimolecular system (Fig. 2) grows from the origin, contains additional nodes, so-called virtual atoms and virtual bonds, and covers all the molecules in the system. The choice of internal coordinates of four kinds, viz. bond length (**b**), bond angle (ω), phase angle (Φ) for a secondary branch of the tree, and torsion angle (φ) for the main branch, rather than that of three kinds, viz. **b**, ω, and an independent torsion for each branch of the tree (φ′), is based on the idea of separating hard (**b**), intermediate (ω, Φ), and soft (φ) degrees of freedom between different independent variables.

A wide variety of models can be constructed by an appropriate selection of constrained internal coordinates. For example, the idealized geometry model is a particular case with the following set of constraints, $b_i = \text{const}$, $\omega_i = \text{const}$, $\Phi_i = \text{const}$, $\varphi_i^{\text{ring}} = \text{const}$, while in a model for rigid-body docking all but six variable parameters specifying the position of the second molecule are fixed (see Φ_{11}, ω_{11}, $\mathbf{b}_{11}$, φ_{12}, ω_{12}, φ_{13} in Fig. 2). Appropriate fixation schemes can be proposed for flexible docking, loop modeling, side-chain placement, and other modeling tasks.

Energy and penalty functions dependent upon interatomic distances and their analytical derivatives with respect to the four types of variables of the ICM tree can be efficiently calculated [51,52,40]. Moreover, general equations of internal coordinate molecular dynamics can also be derived and solved numerically [51–53,11]. Internal coordinate molecular dynamics allows the use of large time steps of integration [9,11] and higher simulation temperatures without breaking the covalent connectivity. Significantly, computational effort is reduced for highly constrained ICM models, i.e. additional constraints in the ICM representation lead to fewer calculations, while additional constraints in Cartesian molecular dynamics under the SHAKE algorithm [54,3] result in more calculations.

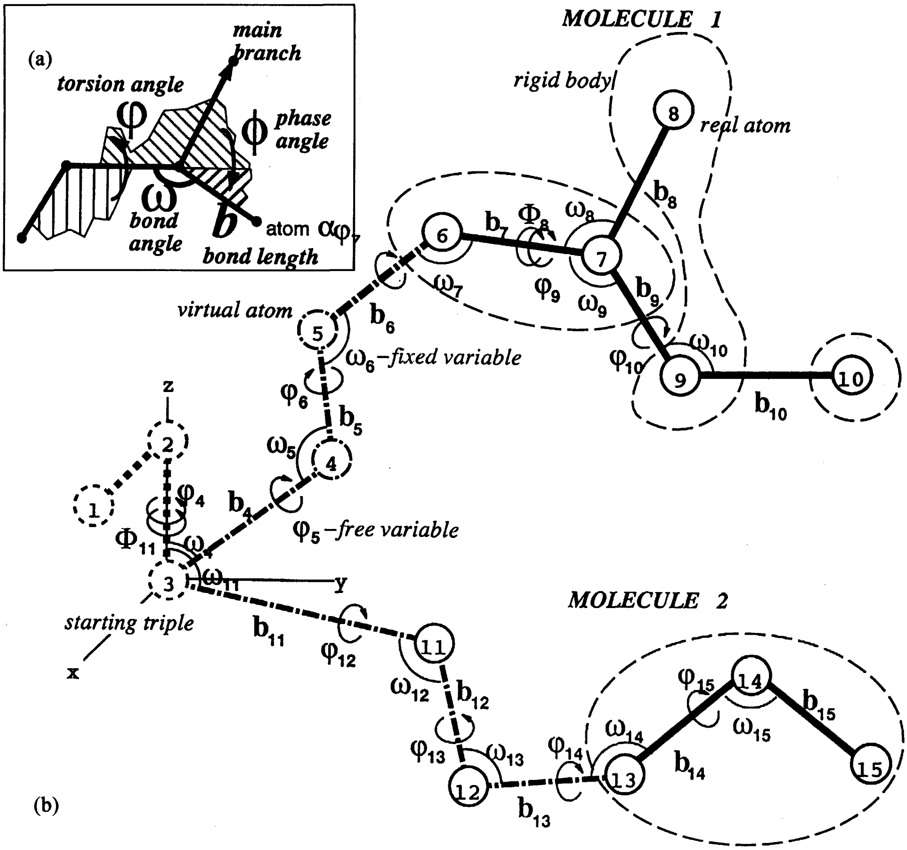

(a)
main branch
torsion angle
φ
phase angle
ϕ
ω
bond angle
b
bond length
atom α_φ
MOLECULE 1
rigid body
real atom
virtual atom
ω_6-fixed variable
φ_5-free variable
starting triple
z
y
x
MOLECULE 2
b_4 b_5 b_6 b_7 b_8 b_9 b_{10} b_{11} b_{12} b_{13} b_{14} b_{15}
φ_4 φ_6 φ_9 φ_{10} φ_{12} φ_{13} φ_{14} φ_{15}
ω_5 ω_7 ω_8 ω_9 ω_{10} ω_{11} ω_{12} ω_{13} ω_{14} ω_{15}
Φ_8 Φ_{11}
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(b)

Fig. 2. The ICM topological tree is a regular connected graph growing from the origin. It contains no cycles and passes through real and virtual atoms and real and virtual bonds. Both molecular position and intramolecular geometry are defined uniformly by a set of free ('unfixed') bond lengths, bond angles, torsion angles, and phase angles. Any subset of these parameters can be fixed, thus leading to rigid bodies.

Accuracy of energy evaluation is crucial

Proteins do not live up to our dreams about them because most protein topologies are ill-behaved, i.e. the free energy of a certain conformation is not a monotonous function of structural resemblance to the native structure. Our dreams and reality are summarized by Table 1.

One can propose two conciliatory remarks, but unfortunately they do not offer a solution to the problem. First, *unrelated mess* may be enriched with the correct elements by, say, a factor of 100 [55], but 1/100 of infinity is still infinity. Second, there are truly well-behaved, 'dream' topologies exhibiting unusual simplicity of sequence

Table 1 *Dreams and reality*

Model	Description	Dream achievement	Reality
M1	Rough simple model (black and white balls)	5 Å model, principal topology	Unrelated mess 1
M2	An improved model: detailed backbone, simple side chains	4 Å model, correct topology	Unrelated mess 2
M3	Further improvement: side chains almost complete	3 Å model, almost perfect	Unrelated mess 3
M4	All-atom force field used to refine a 3 Å model	1 Å model, experimental accuracy	3 Å model becomes a 4 Å model
M5, M6...	Solvation, electrostatics, entropy, polarization, ions	?	?

and folding pattern (like ROP protein [25,56,57]), but they are the exception rather than the rule. For example, if we have two helices with a strong sequence signal (say, every seventh residue is leucine, many alanines, etc.) and the two helices are connected by a short linker, it is definitely a well-behaved topology, and probably even the M1 model (see Table 1) with white and black balls for hydrophobic and hydrophilic residues will be sufficient. However, if the linker is longer or these two helices are disconnected, the situation suddenly becomes much worse since we have to decide if the two helices are parallel, antiparallel, staggered, crossed, etc., and an approximation of a higher level than M5 may become necessary.

A simple explanation of the described behavior which some of us discovered with disappointment and anger is that it is just impossible to get a correct topology as the best energy structure, or one of several low-energy structures, until a certain accuracy is reached. Paraphrasing a notorious Californian criminal law, 'three strikes and you are out', we might say: 'one kilocalorie per residue error and you are out of the business of structure prediction'. Although you may still be in the protein folding business, finding resemblance between your minimal energy folds and the real structure in most cases will require a good deal of imagination and ingenious presentation skills. Why is this? Is it always true and what is the way out?

Let us first admit that there is an astronomically large number of 'protein-like' arrangements available to a polypeptide chain even after the compactness requirement and the buried charge prohibition are imposed. This number is smaller in a discrete conformational space, e.g. on a lattice, and larger in a continuous conformational space even if only local minima are considered. Now, the question is how large the gap is (ΔE) between energies of these 'unrelated protein-like folds' and the energy of the correct fold, and, consequently, what is the required accuracy of the energy calculations.

The necessary model and energy accuracy should depend on the energy gap ΔE separating the correct structure from what we called 'unrelated mess' [38,58,59]. Namely, if our energy evaluation has an error close to ΔE, any of the unrelated conformations may become the lowest energy, while the correct fold will have energy in the 'messy' part of the energy spectrum. What is even worse, the number of false positives will grow *exponentially* with the error. If the error is distributed randomly between residues, the average acceptable error per residue can be estimated as $\Delta E/N^{1/2}$, where N is the number of residues and ΔE is the total error limit. It means that for a 100-residue protein the acceptable error is about *1 kcal/mol*, if ΔE is about 10 kcal/mol. Needless to say, the simplified approximations, both lattice and off-lattice, could hardly achieve this accuracy and will, therefore, fail for most protein topologies.

Low-energy alternative fold (LEAF) hypothesis

We picked $\Delta E = 10$ kcal/mol [58] because that is a typical free energy difference between the native fold and the unfolded state $\Delta E_{F\leftrightarrow U}$ or molten globule. However, the unfolded state (and to some extent the molten globule) is a high-entropy state of many random coil conformations. On this ground, Lattman and Rose [60] argued that sequence-fold *specificity* persists beyond energetic *stability*. This suggests that the energy gap is much larger if high-entropy conformational states are ignored ($\Delta E \gg \Delta E_{F\leftrightarrow U}$), or, simply, that the compact unrelated mess is much further away from the correct fold on the energy scale. If this is true, then the low-accuracy models do have a chance.

Here we argue that this relationship is probably not true for the majority of proteins, with the exception of well-behaved, low-complexity topologies; and there are indeed compact *low-energy alternative folds*, further referred to as LEAFs, that come close to the 10 kcal/mol baseline above the correct fold (Fig. 3).

What are the reasons supporting the LEAF hypothesis other than statistical inevitability due to a huge number of possible conformations? First of all, residues usually can trade entropy for enthalpy (i.e. pack and lose freedom) with a near zero balance. Second, side-chain packing is not a 'jigsaw puzzle' [50] as was believed previously [61], it is much less specific, and protein-like packing can be achieved in a variety of conformations. Third, local conformational preferences of polypeptide chain fragments are rather weak, e.g. it was experimentally shown [62,63] that the same 11-residue fragment may adopt different secondary structures in different structural environments even within the same protein. Fourth, sometimes small sequence modifications can cause large structural rearrangements [64].

The LEAF conformations were not observed experimentally. Theoretical simulations could, in principle, identify such structures. They may appear as false positives in detailed simulations of long peptides and small proteins, provided an accurate free energy function is considered, conformational sampling is sufficient and convergence is achieved.

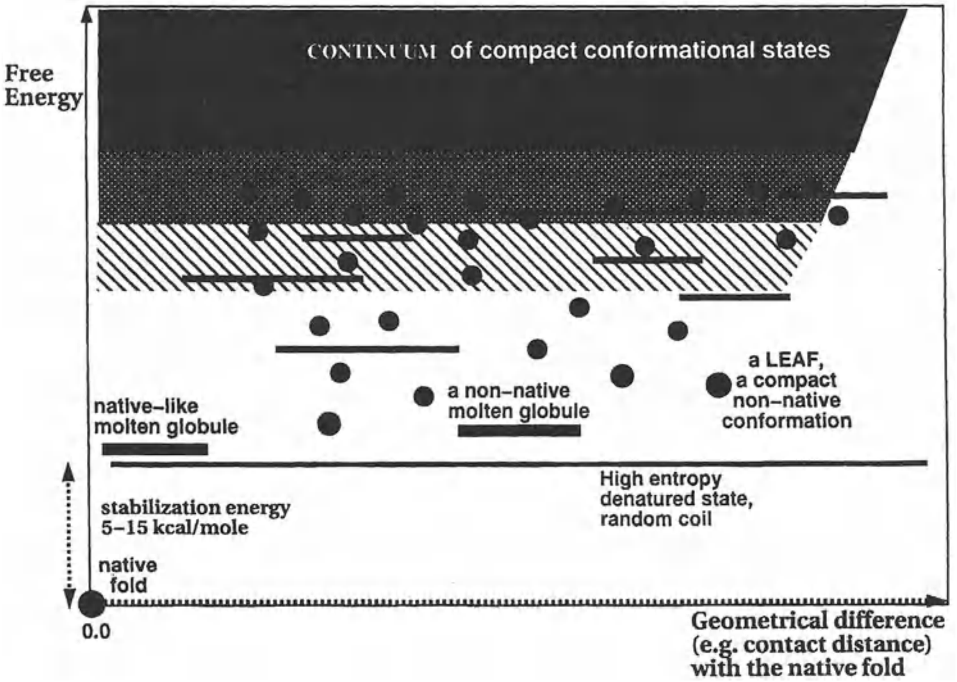

CONTINUUM of compact conformational states
Free Energy
native-like molten globule
a non-native molten globule
a LEAF, a compact non-native conformation
stabilization energy 5-15 kcal/mole
High entropy denatured state, random coil
native fold
0.0
Geometrical difference (e.g. contact distance) with the native fold

Fig. 3. Energy diagram illustrating a hypothetical distribution of different conformational states of a protein (LEAF hypothesis). The width illustrates the entropic component of a state. The widest band is the random coil state, the narrower bands are molten globules. The small squares are alternative compact folds. The near-native molten globule may be lower or higher than the random coil state, depending on the experimental conditions. Our hypothesis is that most of the protein topologies are ill-behaved, and a few amino acid changes can cause the transformation of one protein topology into a different one.

Let us point to two implications of the LEAF hypothesis. First, if the hypothesis is true, a few amino acid mutations that stabilize an alternative fold, and possibly destabilize the original fold, might be sufficient to cause a transformation to a different protein topology, if we could discover what this alternative fold is. Second, the LEAF hypothesis predicts a relative ease of divergent evolution and transformation of protein architectures. Alternative conformations could probably be detected experimentally [65], if their free energy is really close to that of the denatured state, otherwise the detection is problematic. Experimental demonstration of a conformational transformation of a protein into an alternative fold after several mutations is a realistic though challenging task since we need to know the alternative conformation in order to suggest stabilizing sequence modifications. The existence of compact folds only 10–20 kcal/mol away from the native fold would impose very strict accuracy requirements on models for theoretical predictions of protein structure. This energy difference and a corresponding value of 1 kcal/mol/residue are the *upper* estimates of the average accuracy, since the error may fluctuate from fold to fold, while the number of alternative folds grows exponentially with energy.

Energy function for protein structure prediction

In searching for the global free energy minimum of a polypeptide chain, one has to calculate the energy of a large number of trial conformations. This imposes an additional practical limitation on the time of each energy calculation. Therefore, the evaluation of free energy for a trial conformation should be both accurate and fast (Fig. 4). Can the speed and the accuracy requirements both be satisfied? Where does the optimum lie?

Explicit water molecules, flexible bond lengths and bond angles, and high-accuracy calculations of the electrostatic solvation (i.e. Refs. 66–69) are still too computationally expensive to be used in large-scale conformational sampling algorithms. On the

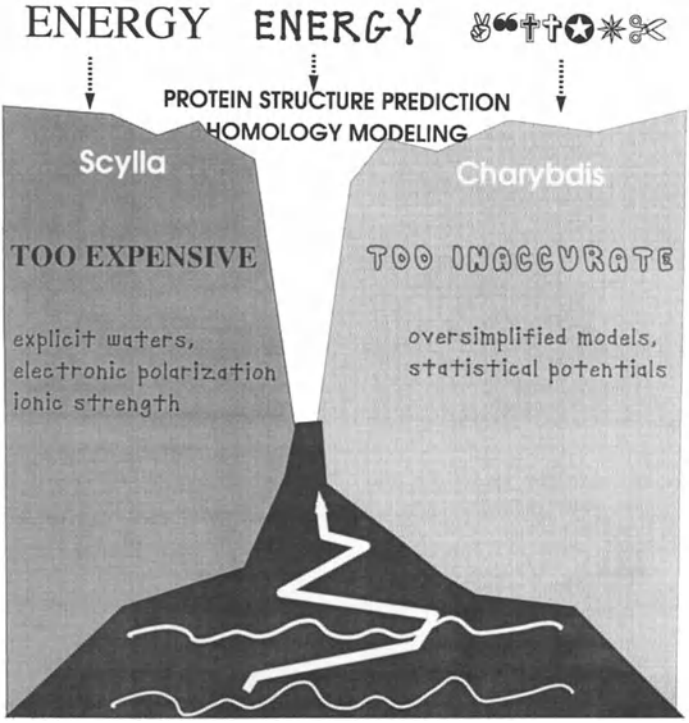


Fig. 4. A cartoon comparing the quest for the optimal approximation for protein structure prediction with the Greek legend about Odysseus trying to find a narrow passage between two monsters in a stormy sea.

other hand, most of the simplified and lattice representations do not reach the required accuracy. Inaccuracies which inevitably accompany simplification of residue representation are further aggravated by incorrect principles of derivation, adjustment and testing of the potentials. It is generally understood that the molecular representation and geometrical parameters (e.g. radii, lattice type and parameters) should be adjusted to reproduce known molecular geometries (reviewed in Ref. 70). However, there is no consensus on how the functional form of simplified potentials and energy parameters (i.e. well depths) should be derived.

The development of new potentials for protein structure prediction involves three steps: (A) choice of energy terms and functional forms for each term; (B) derivation of parameters for the chosen functions; and (C) testing the potentials. We should clearly distinguish between potentials/models designed to *evaluate* a limited set of stereochemically reasonable structures as in threading (reviewed in Refs. 71 and 72), rough potentials/models for a qualitative understanding of the folding process but *not* for structure prediction [73,26,28], and simplified potentials/models designed to *generate* and *predict* protein and peptide structures [74,24,27,29,32,30].

After 20 years of attempts to create simplified potentials of the last type, the question about fruitfulness of these efforts is still open. Yue and Dill concluded the abstract of their recent paper [32] with the following sentence: "Thus, the lowest energy states of very simple energy functions may predict the native structures of globular proteins." Here we argue again [38] that potentials/models for structure generation and prediction should satisfy the strictest requirements which are outlined below.

Requirements for step A: A set of terms and the functional forms of the terms should be justified by the laws of physics. Continuous dependencies should not be replaced by discrete two-level functions; solvation and entropic terms should be included. Extraction of a functional form from the protein database statistics will lead to a wrong functional dependence. Example: the functional form of van der Waals distance dependence derived from statistics of interatomic distances in crystals will be a discrete set of delta functions rather than a continuous Lennard-Jones curve. If the functional form of a potential is wrong, no subsequent adjustment of the parameters (step B) to experimental data will save the potential.

Requirements for step B: Energy parameters for physically justified functions should be *directly* adjusted to experimentally measured free energy differences for a variety of conformational and environmental states, e.g. transfer free energies [75,44], torsion barriers, energies of vaporization, stabilization energy changes due to mutations [76], etc. Similar energy differences must be calculated with less than 1 kcal/mol/residue accuracy in the course of the energy optimization procedure. Energy parameters should not be derived from statistics of residue contacts [70]. The derivation of energy parameters from a *set of decoys* (i.e. Ref. 55) may lead to the 'Plato's man' effect [38].

Requirements for step C: Furthermore, a fixed set of decoys is not the best standard *test* for different energy functions [77] for protein folding simulations, because both the near-native structure in the set and so-called good decoys *depend on the model and* *the energy function.* All the structures in the set should be re-optimized for each model (e.g. a specific lattice, united atoms, all atoms) and set of potentials being tested. Secondly, a set of decoys will never be large enough to compete with a set of conformations generated by global optimization, especially in continuous configurational space.

The *test by global optimization* is the following: take a known experimental structure and find a 'near-native' conformation by local optimization of the energy function being tested (a broader scale optimization or dynamics simulation with native restraints may be required instead of the local minimization). Then, run a free global optimization with the same energy function starting either from the near-native conformation or any random conformation. Stop if the energy drops below the 'near-native energy'. This will mean that a false positive has been found, and the energy function has failed the test. Since false positives are different for different energy functions and molecular representations, it might be more efficient to generate them by a trial optimization.

Historically, the first attempts to predict protein structure at the atomic level were performed by minimization of vacuum atom–atom potential energy [5,78]. A number of algorithms aimed at inclusion of solvation into simulations are based on the solvent accessible surface or volume for individual atoms [79–82,44]. This approach is purely empirical because, obviously, solvation energy of a charge is not a function of its accessible surface. Wesson and Eisenberg [79] derived solvation energy densities for five classes of atoms on the basis of 18 experimental vapor–water transfer energies for side-chain analogues [83]. The accuracy of the surface-based atomic solvation model can further be improved by separating carbons into aliphatic and aromatic classes. In addition, the Wolfenden et al. [83] experimental set contains only data for neutralized side chains of these residues, and volume corrections are not really required for this set of compounds [84]. The atomic surface densities for the charged groups of Lys, Arg, Asp, and Glu are larger and can be derived from additional experimental data for charged solutes [85]. The necessary improvements were implemented in our new set of solvation parameters (see Table 2), which were used in calculations described below.

Although this type of solvation energy function is not justified by any physical model, the parameters of this function can be adjusted to reproduce reasonably well the experimentally observed differences between two extreme states: maximally exposed and deeply buried. However, in the intermediate burial states of atoms as well as in the presence of other charges nearby (e.g. Ref. 86), the error will be comparable with the solvation energy value.

In an alternative, more physical approach, the electrostatic contribution to solvation is separated from the surface-dependent solvation contribution, and the electrostatic component is evaluated by an approximate solution of the Poisson or Poisson–Boltzmann equation for a set of fixed charges immersed in an arbitrarily shaped protein [87–90,42,68]. The methods vary in accuracy and speed from the most accurate boundary element method [91,88,92,67,69] to one which is the least accurate but includes the fastest image charge approximation [93–96,42].

Table 2 *Solvation parameters*

σ (cal/(mol Å^2))	Radius (Å)	Atom type
10	1.95	C aliphatic
– 9	1.8	C aromatic
– 163	1.7	N uncharged
– 280	1.7	N + N_ζ Lys +
– 220	1.7	$N_{\eta 1}$, $N_{\eta 2}$ in Arg +
– 114	1.6	O hydroxyl
– 64	.1.4	O carbonyl
– 280	1.4	O^- Glu, Asp
– 174	1.4	O in COOH
– 22	2.0	S in SH
– 92	1.85	S in Met or S-S

The surface and entropic terms should also be added to the full vacuum force field and electrostatic solvation in order to achieve sufficient accuracy in folding simulations [38,42]. Both terms can contribute up to 1–2 kcal/mol/residue to the energy difference between different conformations. The two terms can compensate for each other [38], e.g. hydrophobic free energy gained upon burial of an aliphatic side chain is lost in entropy. The surface term can be made proportional to the overall solvent accessible surface with a coefficient from 5 to 15 cal/Å^2 (vapor–solution transfer experiments) [81,97–100], or can be divided into contributions from different atom types to account for differences in hydrophobic effect for different groups due to higher order electrostatic effects [101] and specific geometry. For example, the surface tension for aliphatic groups is higher than for more hydrophilic aromatic groups, and this difference is impossible to assign to partial atomic charges. In the following sections, the surface term separated from electrostatic solvation is calculated as a product of the solvent accessible surface by 12 cal/Å^2. Two other terms, electrostatics and side-chain entropy, are described below in more detail.

Affordable solvation electrostatics

The distance-dependent dielectric constant in Coulomb's law has little relevance to the free energy of water molecules polarized in the electric field of a solute. If a solute has a single charged atom, the calculated energy will not be dependent upon the charge position and burial, a result which is obviously wrong. On the other hand, numerical solutions of the Poisson equation [102,66,103,68,90], more sophisticated models [104] with site-specific protein dielectric constant [105–107], or explicit solvation models [108] are too computationally demanding to be incorporated into a simulation procedure, although some of the approaches can be used for re-evaluation of the best solutions [109].

An approximate solution of the Poisson equation can be obtained on the basis of the image charge approximation [93,95]. The MIMEL method (Modified IMage ELectrostatics) [42] is a sufficiently accurate and fast implementation of the image charge approach. It uses an analytical correction to the image charge formulae and a robust algorithm for calculating effective distances between protein charges and their effective dielectric boundaries in the case of an arbitrarily shaped protein.

Electrostatic energy in the MIMEL approximation is represented by the following formula (see Fig. 5):

$$E_{MIMEL} = \sum_{q_i,q_j,i<j} \frac{Cq_iq_j}{\varepsilon_p r_{ij}} + \frac{1}{2}\sum_{q_i,q_k^{im}} \frac{Cq_iq_k^{im}}{\varepsilon_p r_{ik}} - \frac{1}{2}\frac{C(q^{total})^2(\varepsilon_w - \varepsilon_p)}{R\,\varepsilon_w(\varepsilon_w + \varepsilon_p)} \tag{1}$$

where r_{ij} is the distance between charges i and j, the image charge

$$q_i^{im} = -\frac{(\varepsilon_w - \varepsilon_p)}{(\varepsilon_w + \varepsilon_p)}\frac{R}{x_i}q_i$$

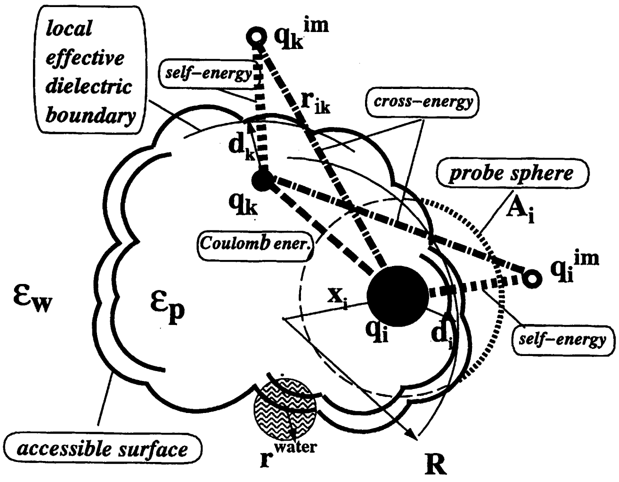


Fig. 5. Protein in water. Accessible surface A defined by the local curvatures (radius R) can be used to define depth d of the charge. Atom i has van der Waals radius r_i^{vw}. To calculate the protein solvent-accessible surface, increased radii $r_i = r_i^{vw} + r^{water}$ are used. The accessible surface A_i of the probe sphere is used to assess the depth d_i. Distances d_i are later corrected to move the effective dielectric boundary from the protein solvent-accessible surface closer to the molecular surface. Derivation of the final set of distances $\tilde{d}_i$ between the charge and the effective dielectric boundary in order to satisfy two conditions: for large positive d_i the distance should be decreased by δr, whereas for negative distances the asymptotic value of $\tilde{d}_i$ should be such that interaction with the image charge reproduces the Born energy of the charge of $r_i - \delta r$ radius. Simple linear functions satisfying the above conditions for the initial dielectric boundary (bold dotted line) and the corrected one (bold solid line) are shown. $\delta r = r^{water}$ brings the effective dielectric boundary close to the molecular surface.

The first sum in Eq. 1 represents the Coulomb energy, the second sum contains contributions from interactions of charges with their own images (self-energy) as well as interactions of charges with other images (cross-energy), and the third term is the correction term depending on the net charge of all real charges in the system q^{total}.

To calculate the electrostatic free energy for a real protein, the interaction energy between two charges i and j and two corresponding image charges should be expressed in terms of two depths d_i and d_j of the charges from the protein surface, their interatomic distance r_{ij}, and effective protein radius R. Distances between charges and their effective dielectric boundary are found via the exposed fraction of the surface of a large probe sphere (about 5 Å) around each charge. The probe sphere algorithm gives a locally averaged estimate of the effective distance. Comparison of the MIMEL energies evaluated for a set of model objects and a series of proteins with energies calculated by the DelPhi program [110,87] solving the Poisson equation numerically, demonstrated a high accuracy of the MIMEL method. The MIMEL approximation was successfully used in a variety of global optimization tasks [42,111–113].

Entropy

The entropic contribution to free energy differences consists of entropy changes of water and configurational entropy of the polypeptide chain. The solvent entropy is an integral part of the solvation energy and does not require special additional consideration. However, this is not the case with the configurational entropy of a polypeptide chain, which is usually ignored in evaluations of the free energy of a trial conformation in the course of conformational sampling.

Configurational entropy includes vibrational entropy in the vicinity of a local minimum [114,115] and the entropy due to the presence of several alternative minima with close energies [116,117]. The main-chain entropy is an important factor in the overall free energy balance between the folded and the random coil states. However, the main-chain entropies of compact conformational states do not differ as much as side-chain entropies do. The reason is that some side chains in compact conformations will still be exposed and flexible, and this set of exposed side chains is fold-specific, while the backbone in both folds is restrained to one local minimum. Therefore, it is the relative side-chain entropy which has to be included first in the globally optimized free energy function.

To calculate the side-chain entropy contribution to the free energy function, we need to know the entropy of an exposed side chain in a random coil, the entropy of a completely buried side chain, and we need an algorithm to quickly evaluate the entropy in intermediate cases. The entropy of the random coil state in which a side chain is exposed and may adopt different rotamer conformations can be calculated under the assumption of discrete rotamer states with the Boltzmann formula: $S_U = -R\sum_i p_i \ln p_i$, where p_i is a probability of rotamer i, and R is the gas constant [118,38,119]. If N individual rotamers have equal probability, the expressions can be rewritten as $S = -R\sum_i \ln N$ [116,120,117]. A hidden assumption of the discrete approximation is that the buried state, to which we assign zero entropy, can be considered as one of the rotamers, i.e. widths of the energy minima are the same in folded and unfolded states [121]. Probabilities p_i for each side-chain rotamer can be calculated in explicit simulation [118] or derived from the analysis of rotamer distributions in known protein structures, and a scale of entropies of exposed side chains can be compiled [119,38,42,122,123]. In the derivation of these numbers it is important to account for side-chain symmetry [38].

The question is: how may the side-chain entropy, a property of an ensemble of rotamers, be assigned to a single trial conformation containing buried, exposed and partially exposed side chains? One could explicitly calculate energies of all the possible rotamers given the backbone conformation, but running a 'microsimulation' nested in our large-scale global sampling/optimization procedure to evaluate new p_i for each side chain is too costly.

The number of states available for a given side chain can be related to its solvent accessibility [42]. The side-chain entropy can be approximately related to solvent accessible area [42] in the following way: if the side chain is buried, $S = 0$; if the side chain is maximally exposed with area A_U, $S = S_U$; and if the side chain is partially buried, the entropy can be approximated by a linear dependence, $S = S_U(A/A_U)$. This empirical approach is a definite improvement over a simple threshold rule ($S = 0$ if $A < A_{cutoff}$; $S = S_U$ if $A > A_{cutoff}$), although there is no true physical dependence of configurational entropy on solvent accessible surface. The approach is also convenient and efficient computationally since solvent accessible areas are required anyway to evaluate the surface (hydrophobic) energy. The side-chain entropy term of the free energy function was used in a number of structure prediction calculations [42,111,124,113,125] and, since entropic contributions to free energy may fluctuate up to 2 kcal/mol/residue, omission of the term may lead to different backbone geometries of peptides (see a test on a nine-residue peptide below) and protein loops.

Global optimization

If the first challenge of protein structure prediction is to develop an accurate and fast approximation of the free energy function, the second challenge is to find the global minimum of this function. Molecular dynamics simulations in both Cartesian [4] and torsion spaces [9] are capable of surpassing kT-large barriers between local minima and, therefore, can be used as global optimization procedures, although not the most efficient ones. Limitations in efficiency stem from the continuity of an MD trajectory.

A number of global optimization procedures including Monte Carlo minimization [48] and the diffusion equation method [37] have been developed by Scheraga and co-workers (reviewed in Ref. 41). The diffusion equation method (DEM) and a related, more general packet annealing method [126] are promising, but these methods have to deal with two critical issues related to the accuracy of the energy function. First, all the energy terms should be approximated by Gaussians, while the energy function has borderline accuracy even without this requirement. Second, in DEM the number of alternative pathways that one needs to store may be exceedingly large, because at each time point the direction of local minimization is influenced by a certain energy error and, potentially, excessive smoothing of the energy landscape. The first problem has been recently addressed [44] by the introduction of a volume-dependent solvation energy term in the Gaussian form.

Running several simulation trajectories in parallel and exchanging information between them constitutes another idea for global optimization [127–129]. Genetic algorithms have been applied to optimize the energy of polypeptides in detailed atomic representations [39,43] (see also Ref. 129, a recent review). Similarity between global minima of homologous sequences can also be used as a restraint in a simulation [128].

The heart of the problem – frequently overlooked – is how the elementary conformational change is generated during sampling. It is much more important than whether it is a simulated annealing or a constant temperature run, or whether several simulations are independent or exchange information between trajectories.

In most of the global optimization procedures, two types of conformational changes are generated: (i) local, quasi-continuous, within one minimum, and (ii) global, large-scale, between minima. Local changes may be generated in the course of a molecular dynamics simulation [130], a local energy minimization [131,48], or a local Monte Carlo procedure [132] based on the harmonic approximation of the energy landscape around the current minimum. Local moves are essential for finding the nearest local minimum or searching around the starting conformation. However, a powerful global sampling procedure requires an efficient algorithm to generate large conformational changes.

Random global moves reported in the literature include the following: evenly distributed random change of one randomly chosen torsion angle and subsequent minimization [48]; reorientation of peptide planes according to the current local electrostatic field in the EDMC method [133]; discrete moves on a lattice [134]; random torsion by random value ('mutations') and random 'crossovers' in genetic algorithms [39]; and loop deformation moves [95,52,134–137].

A Monte Carlo step can be made much more effective if local conformational preferences are used [38,138,42]. The procedure requires the following steps: (i) identification of coupled groups of several torsion angles (say, φ, ψ and χ^1) of a residue, or all torsion angles of a side chain; (ii) derivation of *a priori* continuous probability distributions in identified subspaces; and (iii) generation of global rearrangement by picking one or several subspaces and changing the torsions involved according to the derived probability distributions. The changes are performed independently on current values of coordinates in the subspace, i.e. the procedure makes absolute changes in local subspaces instead of the incremental changes. Detailed mathematical justification of such a scheme in continuous space and the first use of these 'probability-biased' random steps in an MC global optimization is given in Ref. 42. This biased probability Monte Carlo (BPMC) principle combined with energy minimization of the whole structure after each step clearly outperforms the Monte Carlo minimization procedure with the unbiased random step [48].

The best random step distribution function was derived to be exactly equal to the expected probability distribution [42], i.e., in a discrete case, if you expect that one rotamer is 9 times more frequent than the other, it should be sampled 9 times more frequently. The best function can be derived from an analytical expression of optimization efficiency, further referred to as optimization efficiency functional. This functional depends on local expected probability distributions and a set of unknown random step distribution functions. Unfortunately, the optimization efficiency functional can only be written analytically for a very simple model of the optimization procedure. A different formulation of the optimization efficiency functional leads to a square root of the expected probability distribution for the best random step distribution. This means that if you expect one rotamer to be 9 times more frequent than the other one, the first rotamer should be sampled only 3 times more frequently (Zhou and Abagyan, unpublished). The derivation of an analytical functional representing the efficiency of a particular global optimization procedure more accurately is a challenge. However, the proposed scheme is already a step forward, and all further recommendations for efficient random moves can always be tested empirically.

Expected probability distributions for the backbone and side-chain torsion angles can be derived from protein database statistics [138,139,42], or calculated for short fragments with a specific amino acid sequence in a preliminary simulation. These random moves are suitable for side-chain rearrangements [139,111,125] and free backbone movements [42,125].

Two special types of random moves are required to dock two molecules and to predict loop conformations. In docking simulations, positions of one of the molecules can be changed with the pseudo-Brownian random move, a combination of random rotation and translation [40]. Simulations of protein loops require a combination of the BPMC moves with a loop closure algorithm to keep the loop ends unchanged [140,112,125]. These two types of random moves are also followed by the local energy minimizations.

The general scheme of the ICM global optimization for the arbitrarily constrained ICM model, which may include several molecules and flexible loops, is shown in Fig. 6. It includes the following procedures. First, a group of coupled variables is randomly selected. Second, a random move appropriate for these variables is performed: (i) if precalculated local probability distributions are assigned, the biased probability move is performed; (ii) if these variables determine the position of the whole molecule, the pseudo-Brownian move is performed; (iii) if the variables belong to the backbone of a flexible loop, a loop closure algorithm is additionally applied; (iv) otherwise a conventional random move is performed. Third, a local energy minimization with respect to all free variables of the model is performed. To take computationally expensive or nondifferentiable energy terms (such as solvation electrostatics and side-chain entropy) into consideration, these terms are added to the total energy at the end of minimization (the 'double-energy' scheme). Fourth, the Metropolis criterion [141] is applied according to the total energy resulting from the previous step. The accepted conformation is compared to the current conformational stack.

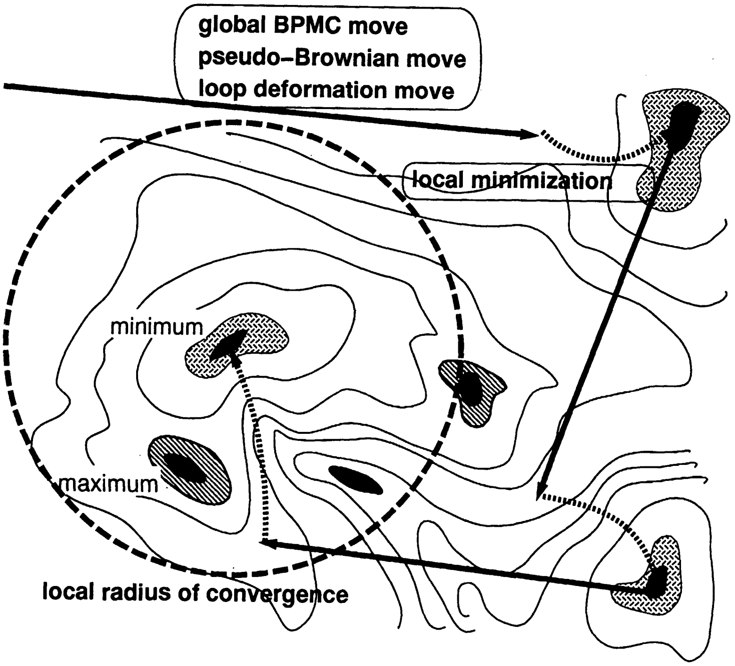


Fig. 6. Schematic representation of the ICM global energy optimization procedure. A hypothetical multidimensional energy landscape is shown. Three different types of random moves are possible. Biased probability moves *are being performed according to torsion energy distribution derived from a set of known three-dimensional protein structures.* Pseudo-Brownian moves *are used to position randomly molecules with respect to each other according to a Gaussian distribution of translational and rotational degrees of freedom. Local deformations and loop closure algorithms are used to perform* loop deformation moves *to satisfy boundary conditions (unchanged positions of a loop's ends). Solvation and side-chain entropy energy terms are evaluated during the Monte Carlo step of the protocol, and omitted during the local minimization step (double-energy scheme, see text).*

Conformational stack is a representative collection of low-energy conformations visited during the global optimization search. Each time a new conformation is accepted, it is quantitatively compared to all those already residing in the stack. Conformations are considered similar if their structural difference does not exceed a certain similarity threshold specified by either coordinate or angular rms difference. If a new conformation is not similar to all those already accumulated in the stack (or the stack is empty at the beginning of the search), a new slot is created where the new conformation is placed. If the new conformation is within the value of the similarity threshold of any conformation already in the stack, it substitutes for the last one if its energy is lower, and is disregarded otherwise. The size of the conformational stack defines the maximal number of conformations simultaneously residing in the stack, and should be specified before the calculations. If the stack is full, but the search is continuing and the accepted conformation differs from all those already in the stack, the new conformation replaces that with the highest energy.

The three parts of the ICM molecular modeling engine, viz. (i) molecular representation and an arbitrary set of constraints, (ii) energy function, and (iii) a set of efficient conformational rearrangements, allow one to address a wide variety of large-scale structure prediction problems, such as peptide folding [42], loop prediction and protein design [140,124], modeling by homology [112], soft protein–protein docking [111,113], and domain rearrangements [125].

Ab initio peptide structure prediction

Development of the ICM optimization procedure using global probability-biased random steps allowed identification of the global minimum of a detailed energy function, including MIMEL solvation electrostatics, surface and entropy terms, from any random starting conformation for large peptides. In the first series of test calculations, 12- and 16-residue peptides converged to the same set of lowest energy conformations (helical in both cases) in a free BPMC simulation from several random starting conformations [42]. Convergence for the 16-residue peptide was achieved after several hours of BPMC simulation at 600 K on an SGI workstation. It became clear that the optimal structure is sensitive to the amino acid sequence and to the set of energy terms added to the vacuum force field. Omission of the additional energy terms led to deformed helices and nonhelical conformations, and a sequence modification disrupting the helix was suggested based on BPMC simulations.

Finding a protein-like peptide adopting a nonhelical conformation in solution to test a global optimization procedure has always been a problem. Fragments of complete proteins can be used with caveats [43]. We were fortunate to find two examples of short nonhelical peptides with an experimentally characterized structure in solution, a nine-residue peptide and a 23-residue peptide [142,143].

A linear nine-residue peptide (YQNPDGSQA) was shown by NMR to form a β-hairpin in aqueous solution [142]. We performed a series of BPMC simulations from different random starting conformations with the same energy terms (including MIMEL solvation electrostatics, surface and entropic terms) and a BPMC setup as for the helix-forming 12-residue and 16-residue peptides simulated previously [42]. The BPMC simulations at 600 K converged in about 1 h. The lowest energy structure was the experimentally found β-hairpin (Fig. 7a). Conformations with higher energy which were accumulated in a conformational stack [144] retained some β-hairpin features in at least a 4 kcal/mol energy range. We also performed the same simulation without the entropy term to analyze its importance. The lowest energy structure generated without the side-chain entropy was different (Fig. 7b) and less similar to the experimental structure. This illustrates the importance of the entropic term even for a small peptide.

A substantially longer 23-residue peptide was designed to adopt a β-β-α architecture in aqueous solution without bound metal ions or disulfide bridges [143]. The

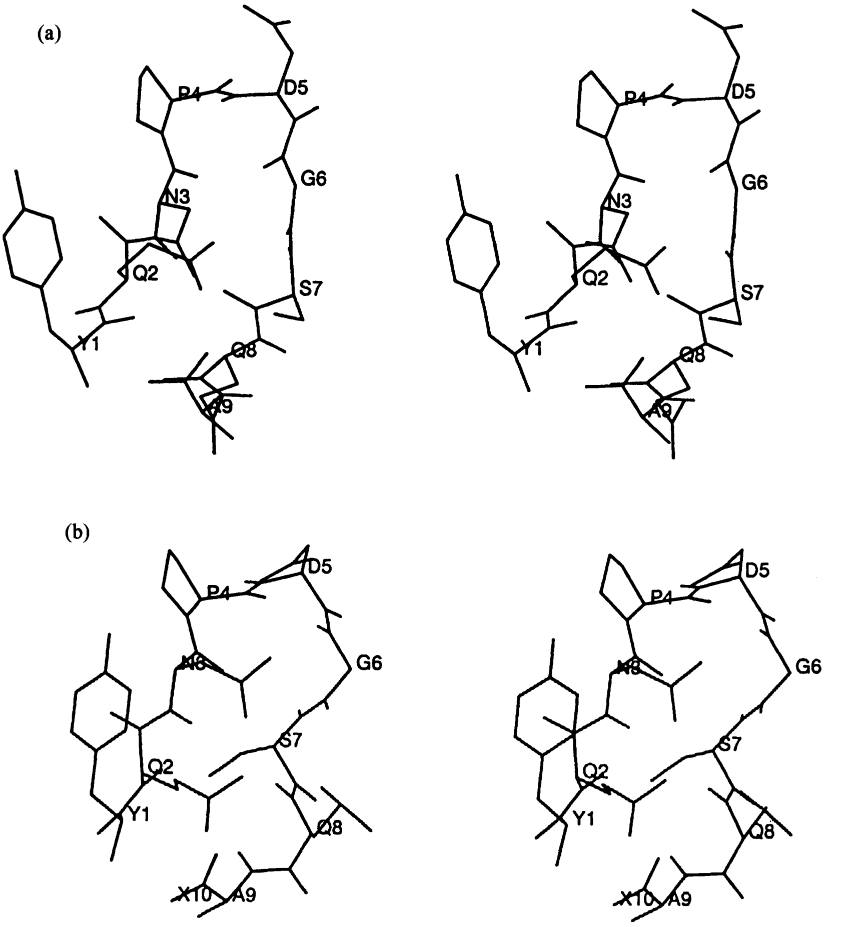


Fig. 7. The lowest energy conformation of the nine-residue peptide [142]. (a) The energy function includes the side-chain entropy term; the best conformation is very similar to the experimental structure [142]. (b) Exclusion of the side-chain entropy term from the energy function leads to a different minimum.

designed topology was confirmed by NMR. Several BPMC simulations at 600 K converged after about 10 million energy evaluations. To reduce the number of variables, all the peptide ω angles were fixed at 180°. Simulations were performed with solvation energy represented by the MIMEL energy and the surface term, as well as with the atomic surface based solvation term described above. The stack of low-energy conformations within 6 kcal/mol from the lowest energy conformation was generated as described in Ref. 144 with the atomic solvation term and with a 15° torsion rms deviation similarity threshold, and is shown in Fig. 8a. All of the stack

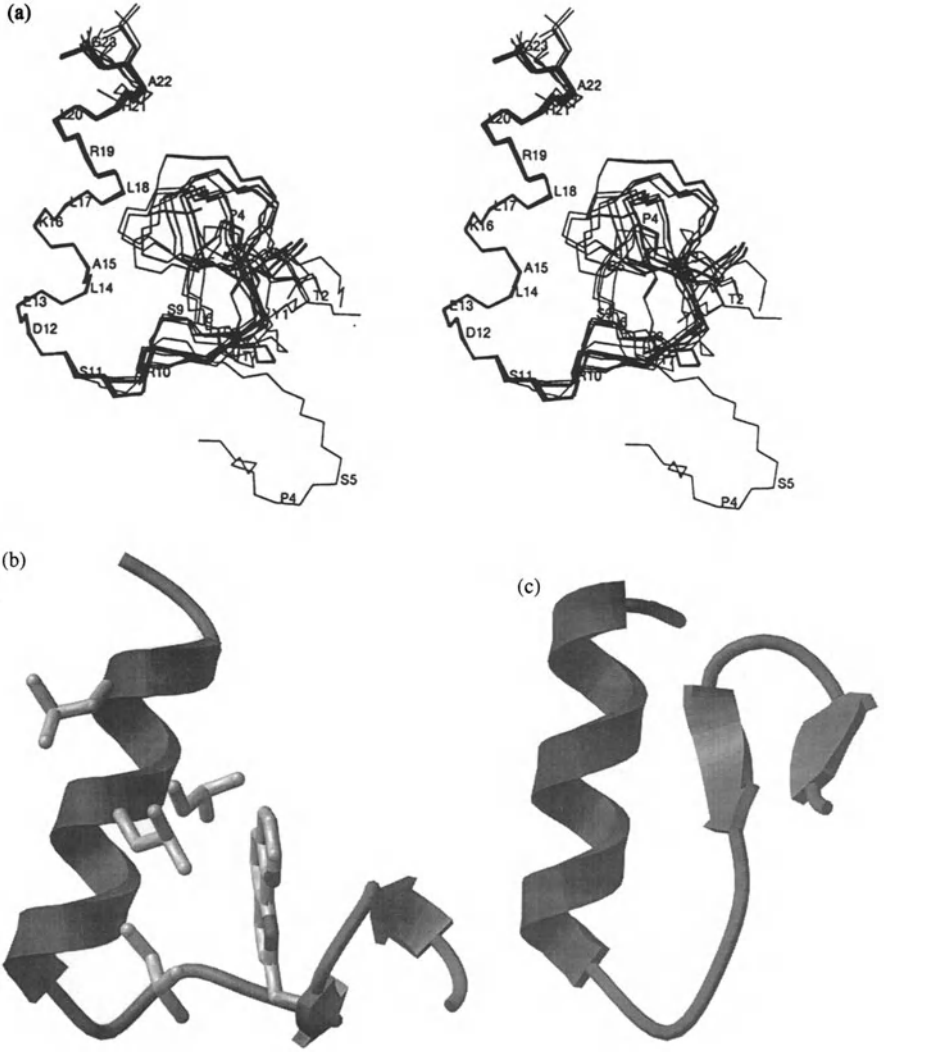


Fig. 8. Conformations of the 23-residue peptide predicted by the ICM global energy optimization. The energy function includes side-chain entropy and solvation terms. (a) All conformations from the stack within 6 kcal/mol from the lowest energy; the solvation term is calculated via accessible surfaces and atomic solvation energy densities. (b) The third energy structure from a set shown in (a). (c) The lowest energy conformation; the solvation term is calculated as the MIMEL energy plus constant surface energy term.

conformations have an α-helix formed at the C-terminus, and almost all of them have the helix interrupted in the same position. All stack conformations also had a β-hairpin formed at the N-terminus, but the orientation of the hairpin varied. The third lowest energy conformation having the characteristic β-β-α architecture is shown in

Fig. 8b. A similar architecture, but a different packing of the β-hairpin against the α-helix, was found in the lowest energy conformation in the simulation MIMEL and constant surface tension (Fig. 8c).

We conclude that even for a molecule with over 100 free torsion angles and 400 atoms, the essential topological features can be predicted in a true *ab initio* simulation without any experimental restraints and convergence can be achieved. Local structural features such as an α-helix starting from a certain residue and a β-hairpin at the N-terminus can be predicted with higher reliability than the exact packing of these two elements. This computational experiment implies that, even though the general topology can be successfully predicted and corresponds to the lowest energy conformation, the accuracy of the model and the energy function is dangerously close to its limit, and the accuracy should be further improved in order to predict the three-dimensional structure of larger peptides and proteins.

Domain rearrangements

ICM was applied to the prediction of large-scale movements of protein domains from a single three-dimensional structure. Multidomain proteins often exist in different conformational forms, depending on crystal packing, bound ligands, etc. The prediction of protein domain motions at atomic resolution is a serious theoretical and computational problem due to the large time scale of these motions and the large size of multidomain proteins. Both obstacles make the problem barely tractable by traditional computational methods. The ICM approach relies on the observation that only a relatively small portion of a protein including interdomain linker and side chains at the domain interface undergoes conformational changes upon domain rearrangements. An interdomain linker is automatically identified, and torsion angles belonging to the linker and side chains at the potential domain interface are set free. These torsions are sampled in a large-scale global search for low-energy conformations. A special procedure has been developed to generate deformations of the double-stranded interdomain linker preserving the chain continuity.

Domain motion modeling was performed for two-domain structures of Bence-Jones protein (with a single-stranded interdomain linker) and lysine/arginine/ornithine-binding (LAO) protein (with a double-stranded linker). For each protein, two sets of low-energy conformations were generated starting from the crystallographically determined 'closed' and 'open' forms. Two sets of Bence-Jones protein conformations substantially overlapped and revealed a high diversity of possible relative domain positions. In the case of LAO protein, the concerted changes in the double-stranded linker were observed, and the two sets of the generated conformations overlapped (Fig. 9). Interestingly, for the LAO protein one of the low-energy conformations generated from the closed form was only 2.2 Å apart from the open structure. The obtained results indicate that the method can generate a series of low-energy conformations representing possible domain arrangements within a reasonable computer time. This may be helpful for predicting the scope of possible

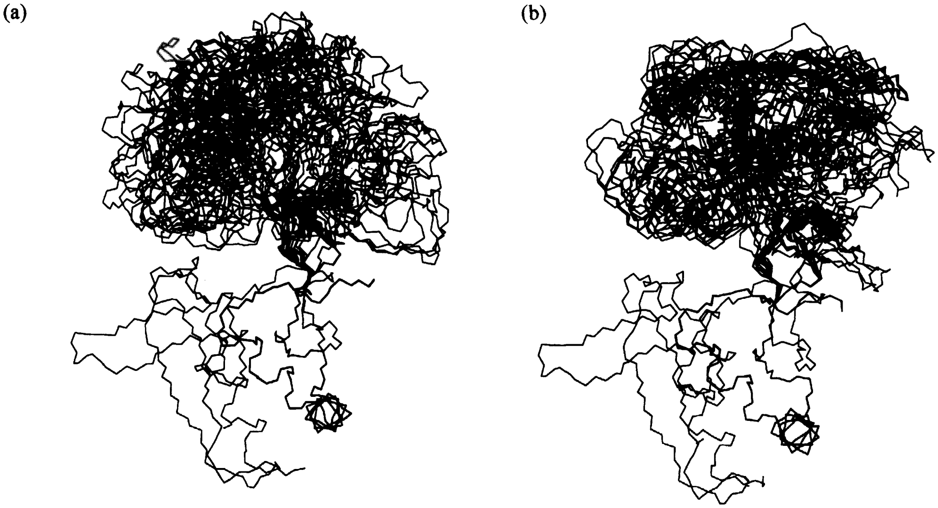


Fig. 9. Backbone display of different domain arrangements in the low-energy conformations of the LAO protein generated starting from the closed form (PDB code 1lst) (a) and from the open form (2lao) (b). The conformations of the larger domain are superimposed.

domain rearrangements of a multidomain protein on the basis of only one known conformation.

Docking

The prediction of association of two large molecules requires a scoring function to evaluate a trial configuration of the two molecules and the algorithm searching for the global minimum of this function, and a number of procedures were developed over the years [145–159,40,111] (see also Ref. 160, an excellent recent review). The ideal docking algorithm has to deal with the flexibility of both receptor and ligand. However, in many protein–protein docking problems simple energy/scoring functions and rigid models are sufficient and successful, as opposed to problems of peptide folding. The reason of larger tolerance of the docking procedures to simplifications of shape and energy is a smaller dimensionality of the problem (the ligand position with respect to a receptor is described by only six variables), which significantly reduces the total number of configurations and, accordingly, the number of potential false positives.

ICM can be easily configured for either rigid or soft docking because of its ability to consider arbitrarily constrained multimolecular trees [40] (Fig. 2) and search the global minimum of a detailed energy function including solvation and side-chain entropy with respect to the set of free variables. Soft docking of a protein or a flexible ligand to a receptor, or global refinement of initial docking solutions generated by

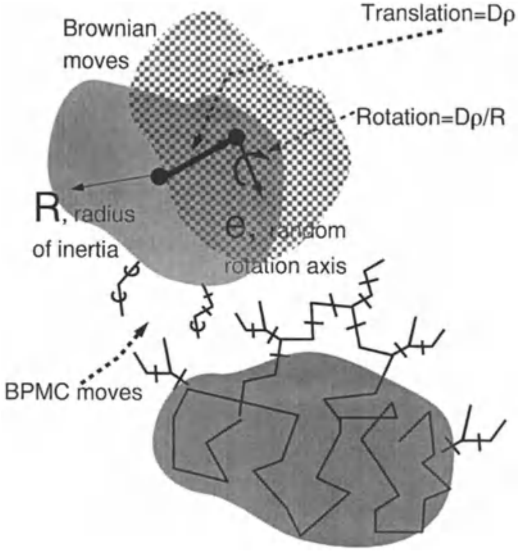


Fig. 10. Schematic diagram of random moves in the ICM docking procedure. In a pseudo-Brownian move, D is the move amplitude (2 Å), ρ is a random number between 0 and 1, and R is the radius of inertia. Side-chain torsion angles are modified according to the rotamer probability distributions.

another method [111], usually requires a detailed atomic model, while the initial rigid-body docking may benefit from a more simplified molecular model. A set of variables in the ICM tree for flexible docking includes six positional variables to allow free movements of the ligand molecule and a set of side-chain torsion angles at the protein surface. The pseudo-Brownian random moves and probability-biased moves are applied to the positional variables and side-chain torsion angles, respectively (Fig. 10). Each random step is combined with local minimization. The globally optimized function includes ECEPP/3 energy plus electrostatic solvation, surface energy and side-chain entropies.

First, the ICM docking technique with flexible side chains was tested on the *ab initio* prediction of association of two GCN4 helices [40]. It was shown that sampling is sufficient and all possible arrangements are found, and the correct parallel arrangement of two helices has the lowest energy, which is 5 kcal/mol lower than crossed, staggered and antiparallel arrangements. The second test was flexible docking of the uncomplexed lysozyme [161] and the HyHel5 antibody [162] in a detailed simulation [111]. ICM global optimization was performed from 120 starting conformations. The best 30 conformations were refined by the global BPMC optimization of all the surface side chains. It turned out that such a global refinement of the interface both improved the geometrical accuracy of the prediction (the rms difference with the X-ray structure was reduced from 5.46 to 1.57 Å) and, more importantly, increased the energy gap between the correct solution and the first false solution from 4.3 to 19 kcal/mol, thus making the prediction more significant and reliable.

The last example is the successful blind prediction of association of β-lactamase and its protein inhibitor [160,113]. Similar to the lysozyme-antibody docking, it was effected by the pseudo-Brownian Monte Carlo procedure [40] from multiple starting conformations of the inhibitor. Simplified description of two molecules was used for the initial docking. A stack of about 30 best conformations was then refined in full atomic detail by the biased probability Monte Carlo simulation [111] which globally optimizes molecular positions and conformations of the interfacial residues. Three lowest energy structures with association energies of − 29.3, −11.2, and − 6.8 kcal/mol were submitted (Fig. 11). All the other solutions had energies higher than + 5 kcal/mol. As in the lysozyme-antibody case, the refinement increased the gap between the near-native and the false solutions. The lowest energy structure turned out to be 1.9 Å backbone rms from the correct solution [160].

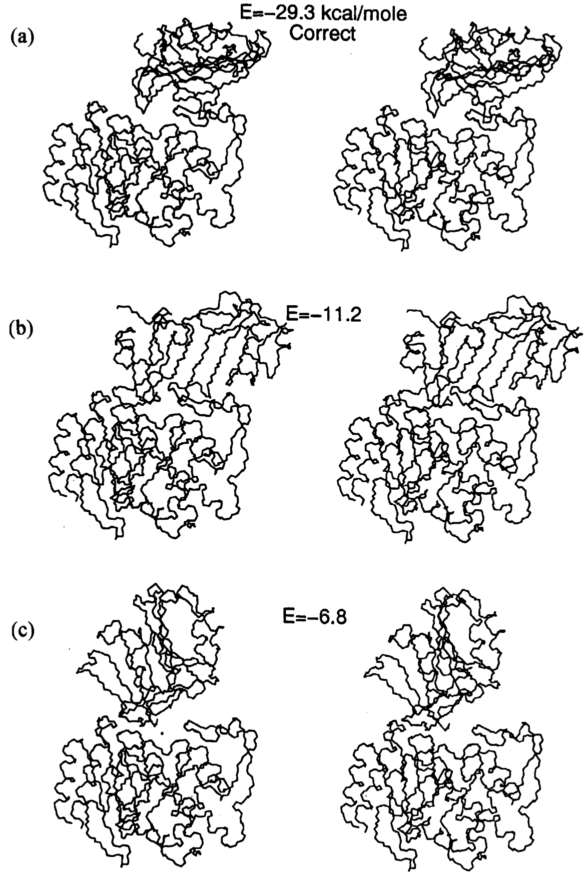


Fig. 11. Stereo diagrams of the three lowest energy conformations of β-lactamase and its inhibitor after the global positional and side-chain refinement [113].

Loop modeling and design

To predict conformations of protein loops by the ICM method, we need the following global optimization setup: (i) only torsion angles of the loop residues and side chains in the loop vicinity are free, the rest are fixed; (ii) a random move deforming the loop without moving its ends is applied along with the BPMC moves for the surrounding side chains; and (iii) all the energy terms and the optimization procedure are the same as in a peptide folding simulation.

The ICM method was applied to redesign a 15-residue loop-3 of trypanosomal triosephosphate isomerase (TIM) dimer [140,124,164]. The purpose of the design was conversion of the dimer into a monomer (Fig. 12a). Several polypeptide chain fragments of different amino acid sequence and length were tried and the ICM global

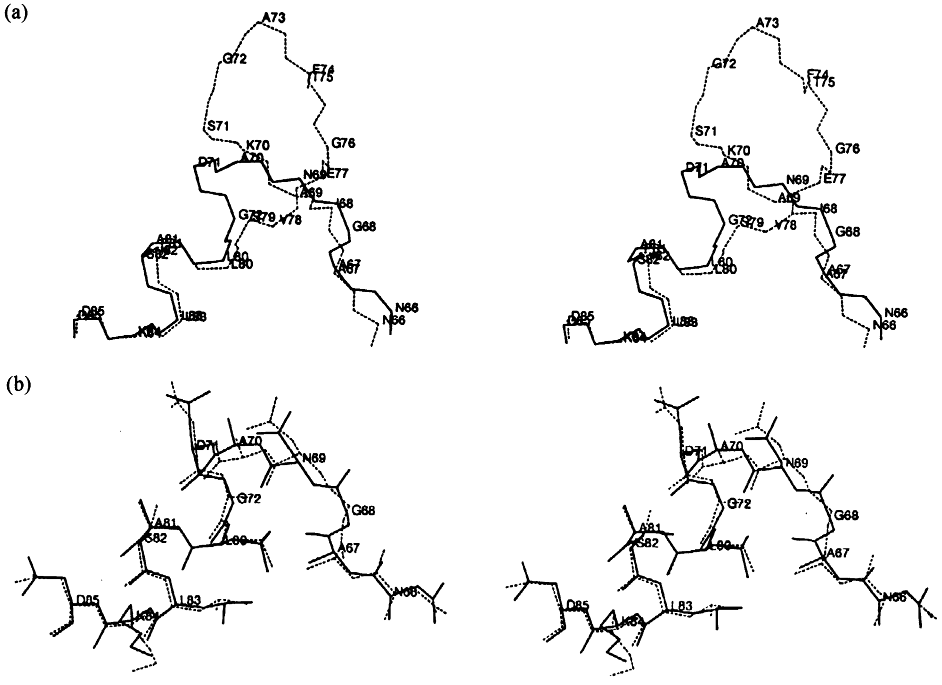


Fig. 12. Loop design for trypanosomal triosephosphate isomerase. (a) Sixteen-residue loop-3 of the original structure (PDB code 5tim) (dotted line) was replaced by an eight-residue fragment (solid line). Its amino acid sequence was designed and modeled by the BPMC procedure to guarantee a conformation free of energy strain. (b) The protein with the proposed sequence was synthesized, and its three-dimensional structure was determined by X-ray crystallography. Modeled (solid line) and experimental (dotted line) conformations were very similar, with the rms deviation equal to 0.4 Å.

energy optimization was performed for each of them. An eight-residue connection with sequence GNADALAS replacing the native sequence IAKSGAFTGEVSLPI between positions 68 and 82 was predicted to fold into a strainless loop with an additional helical turn at the N-terminus of helix 3.

The modified polypeptide was synthesized and the first suggested variant was experimentally proven to form a stable, monomeric structure with TIM activity [124]. Subsequent crystallographic studies of the redesigned protein, referred to as monoTIM, demonstrated that the protein retains the characteristic TIM-barrel fold and that the new loop was correctly predicted with a main-chain atom rms difference of 0.4 Å for the loop residues [140] (Fig. 12b).

Interestingly, two other loops of the original dimer interface, loop-1 and loop-4, were found to change their conformational state [163]. Loop-1 became disordered, which in turn influenced the ability of an active site residue Lys[13] to reach the substrate. This inspired another design project in which loop-1 was rigidified [164]. This loop design was the second blind test of the ICM loop prediction algorithm. A design scheme similar to the scheme used previously for loop-3 was employed. A series of ICM simulations suggested shortening of the eight-residue loop by one residue as well as some modifications of the sequence. The predicted structure was deposited in PDB and the crystallographic structure was determined. The experimental structure confirmed that the loop became rigid and was predicted correctly. The direct superposition of the lowest energy structure of the proposed loop KSGSPDS to the crystallographic structure results in an rms difference of 0.5 Å for the 28 main-chain atoms.

In summary, two (out of two) blind predictions aimed at designing loop-1 (eight residues) and loop-3 (seven residues) in triosephosphate isomerase were successful examples of the ICM *ab initio* loop prediction technique. Let us note, however, that these were not single loop predictions, but rather a series of iterative sequence modifications followed by structure predictions. In this setup, even slightly wrong initial predictions can be stabilized by further sequence modifications. Loop predictions in modeling by homology are much more challenging because a sequence cannot be adjusted and, more importantly, because the structural environment (loop ends and surrounding residues) of the loop on a homologous template may be strongly distorted with respect to the true environment [112].

Conclusions

1. Constraining covalent geometry results in a drastic increase in smoothness of the energy landscape. Consequently, the radius of convergence and efficiency of local minimization are increased. Peptide folding, loop prediction, and flexible docking can be efficiently formulated as a global energy optimization problem in a *subset* of internal coordinates.

2. Many compact nonnative conformations of proteins with energies higher than, but close to, the energy of the high-entropy unfolded state may exist (the *LEAF*

hypothesis). These structures can be stabilized by even a small number of mutations and these transformations may play a role in the evolution of protein topologies.

3. The existence of LEAF conformations would impose a limit of about 1 *kcal/mol/residue* on the *accuracy* of energy evaluations in the course of global energy optimization. Theoretical simulations of peptides with experimentally known conformations confirm this accuracy limit.

4. *Electrostatic solvation energy, surface energy and the side-chain entropy* term must be included in the optimized energy function. Omission of any of these terms may lead to an impermissible level of energy error. Algorithms for the fast evaluation of these terms have been developed.

5. A number of peptides up to 23 residues long having different experimentally characterized topologies (β-hairpin, α-helix, β–β–α-fold) can be predicted *ab initio* in a detailed full-atom ICM global optimization of the same energy function, including solvation and entropy terms. Exclusion of the entropic term led to significant deformations of the β-hairpin.

6. Although the essential features of the β–β–α-fold were reproduced, the accuracy was insufficient to predict the packing of two secondary structure elements unambiguously. This implies that *even better accuracy is necessary* for larger molecules.

7. Large-scale domain rearrangements can be simulated with limited success from a single starting conformation by the ICM global energy optimization of the interdomain linker and the interfacial side-chain torsion angles. The correct identification of the essential degrees of freedom, their number, and the validity of the underlying assumptions (no changes of the intradomain structure) are critical.

8. The association of two protein molecules can be predicted *ab initio* by the pseudo-Brownian Monte Carlo minimization procedure. The refinement of potential docking solutions by global energy optimization of the interfacial side chains results in a more reliable discrimination between correct and incorrect solutions.

9. Loop prediction and design: in two out of two cases of blind prediction in the course of loop design by the ICM method, the lowest energy conformations were practically identical to the conformations determined later by X-ray crystallography. Reliable loop prediction in models by homology is a much more difficult problem.

Acknowledgements

The author thanks Alexey Finkelstein for interesting discussions, Maxim Totrov for many discussions and help with simulations of the 23-residue peptide, and Vladimir Maiorov, Jonathan Weider, and Tim Cardozo for their help in preparation of the manuscript.

References

1. Anfinsen, C.B., Science, 181(1973)223.
2. Anfinsen, C.B. and Scheraga, H.A., Adv. Protein Chem., 29(1975)209.
3. Van Gunsteren, W.F. and Berendsen, H.J.C., Mol. Phys., 34(1977)1311.

4. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
5. Levitt, M., J. Mol. Biol., 168(1983)595.
6. Van Gunsteren, W.F., Berendsen, H.J.C., Hermans, J., Hol, W.G.J. and Postma, J.P.M., Proc. Natl. Acad. Sci. USA, 80(1983)4315.
7. Bruccoleri, R.E. and Karplus, M.A., Biopolymers, 29(1990)1847.
8. Van Gunsteren, W.F. and Berendsen, H.J.C., Angew. Chem., Int. Ed. Engl., 29(1990)992.
9. Mazur, A.K., Dorofeev, V.E. and Abagyan, R.A., J. Comput. Phys., 92(1991)261.
10. Tobias, D.J., Mertz, J.E. and Brooks, C.L., Biochemistry, 30(1991)6054.
11. Dorofeyev, V.E. and Mazur, A.K., J. Biomol. Struct. Dyn., 10(1993)143.
12. Levitt, M. and Warshell, A., Nature, 253(1975)694.
13. Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D., J. Mol. Biol., 106(1976)983.
14. Go, N. and Taketomi, H., Proc. Natl. Acad. Sci. USA, 75(1978)559.
15. Miyazawa, S. and Jernigan, R.L., Macromolecules, 18(1985)534.
16. Wilson, C. and Doniach, S., Proteins, 6(1989)193.
17. Sippl, M.J., J. Mol. Biol., 213(1990)859.
18. Friedrichs, M.S. and Wolynes, P.G., Science, 246(1990)371.
19. Chan, H.S. and Dill, K.A., Annu. Rev. Biophys. Biophys. Chem., 20(1991)447.
20. Finkelstein, A.V. and Reva, B., Nature, 351(1991)497.
21. Hinds, D.A. and Levitt, M., Proc. Natl. Acad. Sci. USA, 89(1992)2536.
22. Brower, R.C., Vasmatzis, G., Silverman, M. and Delisi, C., Biopolymers, 33(1993)329.
23. Madej, T. and Mossing, M.C., J. Mol. Biol., 233(1993)480.
24. Wallqvist, A. and Ullner, M., Proteins, 18(1994)267.
25. Monge, A., Friesner, R.A. and Honig, B., Proc. Natl. Acad. Sci. USA, 91(1994)5027.
26. Sali, A., Shakhnovich, E. and Karplus, M., Nature, 369(1994)248.
27. Kolinski, A. and Skolnick, J., Proteins, 18(1994)338.
28. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S., Protein Sci., 4(1995)561.
29. Srinivasan, R. and Rose, G.D., Proteins, 22(1995)81.
30. Hao, M.-H. and Scheraga, H.A., J. Phys. Chem., 100(1996)14540.
31. Li, H., Helling, R., Tang, C. and Wingreen, N., Science, 273(1996)666.
32. Yue, K. and Dill, K.A., Protein Sci., 5(1996)254.
33. Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A., J. Phys. Chem., 79(1975)2361.
34. Nemethy, G., Pottle, M.S. and Scheraga, H.A., J. Phys. Chem., 87(1983)1883.
35. Ripoll, D.R. and Scheraga, H.A., Biopolymers, 27(1988)1283.
36. Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S. and Scheraga, H.A., J. Phys. Chem., 96(1992)6472.
37. Kostrowicki, J. and Scheraga, H.A., J. Phys. Chem., 96(1992)7442.
38. Abagyan, R.A., FEBS Lett., 325(1993)17.
39. Unger, R. and Moult, J., J. Mol. Biol., 231(1993)75.
40. Abagyan, R.A., Totrov, M.M. and Kuznetsov, D.A., J. Comput. Chem., 15(1994)488.
41. Vasquez, M., Nemethy, G. and Scheraga, H.A., Chem. Rev., 94(1994)2183.
42. Abagyan, R.A. and Totrov, M.M., J. Mol. Biol., 235(1994)983.
43. Pedersen, J.T. and Moult, J., Proteins, 23(1995)454.
44. Augspurger, J.D. and Scheraga, H.A., J. Comput. Chem., 17(1996)1549.

45. Mertz, J.E., Tobias, D.J., Brooks III, C.L. and Singh, U.C., J. Comput. Chem., 12(1991)1270.
46. Board Jr., J.A., Causey, J.W., Leathrum Jr., J.F., Windemuth, A. and Schulten, K., Chem. Phys. Lett., 198(1992)89.
47. Rykunov, D.S., Reva, B.A. and Finkelstein, A.V., Proteins, 22(1995)100.
48. Li, Z. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 84(1987)6611.
49. Holak, T.A., Gondol, D., Otlewski, J. and Wilusz, T., J. Mol. Biol., 210(1989)635.
50. Eisenmenger, F., Argos, P. and Abagyan, R.A., J. Mol. Biol., 231(1993)849.
51. Mazur, A.K. and Abagyan, R.A., J. Biomol. Struct. Dyn., 6(1989)815.
52. Abagyan, R.A. and Mazur, A.K., J. Biomol. Struct. Dyn., 6(1989)833.
53. Rice, L.M. and Brunger, A.T., Proteins, 19(1994)277.
54. Ryckaert, J.P., Ciccotti, A.M. and Berendsen, H.J.C., J. Comput. Phys., 23(1977)327.
55. Park, B. and Levitt, M., J. Mol. Biol., 258(1996)367.
56. Harris, N.L., Presnell, S.R. and Cohen, F.E., J. Mol. Biol., 236(1994)1356.
57. Kolinski, A. and Skolnick, J., Proteins, 18(1994)353.
58. Honig, B. and Yang, A.S., Adv. Protein Chem., 46(1995)27.
59. Finkelstein, A.V., Gutin, A.M. and Badretdinov, A.Y., Proteins, 23(1995)151.
60. Lattman, E.E. and Rose, G.D., Proc. Natl. Acad. Sci. USA, 90(1993)439.
61. Ponder, J.W. and Richards, F.M., J. Mol. Biol., 193(1987)775.
62. Kabsch, W. and Sander, C., Proc. Natl. Acad. Sci. USA, 81(1984)1075.
63. Minor, Jr., D.L. and Kim, P.S., Nature, 380(1996)730.
64. Flanagan, J.M., Kataoka, M., Fujisawa, T. and Engelman, D.M., Biochemistry, 32(1993)10359.
65. Shortle, D.R., Curr. Opin. Struct. Biol., 6(1996)24.
66. Davis, M.E. and McCammon, J.A., Chem. Rev., 90(1990)509.
67. Bharadwaj, A., Windemuth, A., Sridharan, S., Honig, B. and Nicholls, A., J. Comput. Chem., 16(1995)898.
68. Honig, B. and Nicholls, A., Science, 268(1995)1144.
69. Purisima, E.O. and Nilar, S.H., J. Comput. Chem., 16(1995)864.
70. Jernigan, R.L. and Bahar, I., Curr. Opin. Struct. Biol., 6(1996)195.
71. Wodak, S.J. and Rooman, M.J., Curr. Opin. Struct. Biol., 3(1993)247.
72. Jones, D.T. and Thornton, J.M., Curr. Opin. Struct. Biol., 6(1996)210.
73. Shakhnovich, E.I. and Gutin, A.M., Nature, 346(1990)773.
74. Levitt, M., J. Mol. Biol., 104(1976)59.
75. Makhatadze, G.I. and Privalov, P.L., Biophys. Chem., 51(1994)291.
76. Shortle, D., Q. Rev. Biophys., 25(1992)205.
77. Braxenthaler, M., Samudrala, R., Pedersen, J.T. and Moult, J., Proc. CASP2, 1(1996)5.
78. Purisima, E.O. and Scheraga, H.A., J. Mol. Biol., 196(1987)697.
79. Wesson, L. and Eisenberg, D., Protein Sci., 1(1992)227.
80. Williams, R.L., Vila, J., Perrot, G. and Scheraga, H.A., Proteins, 14(1992)110.
81. Rashin, A.A., Prog. Biophys. Mol. Biol., 60(1993)73.
82. Juffer, A.H., Eisenhaber, F., Hubbard, S.J., Walther, D. and Argos, P., Protein Sci., 4(1995)2499.
83. Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C., Biochemistry, 20(1981)849.
84. Chan, H.S. and Dill, K.A., Biopolymers, 101(1994)7007.
85. Pearson, R.G., J. Am. Chem. Soc., 108(1986)6109.

86. Hempel, J.C., Fine, R.M., Hassan, M., Ghoul, W., Guaragna, A., Koerber, S.C., Li, Z. and Hagler, A.T., Biopolymers, 36(1995)283.
87. Nicholls, A. and Honig, B., J. Comput. Chem., 12(1991)435.
88. Juffer, A.H., Botta, E.F.F., van Keulen, B.A.M., van der Ploeg, A. and Berendsen, H.J.C., J. Comput. Phys., 97(1991)144.
89. Sitkoff, D., Sharp, K.A. and Honig, B., J. Phys. Chem., 98(1994)1978.
90. Bruccoleri, R.E., Novotny, J., Davis, M.E. and Sharp, K.A., J. Comput. Chem., 18(1997)268.
91. Zauhar, R.J. and Morgan, R.S., J. Mol. Biol., 186(1985)815.
92. Zauhar, R.J. and Varnek, A., J. Comput. Chem., 17(1996)864.
93. Friedman, H.L., Mol. Phys., 29(1975)1533.
94. Imoto, T., Biophys. J., 44(1983)293.
95. Moult, J. and James, M.N.G., Proteins Struct. Funct. Genet., 1(1986)146.
96. Schaefer, M. and Froemmel, C., J. Mol. Biol., 216(1990)1045.
97. Ben-Naim, A. and Marcus, Y., J. Chem. Phys., 81(1984)2016.
98. Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T., J. Am. Chem. Soc., 112(1990)6127.
99. Smith, K.C. and Honig, B., Proteins, 18(1994)119.
100. Eisenhaber, F., Protein Sci., 5(1996)1676.
101. Dougherty, D.A., Science, 271(1996)5246.
102. Warshel, A. and Russell, S.T., Q. Rev. Biophys., 17(1984)283.
103. Sharp, K.A. and Honig, B., J. Phys. Chem., 94 (1990)7684.
104. Banks, J., Brower, R.C. and Ma, J., Biopolymers, 35(1995)331.
105. Simonson, T., Perahia, D. and Bricogne, G., J. Mol. Biol., 218(1991)859.
106. Simonson, T. and Perahia, D., Proc. Natl. Acad. Sci. USA, 92(1995)1082.
107. Demchuk, E. and Wade, R.C., J. Phys. Chem., 100(1996)17373.
108. Brooks III, C.L. and Karplus, M., Methods Enzymol., 127(1986)369.
109. Zacharias, M., Luty, B.A., Davis, M.E. and McCammon, J.A., J. Mol. Biol., 238(1994)455.
110. Gilson, M.K. and Honig, B., Nature, 330(1987)84.
111. Totrov, M.M. and Abagyan, R.A., Nat. Struct. Biol., 1(1994)259.
112. Cardozo, T., Totrov, M. and Abagyan, R., Proteins Struct. Funct. Genet., 23(1995)403.
113. Strynadka, N.C.J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B.K., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. and James, M.N.G., Nat. Struct. Biol., 3(1996)233.
114. Go, N. and Scheraga, H.A., J. Chem. Phys., 51(1969)4751.
115. Karplus, M., Ichiye, T. and Pettitt, B.M., Biophys. J., 52(1987)1083.
116. Nemethy, G., Leach, S.J. and Scheraga, H.A., J. Phys. Chem., 70(1966)998.
117. Finkelstein, A.V. and Janin, J., Protein Eng., 3(1989)1.
118. Creamer, T.P. and Rose, G.D., Proc. Natl. Acad. Sci. USA, 89(1992)5937.
119. Pickett, S.D. and Sternberg, M.J., J. Mol. Biol., 231(1993)825.
120. Novotny, J., Bruccoleri, R.E. and Saul, F.A., Biochemistry, 28(1989)4735.
121. Doig, A.J. and Sternberg, M.J., Protein Sci., 4(1995)2247.
122. Lee, K.H., Xie, D., Freire, E. and Amzel, L.M., Proteins, 20(1994)68.
123. Koehl, P. and Delarue, M., J. Mol. Biol., 230(1994)249.
124. Borchert, T.V., Abagyan, R.A., Jaenicke, R. and Wierenga, R.K., Proc. Natl. Acad. Sci. USA, 91(1994)1515.
125. Maiorov, V.N. and Abagyan, R.A., Proteins, in press.

126. Shalloway, D., In Floudas, C.A. and Pardalos, P.M. (Eds.) Recent Advances in Global Optimization, Vol. 1, Princeton University Press, Princeton, NJ, 1991, pp. 433–648.
127. Clearwater, S.H., Huberman, B.A. and Hogg, T., Science, 254(1991)1181.
128. Kaesar, C. and Elber, R., J. Phys. Chem., 99(1995)11550.
129. Pedersen, J. and Moult, J., Curr. Opin. Struct. Biol., 6(1996)227.
130. McCammon, J.A., Gelin, B.R. and Karplus, M., Nature, 267(1977)585.
131. Powell, M.J.D., Math. Programming, 12(1977)241.
132. Noguti, T. and Go, N., Biopolymers, 24(1985)527.
133. Ripoll, D.R. and Scheraga, H.A., J. Protein Chem., 8(1989)263.
134. Skolnick, J. and Kolinski, A., Science, 250(1990)1121.
135. Go, N. and Scheraga, H.A., Macromolecules, 3(1970)178.
136. Ring, C.S. and Cohen, F.E., Isr. J. Chem., 34(1994)245.
137. Elofsson, A., Le Grand, S.M. and Eisenberg, D., Proteins, 23(1995)73.
138. Kang, H.S., Kurochkina, N.A. and Lee, B., J. Mol. Biol., 229(1993)448.
139. Dunbrack, R.L. and Karplus, M., J. Mol. Biol., 230(1993)543.
140. Borchert, T.V., Abagyan, R.A., Kishan, K.V.R., Zeelen, J.Ph. and Wierenga, R.K., Structure, 1(1993)205.
141. Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H. and Teller, E., J. Chem. Phys., 21(1953)1087.
142. Blanco, F.J., Jimenez, M.A., Herranz, J., Rico, M., Santoro, J. and Nieto, J., J. Am. Chem. Soc., 115(1993)5887.
143. Struthers, M.D., Cheng, R.P. and Imperiali, B., Science, 271(1996)342.
144. Abagyan, R.A. and Argos, P., J. Mol. Biol., 225(1992)519.
145. Wodak, S.J. and Janin, J., J. Mol. Biol., 124(1978)323.
146. Kuntz, I.D. and Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E., J. Mol. Biol., 161(1982)269.
147. Connolly, M.L., Biopolymers, 25(1986)1229.
148. Warwicker, J., J. Mol. Biol., 206(1989)381.
149. Goodsell, A.S., and Olson, A.J., Proteins Struct. Funct. Genet., 8(1990)195.
150. Cherfils, J., Duquerroy, S. and Janin, J., Proteins, 11(1991)271.
151. Jiang, F. and Kim, S.-H., J. Mol. Biol., 219(1991)79.
152. Chou, K.-C. and Carlacci, L., Protein Eng., 4(1991)661.
153. Bacon, D.J. and Moult, J., J. Mol. Biol., 225(1992)849.
154. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A., Proc. Natl. Acad. Sci. USA, 89(1992)2195.
155. Walls, P.H. and Sternberg, M.J.E., J. Mol. Biol., 228(1992)277.
156. Pellegrini, M. and Doniah, S., Proteins, 15(1993)436.
157. Vakser, I.A., Protein Eng., 8(1995)371.
158. Fischer, D., Lin, S.L., Wolfson, H.L. and Nussinov, R., J. Mol. Biol., 248(1995)459.
159. Goodsell, D.S., Morris, G.M. and Olson, A.J., J. Mol. Recog., 9(1996)1.
160. Janin, J., Prog. Biophys. Mol. Biol., 64(1995)145.
161. Diamond, R., J. Mol. Biol., 82(1974)371.
162. Sheriff, S., Silverton, E.W., Padlan, E.A., Cohen, G.H., Smith-Gill, S.J., Finzel, B.C. and Davies, D.R., Proc. Natl. Acad. Sci. USA, 84(1987)8075.
163. Borchert, T.V., Kishan, K.V.R., Zeelen, J.Ph., Schliebs, W., Thanki, N., Abagyan, R.A., Jaenicke, R. and Wierenga, R.K., Structure, 3(1995)669.
164. Thanki, N., Zeelen, J.Ph., Mathieu, M., Jaenicke, R., Abagyan, R.A., Wierenga, R.K. and Schliebs, W., Protein Eng., 10(1997)159.

Monte Carlo lattice dynamics and the prediction of protein folds

Jeffrey Skolnick[a] **and Andrzej Kolinski**[a,b]
[a]*Department of Molecular Biology, The Scripps Research Institute,*
10666 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.
[b]*Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland*

Introduction

The general solution to the protein folding problem demands that two very difficult problems be concomitantly solved [1]. An energy function whose global minimum is in the native conformation of the protein must be developed [2]. Simultaneously, an efficient strategy to search through the myriad of local energy minima for the desired global minimum must be formulated [3]. One way to attack both problems is to reduce the complexity of the model being considered [4]. Rather than treat the model at the level of atomic detail, the representation of the protein can be simplified. Various extents of this simplification have been explored. They range from highly simplified models that treat the native conformation of proteins as points on a small cube to high coordination lattice models that describe the native conformation of proteins with high geometric fidelity [5–14]. While highly idealized models have been useful in providing a number of qualitative insights into some general features of protein folding [10,15,16], they cannot be used to fold a real protein. This chapter focuses on results from high coordination lattice models of proteins that have been developed over the past several years and which are complementary to the simplified model studies [17–31]. These high coordination models not only provide insights into the thermodynamics of the protein folding process [25], but in a number of cases can predict the native conformation of a number of proteins at the level of 2–4 Å root-mean-square deviation (rms) from native [20,21].

The outline of this chapter is as follows. We begin with a discussion of the geometric model of a protein and the interaction scheme. In particular, we focus on the reasons why the various contributions to the potential are included and describe what happens if the individual terms are considered in isolation. We then describe the two types of Monte Carlo sampling schemes that have been employed, namely classical Metropolis Monte Carlo (MMC) [32] and the novel entropy sampling Monte Carlo technique (ESMC) due to Hao and Scheraga [33–35]. Next, results from the folding of some idealized protein sequences are presented [18,23,25]. These studies enable an exploration of the possible origins of the cooperativity of the protein folding process. We then summarize results on the folding of a number of small globular proteins [20,36], as well as some predictions for protein redesign [29,30]. Subsequently, a novel algorithm for the prediction of the locations where the protein chain reverses global direction, i.e. 'U'-turns and the dominant secondary structure found in the regions between the U-turns, is described [27]. This is followed by a review of folding results when a relatively small number of tertiary constraints (which may come from NMR experiments) are provided to the model [26]. Then, results from the *de novo* folding of the GCN4 leucine zipper, which adopts a dimeric coiled coil in solution, are summarized [21]. Next, an overview of the general formalism designed to predict the state of association of coiled coils [22,31], and comparison with experimental data on a variety of sequences, are presented [37,38]. We conclude with a discussion of the weaknesses of the present generation of lattice models and a perspective on the outlook for future progress.

Lattice models of proteins

As indicated in Fig. 1, the C^α coordinates of the protein backbone are confined to a set of lattice points which reside on an underlying cubic lattice, whose lattice spacing a = 1.22 Å [19]. Successive C^α atoms are connected by virtual bond vectors $a \cdot \mathbf{v}$, with $\{\mathbf{v}\} = \{(\pm 3, \pm 1, \pm 1), \ldots, (\pm 3, \pm 1, 0), \ldots, (\pm 3, 0, 0), \ldots, (\pm 2, \pm 2, \pm 1), \ldots, (\pm 2, \pm 2, 0), \ldots\}$. On considering all possible permutations of the coordinates, $\{\mathbf{v}\}$

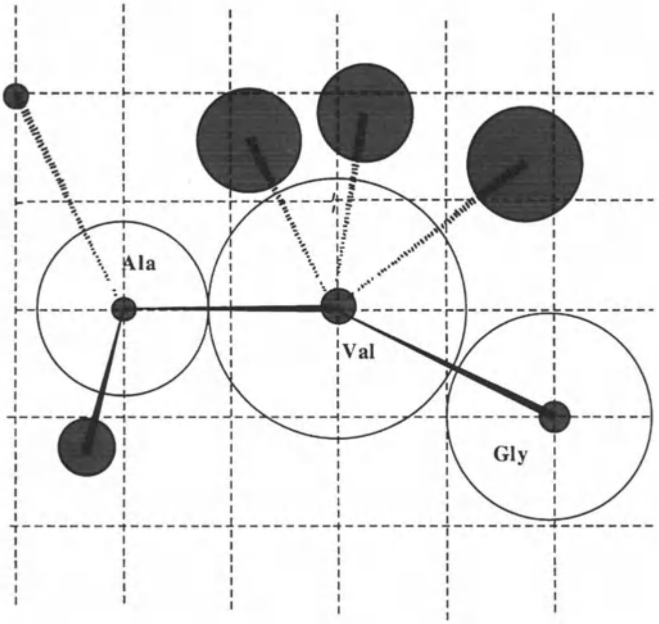


Fig. 1. Schematic representation of the geometry of the protein model. The C^α vertices are confined to high coordination lattice points. The side-chain centers of mass are located off-lattice. Ala, Pro, and Gly have a single rotamer for a given backbone virtual bond angle. All the other residues have multiple rotamers.

contains 90 basis vectors; thus, we refer to it as the 90-neighbor lattice. However, when the virtual bond angles are restricted to realistic values, the number of possible continuations of the C^{α} trace, given a pair of preceding C^{α}'s, is about 30. Thus, the intrinsic conformational entropy of the backbone is comparable to real proteins. The geometric accuracy of the C^{α} representation is in the range of 0.6–0.7 Å rms with respect to high-resolution PDB structures [39]. This is true regardless of protein size and orientation of the protein on the lattice. The fact that space is essentially isotropic and that all structures can be represented at comparable geometric resolution is the reason why this high coordination lattice is used.

Side chains are represented as a set of pseudoatoms located at the side-chain center of mass. For all amino acids except Gly, Pro and Ala, there are multiple rotamers. These rotamers are chosen so that the center of mass of a side chain in real proteins will be no farther than 1 Å from another member of the rotamer library. The side-chain rotamers are not confined to lattice points; however, the C^{α} backbone defines the reference frame for the rotamer coordinates. In a similar fashion, it is possible to rapidly and quite accurately reconstruct the peptide backbone and C^{β} atoms given a set of three virtual backbone bond vectors; the former can be used in an explicit atom hydrogen bond scheme [28,40]. More generally, many geometric and energetic quantities can be rapidly accessed from the set of virtual backbone vectors that define the instantaneous conformation of the chain. Thus, many quantities can be precalculated in advance. This allows for a two-order-of-magnitude speed-up over the corresponding model described in a continuous space representation [41]. The possibility of such a speed-up is absolutely essential to be able to adequately explore conformational space and is the principal reason why lattice models are used.

Interaction scheme

The key aspect of any successful model for protein folding is the nature of the terms that define the potential. Recently, it has once again become very popular to consider a very simple interaction set [42,43]; for example, all hydrophobic residues are treated as having the same interactions. While this simplicity is appealing, it belies the fact that such an approach generates many essentially isoenergetic chain conformations, many of which are geometrically unlike folded proteins. Typically, such an approach results in native-like states being in the best several hundred structures as ranked according to their energy [43]. While this selection may be somewhat better than random, in practice, there are far too many conformations to be of practical use. If, for example, one could predict a handful of different topologies, then such topologies might be differentiated experimentally. However, when there are several hundred possible answers, it is very unlikely that the correct fold can be fished out from the myriad of possibilities.

There is another reason why such a simplistic approach will not work. Consider the recent studies of Harbury et al. [37] on GCN4 leucine zippers and a number of mutants. They mutated the residues in the core to various combinations of Val, Leu, and Ile. Depending on the identity and location in the sequence of the hydrophobic residues, the equilibrium shifted from dimers to trimers and then to tetramers. Since for these sequences the residues in the core are always hydrophobic, an interaction scheme based on just two types of residues, hydrophobic and hydrophilic, could not possibly predict the state of association. More generally, it is possible to build structures of different topologies that have the same pair interaction as assessed by the number of interacting hydrophils and hydrophobes. While complexity for complexity's sake is to be avoided, it is precisely to reduce the number of possible low-energy topologies that more complicated interaction schemes have been developed.

Contributions to the potential

In what follows, we describe the qualitative features of the interaction scheme which we have developed. The origin of these models goes back to very simple HP-type models where there are only two kinds of residues, polar and nonpolar [4,6–8,14,44–50], but additional complexity has been added to reproduce essential features of the physics which would be absent if such terms were excluded [18,19,21,23,36]. We wish however to emphasize that the force field is constantly being improved and modified in order to enable us to fold a broader class of proteins; it reflects the ongoing process of our increased understanding of interactions in proteins. Each time the potential changes, we go back and repeat the folding simulations on those proteins already folded so as to ensure that the 'improvements' permit an ever-increasing set of proteins to be folded.

The potential must be designed so as to capture both generic and sequence-specific features of proteins. The nature of the individual contributions is listed below.

Hydrogen bonds

The most important generic term involves hydrogen bonds. Whether the relative intraprotein hydrogen bond free energy is less favorable or more favorable than that of hydrogen bonds to water is not the crux of the effect. What is most salient is that the presence of unsatisfied hydrogen bonds within a protein is energetically very unfavorable. Since hydrogen bonds are both distance- and orientation-dependent, they are an extremely important structural regularizing term. They serve to greatly restrict the manifold of accessible compact conformations.

Two versions of hydrogen bonds have been implemented. One is C^{α}-based [19] and the other uses an explicit backbone amide hydrogen and carbonyl oxygen representation [28]. Both reproduce about 90% of the hydrogen bonds as assigned by Kabsch and Sander [51]. The former is very much in the spirit of Levitt and Greer [52], whereas the latter was introduced to improve the hydrogen bonding in β-structures. In the absence of other contributions to the potential, at low temperature, they tend to generate helices punctuated by breaks where the prolines are located [53]. The choice of helices over β-states is due to entropic reasons.

Intrinsic secondary preferences

Next, there are amino acid pair-specific contributions that reflect the statistical preference of individual amino acids to adopt a given type of secondary structure. Both cooperative and noncooperative versions of this potential have been used. Basically, similar behavior is observed, but the former version yields a better defined interface between secondary structural elements [19,21,23,24,29,30,36]. This contribution to the energy of the folded state is typically about 20–25% of the total. When used alone, this term produces fragments of secondary structure, with a very diffuse and continuous conformational transition. When combined with terms that account for the generic stiffness of polypeptide backbones, the accuracy of secondary structure prediction is comparable to more standard methods, i.e., depending on the sequence, between 50 and 70% of the residues are correctly assigned [24]. Finally, by providing for a small, but nonnegligible, amount of secondary structure in the denatured state, these terms assist in the early states of folding and also act to reduce the configurational entropy of compact states. On average, they determine which type of secondary structure a protein adopts, but they can be overridden by tertiary interactions [20].

Burial as a one-body term

The next class of terms reflects the individual preference of a given residue to be buried or exposed. One-body burial terms serve to generate compact structures where on average the hydrophobic residues are in the interior and the hydrophilic residues are exposed. But, they generate nonspecific side-chain packing arrangements, and multiple topologies can have an essentially identical burial energy. In the absence of other terms in the interaction scheme, when reasonable coordination number lattices are used and compact structures at protein-like densities are generated, contrary to the hypothesis of Dill and co-workers [20], secondary structure is not enhanced on compaction [17,53–56]. Many of these random conformations would in reality have very high energies because these conformations would not be hydrogen bonded.

For single-domain globular proteins, a centrosymmetric potential that describes the tendency for an amino acid side-chain side to be located at a given relative distance from the center of mass of the protein has been used [19,57]. This formulation offers the advantage that it can accommodate the fact that residues such as tyrosine prefer to be located near the protein surface. However, it suffers from the disadvantages that there may be problems if the protein is very asymmetric, it has trouble differentiating edge from interior strands in β-structures, and it cannot be applied to multimeric or multidomain proteins. To address these concerns, a potential based on the number of side-chain contacts has also been introduced [21]. Whenever the environment of a residue exceeds the contact threshold, it is counted as buried. A possible problem with this approach is that the situation can arise where a substantial amount of the surface is actually exposed, but where the contacts are clustered over a relatively small portion of the surface. Improvement in the formulation of the burial potential is clearly necessary.

Pair potentials

These potentials of mean force help to select out the preferred topology and are operative in the molten globule and the native state. However, they do not provide a sufficient energetic separation between the native conformation and alternative higher energy structures, many of which have the native fold but different side-chain packing arrangements. Thus, they do not yield a unique native state with long-lived side-chain contacts. The best pair potentials of mean force have attractive or neutral interactions between hydrophils and attractive interactions between hydrophobes [23]; obviously, it is essential that hydrophilic and hydrophobic residues experience net repulsive interactions. When a pair interaction scale in which hydrophilic residues are repulsive is applied to a β-protein, then highly curved β-sheets are generated so as to minimize the number of hydrophilic–hydrophilic contacts. At a bare minimum, in order for twisted, quasiplanar β-sheets to be stable, the hydrophilic pair interaction should be no worse than neutral. However, in those scales where pairs of hydrophilic residues are attractive as are pairs of hydrophobic residues, then this contribution by itself cannot create the phase segregation where hydrophobic residues are on average found in the protein interior.

Many investigators, ourselves included, have developed statistical potentials of mean force between pairs of residues i and j obtained from expressions of the type [19,23,58,59]:

$$\varepsilon_{i,j} = -kT\ln(n_{obs}(i,j)/n_{exp}(i,j)) \qquad (1)$$

with $n_{obs}(i, j)$ the observed number of contacts between pairs of amino acids i and j. k is Boltzmann's constant and T is the absolute temperature. This quantity is directly obtained from a set of Protein Data Bank protein structures [60]. Here, $n_{exp}(i, j)$ is the expected number of contacts if interactions between i and j are random. It is in this term that the difference between all statistical contact potentials resides [59].

To date, the most sensitive residue-based pair potentials have been derived assuming that the quasichemical approximation holds for groups of heavy atoms; then, the average residue interaction based on the interaction between such groups is calculated [23]. The problem with the quasichemical approximation is that it ignores chain connectivity and the presence of regular secondary structure. Recently, a more general approach which includes these effects has been developed. Even at the level of interacting residues, it is the best inverse folding pair potential derived to date by our group [61].

Effective multibody interactions

If one defines a contact as occurring when any pair of heavy atoms is less than 4.2 Å apart, then interacting supersecondary structural elements in globular proteins exhibit well-defined side-chain contact patterns [18,19,62]. Typical helix-to-helix and beta-to-beta packing patterns are shown in Figs. 2A and B, respectively. Furthermore, it is possible to define a set of side-chain center of mass contact distance thresholds so that 82% of the heavy-atom contacts are recovered with a Matthews coefficient of

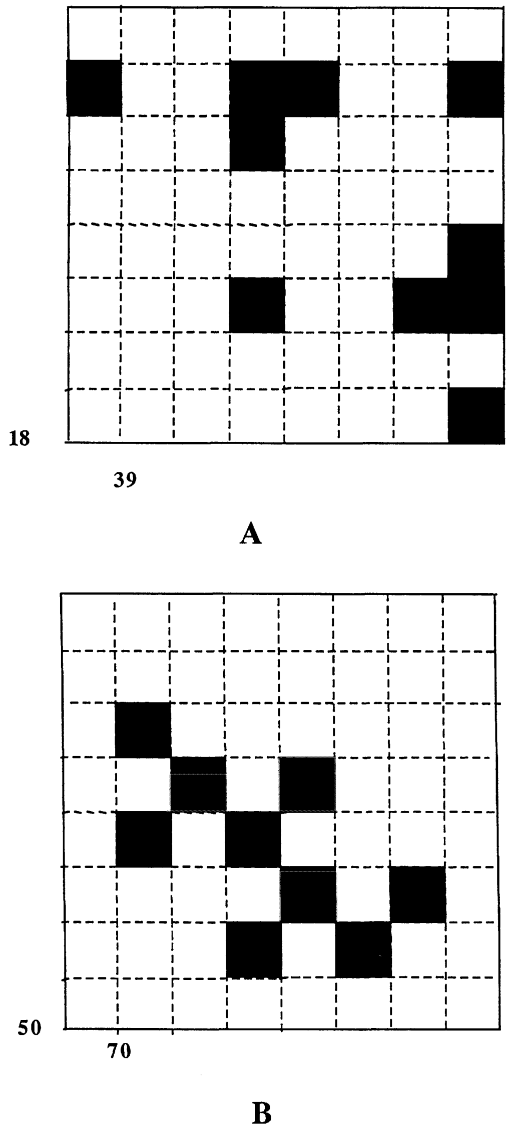


Fig. 2. Representative side-chain packing contact maps for an interacting pair of (A) antiparallel helices and (B) β-strands.

0.85 [61]. Thus, contact maps in the single ball side-chain description can essentially recover the heavy-atom contact description. In the absence of higher order multibody terms beyond pair contributions, we find that the predicted packing patterns of the resulting ensemble of structures exhibit essentially random overlap with the native state. However, the overlap with native contacts of the lowest energy structures is substantial [25]. Unfortunately, these very low energy states are rarely populated, and the folding transition is only very weakly cooperative. Furthermore, the models have much in common with the molten globule state of proteins [63–66]. They have substantial native-state secondary structure, but there is no fixation of tertiary contacts, and the manifold of structures tend to be swollen relative to the native state [18,25]. This can be rationalized as follows. Both one-body and pair interaction terms lack sufficient specificity to produce a native conformation that has a substantial energy gap with respect to other relatively nearby conformations. This results in an almost continuous transition from the unfolded state.

The higher order multibody component of the potential is only important when one has dense compact states; it permits, but does not require, side-chain fixation. Furthermore, without such terms, microphase separation of the side chains results, with an unphysical number and pattern of side-chain contacts. However, a key question is whether such potentials are physical or arise simply because reduced models are considered. The presence of reduced models certainly suggests that to some extent one must modify the interactions to reproduce the finer details of side-chain packing. However, even in molecular dynamics simulations of full-atom protein models, on starting from the crystal structure, the simulations tend to diffuse native side-chain packing towards a more liquid state [67]. Thus, the problem with extant potentials may be much deeper. Finally, we note that the potentials we are using are potentials of mean force. It is a well-known result from the statistical mechanics of small-molecule liquids that higher order correlation functions (for example, the three-body radial distribution function) are not simply factorizable into lower order distribution functions (this approximation is the Kirkwood superposition approximation), even if the naked potential is pairwise additive [68].

In order to introduce the possibility of side-chain fixation, Kolinski et al. [18,25,29,30,53,62] examined two classes of multibody terms. The first is of the form

$$E_4 = \sum(\varepsilon_{ij} + \varepsilon_{i+k,\,j+n})C_{ij}C_{i+k,\,j+n} \tag{2}$$

with $|k| = |n|$. $C_{ij} = 1$ when side groups i and j are in contact; otherwise $C_{ij} = 0$. $\varepsilon_{i,j}$ is the pair potential between amino acids i and j. We have considered models with $n = 3$ and 4. As indicated in Figs. 2A and B, such patterns are typical of both helix-to-helix and beta-to-beta side-chain packing patterns. We have also included $n = 1$ terms which are typical of beta-to-beta contacts. A second implementation of the multibody potential involves the use of a neural network to recognize whether or not 7-residue by 7-residue subfragments of dense regions of contact maps are native-like [62]. Such a formulation can in many cases recognize misfolded proteins based on contact maps alone. It offers the advantage that many more kinds of contact patterns are considered than are possible based on Eq. 2. On the other hand, the neural network does not consider the identity of participating amino acids. Although it was later modified to include the average pair potential of such interacting subfragments [29], it ignores the effect of side-chain size and may result in nonphysical packing arrangements.

Synergism of the contributions to the potential

Based on a large variety of simulations, we conclude that there is no single dominant interaction responsible for protein folding. In agreement with Hao and Scheraga [33–35], we conclude that the contribution to the stability of a protein due to interactions reflecting intrinsic secondary structure propensities (local hydrogen bonding plus local conformational preferences) is roughly equal to that of tertiary interactions [18,23,53]. Hydrogen bonding acts to restrict the manifold of compact states to those which are almost maximally hydrogen bonded, thereby reducing the conformational entropy of compact states. Similarly, intrinsic secondary structural preferences, although inherently weak, bias the system towards the secondary structure found in the native state. Of course, these can be overridden by tertiary interactions. Hydrophobic interactions create the average phase segregation of the amino acids. That is, in a typical protein roughly 75% of all hydrophobic residues are buried. However, since all hydrophobic residues are not buried, this argues that there are interactions (e.g. the resulting compact structure might not be hydrogen bonded) that oppose the burial of all hydrophobic residues. Pair interactions help reduce the configurational entropy of compact states by acting to break the degeneracy of compact structures and may serve to destabilize alternative conformations as well as to stabilize the native fold. Finally, higher order packing interactions might be responsible for the fixation of structure on passage from the molten globule to the native state [25].

Our simulations argue that a protein is a system under tension in the sense that while the system is in a global free-energy minimum, this minimum arises as a compromise between all the above terms [53]. The native conformation lies in that portion of conformational space consistent with the interplay of the interactions that comprise a globular protein. By eliminating any given class of terms, an important physical feature of a protein is removed. Thus, in these models, there is no single dominant term driving protein folding; rather the stability of the native state arises from the consensus and interplay of a number of terms representing different physical effects.

Monte Carlo sampling schemes

The well-known Metropolis Monte Carlo (MMC) procedure randomly samples conformational space according to the Boltzmann distribution of (distinguishable) conformations [32]:

$$P_i = \exp(-E_i/kT) \tag{3}$$

In order to generate this distribution, the transition probability $p_{i,j}$ from an 'old' conformation i to a 'new' conformation j (for the asymmetric scheme) is controlled by the energy difference $\Delta E_{ij} = E_j - E_i$ via

$$p_{i,j} = \min\{1, \exp(-\Delta E_{ij}/kT)\} \tag{4}$$

Obviously, this technique is very sensitive to the presence of energy barriers. To ensure adequate sampling, typically a collection of elemental backbone moves involving end moves, and collective motions of two to four bonds are randomly performed. In addition, small-distance motions of a large, randomly selected part of the chain are employed. Side chains can also independently move. The key to a successful dynamic Monte Carlo protocol is to include a sufficiently large move set so that no element of structure is artificially frozen in space.

To enhance the sampling efficiency, Hao and Scheraga [33–35] have employed the entropy sampling Monte Carlo method (ESMC) in their study of simplified protein models. ESMC was originally proposed by Lee [69] in the context of a simple Ising model and is closely related to the multicanonical MC technique of Berg and Neuhaus [70]. Since the formulation of Hao and Scheraga is the most straightforward and has been applied to both simplified and higher resolution models, we briefly review their approach.

Unlike MMC, ESMC generates an artificial distribution of states that is controlled by the conformational entropy as a function of the energy of a particular conformation E_i:

$$P_i^{ESMC} = \exp(-S(E_i)/k) \tag{5}$$

The transition probability can be formally written as

$$p_{i,j}^{ESMC} = \min\{1, \exp(-\Delta S_{i,j}/k) \tag{6}$$

with $\Delta S_{i,j}$ being the entropy difference between energy levels i and j, respectively.

At the beginning of the simulation, the entropy is not known. However, from a density-of-states energy histogram, H(E), an estimate, J(E), for the entropy S(E) can be iteratively generated. The kth iteration consists of an ESMC simulation run with S(E) approximated by $J_{k-1}(E)$. Here,

$$J_k(E) = J_{k-1}(E) + \ln(\max\{1, H_k(E)\}) \tag{7}$$

After a sufficient number of runs, all the states are sampled with the same frequency. Then, the histogram of H(E) becomes flat, and the curve of J(E) + constant approaches the true S(E) curve.

Folding protocol

For each sequence considered, starting from arbitrary random conformations, a series of independent simulated annealing experiments are performed. In many cases, at least 10, and more recently at least 20, independent simulations are performed, and the resulting minimum-energy structures are clustered according to global topology. If the dispersion in topologies is large, then the sequence is viewed as being nonfoldable using that generation of the model and its associated potentials. For those sequences that produce a handful (less than four topologies), then each of the topologies is subjected to an isothermal stability run. In a number of cases, the topologies which result are the native fold and its topological mirror image. For example, as shown in Fig. 3, there are left- and right-turning four-helix bundles. In both cases, the helices are right-handed, but the chirality of the topology is reversed. The structure with the lowest average and minimum energy is assigned to be the predicted native state. The resulting lattice model with side chains is then pulled off-lattice, and the backbone and side chains are reconstructed using the procedure described in Ref. 21. To date, the reduced and full atom models are completely compatible [21,29,30,36].

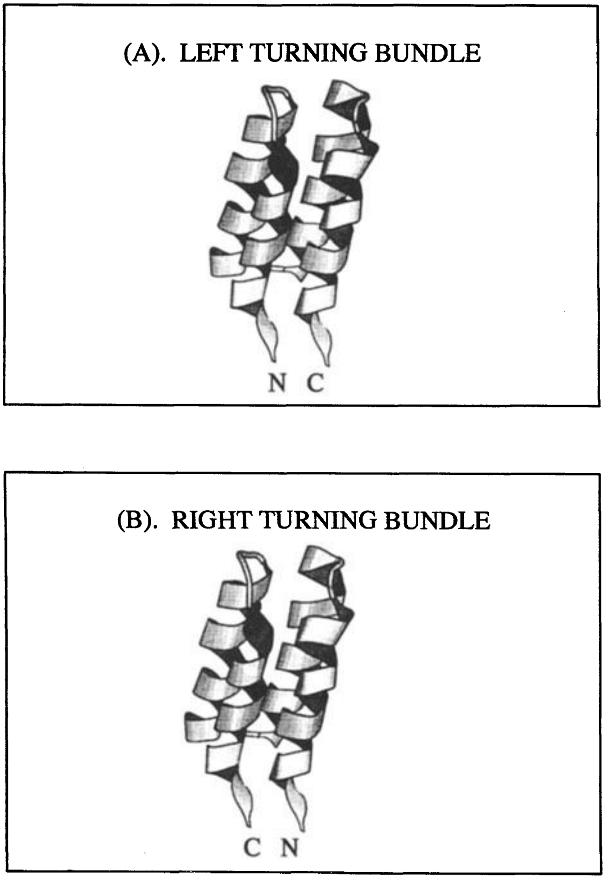


Fig. 3. Schematic illustration of (A) left- and (B) right-turning four-helix bundles.

Folding of exaggerated helical protein sequences

Using a lower coordination lattice, an early set of *de novo* simulations [18] (i.e. folding without any encoded knowledge of the native conformation) was performed on two 73-residue sequences designed by DeGrado and co-workers [71,72]. The first sequence contained an all-leucine core. In excellent agreement with experiment, this sequence is predicted to form a thermodynamically very stable four-helix bundle, but one with nonunique side-chain packing. The simulated sequence had many of the properties of a molten globule state. It had substantial secondary structure, and its average mean-square radius of gyration was about 15% larger than that found in the native state of a redesigned sequence (see below). Moreover, the simulations predicted that the right- and left-handed four-helix bundles should be isoenergetic. This prediction was subsequently confirmed by experiment [73]. Finally, within each topology, the molecule migrates among a few distinct families of structures which share the same global topology, but which differ in the identity of the residues which stabilize them.

A second sequence designed by DeGrado had 14 amino acid substitutions in the hydrophobic core [72]. In contrast to the first sequence, due to sequence heterogeneity in the hydrophobic core, differential pair interactions break the degeneracy of the various structures, and this molecule is predicted to prefer the right-turning, four-helix bundle topology. Following rapid assembly using standard MMC to a four-helix bundle topology, the molecule slowly relaxed to a more compact structure which is unique at the level of resolution of this class of models. Moreover, since the energy monotonically decayed as a function of time during the relaxation process, this implied the existence of entropic barriers between the compact molten globule-like state and the predicted native conformation. Fixation of the side chains was observed to occur when higher order multibody terms of the type given by Eq. 2 are included in the model, but it does not happen if such terms are deleted. These simulations pointed out the importance of including a cooperative protein-like interaction scheme into the potential used in folding.

Factors responsible for the uniqueness of the native structure

The full lattice model described above was used to explore the requirements for the *de novo* folding from an arbitrary random conformation of idealized sequences of four- and six-stranded β-barrels [23,25]. Of particular interest is the design of a putative 45-residue, six-stranded β-barrel which adopts the schematic topology shown in Fig. 4A. Simulations using MMC were used to test various possible conjectures about the factors responsible for the structural uniqueness of the native state [23]. Among these were the relative importance of generic hydrophilic/hydrophobic amino acid patterns, and the possible role of polar amino acids in destabilizing misfolded conformations [37].

A simple alternating pattern of valines and serines in the putative β-strand regions, when punctuated by appropriate turn-forming residues, is found to produce

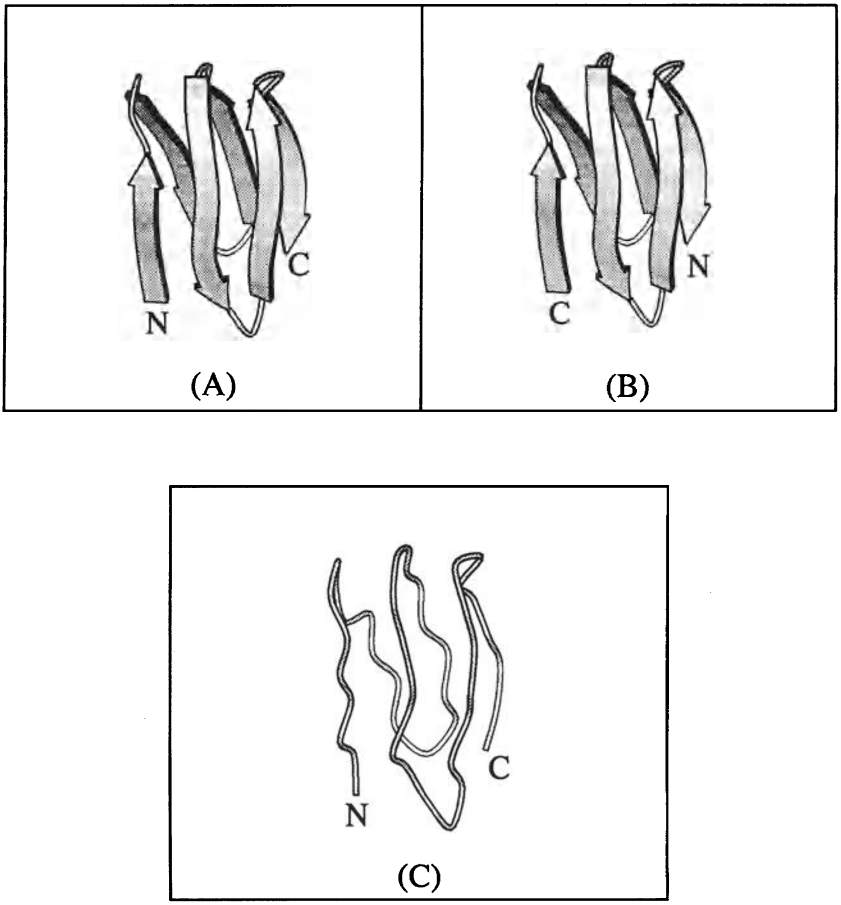


Fig. 4. Schematic illustration of (A) the desired six-stranded β-barrel and (B) the mirror image barrel. The predicted C^α trace of the native structure of the designed sequence betamod is shown in (C).

a manifold of six-stranded β-barrels having different topologies. This implies that a simple HP (nonpolar/polar) model is not sufficient to yield a structurally unique native state when systems having conformational entropy on the order of that of real proteins are considered. Furthermore, the packing of the resulting hydrophobic core is very diffuse. Thus, to enhance the stability of the hydrophobic core and to partially break the degeneracy of the various topologies, four Phe residues were introduced into the sequence. This reduced the number of observed topologies; however, the topology was not uniquely defined. Substitution of Asp for Ser residues was done at positions designed to destabilize incorrect topologies. The resulting sequence adopted the desired as well as the mirror image topology shown in Fig. 4B. Analysis of the energetic contributions indicated that the packing interactions favored the desired fold, but that the residues introduced in the turns favored the mirror image topology.

Substitution with Gly linkers resulted in the desired native fold becoming the most stable topology. The resulting designed sequence, called betamod, is given by

GVDVDV-GGG-**VDVDV**-GGG-**FRFRV**-GGG-**VRFRF**-GG-**VDVDV**-
GGG-**VDVDV**

The residues in bold indicate the location of the putative β-strands. Strands 1, 4, and 5 form the first β-sheet, while strands 2, 3, and 6 form the second sheet. The loop/turn regions are composed of flexible Gly connectors. A representative conformation obtained from the simulations is shown in Fig. 4C.

A question remains as to whether this sequence would in reality adopt a unique native conformation, a molten globule state or would not fold at all. Thus, experimental examination of this sequence is currently underway in the laboratory of Dr. Derek Woolfson [74]. In the interim, the results of these simulations suggest that these models might prove to be useful tools in protein design.

Origin of the cooperativity of protein folding

A very important question is whether simplified models can reproduce the thermodynamic behavior of proteins. Experimentally, in a number of proteins, the cooperativity in protein folding arises on passage from the molten globule state to the native conformation [63,64,75]. Such molten globules or compact intermediates have a volume which is about 50% larger than native, a substantial amount of native secondary structure, but diffuse tertiary contacts. These observations suggest that the fixation of side chains accompanying the transition to the native conformation is involved in the cooperativity of protein folding.

To investigate the possibility of a first-order transition in protein folding, Hao and Scheraga [33] employed the ESMC method to examine the folding thermodynamics of a 38-residue protein confined to the 210-lattice introduced by Kolinski et al. [14] and Skolnick and Kolinski [76]. Subsequently, they examined the sequence requirements for an all or none transition [34,35]. They conclude that designed or optimized sequences exhibit a cooperative folding transition which is long-range (i.e. involves tertiary interactions), whereas random sequences fold to compact states by what is an essentially continuous transition. These are very important studies, because they show for the first time in a nontrivial model that adoption of a unique low-energy state depends on the interplay of long- and short-range interactions. These model proteins included a local conformational bias for native-like secondary structure and a single side-chain rotamer for each residue.

Subsequently, Kolinski et al. [25] employed the ESMC method to investigate the folding thermodynamics of betamod. These studies build on the Hao–Scheraga work in the following ways. Now, a much higher coordination lattice is used, there is no target bias for the native state's secondary structure, and multiple side-chain rotamers are present so that the possibility of side-chain fixation exists. These three differences result in a model which has a considerably higher entropy in the compact state.

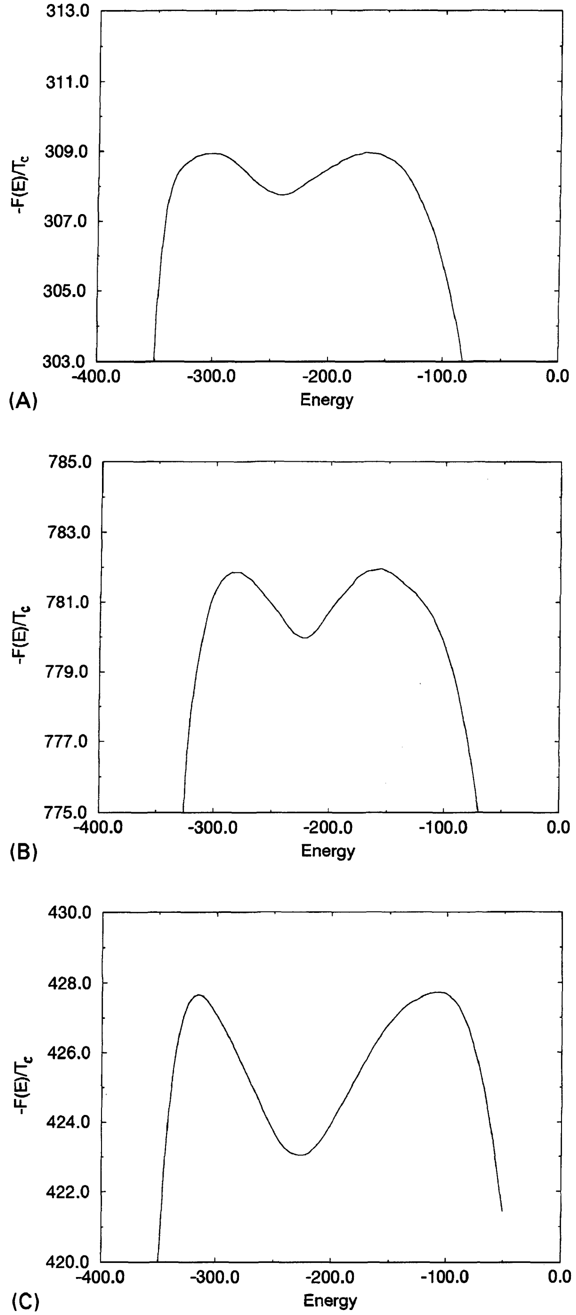


Fig. 5. Plots of the free energy, F/T_c, versus energy, E, for models I–III at the transition midpoint temperature, T_c.

Three distinct versions of tertiary interactions were considered. In the first, model I, only pair potentials are used. Model II also includes n = 3 and n = 4 type terms given by Eq. 2. Model III extends model II to include beta-type n = 1 terms. The scale factors for the pair and multibody interactions have been adjusted so that the tertiary interaction energy in the putative native fold of all three models is essentially the same.

As shown in Fig. 5, where the reduced free energy, F/T_c, versus energy, E, is plotted at the folding transition temperature, T_c, qualitatively different behavior is seen on passage from model I to model III. Model I, lacking high-order multibody interactions, essentially has a continuous thermodynamic transition. With the inclusion of higher order multibody packing interactions, the conformational transition becomes all or none. Interestingly, the lowest energy states in all three models correspond to the same manifold of structures (i.e. structures which are unique at the level of resolution of the lattice models) and correspond to the native fold shown in Fig. 4C. What differs in the three models is the separation of the low-energy native-like state from the manifold of other conformations that contribute to the partition function.

Nature of the transition state

The nature of the conformations located at the free energy versus energy maximum, viz. the transition state, was examined. In models II and III, which exhibit two-state

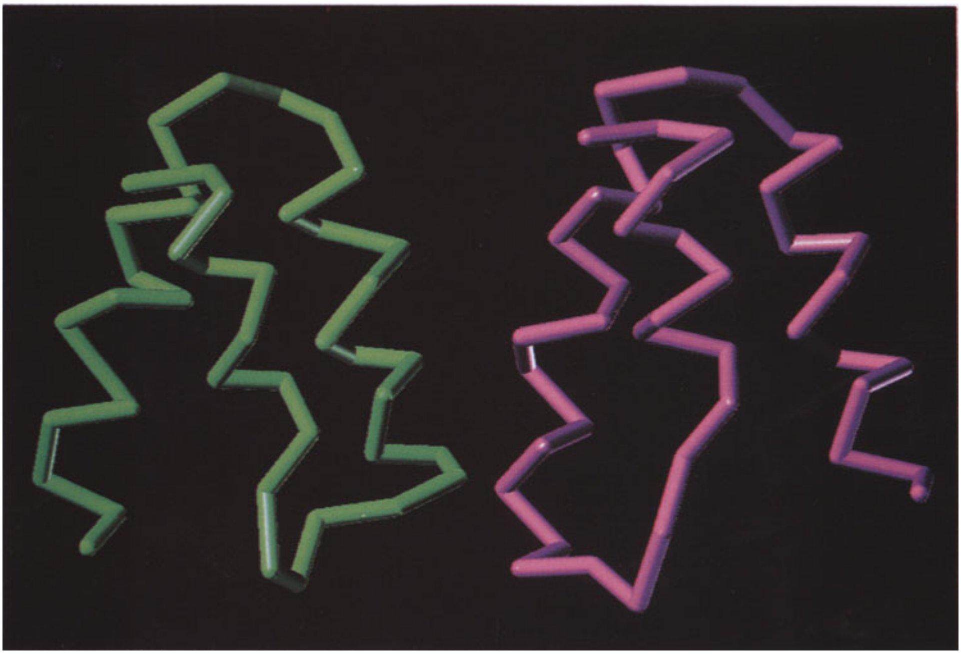


Fig. 6. Predicted native state and corresponding mirror image topology of the B domain of protein A shown in green and magenta, respectively.

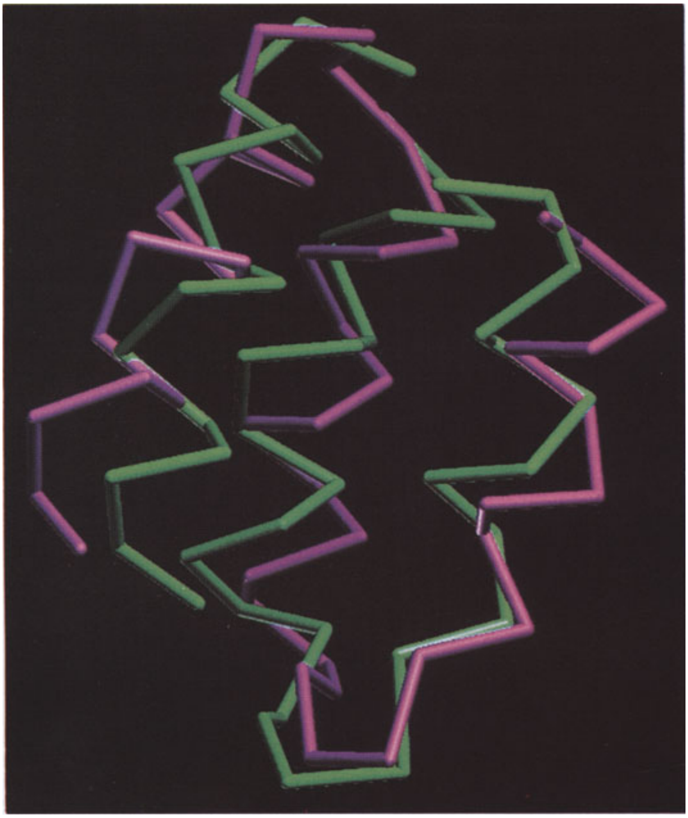


Fig. 7. Predicted backbone trace of protein A in magenta, superimposed on the experimental NMR solution structure in green.

thermodynamics, the transition state is comprised of structures having about 60% of the native state's secondary structure, about 50% of the side-chain contacts which are native, and a volume which is about 50% larger than native. This description of the transition state supports Kuwajima's [63] critical substructure model, where the activated state has a partial amount of native secondary structure, there are a subset of native contacts and the molecule is swollen relative to the native conformation. Such a range of physical properties has also been experimentally observed in a number of systems including α-lactalbumin and calcium-binding parvalbumin [77]. These simulations suggest that cooperative many-body interactions involving protein side-chains are the dominant factor responsible for the cooperativity of protein folding in models where the side-chains have internal degrees of freedom and perhaps in real proteins as well.

Folding of domains of protein A

In solution, the B domain of protein A is a 55-residue protein in which residues 10–55 adopt a three-helix bundle geometry [78]. Because of its structural simplicity and small size, this protein is a natural testing ground for *de novo* prediction methods, and a variety of generations of the model have been applied to this molecule [20,29,36]. Initially, folding of this molecule was attempted on a coarser lattice, followed by refinement on the 90-neighbor (finer) lattice described above [20,36]. Based on both the average and minimum energy, the correct topology is chosen over the mirror image, both of which are depicted in Fig. 6 in green and magenta, respectively. For this sequence, all contributions to the energy favor the native state. The resulting structures have an rms from native for residues 13–55 of 3.3 Å. A typical predicted conformation superimposed on the solution NMR structure of Gouda et al. [78] is shown in Fig. 7. Subsequent refinement of the model showed that the folding on a coarser lattice followed by refinement on the finer lattice (which permits better helix-to-helix packing) is unnecessary; rather, direct folding on the finer lattice is a more straightforward and simpler procedure [20,29,36].

In the original simulations, folding tended to occur by the preferential formation of the N-terminal hairpin, followed by assembly of the final helix [20]. Subsequent simulations on a more refined model suggest that the C-terminal hairpin assembly is more likely. There is also an indication that the C-terminal helix of the B domain of protein A may be stable in solution. There is some experimental indication that this might be the case [79]. In agreement with experiment, the simulations predict that the folded state is native-like with very long-lived side-chain contacts.

The B domain is but one of five highly homologous, extracellular domains of protein A designated as E, D, A, B, and C, respectively. These five domains all bind to immunoglobulin. Thus, it is a reasonable conjecture that all have the same solution structure. In addition, Montelione and co-workers have determined the NMR solution structure of the Z domain of protein A, which differs from the B domain by the single-point mutation G30A [80,81]. They find that it has a very similar fold to that of the B domain. In order to investigate the ability of the folding algorithm to fold homologous sequences, the folding of all five wild-type domains and the Z domain was successfully undertaken. In all cases, the native topology is energetically favored over the mirror image topology, with the C domain exhibiting the smallest energetic preference for the native over the mirror image fold. For all six sequences, the final structures are within 3.5 Å rms of the predicted B domain conformation.

Redesign of protein A to adopt the mirror image topology

A key question in understanding the principles of protein folding is the origin of the preference for a given topology as opposed to the topological mirror image. Two viewpoints have emerged: in one, the topology is dictated by the packing interactions in the hydrophobic core [82,83], and in the other the turns play a role in dictating the preferred topology [84,85]. To examine these questions, we attempted to redesign the sequences of the B and A domains of protein A so that they adopt the mirror topology shown in Fig. 6 in magenta. Both multiple mutations in the hydrophobic core and in the turn regions between helices I and II were made. To scan a large number of mutations to search for sequences that favor the mirror image over the native fold, a sieve method was developed. Modifications in the hydrophobic core were made in three groups, each involving point mutations at six sites. At each site, the native residue was replaced by Ala, Val, Ile, Leu and Phe. Thus, 15 625 mutations were examined for each group of mutation sites. For those sequences which survived the sieve procedure, none was found to prefer the mirror image over the native fold. Therefore, these results are consistent with the idea that the fine details of hydrophobic packing do not constitute the sole driving force for the folding process, but may stabilize an already acquired motif [86].

Next, two-point mutations in the turn connecting helices I and II were examined. With the exception of glycine, proline and cysteine, all possible mutations of Asn^{22} and Asn^{24} were allowed. Again applying the sieve procedure, most mutations in the turn regions do not disrupt the preference for the native fold. This is qualitatively consistent with experiments which indicate that, in general, turns can be modified without qualitatively changing the global fold [87,88]. However, about 11 of the 289 mutants, i.e. about 4%, resulted in sequences having varying degrees of preference for the mirror image topology.

The most promising N22R and N24M double mutant was subject to further evaluation. While the probability of finding an Arg in the $i+2$ position of the turn is relatively high, Met is rarely seen in the N-caps of helices [89]. To confirm that the RM mutant is foldable (at least in computro), a series of 10 independent MMC folding simulations were undertaken. In 6 of 10 simulations, the mirror image topology is obtained, with the remainder adopting the native fold. The RM mutation modifies the intrinsic secondary structure preferences so that in contrast to the wild-type they now favor the mirror image turn. This tendency is further augmented by the burial of M in the mirror image, but not in the native fold. This produces a net favorable pair interaction for the mirror image topology. In other words, the predicted preference for the image topology arises from the favorable juxtaposition of intrinsic secondary preferences and tertiary interactions [85]. These models indicate that such a juxtaposition is what is responsible for the adoption of a particular fold. To ensure that the predicted lattice models are consistent with atomic resolution models, all-atom models were constructed and were found to be in complete qualitative agreement. The experimental test of this prediction is now underway in Dr. Peter Wright's [79] laboratory here at Scripps. Finally, to examine the robustness of the RM mutation, it was applied to the A domain of protein A. The simulations predict that this sequence should also be a likely candidate for adopting the mirror image topology.

Effect of amino acid order on folding

By reading the sequence of a naturally occurring protein backwards, i.e. generating a retroprotein, a sequence of the same composition and hydrophobicity as the wild-type protein results [90]. However, proteins are chiral systems, and there have been a number of conjectures as to whether retroproteins will fold, and, if so, what topology will they adopt. Some authors have gone so far as to suggest that a retroprotein might adopt the mirror image structure, including left-handed helices [91]. While this conjecture is unlikely, there are a number of consequences accompanying the retroinversion of a sequence. All prediction methods based on composition will predict the same structural class for the wild type and retroprotein [92]. However, if the positions of the helices and turns remain the same as in the wild type, then in general the locations of the capping [93,94] and turn residues will not be optimal [89]. At a minimum, one might expect some rearrangement of the secondary structural elements. If side-chain packing and/or the distribution of nonpolar side chains is a dominant factor in determining the global fold, then one might expect the retroprotein to adopt the same topology as that of the native protein. There is always the possibility that an entirely different fold might be adopted or the retroprotein might not fold at all.

Because of the robustness of the lattice folding algorithm as applied to protein A, the retrosequence of the B domain was generated and subjected to a series of 15 folding experiments [30]. The results strongly suggest that the predicted native state of retroprotein A, shown in Fig. 8, is a three-helix bundle of the same topology as the wild-type sequence. It is important to emphasize here that the prediction that the retrosequence adopts the same topology as the wild-type sequence may be due to

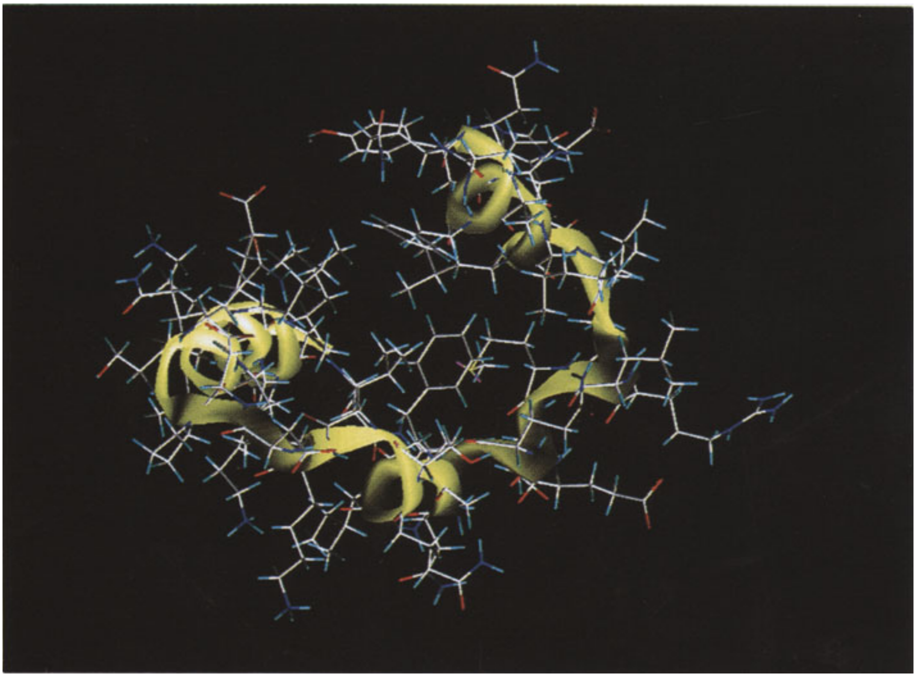


Fig. 8. Predicted all-atom model of the retrosequence of the B domain of protein A. The ribbon tube depicts the position of the backbone atoms and clearly indicates a three-helix bundle topology.

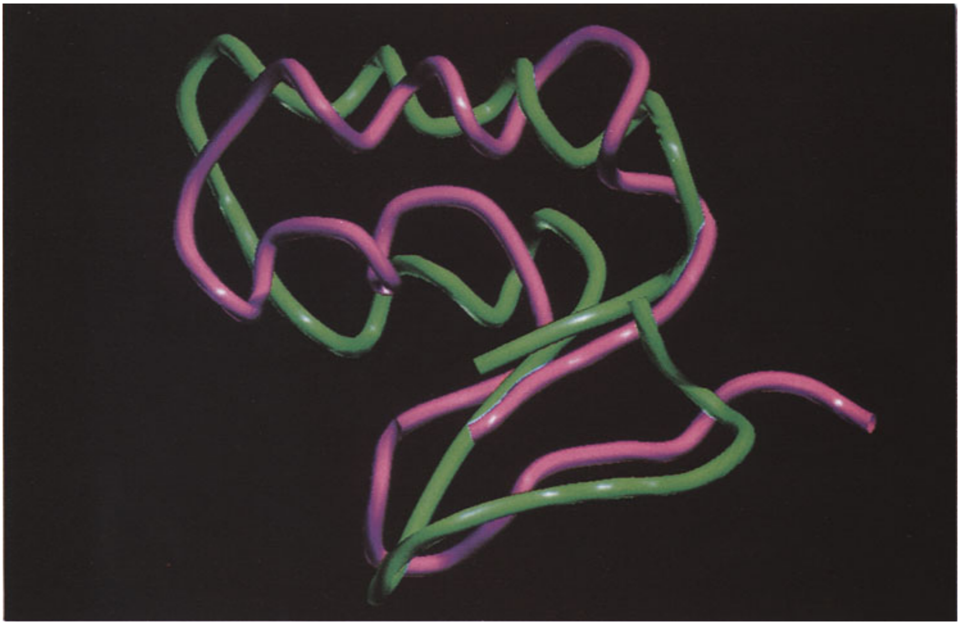


Fig. 9. Predicted C^{α} trace of the native conformation of crambin, in magenta, superimposed on the crystal structure, in green.

the high symmetry of the three-helix bundle fold, and it is very likely that this result is not true in general.

To accommodate the local secondary structural propensities, the secondary structural elements shift their positions with respect to the B domain. Among the most salient changes is the shift in the location of the C-terminal turn. This adjustment in position allows for the third helix to have N-cap residues that are favored. Furthermore, the predicted structure retains many of the hydrophobic core contacts as in the B domain. This suggests that hydrophobic interactions exert an important influence on driving the system to adopt a three-helix bundle topology. However, pair interactions alone are isoenergetic in the native and mirror image fold. What drives the system to favor the native fold is the difference in burial energies of the two topologies. This implies that in this case burial preferences select out the native over the mirror image topology.

Subsequently, atomic models were built from the lattice structures. In all cases, the hydrophobic core is well packed. Depending on the starting structure, the N- and C-terminal ends of the helices vary by about one residue. Overall, the lattice and all-atom models are consistent. Encouraged by these results, the structure of this sequence is now being determined by Dr. Chi-Huey Wong's group [95] at the Scripps Research Institute.

Folding of ROP monomer

The native structure of wild-type ROP is a dimer consisting of two antiparallel helical hairpins arranged in a coiled-coil geometry [96]. Sander and co-workers have redesigned this molecule to form a 120-residue, monomeric, left-turning, four-helix bundle [97]. Subsequently, Regan et al. have also redesigned the dimer to form a monomer using glycine linkers of various lengths [88]. In simulations done to date, the original Sander sequence has been used; work is in progress to fold those sequences designed by Regan et al. In the earlier simulations, folding commenced on a coarser lattice from random geometries [19,20]. A very strong preference for the designed, left-turning bundle was indicated. As in the case of protein A, the resulting low-energy structures were then projected onto the 90-neighbor lattice and refined. The predicted structures have a C^{α} rms ranging from 2.6 to 4.2 Å with respect to the set of equivalent residues in the ROP dimer crystal structure. What is striking is that the simulations predict that the molecule has less supertwist than is found in the ROP dimer structure. Whether these predictions are true or not awaits the experimental determination of the ROP crystal structure.

The folding simulations predict the existence of late, presumably molten globule, folding intermediates that are present prior to the formation of the native state. These metastable intermediates have the same global fold as native, but their radius of gyration is about 5% larger. Similar chain expansions have been observed in an apomyoglobin folding intermediate [98]. The secondary structure is essentially identical to native but there are much larger fluctuations in the turn regions and at the chain ends. In the molten globule, none of the side-chain contacts survives for the entire simulation run, while in the native state there are many such long-lived contacts. Furthermore, in the molten globule, the side-chain contact patterns are more diffuse, which is consistent with the observation that the helices are sloshing back and forth against each other. Detailed analysis of the dynamics of the molten globule state indicated that it is very liquid-like, and has much in common with the dynamics of a gel. In contrast, the native conformation is much less mobile, with the displacements (apart from global diffusion) limited to relatively small-scale motions.

Folding of crambin and the use of predicted secondary bias to enhance folding efficiency

To address the concern of whether the model can predict the tertiary structure of α/β proteins, the folding of crambin was undertaken [19,20]. This 46-residue protein has a native state comprising a helical hairpin and a three-stranded antiparallel β-sheet. It also contains three disulfide cross-links [99]. The simulations do not assume anything about a specific cross-link pattern, but rather that cystines of some sort are present. When straightforward folding from the random state was undertaken, the correct topology and disulfide pattern is always recovered, but in most cases the secondary structure, especially in the putative helical regions, is highly distorted.

To alleviate this problem, higher temperature simulations were undertaken where the S-S bond dissociation rate is sufficiently high. Then, statistics about secondary structure preferences (helix/turn or extended/loop) are collected. This prescreening predicts the location of the N-terminal helix with a shift of two to three residues towards the amino terminus, but the prediction for the second helix is more accurate. This stands in contrast to most standard secondary structure prediction methods which mostly predict β-strands in these regions [100,101]. With an approximate prediction of the helical regions in hand, a small energetic bias (proportional to the helicity of a given residue at the higher temperature) is added to the model. In about 50% of the folding simulations, low-energy conformations having a helical hairpin whose C^{α} rms from native is about 4 Å are predicted. The other conformations preserve the global topology of the native fold, but are 20% higher in energy. Subsequent refinement at low temperature produces structures whose average C^{α} rms is below 4 Å. For residues 3–42, the average coordinate rms is 3.6 Å, with a distance rms of 2.6 Å. As is evident from Fig. 9, which shows the predicted structure in magenta superimposed on the backbone of the crystal structure, in green, while the global fold is well reproduced, there are slight shifts in the position of one of the helices and the conformation of residues 43–46 is incorrect.

This protocol has also been applied to the folding of protein A, with comparable results obtained as when the predicted secondary structure bias is not incorporated into the folding algorithm. In addition, the protocol has been employed to predict the tertiary structure of the V-3 loop of gp-120 [102]. The resulting conformation is suggested to consist of three β-strands and a small C-terminal helix. The results of these simulations suggest that either experimental or predicted secondary structural constraints can be incorporated into the folding algorithm. Such predicted biases can greatly speed up the folding process. However, care has to be taken to ensure that if the prediction is uncertain, it can be overridden by other interactions.

Method for the prediction of surface U-turns and transglobular connections in small proteins

A knowledge of the locations where the chain changes its global direction, i.e. the U-turns, and of the dominant secondary structure of the intervening transglobular regions, i.e. the blocks, represents very useful information for a folding algorithm [27]. Thus, a simple method for predicting these building blocks in small single-domain proteins has been developed. Such an approach is complementary to more standard secondary structure prediction schemes [103]. Here, global rather than local information is desired; the structural assignments depend on the conformation of the entire chain. For example, if a given region favors helix, this tendency can be overridden because another part of the chain has a lower energy if it is helical and the first helical region is shifted to form a turn.

Table 1 *Summary of prediction statistics for the blocks and U-turns algorithm*

Protein name[a]	Surface U-turn prediction accuracy[b]	Errors of U-turn locations[c]	Secondary structure block prediction accuracy[d]	Comments on wrong assignment
1gbl	4/4	0-2/2#-3-0	5/5	–
proA	2/2	2-3	3/3	–
1fas	5/5	2-1-0-2-0	5/5	Terminal coiled assigned β, inserted β without a turn
1pou	3/3 1 over	4-3-2-0	4/4	Extended coil inserted
1tlk	7/7	5-3-2/1-1-2-4-7	7/8	Turn inserted into the C-terminal β-strand
1ris	5/5	3-5-6-3/4-4	5/6	Second β-strand predicted helical
1lpt	4/4	0-5-2-0	4/4	Shifted turns, hairpin-like C-terminus predicted as β
1ten	6/7	1/1-3-0-3-2-2/1	7/8	Shifted turns, one β-strand missed
1mjc	5/5	1-2-0-2-0	6/6	Long central coil added as β
–	41/42 = 98% 1 over	–	46/49 = 94%	Including overpredicted turn in 1tlk

[a] PDB descriptor.

[b] The ratio of the correctly predicted number of surface U-turns to the actual number in the protein. A turn is said to be correctly predicted if its boundaries at least partially overlap with the actual turn location. 'Over' indicates that an additional U-turn(s) is predicted which does not occur in the protein structure.

[c] i/j means that i residues of the preceding block and j residues of the following block have been incorrectly assigned as a part of the surface loop/turn. Otherwise, the number of overassigned residues of one of the transglobular blocks is given.

[d] The ratio of the correctly predicted number of secondary structure blocks to the actual number of surface turns in the protein. The secondary structure of a given block is said to be correctly predicted when the secondary structure of the three central residues agrees with the experimental structure.

The method consists of five basic steps:

(1) Estimate the radius of gyration of the protein. This imposes restrictions on the maximum and minimum length of the extended and helical fragments that can fit into the globule.

(2) The locations of the U-turns are randomly chosen and hairpins appropriate to the chain division are pulled from a database of protein structures. While a lattice realization of the structures is used for computational convenience, in principle, the approach is completely general.

(3) The energy, consisting of local secondary preferences, a centrosymmetric burial term, and a term which reflects the orientation of the hydrophobic face with respect to the core, is calculated.

(4) The division process is repeated many times.

(5) At the end of the selection process, the statistics on the set of lowest energy structures are performed. The location of the predicted U-turns is established, and the dominant secondary structure in the three central residues between U-turns is used to assign the secondary structure of the entire block.

Application has been made to a set of test proteins, and part of the results are summarized in Table 1. At least for the testing set, the method is quite accurate, with over 90% of the U-turns and blocks correctly predicted. In six of the nine test sequences, the number of U-turns and the secondary structure of the blocks are correctly predicted. These encouraging results suggest that the blocks and U-turns algorithm holds considerable promise in providing important information for three-dimensional modeling procedures. When successful, it provides sufficient information to propose a relatively small number of low-resolution alternative folds. Furthermore, it can be used as a filter or constraint in inverse folding algorithms either to predict the global topology or, in a potentially more powerful application, to predict the conformation of hairpin fragments. At present, when inverse folding algorithms are used to predict the structure of 15–20-residue pieces of a protein, mixed in with the correct, low-energy structures are a variety of comparable energy false positives. The blocks and U-turns algorithm can be used to filter out such false positives. Preliminary application of this combined approach has yielded promising results.

Folding with a small number of long-range restraints

A number of investigators have examined the problem of determining a low to moderate resolution protein structure given a relatively small number of distance restraints and some knowledge of the secondary structure [26]. The ability to predict such structures would aid in the early stages of NMR structural refinement when secondary structure information and a limited number of distance restraints are known. In contrast, when a large number of restraints are available, then the use of distance geometry or distance geometry supplemented by molecular dynamics are the methods of choice [104]. Here, we describe results when the lattice model of protein folding is supplemented by a rough knowledge of secondary structure and some tertiary constraint information. Such an investigation can also clarify whether the present realization of the model is basically correct, but is simply in need of further refinement, or more substantial problems with the model exist which would require its fundamental reformulation.

From random extended states, the folding of L7/L12 ribosomal protein, 1ctf, protein G, 1gb1, and thioredoxin, 2trx, all of which are α/β proteins, plastocyanin, 1pcy, which is an all-beta protein, and the helical protein, sperm whale myoglobin, 1mba, were undertaken [60]. After a simulated annealing run, the resulting final conformation is subject to isothermal refinement. At least five, and in many cases 20, independent folding/refinement runs were performed. For the sake of brevity, the simulation results for the run having the lowest average energy are presented in Table 2.

For the three α/β proteins considered, it is apparent that reasonable structures are obtained when there is on the order of one long-range constraint every seven residues. Similar results are found for helical proteins. Reflecting inherent problems in the model, this class of models requires a greater number of restraints for β-proteins. For the β-protein plastocyanin, one tertiary constraint every four residues is required.

These results should be compared to those of Smith-Brown et al. [105]. To obtain results of comparable accuracy to plastocyanin folded with 46 restraints, they required 90 restraints to fold a variable light domain of human immunoglobulin, 3Fab [60]. Similarly, Smith-Brown et al. require 147 constraints to obtain a structure that is 3.18 Å rms from the native conformation of flavodoxin [60]. In contrast, preliminary results on the folding of flavodoxin with 35 tertiary restraints indicate that structures on the level of 4 Å rms are obtained.

Aszodi and co-workers [106] have applied a distance geometry protocol where the secondary structure is known and where correct constraints are supplemented by predicted interresidue distances based on multiple sequence alignments. They refold thioredoxin with about 48 restraints to structures whose rms is about 5.0 Å. However, for a smaller number of restraints, the structures are almost random, having an rms from native on the order of 10 Å. In contrast, with just 15 restraints, structures on the

Table 2 *Results from NMR docking simulations with a limited number of tertiary restraints*

Protein	Number of residues	Number of tertiary restraints	Average coordinate rms[a]	Average distance rms[a]
1gb1	56	8	3.81	2.72
1ctf	68	8	4.27	3.13
2trx	108	15	6.60	4.77
1pcy	99	46	3.32	2.46
1pcy	99	25	6.22	3.93
1mba	146	20	5.52	3.85

[a] Root-mean-square deviation of the C^{α} coordinates in Å.

level of 6.6 Å rms are obtained here. Moreover, in the case of helical proteins such as 1mba, the docking algorithms can assemble the approximate topology. This suggests that the present lattice-based approach could be used to generate low to moderate resolution structures from a rather small number of restraints. However, the present realization needs improvement. Short-range restraints are incorporated as a very soft energetic bias to helix, turn or extended conformations, as appropriate. Tertiary restraints are also very loosely defined as operating on the level of side-chain contacts. Better restraints that actually include the information contained in 2D and 3D NMR experiments should be implemented. Similarly, disulfide bond restraints, which have a very specific geometry, could also be included. Thus, the results from this very simple realization of tertiary restraints could possibly be improved by better use of experimental data.

Folding of the GCN4 leucine zipper

Because of their sequential and structural simplicity and biological importance, coiled coils are natural test systems for protein folding and multimer assembly algorithms. The simplest realization of the coiled-coil motif consists of two α-helices wrapped around each other with a left-handed supertwist. Furthermore, coiled-coil sequences are characterized by a quasirepeating heptad of residues designated by the letters a–g, where positions a and d occur in the coiled-coil interface [107]. A particularly well characterized coiled coil is the leucine zipper of the transcriptional activator, GCN4 [108,109]. Each protein chain contains 33 residues, and it has a high-resolution crystal structure. To predict the GCN4 quaternary structure, Nilges and Brunger [110,111] assumed an initial conformation consisting of an idealized, parallel coiled coil. They were able to refine the structure from an initial 3.1 Å rms on the backbone atoms to a level of 1.26 Å rms for the backbone atoms and 1.75 Å for all heavy atoms in the dimerization interface. To accomplish this refinement, they used molecular dynamics supplemented by imposed helical backbone hydrogen bond restraints and a number of distance restraints.

Vieth and co-workers [21] have employed a hierarchical approach to fold the GCN4 leucine zipper from two chains that were initially in random conformations. No information about the global fold is assumed other than that there are two chains in a box. Thus, the possibility of higher order multimer formation was not considered there. First, a high-resolution lattice model is employed to assemble the topology. The lowest energy lattice structures have an rms from the C^{α} trace of the crystal structure ranging from 2.3 to 3.7 Å. Then, using these structures, detailed atomic models were built and relaxed using CHARMM all-atom building and MD-based simulated annealing in explicit water [112]. The average structure built from the entire family of five independently refined conformations has an rms deviation of 0.8 Å for the backbone atoms, 1.31 Å for the heavy atoms in the dimerization interface, and 2.29 Å for all heavy atoms. Figure 10 shows tube diagrams of the backbones of the five refined structures along with the crystal structure, which is shown in magenta. The

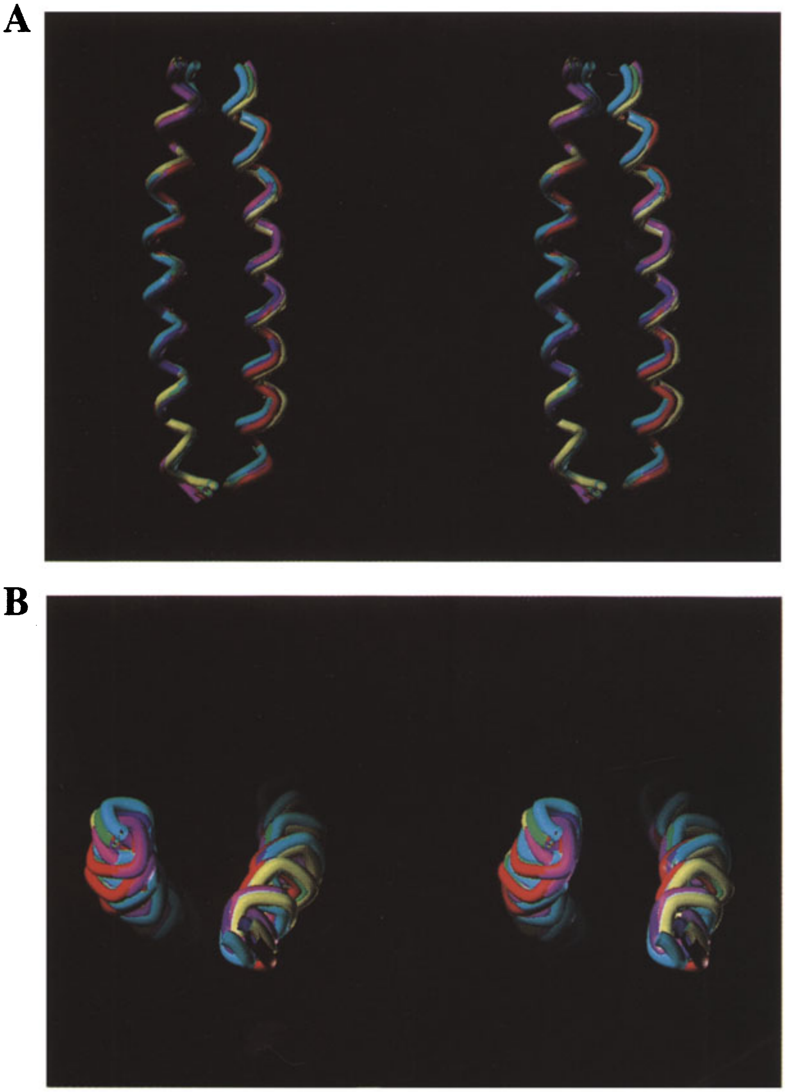


Fig. 10. Side and top views of the tube diagrams of five refined, predicted structures of the GCN4 leucine zipper, shown in red, yellow, green, cyan and white, along with the crystal structure, shown in magenta. Basically, there is one fused, six-color tube.

predicted positions of the side chains in the dimerization interface are essentially unique, but much greater variation is found in the positions of the surface residues.

These simulations also suggested a possible mechanism of the GCN4 leucine zipper coiled-coil assembly. Folding commences from the collision of two short helical stretches, generally located at the ends of the chain. These interacting helical stretches then propagate along the molecule. After small adjustments in registration by an inch worm type mechanism, the final, parallel in register coiled-coil dimer forms. Among the last regions to lock into place are the Asn^{16} residues which are located in the dimerization interface and which are regions of predicted low intrinsic stability. Although a detailed study has not yet been performed, there is some indication that assembly from the N-termini is preferred.

Method for predicting the state of association of proteins

A limitation of the above calculation is the assumption that the oligomerization state has no higher order dimers. However, coiled coils can associate to tetramers or even higher order aggregates [37,113,114]. Due to computer time limitations, the straightforward simulation of multimer equilibria is far beyond contemporary computer resources. Furthermore, if MMC is used, then the folding process must be repeated tens, if not hundreds, of times to be statistically significant. Thus, to predict the state of association, we developed a methodology that estimates the equilibrium constants among a spectrum of assumed parallel and antiparallel oligomers [22,115]. Subsequently, a more refined, lattice-based method was developed that also permits the monomeric state to be included [31]. Since the second method gives essentially the same results as the original technique in the regime where the two approaches overlap, we summarize the results from this more general approach.

In order to calculate the equilibrium constant [116], the internal partition function, Z_{int}, is required. For the denatured state, to estimate Z_{int}, we developed a transfer matrix treatment that includes all interactions within five-residue fragments [31,76]. The disadvantage of this approach is that it ignores longer range interactions. Dimers and higher order multimers are treated somewhat differently. First, we note that, in general, P(E), the probability of being in an energy level E, is related to Z_{int} by

$$Z_{\mathrm{int}} = N(E)\exp(-\beta E)/P(E) \tag{8}$$

Here, N(E) is the degeneracy of energy level E and $\beta = 1/kT$. Since E and P(E) are readily obtained from an MMC simulation, the remaining problem is to determine N(E). It may be estimated to within a constant by ESMC [33,34]; however, here we use a quasianalytic method. The basic idea is to focus on the most probable energy state. The Monte Carlo simulation provides the set of three consecutive C^{α} virtual bonds that are sampled by the ensemble of structures having the most probable energy, $\bar{E}$. $N(\bar{E})$ is obtained as the transfer matrix product of all such sets of three-bond vectors that are appropriately combined.

Initially, this approach was applied to calculate the monomer–dimer equilibrium in the GCN4 leucine zipper [108] and its fragments [117]. The results of the calculations and the comparison with experiment for the predicted dominant species are shown in Table 3. For the three cases for which experimental data are available, the prediction agrees with experiment. Examination of the stability of the GCN4 fragments indicates

Table 3 *Comparison with experiment of the predicted dominant species for the GCN4 leucine zipper*

Protein	Predicted dominant species	Experimental dominant species
GCN4 wild-type	2	2
GCN4 8–30	2	2
GCN4 11–33	1	1
GCN 4–26	2	Not yet measured

that the stability of a given fragment cannot be estimated from the stability of the parent molecule. A similar situation is obtained for other coiled coils such as tropomyosin. The origin of the lack of stability of the 11–33 fragment arises from the difficulty in burying Asn^{16} in a helical conformation in the core. In the 11–33 fragment, there would be a single helical turn at the N-terminus before the Asn. Since it is not stable enough to force the hydrophilic Asn to be helical, the entire fragment becomes disordered. This effect is due to loop entropy, which in coiled coils acts to prohibit random coiled conformations between interacting helical stretches [118]. In this and in all the systems studied, the simulations suggest that coiled coils are highly cooperative with many of the observed phenomena caused by nonadditive effects.

Next, we examined the stability of Fos and Jun coiled coils. In partial agreement with experiment, Fos without a GCG linker is predicted to be monomeric, whereas at high concentration both monomer and dimers are present [38]. In an equimolar mixture of cross-linked Fos and Jun homodimers, as in the experimental system, the simulation predicts that Fos heterodimers should preferentially form. The calculations suggest that the presence of Thr and Lys in the interfacial region of Fos homodimers gives rise to the relative instability of Fos homodimers. When an equimolar system of Fos and Jun are present, the system can lower its overall free energy by forming Fos–Jun heterodimers.

Coiled coils can also provide insights into the factors driving the formation of quaternary structure. In an elegant study, Harbury et al. [37] simultaneously replaced all four a and d residues of the GCN4 leucine zipper by Leu, Val, and Ile. In Table 4, the theory is compared with experiment. In five of eight cases, the simulations and experiment are in agreement over the entire concentration range, and, in another case, agreement is found over a portion of the experimental range. These calculations suggest that intrinsic secondary structural preferences and configurational entropy favor lower order species, while quaternary interactions favor higher order species. This conjectured origin of multimer stability is inconsistent with the suggestion of Harbury et al. [37] that the selection of a given species is due to the requirement that the lowest energy side-chain rotamer selects the particular interchain packing geometry. Such a level of detail is beyond the lattice models where the side chains are represented by soft core balls.

Table 4 *Comparison of the predicted and experimentally measured dominant species of GCN4 and seven mutants*

Residues at positions		Dominant species from experiment	Dominant species from simulation
a	d		
GCN4 wild-type		2	2
I	L	2	3
I	I	3	3
L	I	4	4
V	I	?	3
L	V	3	3
V	L	2,3	2
L	L	3	3

Weaknesses of the lattice models

While this chapter has presented a number of examples where folding of a protein from sequence alone has been achieved, the full solution of the protein folding problem is not in hand. There still remain problems with the potential. While the current generation can differentiate grossly incorrect folds from native, in many cases it is very difficult to differentiate topologies having substantial similarity to the native fold. These close topological cousins have essentially the same burial energy as native and differ by a relatively small number of side-chain contacts and differences in secondary structure. To some extent, this is a physical effect. Even if two topologies differ on average by 10kT in energy but have comparable configurational entropy, then the lower energy fold will be thermodynamically very favored [22,115]. However, the accurate portrayal of this difference is nontrivial. Furthermore, due to the representation of side chains as soft balls with a single interaction center and a relatively wide interaction basin, the high coordination lattice models overestimate the protein's entropy. Such an excess entropy also results in an increase in the backbone's flexibility. This is probably a major cause of the difficulty the models have with the folding of naturally occurring β-proteins. Part of this effect is inevitable in any reduced protein model. Clearly, improved side-chain representations are necessary. Another problem concerns adequate conformational sampling. Given that close topologies have overlapping energy spectra, a sufficient number of simulations must be done to ensure that the predicted low-energy structure is well characterized. At present, this is possible only for simple folds lacking reversals in chain direction. ESMC may be helpful in this regard, but such calculations can also be very expensive [25,33–35]. Thus, the model representation, potential and sampling protocols all require improvement.

Conclusions

In this chapter we have described the folding of protein A and seven homologous sequences, the putative retrosequence of protein A, two sequences designed by DeGrado, a putative ROP monomer, crambin, the V-3 loop of gp-120, and the GCN4 leucine zipper. For many of these sequences, structures which are in reasonable agreement with experiment have been predicted, and the remainder stand as predictions to be tested by experiment. Furthermore, assuming that the native state is located in a collection of parallel and antiparallel dimers, trimers and tetramers, the quaternary structure of GCN4, two of its fragments, five of eight wild-type mutants, Fos, Jun, and Fos–Jun heterodimers have been successfully predicted. In addition, the 4–26 fragment of GCN4 is predicted to be dimeric. The simulations argue that the native structure is a compromise among numerous contributions to the potential. The different terms such as hydrophobic interactions, hydrogen bonding and cooperative side-chain packing interactions give rise to different aspects of protein-like behavior. The results to date suggest that progress is being made in the *de novo* prediction of protein structure. Future advances are likely to result from better, more specific energy functions, better model realizations, and combined approaches such as the use of inverse folding to predict the structure of fragments followed by their assembly using reduced models such as have been discussed here. Overall, the prospect for future progress in the protein folding problem remains bright.

Acknowledgements

This work was supported by NIH Grant Nos. GM-37408, GM-38794, TW-00418, and by the University of Warsaw, BST-532-34/96. A.K. is an International Research Scholar of the Howard Hughes Medical Institute (Grant #75195-543402). The contributions of Drs. C.L. Brooks III, A. Godzik, M. Milik, M. Vieth, and K. Olszewski to the work described here are gratefully acknowledged.

References

1. Jernigan, R.L., Curr. Opin. Struct. Biol., 2(1992)248.
2. Anfinsen, C.B., Science, 181(1973)223.
3. Ripoll, D.R., Piela, L., Velasquez, M. and Scheraga, H.A., Proteins, 10(1991)188.
4. Skolnick, J. and Kolinski, A., Science, 250(1990)1121.
5. Go, N. and Taketomi, H., Proc. Natl. Acad. Sci. USA, 75(1978)559.
6. Go, N., Abe, H., Mizuno, H. and Taketomi, H., Protein Folding, Elsevier, Amsterdam, 1980, p. 167.
7. Kolinski, A., Skolnick, J. and Yaris, R., Biopolymers, 26(1987)937.
8. Skolnick, J. and Kolinski, A., Annu. Rev. Phys. Chem., 40(1989)207.
9. Godzik, A., Kolinski, A. and Skolnick, J., J. Comput.-Aided Mol. Design, 7(1993)397.
10. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. and Wolynes, P.G., Proteins, 21(1995)167.

11. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S., Protein Sci., 4(1995)561.
12. Sali, A., Shakhnovich, E. and Karplus, M., J. Mol. Biol., 235(1994)1614.
13. Sali, A., Shakhnovich, E. and Karplus, M., Nature, 369(1994)248.
14. Kolinski, A., Milik, M. and Skolnick, J., J. Chem. Phys., 94(1991)3978.
15. Shakhnovich, E.I. and Gutin, A.M., Proc. Natl. Acad. Sci. USA, 90(1993)7195.
16. Socci, N.D. and Onuchic, J.N., J. Chem. Phys., 100(1994)1519.
17. Kolinski, A. and Skolnick, J., J. Phys. Chem., 97(1992)9412.
18. Kolinski, A., Godzik, A. and Skolnick, J., J. Chem. Phys., 98(1993)7420.
19. Kolinski, A. and Skolnick, J., Proteins, 18(1994)338.
20. Kolinski, A. and Skolnick, J., Proteins, 18(1994)353.
21. Vieth, M., Kolinski, A., Brooks III, C.L. and Skolnick, J., J. Mol. Biol., 237(1994)361.
22. Vieth, M., Kolinski, A., Brooks III, C.L. and Skolnick, J., J. Mol. Biol., 251(1995)448.
23. Kolinski, A., Galazka, W. and Skolnick, J., J. Chem. Phys., 103(1995)10286.
24. Kolinski, A., Milik, M., Rycombel, J. and Skolnick, J., J. Chem. Phys., 103(1995)4312.
25. Kolinski, A., Galazka, W. and Skolnick, J., Proteins (1997) in press.
26. Skolnick, J., Kolinski, A. and Ortiz, A.R., J. Mol. Biol., 265(1997)217.
27. Kolinski, A., Skolnick, J., Godzik, A. and Hu, N.P., Proteins, 27(1997)290.
28. Milik, M., Kolinski, A. and Skolnick, J., J. Comput. Chem., 18(1997)80.
29. Olszewski, K.A., Kolinski, A. and Skolnick, J., Proteins, 25(1996)286.
30. Olszewski, K.A., Kolinski, A. and Skolnick, J., Protein Eng., 9(1996)5.
31. Vieth, M., Kolinski, A. and Skolnick, J., Biochemistry, 35(1996)955.
32. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., J. Chem. Phys., 51(1953)1087.
33. Hao, M.-H. and Scheraga, H.A., J. Phys. Chem., 98(1994)4940.
34. Hao, M.-H. and Scheraga, H.A., J. Phys. Chem., 98(1994)9882.
35. Hao, M.-H. and Scheraga, H.A., J. Chem. Phys., 102(1995)1334.
36. Skolnick, J., Kolinski, A., Brooks III, C.L., Godzik, A. and Rey, A., Curr. Biol., 3(1993)414.
37. Harbury, P.B., Zhang, T., Kim, P.S. and Alber, T., Science, 262(1993)1401.
38. O'Shea, E.K., Rutkowski, R., Stafford III, W.F. and Kim, P.S., Science, 245(1989)646.
39. Godzik, A., Kolinski, A. and Skolnick, J., J. Comput. Chem., 14(1993)1194.
40. Rey, A. and Skolnick, J., J. Comput. Chem., 13(1992)443.
41. Rey, A. and Skolnick, J., Proteins, 16(1993)8.
42. Srinivasan, R. and Rose, G.D., Proteins, 22(1995)81.
43. Dill, K.A. and Yue, K., Protein Sci., 5(1996)254.
44. Kolinski, A., Skolnick, J. and Yaris, R., J. Chem. Phys., 85(1986)3585.
45. Kolinski, A., Skolnick, J. and Yaris, R., Macromolecules, 20(1987)438.
46. Kolinski, A. and Skolnick, J., Proc. Natl. Acad. Sci. USA, 83(1986)7267.
47. Skolnick, J. and Kolinski, A., J. Mol. Biol., 212(1990)787.
48. Skolnick, J., Kolinski, A. and Yaris, R., Proc. Natl. Acad. Sci. USA, 86(1989)1229.
49. Skolnick, J., Kolinski, A. and Yaris, R., Biopolymers, 28(1989)1059.
50. Skolnick, J., Kolinski, A. and Yaris, R., Proc. Natl. Acad. Sci. USA, 85(1988)5057.
51. Kabsch, W. and Sander, C., Biopolymers, 22(1983)2577.
52. Levitt, M. and Greer, J., J. Mol. Biol., 114(1977)181.
53. Olszewski, K.A., Kolinski, A. and Skolnick, J., Protein Eng., 9(1996)5.
54. Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 219(1991)109.

55. Hunt, G.N., Gregoret, L.M. and Cohen, F.E., J. Mol. Biol., 241(1994)214.
56. Hao, M.-H., Rackovsky, S., Liwo, A., Pincus, M.R. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 89(1992)6614.
57. Nikishawa, K. and Ooi, T., Biochemistry, 100(1986)1043.
58. Miyazawa, S. and Jernigan, R.L., Macromolecules, 18(1985)534.
59. Godzik, A., Kolinski, A. and Skolnick, J., Protein Sci., 4(1995)2107.
60. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112(1977)535.
61. Skolnick, J., Jaroszewski, L., Kolinski, A. and Godzik, A., Protein Sci., 6(1997)676.
62. Milik, M., Kolinski, A. and Skolnick, J., Protein Eng., 8(1995)225.
63. Kuwajima, K., Proteins, 6(1989)87.
64. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E. and Razgulyaev, O.I., FEBS Lett., 262(1990)20.
65. Brooks, C.L., Curr. Opin. Struct. Biol., 3(1993)92.
66. Skolnick, J., Kolinski, A. and Godzik, A., Proc. Natl. Acad. Sci. USA, 90(1993)2099.
67. Eloffson, A. and Nilsson, L., J. Mol. Biol., 223(1993)766.
68. Barker, J.A. and Henderson, D., Rev. Mod. Phys., 48(1976)587.
69. Lee, J., Phys. Rev. Lett., 71(1993)211.
70. Berg, B.A. and Neuhaus, T., Phys. Rev. Lett., 68(1991)9.
71. Handel, T. and DeGrado, W.F., Biophys. J., 61(1992)A265.
72. Raleigh, D.P. and DeGrado, W.F., J. Am. Chem. Soc., 114(1992)10079.
73. Handel, T.M., Williams, S.A. and DeGrado, W.F., Science, 261(1993)879.
74. Woolfson, D.N. (1996) personal communication.
75. Ptitsyn, O.B., J. Protein Chem., 6(1987)273.
76. Skolnick, J. and Kolinski, A., J. Mol. Biol., 221(1991)499.
77. Kuwajima, K., Mitani, M. and Sugai, S., J. Mol. Biol., 206(1989)547.
78. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. and Schimada, I., Biochemistry, 40(1992)9665.
79. Wright, P.E. (1996) personal communication.
80. Nilsson, B., Moks, T., Jansson, B., Abrahamsen, L., Elmblad, A., Holmgren, E., Henrichson, C. and Jones, T.A., Protein Eng., 1(1987)107.
81. Lyons, B.A., Tashiro, M., Cedergren, L. and Montelione, G.T., Biochemistry, 32(1993)7839.
82. Bowie, J.U., Reidhaar, O.J.F., Lim, W.A. and Sauer, R.T., Science, 247(1990)1306.
83. Rose, G.D. and Wolfenden, R., Annu. Rev. Biophys. Biomol. Struct., 22(1993)381.
84. Dyson, J.H. and Wright, P.E., Curr. Biol., 3(1993)60.
85. Chou, K.C., Maggiora, G. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 89(1992)7315.
86. Behe, M.J., Lattman, E.E. and Rose, G.D., Proc. Natl. Acad. Sci. USA, 88(1991)4195.
87. Brunet, A.P., Huang, E.S., Huffine, M.E., Loeb, J.E., Weltman, R.J. and Hecht, M.H., Nature, 364(1993)355.
88. Predki, P.F. and Regan, L.R., Biochemistry, 34(1995)9834.
89. Wilmot, C.M. and Thornton, J.M., J. Mol. Biol., 203(1988)221.
90. Goodman, M. and Chorev, M., Acc. Chem. Res., 12(1979)1.
91. Guptasarma, P., FEBS Lett., 310(1992)205.
92. Chou, K.C., Proteins, 21(1995)319.
93. Presta, L.G. and Rose, G.D., Science, 240(1988)1632.
94. Richardson, J.S. and Richardson, D.C., Science, 240(1988)1648.

95. Wong, C.H. (1996) personal communication.
96. Banner, D.W., Kokkinidis, M. and Tsernoglou, D., J. Mol. Biol., 196(1987)657.
97. Sander, C. (1993) personal communication.
98. Eliezer, D., Jennings, P.A., Wright, P.E., Doniach, S., Hodgson, K.O. and Tsuruta, H., Science, 270(1995)487.
99. Hendrickson, W.A. and Teeter, M.M., Nature, 290(1981)107.
100. Holley, L.H. and Karplus, M., Proc. Natl. Acad. Sci. USA, 86(1989)152.
101. Garnier, J., Osguthorpe, D.J. and Robson, B., J. Mol. Biol., 120(1978)97.
102. LaRosa, G.J., Davide, J.P., Weinhold, K., Waterbury, J.A., Profy, A.T., Lewis, J.A., Langlois, A.J., Dreesman, G.R., Boswell, R.N., Shadduck, P., Holley, L.M., Karplus, M., Bolognesi, D.P., Matthews, T.J., Emini, E.A. and Putney, S.D., Science, 249(1990)932.
103. Rost, B. and Sander, C., J. Mol. Biol., 232(1993)584.
104. Havel, T.F., Prog. Biophys. Mol. Biol., 56(1991)43.
105. Smith-Brown, M.J., Kominos, D. and Levy, R.M., Protein Eng., 6(1993)605.
106. Aszodi, A., Gradwell, M.J. and Taylor, W.R., J. Mol. Biol., 248(1995)308.
107. McLachlan, A.D. and Stewart, M., J. Mol. Biol., 98(1975)298.
108. O'Shea, E.K., Klemm, J.D., Kim, P.S. and Alber, T., Science, 254(1991)539.
109. Ellenberger, T.E., Brandl, C.J., Struhl, K. and Harrison, S.C., Cell, 71(1992)38.
110. Nilges, M. and Brunger, A.T., Proteins, 15(1993)133.
111. Nilges, M. and Brunger, A.T., Protein Eng., 4(1991)649.
112. Brooks, B.R., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
113. Harbury, P.B., Kim, P.S. and Alber, T., Nature, 371(1994)80.
114. Chmielewski, J., J. Am. Chem. Soc., 116(1994)6451.
115. Vieth, M., Kolinski, A. and Skolnick, J., J. Chem. Phys., 102(1995)6189.
116. McQuarrie, A.D., Statistical Mechanics, Harper & Row, New York, NY, 1976.
117. Lumb, K.J., Carr, C.M. and Kim, P.S., Biochemistry, 33(1994)7361.
118. Skolnick, J., Macromolecules, 17(1984)645.

Part VI
Structure-based design

Computational tools for structure-based design

Stuart M. Green and A. Peter Johnson
Institute for Computer Applications in Molecular Sciences, School of Chemistry,
University of Leeds, Leeds LS2 9JT, U.K.

1. Introduction

The drug discovery and design process is an awkward mixture of chance and skill; hence, any technique that improves the odds or simplifies the procedure is rapidly exploited and developed by the medicinal chemistry community. A range of computational techniques including molecular modelling have been widely applied, resulting in structure-based drug design. This design process is reliant on structural information on either the receptor of the biological target or from a series of ligands known to bind at the receptor. It has been highly successful and leads to the generation of highly potent inhibitors for a range of biological targets including HIV-1 protease [1–11], thymidylate synthase [12–19], carbonic anhydrase [6,20,21], matrix metalloproteinase [22–24], FKBP12 [25–28], thrombin [29–36] and influenza-virus sialidase [37,38].

Once our biological target has been selected, the design strategy can be divided into three stages:

1. Analysis of the structural information of the receptor (derived via X-ray crystallography, NMR or homology) or the pharmacophore (derived from a series of ligands known to bind at the specified receptor). This yields our active site model.
2. Satisfaction of some or all of the binding requirements of the active site by placing appropriate chemical functionality (hydrogen bond donors, hydrogen bond acceptors, hydrophobic groups, charged species) in the required locations and constructing a molecular scaffold to hold them in place.
3. Sorting and selecting the designed molecules by estimation of their chemical and biological properties.

This document will review the various computational procedures developed to address these design stages.

2. Generating the active site model

2.1. Known receptor structure

The advances in molecular biology, protein crystallography and NMR have led to the solution of protein structures at an ever-increasing rate (> 4600 protein structures in the Brookhaven Protein Databank (PDB) [39] at the time of writing). If we are fortunate enough to capture a ligand binding to our receptor at the time of structural analysis, we can easily identify the active site and the interacting residues. Otherwise, we can use experimental data (site-directed mutagenesis, spectroscopic analysis, evolutionary considerations) to track down the active site. If neither piece of information is available, simple visual inspection via molecular graphics may focus our attention to a suitable pocket, cleft, or depression on the protein surface.

2.1.1. Locating the active site

Peters et al. [40] have recently described an automated procedure based purely on computational geometry. The method generates two molecular envelopes: one preserving the global shape of the protein, and the other at a higher resolution describing the molecular detail. By analysing the differences between these two envelopes, they were able to cluster sets of neighbouring atoms to various pockets on the protein surface. They claim a 95% success rate for locating binding sites in 75 different families of proteins. Bayada and Johnson [41] have developed a method (CANGAROO, part of the SPROUT system) based on the local curvature of the solvent-accessible surface to locate large inward facing regions (clefts) as potential binding sites.

2.1.2. Characterising the binding requirements

Once we have located our potential active site, we can now analyse it further to deduce the particular types of interactions (hydrogen bond donation and acceptance, hydrophobics and charge/charge) required for successful binding. Goodford [42] has developed a force field to explore these nonbonded interactions – GRID uses probe atoms of varying functionality (amino, carbonyl oxygen, carboxyl oxygen, hydroxyl, methyl, water) which are placed at positions on a regular lattice within the active site. The interaction energy between the probe and the protein residues is then computed. Contouring and visualisation of this potential yields areas of favoured/disfavoured probe interaction. The force field has been extended from a simple Lennard-Jones plus Coulombic potential to include a hydrogen bonding term [43] and to account for the probe forming multiple hydrogen bonds [44,45]. Tomioka and Itai [46] also advocate the use of a grid-based potential in their GREEN program package. Precalculated data from a series of molecular probes are used to represent the binding-site environment and facilitate the realtime evaluation of the protein–ligand interaction energy to permit interactive docking. A similar grid has also been used by Gehlhaar et al. with MCDNLG [47] and by Rotstein and Murcko in their GenStar [48] and GroupBuild [49] methods, respectively.

Miranker and Karplus's [50] procedure (MCSS) locates the interaction sites by flooding the active site with many thousands of copies of randomly orientated functional groups (acetate, methanol, methane, methyl ammonium and water) which are subjected to simultaneous energy minimisation and/or molecular dynamics (MD). The local minima are further explored, to account for the flexibility of the receptor, via grid-searching or constrained minimisation. A flood-fill of atoms followed by MD or Monte Carlo has been applied by Pearlman and Murcko [51] (CONCEPTS) and Bohacek and McMartin [52], respectively. Recently, Pearlman and Murcko [53] have discarded the solely atom-based approach in favour of a mixture of single atoms with molecular fragments.

Alternative to the potential-based methods are those which are based on a set of rules. Danziger and Dean's [54] HSITE program generates a set of hydrogen bond donor/acceptor points based on rules derived from published hydrogen bond surveys. Böhm's [55,56] LUDI program also contains such rules to generate sets of vectors around a donor or acceptor atom of an interacting residue. The orientation of the vectors reflects the directional nature of the potential hydrogen bonds at his site. Hydrophobic aromatic and aliphatic interaction sites are also located, but represented as single points. Clark et al.'s [57] PRO_LIGAND approach also uses such vector-based interaction sites. The HIPPO module of the SPROUT program of Gillet et al. [58] is also rule-based, but with an improved representation of the interaction sites. Rather than the simple point or vector, hemispherical volumes in which the appropriate atoms can be placed are used (Fig. 1). These permit a wider range of bonding situations as multicentred and/or bifurcated sites can be identified at regions of intersection. Covalently bonding and metal ion target sites, along with hydrogen bond donors/acceptors, are also identified by HIPPO.

The simple rules used in the above approaches are derived from statistical analyses of crystallographic databases. These data have been applied more directly by Böhm [55,56], who derived possible interaction sites from distributions of nonbonded contacts contained in the Cambridge Structural Database (CSD) [59]. Klebe [60] has also used the CSD to construct composite crystal-field environments for many different functional groups. In the X-SITE method of Laskowski et al. [61], favourable interaction regions are derived from an analysis of the spatial distributions of atomic contact preferences of a data set of 83 nonhomologous high-resolution protein structures taken from the PDB [39]. Contact preferences for a variety of atom types are obtained relative to a coordinate reference frame defined by a triplet of bonded atoms. Given our active site, the bonding triplets are found and the appropriate distribution is transformed to this location. By combining each distribution for every triplet, favourable interaction regions can be identified.

2.2. Pharmacophores

If we have no accurate structural information regarding our receptor, we can attempt to create a picture (the pharmacophore) of the receptor binding requirements from an analysis of molecules which are known to bind. In attempting such an analysis, we are confronted with two major problems:

1. What are the bound conformations of our possibly highly flexible molecules?
2. How do we overlay the molecules in the series?

Once these are solved, we can then explore the aligned molecular family to generate the target interaction sites from the pharmacophoric elements common to all molecules.

2.2.1 Alignment

Marshall's active analog approach (AAA) [62–64] is a conformational searching protocol for generating the pharmacophoric alignment. Pharmacophoric elements

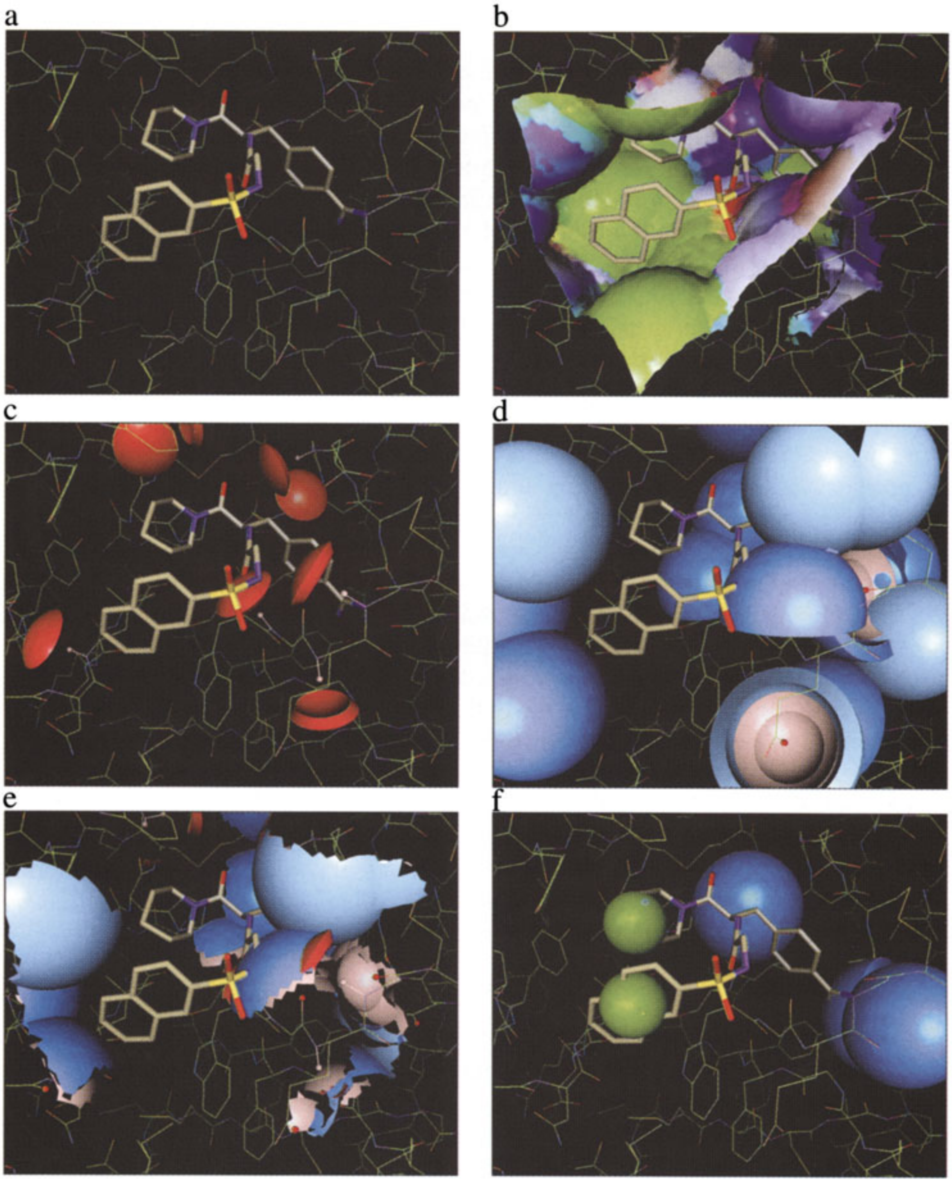


Fig. 1. Target sites in SPROUT: (a) thrombin and inhibitor NAPAP (1ets.pdb); (b) boundary surface of the receptor colour coded by hydrophobicity (green), hydrogen bond acceptance (blue) and donation by the receptor (red), and mixed classification (white); (c) acceptor atom target sites (red); (d) donor atom target sites, white hemispheres are for hydrogens and blue are for parent atoms; (e) acceptor and donor sites which protrude beyond the boundary surface; and (f) a selection of sites chosen as the design query, three donor sites (two of these complement with the amidine function of NAPAP) and two spherical sites (green) located in the hydrophobic regions.

(hydrogen bond acceptors/donors, metal binding functionality, etc.) common to every molecule are first selected and a conformational search is carried out on the least flexible molecule. For every valid conformation, the distance between each of the pharmacophoric elements is stored. These distances are then used to constrain the conformational space searched of the next molecule in the series. Distance sets that produce valid conformations are retained and the process is repeated for every molecule. Hopefully, only a small number of distance sets will remain corresponding to the bound conformers of all the molecules.

Alternatively, a variant of distance geometry can be applied to derive sets of conformations from a defined set of pharmacophoric elements [65]. *Ensemble distance geometry* is essentially the same as standard distance geometry, but with all the molecules under consideration at once. This is achieved by using much larger distance matrices to capture all atoms in all molecules. The bounds matrices are then defined thus: for atoms in the same molecule, the upper and lower bounds are defined in the usual way; the lower bounds for atoms in different molecules are set to zero to permit molecular overlay; the upper bounds for atoms in different molecules are set to a large value except for those atoms which define a pharmacophoric point, then they are given a small tolerance value to force superposition. Consensus molecular dynamics has also been used to elucidate pharmacophores by applying distance constraints between pharmacophoric points of different molecules [66].

The need for preselection of pharmacophoric elements in the above approaches is a great handicap, especially if we do not know which ones are responsible for binding. One solution offered by Martin et al. is the DISCO method [3,67]. All potential pharmacophoric elements are located in every molecule and pairs of distance between elements are matched using clique detection. Flexibility is accounted for to some degree by using a representative class of low-energy conformations of each molecule. Jones et al. [68] have explored the use of a genetic algorithm to encode both the mapping between elements in pairs of molecules and to drive the torsional angles of the flexible bonds.

Molecular features other than the pharmacophoric elements can also be used. Dean's group has demonstrated the use of simulated annealing and cluster analysis to align flexible molecules via their atomic coordinates [69] and molecular surfaces [70].

3. Obtaining structures

Once the model of the active site has been generated, molecules that match it can be found either by scanning a database for ones that fit or suitable structures can be generated *de novo.*

3.1. Database searching

The techniques of three-dimensional database searching using queries based on sets of interatomic distances and atom typing are well established and described in earlier literature [71,72]. Recent developments in the field have included the ability to explore fully the molecular flexibility. Storing a small set of conformations for each structure does not address all of conformational space adequately. Hurst's [73] directed tweak technique quickly drives, via minimisation, the torsions of flexible molecules in the database toward the query. Clark et al. [74] also reported a similar tweaking technique of comparable speed based on genetic algorithms.

Other techniques have been developed which focus on the directionality of bonds rather than atomic location. Bartlett et al.'s [75,76] CAVEAT method attempts to find molecular frameworks and was designed to retrieve, from a database, molecules with specific bonds which match a·bond vector defined in the query. In Ho and Marshall's [77] searching algorithm, FOUNDATION, queries may contain bond direction information, atom type designation, volume specifications, etc. As this method uses clique detection, partial solutions can also be obtained. In the CLIX algorithm of Lawrence and Davis [78], GRID [42] is first used to locate the target sites for a variety of chemical probes. Molecules from the CSD are then exhaustively docked in an attempt to find a coincidence between pairs of functional groups of the molecule and interaction sites in the protein.

3.2. De novo structure generation

Structure generation methods can be broadly divided into two categories: those which follow a deterministic course and those which are stochastic in nature.

3.2.1. Deterministic methods

Moon and Howe's [79] GROW was the first published method for structure generation. Initially, it was developed for the design of peptides using amino acid building blocks. Starting with an acetyl group as seed (placed within the active site via docking or from a known ligand), the structures are built up stepwise via the amide bond. The fragment library consists of natural and unnatural amino acids in many hundreds of their low-energy conformations. Structures are scored at each step of the build-up process (i.e. monopeptide, dipeptide, tripeptide, etc.) via an energy evaluation using a modified form of the AMBER potential with softened van der Waals penetrations. A number (default 10) of the highest scoring structures are retained and used in the next step. Build-up is stopped once the desired length of peptide is obtained.

The LEGEND program [80,81] grows molecules by adding atoms one at a time up to a given molecular size under the influence of the MM2 force field and an electrostatic potential grid. A seed atom is first generated (either automatically or user-defined) at a point where it is capable of forming a hydrogen bond with the receptor. A suitable atom type is then assigned to this point. All subsequent atoms are all assigned at random accordingly:

1. Choose root atom.
2. Select atom and bond type.
3. Generate new atom at appropriate distance from root.
4. Reject if intermolecular energy unfavourable.
5. Mutate carbon to heteroatom if in region of high electrostatic potential.

A range of small fragments (e.g. carbonyl groups, aromatic rings and amides) can also be used. Once a predefined atom limit has been reached, atom growing is terminated and the structure is completed by the addition of extra carbons to make aromatic rings where possible. Hydrogens are added finally and the structure is optimised. Structure generation is repeated until a user-defined number of molecules has been reached.

An early contribution of Rotstein and Murcko [48] describes GenStar, another atomistic structure generation method. Rather than a range of atom types, GenStar uses only sp^3 carbons to propose molecules. After calculating an interaction grid to represent the active site (atoms are not restricted to grid points, the grid is used as a fast means of representing the local environment), seed points are generated proximal to certain enzyme atoms. From each seed, a spherical shell of points is generated with minimum and maximum radii of 3 Å and 5 Å, respectively. This shell has $\approx$ 400 possible points for locating the first atom of the *de novo* molecule. Points are retained at each step based on a scoring procedure detailed in the next paragraph. The second set of possible atom positions is created in a similar manner to the first,but with the shell radii set to the upper and lower bond length bounds. Bond length and angle constraints now apply to the third and a torus-like region of points defines those. All additionally generated atoms are restricted by torsion angle as well; hence, the region of points constructed is a form of notched torus.

After each point generation phase, the majority are eliminated based on intermolecular contact with the enzyme, boundary violation or intramolecular bumping (exceptions are made for ring closure). The remaining points are then scored via their eight neighbouring grid points which describe proximal enzyme atoms. The top 20% of scoring points are found and a final single point, chosen from these at random, is retained (randomness is introduced to overcome the crudity of scoring and to maintain diversity). Occasionally, good scores will arise for points in two different conformational regions. If so, branching may occur from the parent point. Structure generation returns to its parent point if a dead-end is found and branches off if possible. Finally, using the electrostatic potential as a guide, the carbon framework is modified by placing heteroatoms where the opportunity exists for favourable electrostatic interaction, i.e. where there is a potential hydrogen bond.

GroupBuild was the next development of Rotstein and Murcko [49]. This fragment-based approach starts from a *core* fragment, either user-defined, automatically suggested or derived from a portion of a known inhibitor. Each of the hydrogens of the core is replaced in turn with a randomly selected fragment from a predefined library (a variety of atom types are now permitted). If the new fragment is chemically acceptable (defined by a set of bonding rules), the torsion of the new bond is varied through 10 increments and scored. Scoring is similar to GenStar using a predefined grid, but solvation effects can also be included. Once all rotamers have been scored, one from the best 25% is randomly selected, the structure is briefly minimised and generation continues with the enlarged core until the termination conditions are met (e.g. molecular weight, number of atoms, number of fragments, active site full).

In the LUDI package of Böhm [55,56] the interaction sites are defined as mentioned earlier or via GRID located regions. Fragments are taken from a library (600 entries, 5–30 atoms) and fitted in these sites via root-mean-square (rms) superposition. A second library of bridging fragments (1100 entries), with links explicitly defined, is then used to link the fitted fragments in a single step. Only those fragments with a tolerable rms deviation of fit to the interaction sites and those which do not bump into the protein are retained. An electrostatic repulsion check is also made: if heteroatoms of the same polarity are present in both the protein and the fragment, within a certain distance of each other, the fragment is rejected. Retained structures are ranked according to a score based on the number and quality of hydrogen bonds, and the hydrophobic contact surface area, between ligand and receptor.

Using a vector-based representation of the interaction sites, PRO_LIGAND [57] has four phases of structure generation, each with their own fragment library. The first, *placement* stage tries to place a fragment that satisfies an interaction site. Next, the *place-joining* phase attempts to attach a new fragment to both a target site and also one already located at another target site. In a third phase, *place-bridging*, fragments which satisfy a target interaction and can be connected to two previously placed fragments are fitted. Finally, *bridging* can take place where fragments are added simply to join other fragments together. At each stage, fragments are selected randomly from the appropriate database with a random conformation (if flexible) and a random set of fragment linking points (if multiple choices exist) is selected. If the fragment fits, it is docked and clashes are resolved if possible via bond formation. If the fragment forms an unacceptable bond pair (e.g. O-O, N-N) or clashes with the receptor, the fragment is rejected. As new bond growth may have modified the geometry somewhat, the newly joined fragment is then refitted to its interaction sites (rejection for failure). Structure generation may then be repeated to build up a ligand in a depth-first strategy until various criteria are met (i.e. number of interaction sites satisfied, number of atoms in ligand).

Tschinke and Cohen's [82] NEWLEAD program also attempts to connect interaction site vectors. These sites must already have the appropriate functionality placed by the user; NEWLEAD will then attempt to connect pairs of sites (starting with the closest) using spacers in a manner similar to CAVEAT [76]. If suitable spacers cannot be found, single atoms are added or a ring-fusing operation is carried out to extend the fragments. The process is repeated until a single molecule is created from the fragments and spacers. Fragments are rejected on the basis of van der Waals collision.

To exploit the results of a FOUNDATION search [77], Ho and Marshall developed the SPLICE program [83]. This assembles and prunes partial queries, which match different pharmacophoric elements and have overlapping bonds, into novel ligands. HOOK, a procedure developed by Eisen et al. [84], finds molecular frameworks, defined in a database, which can form sensible bonds to the functional fragments that result from the application of MCSS.

SPROUT [58,85–87] is tailored to use the interaction volumes generated as a result of a HIPPO analysis on the active site [58]. Initially, a set of starting templates is docked into the user-selected target volumes (Fig. 2). These are then connected

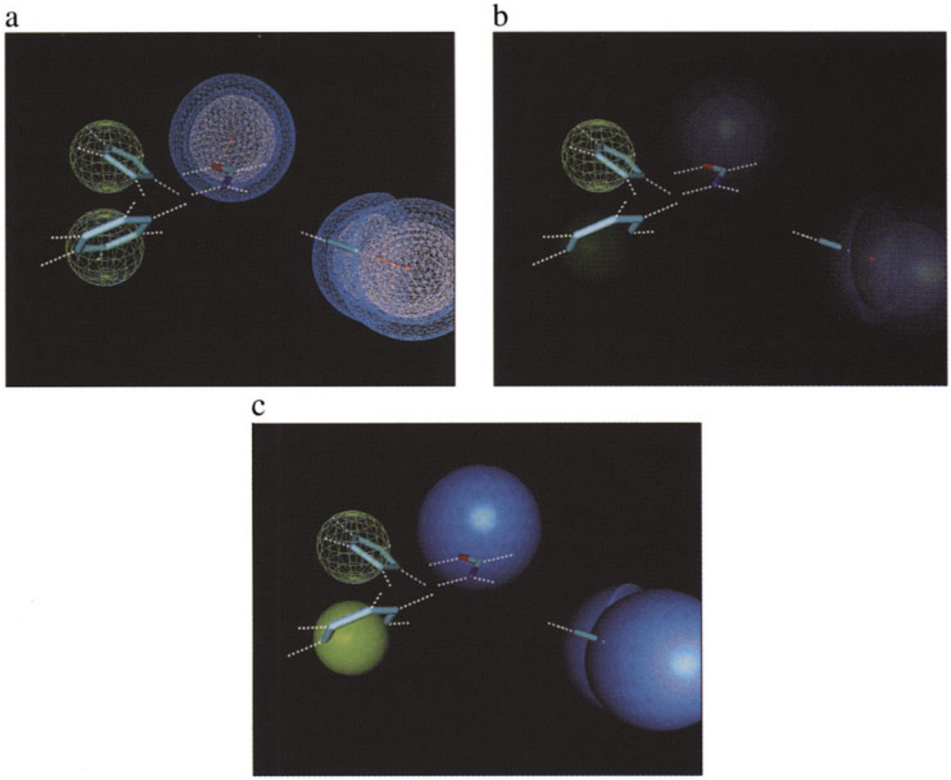


Fig. 2. Fragments docked into SPROUT target sites.

together via one or more spacer templates defined in a library. The templates used represent generalised atoms and chemical bonds, i.e. hybridisation is defined (as this determines the orientation of any new bonds formed to the template) but not atom type. The range of substructures used as templates include acyclic fragments with one to four atoms and three- to seven-membered rings. Commonly occurring conformations of each substructure are also described.

In SPROUT, a highly efficient bidirectional growth procedure is used to connect pairs of target sites. Once partial molecular skeletons that are grown inwards from each of the starting templates are sufficiently large, they are checked for overlapping templates common to both. The partial skeletons are then merged and redocked into the target sites. The skeleton can then be further enlarged via bidirectional growth toward the other starting templates located in their respective interaction volumes. Once a collection of satisfactory molecular skeletons is generated, the generalised atoms are mutated via a rule-based atom substitution procedure in an attempt to complement the electrostatic characteristics of the active site. Such an atom assignment approach has also been extensively explored by Dean and co-workers [88–93], where they use simulated annealing to optimise the atom types of the ligand to parallel the electrostatic, hydrophobic and hydrogen bonding requirements of the active site.

Starting with a pair of fragments in the active site that are to be joined, BUILDER [94,95] finds paths through a previously generated random lattice of atoms to connect the two sites. The elemental type is ignored at this stage with only hybridisation considered. A variety of chemical rules are applied to first exclude undesirable combinations of hybridisation and, secondly, to change the generic atoms of the path to a specific type. The SHAKE [96] algorithm is then applied to adjust bond lengths and angles. Finally, extra atoms are added to form rings along the atomic path or to create extra functionality, i.e. oxygens to carbonyl carbons. The process is then repeated to connect the new linker fragment with others.

3.2.2. Stochastic methods

Pearlman and Murcko's [51] CONCEPTS method is a dynamic algorithm for drug suggestion. A location is picked within the active site and a spherical region centred about this point is uniformly filled with 'particles'. Each particle is randomly offset from its original point and randomly assigned an atom type describing its bonding and electronic characteristics. Holding the protein fixed, the particles are subjected to MD equilibration. The energy function used (a modified form of the AMBER potential) in this simulation includes terms for particle–protein, particle– particle nonbonded and bonded particle interaction. A particle is chosen at random and its type changed according to a predefined probability schedule. Any previously defined connections to this particle are destroyed and new ones are chosen to neighbouring particles based on a probability function dependent on their type, distance and angle. After a number of steps of particle mutation and connection, all unfilled valences are satisfied by selecting unsatisfied particles at random and reapplying the connection scheme but without destroying the current bonds. The system is then relaxed via MD with a penalty function to bias the acceptance procedure away from isolated particles with unfilled valences. A Metropolis Monte Carlo (MC) procedure is then used to assess whether the particular particle mutation is accepted. The process from the mutation step is repeated until a specified number of changes have been made or many consecutive rejections occur.

The ideas established in CONCEPTS have been further developed by Pearlman and Murcko into the CONCERTS [53] method. Rather than single atom 'particles', molecular fragments are used. The fragments are randomly oriented within the active site and subjected to consensus minimisation (the fragments are unaware of each other as in the MCSS procedure [50]). Consensus MD is applied and, at given intervals, attempts are made to join fragments that are suitably oriented along specific bonds to hydrogen. If a connection is made, a new macrofragment is formed and the simulation continues. This connection is not static and may be broken and reformed as the molecules evolve during the course of the simulation.

Bohacek and McMartin's [52] approach starts with a precomputed grid map that classifies the binding zones within the active site. A root atom is chosen and new atoms and fragments are grown from this. Growth points are determined and one is randomly selected to form a connection with another randomly chosen atom or functional group (sp^3 C, O, N, O^-, H, CO, NH, benzene and five-membered unsaturated rings). The new atom or group is placed and a complementarity score is evaluated from the binding zone classification. Subject to van der Waals clashes with the protein, an MC sampling criterion is used to decide whether this atom or group is to be accepted using the complementarity score for the whole molecule as an energy term. If accepted, new growth points are determined and structure generation is repeated from any one of the potential growth points of the molecule. If a distance and bond angle of the growing structure are appropriate, the MC procedure can also spontaneously perform ring closure.

As an alternative to the HOOK method [84] of linking MCSS [50] fragments into larger molecules, Miranker and Karplus [97] describe an automated method for dynamic ligand design, DLD. The active site containing the MCSS fragments is saturated with sp^3 carbons. These form bonds with the fragments and each other under the influence of a generalised potential function. This potential is designed such that bonded systems with the correct geometry have lower energies than discrete species. It has a continuous first derivative permitting optimisation and sampling via a range of techniques. The use of MC sampling and optimisation via simulated annealing is demonstrated.

Gehlhaar et al. [47] propose a Monte Carlo *de novo* ligand generator, MCDNLG. A random collection of atoms is densely packed into the active site. Bonding between atoms occurs when any two atoms are closer than 2.1 Å. This leads to atoms which far exceed their traditional valence. To evolve a ligand from this supermolecule, MC sampling coupled with simulated annealing is applied. At each step a random modification is picked from either change atom occupancy (atoms can be made to disappear and reappear, when disappeared all bonds are lost and take no part in any intramolecular energy evaluation), change atom position, change atom type, change bond type, translate fragment, rotate fragment, rotate a torsion. This modification is applied to a randomly selected atom, bond or fragment, respectively. The energy is evaluated intra- and intermolecularly via a force field. Also added are some heuristic energy terms which encourage the generation of chemically sensible structures. This energy is then used by the MC technique to decide whether the modification is accepted. A gradual temperature decrease, with a steep temperature burst at step 200 000, was used over the 300 000 steps of the annealing protocol to yield low-energy structures.

4. Sorting and selecting

Many of the structure generation programs have the ability to generate many thousands of different molecules that 'fit' the model of the active site. This arises from the combinatorial explosion due to the array of choices offered at each stage of the structure generation process, i.e. choice of building block (fragment and/or atom) to connect from and to, bond type, conformation of new bond and fragment. To reduce the search space down to a more manageable size, it is tempting to prune on the basis of an estimate of the binding energy of a partially grown ligand. However, such an approach is dangerous due to the crudity of such functions and as it may exclude ligands which may bind much better once other structural features are added.

Reducing the number of building blocks is another option – using fragments rather than atoms reduces the number of joining operations needed to build the scaffold used to hold the binding functionality. Also, such fragments tend to be chemically sensible; hence, the structures made from them tend to be chemically reasonable. SPROUT [58] and Builder [95] exploit the molecular skeleton approach favoured by Dean and co-workers [88–93] where element types are assigned only after the complete structure has been generated. Similarly, GenStar [48] restricts itself to single carbon atoms during structure generation.

Even after we have attempted to prune the combinatorial tree, it is still likely that hundreds of structures will remain, still far too many to synthesise. A variety of postprocessing procedures can be applied. Molecules can be simply ranked according to molecular weight, number of atoms, number of hydrogen bonds/lipophilic contacts made with the active site, number of rings, log P, molecular volume, rotatable bonds, etc. SPROUT [98] and PRO_LIGAND [99] both provide clustering tools based on 2D fragment descriptors.

Other means of prioritising ligands involve formulating a score based on a number of features. In PRO_LIGAND [57] a structure's score is evaluated as follows:

$$\begin{aligned}\text{score} = {} & N_{\text{acceptor}} + N_{\text{donor}} + 0.25\,N_{\text{aliphatic}} + 0.25\,N_{\text{aromatic}} \\ & - 0.1\,N_{\text{rotatable}} - 0.1\,N_{\text{asymmetric}} - 2.0\,N_{\text{fragments}}\end{aligned}$$

where N is the number of hydrogen bond acceptors and donors satisfied, aliphatic and aromatic lipophilic sites exploited, rotatable bonds, asymmetric carbons and unconnected fragments, respectively. The weights quoted are the default values and may be altered to suit the given drug design problem.

We can also attempt to predict the binding affinity via simulation (Ajay and Murcko [100] have recently reviewed this field). Free energy perturbation (FEP) methods have shown great promise, but their application is problematic – we need a starting molecule with measured binding affinity and its structure may only be perturbed by a small amount (i.e. a side chain). This cuts against the grain of our drug design goal as we wish to create novel, patentable molecular entities. FEP methods also warrant a high computational burden – something you do not want to attempt on hundreds of structures.

If there are enough data on other compounds which bind to our target receptor, we can try and form a quantitative structure–activity relationship (QSAR) correlating molecule features to the binding affinity. A currently popular technique is 3D QSAR

[101] using the CoMFA method [102], but this is reliant on the alignment of all structures in the molecular series.

Bohacek and McMartin [52] have estimated the potency of some designed thermolysin inhibitors. They formulated a QSAR from nine known inhibitors based on the number of hydrophobic contacts and the number of hydrogen bonds they made with the receptor:

$$\log K_i = 3.16 - 0.42\,N_{\mathrm{hydrophobic}} - 0.39\,N_{\mathrm{hbond}}$$

The estimation of potency was found to be greatly improved upon minimisation of the generated structures. Rather than derive a relationship for a specific protein and its ligands, Böhm [103] has developed a scoring function that takes into account various interactions and is derived from a linear regression analysis of 45 protein–ligand complexes containing 24 different protein families:

$$\Delta G_{\mathrm{bind}} = 5.4 - 4.7 \sum_{\mathrm{hbonds}} f(r,\theta) - 8.3 \sum_{\mathrm{ionic}} g(r,\theta) - 0.17\,A_{\mathrm{lipophilic}} + 1.4\,N_{\mathrm{rotatable}}$$

where ΔG_{bind}, the free energy of binding (in kJ mol^{-1}), is a weighted sum of the number of hydrogen bonds between ligand and receptor (N_{hbond}) (where the function f( ) accounts for hydrogen bonding geometry), the number of ionic interactions (N_{ionic}) (with function g( ) accounting for geometry), the lipophilic contact surface area ($A_{\mathrm{lipophilic}}$) and the number of rotatable bonds ($N_{\mathrm{rotatable}}$). A similar approach used in SPROUT has resulted in the following equation [104]:

$$\log K_i = -0.096 \sum_{\mathrm{hbonds}} f(r,\theta) - 0.015\,A_{\mathrm{phob\text{-}phob}} + 0.008\,A_{\mathrm{phob\text{-}phil}} - 0.002\,E_{\mathrm{vdW}} + 0.136\,N_{\mathrm{rotatable}}$$

where K_i is the binding affinity, an alternative weighted sum describes the effect of hydrogen bonding, $A_{\mathrm{phob\text{-}phob}}$ is the contact surface area between hydrophobic regions of both the protein and ligand, $A_{\mathrm{phob\text{-}phil}}$ is the contact surface area between hydrophobic regions of the protein and hydrophilic regions of the ligand, and *vice versa*; E_{vdW} is a term to assess van der Waals interactions and $N_{\mathrm{rotatable}}$ is the number of rotatable bonds. Despite the absence of an electrostatic term in the current version, the predictivity is at least as good as the Böhm function.

VALIDATE [105] is a hybrid strategy where the ligand interaction energy (steric, electrostatic and induction) is computed via molecular mechanics. These energy terms were combined with a range of descriptors (rotatable bonds, calculated log P, steric fit, measures of complementary/uncomplementary contact between lipophilic/hydrophilic surfaces), and via partial least squares or neural network analysis, a relationship was obtained for a diverse training set of 51 ligand–receptor complexes. The predictivity of the model was established with three different test sets: 14 complexes of protein classes not in the training set, 13 HIV protease inhibitors docked to the HIV-1 protease crystal structure, and 11 thermolysin inhibitors docked to the thermolysin structure.

Jain [106] has also developed an empirically derived scoring function based on 34 diverse protein–ligand complexes. The primary terms of the function arise for hydrophobic and polar complementarity, with additional factors for entropic and solvation effects. From this, Jain has constructed a sufficiently fast continuously differentiable nonlinear function, such that optimisation of alignment/conformation of the ligand within the receptor, based on the predicted affinity, may be readily achieved.

All these empirical scoring functions seem to be reasonably accurate when applied to protein–ligand complexes where an X-ray structure of the complex is available. The fact that the protein–ligand complex is sufficiently stable that it can be isolated probably indicates that there are no strongly unfavourable interactions. Hence, it may prove to be the case that current scoring functions are underestimating the magnitude of certain types of unfavourable interactions and some caution should be exercised in using them as a pruning mechanism in the course of structure generation.

By their very nature, structure generation programs can produce molecules which are new to chemistry. We can reduce the likelihood of creating synthetically troublesome molecules by including rules during the structure generation to exclude or limit the number of spiro joins or the formation of small rings, etc. Even so, the program may produce many hundreds of molecules which need to be prioritised on both their functional and synthetic merits. CAESA [58] is an automated method for estimating the synthetic accessibility of a given molecule. It attempts exhaustive retrosynthetic analysis to take the target structure back to starting materials contained within either an in-house or a chemical supplier's database. This expert system assesses the molecular complexity (based on the stereochemistry, functional groups and topology) of those portions of the target molecule which are not readily derived from a known starting material. A causal network is used to combine the various factors to yield an overall measure of the synthetic ease.

5. Conclusions

Every method described carries its own particular strengths and weaknesses. However, there is one oversimplification that is applied to all methods. In all techniques, we assume there are no conformational changes of the receptor on binding. Our understanding of intermolecular interactions is still too primitive to gain an accurate energetic picture of the binding event. Also, such complex simulations would require far larger computational resources than currently available for routine application. Currently, conformational changes can be monitored by iterating through the drug design cycle: generating structures, synthesis, and testing followed by structural analysis via NMR or X-ray. Our model can then be refined at each step and reapplied to generate new structures.

Structure-based design techniques by their very nature are highly focused toward the receptor structure or pharmacophore and neglect other factors important to the design of bioactive compounds, e.g. transport properties, toxicity and stability. Hence, these techniques must generate as many suggestions as possible and methods need to be developed to analyse these large sets in order to rule out those with unsuitable molecular properties.

Combinatorial chemistry coupled with high-throughput screening has already had a huge influence on the drug discovery process. Coupling this essentially irrational procedure to the elegance of structure-based design could further enhance the lead discovery and optimisation process. Fragment-based methods for *de novo* structure generation could be modified to create molecules that were readily synthesised combinatorially to yield a focused library. Scoring becomes less essential as all molecules can be readily prepared via automated synthesis.

References

1. DesJarlais, R.L., Seibel, G.L., Kuntz, I.D., Furth, P.S., Alvarez, J.C., Ortiz de Montellano, P.R., DeCamp, D.L., Babé, L.M. and Craik, C., Proc. Natl. Acad. Sci. USA, 87(1990)6644.
2. Clare, M., Perspect. Drug Discov. Design, 1(1993)49.
3. Lam, P.Y.S., Jadhav, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C.-H., Weber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-Viitanen, S., Science, 263(1994)380.
4. Kim, E.E., Baker, C.T., Dwyer, M.D., Murcko, M.A., Rao, B.G., Tung, R.D. and Navia, M. A., J. Am. Chem. Soc., 117(1995)1181.
5. Romines, K.R., Watenpaugh, K.D., Howe, W.J., Tomich, P.T., Lovasz, K.D., Morris, J.K., Janakiraman, M.N., Horng, M.-M., Chong, K.-T., Hinshaw, R.R. and Dolak, L.A., J. Med. Chem., 38(1995)4463.
6. Greer, J., Erickson, J.W., Baldwin, J.J. and Varney, M.D., J. Med. Chem., 37(1994)1035.
7. Rao, B.G., Kim, E.E. and Murcko, M.A., J. Comput.-Aided Mol. Design, 10(1996)23.
8. Hodge, C.N., Aldrich, P.E., Bacheler, L.T., Chang, C.H., Eyermann, C.J., Garber, S., Grubb, M., Jackson, D.A., Jadhav, P.K., Korant, B., Lam, P.Y.S., Maurin, M.B., Meek, J.L., Otto, M.J., Rayner, M.M., Reid, C., Sharpe, T.R., Shum, L., Winslow, D.L. and Erickson-Viitanen, S., Chem. Biol., 3(1996)301.
9. Lam, P.Y.S., Ru, Y., Jadhav, P.K., Aldrich, P.E., Delucca, G.V., Eyermann, C.J., Chang, C.H., Emmett, G., Holler, E.R., Daneker, W.F., Li, L.Z., Confalone, P.N., McHugh, R.J., Han, Q., Li, R.H., Markwalder, J.A., Seitz, S.P., Sharpe, T.R., Bacheler, L.T., Rayner, M.M., Klabe, R.M., Shum, L.Y., Winslow, D.L., Kornhauser, D.M., Jackson, D.A., Erickson-Viitanen, S. and Hodge, C.N., J. Med. Chem., 39(1996)3514.
10. Reid, R.C., March, D.R., Dooley, M.J., Bergman, D.A., Abbenante, G. and Fairlie, D.P., J. Am. Chem. Soc., 118(1996)8511.
11. Thairivongs, S., Romero, D.L., Tommasi, R.A., Janakiraman, M., Strobach, J.W., Turner, S.R., Biles, C., Morge, R.R., Johnson, P.D., Aristoff, P.A., Tomich, P.K., Lynn, J.C., Horng, M.-M., Chong, K.-T., Hinshaw, R.R., Howe, W.J., Finzel, B.C. and Watenpaugh, K.D., J. Med. Chem., 39(1996)4630.
12. Appelt, K., Bacquet, R., Bartlett, C., Booth, C., Freer, S., Fuhry, M., Gehring, M., Herrmann, S., Howland, E., Janson, C., Jones, T., Kan, C., Kathardekar, V., Lewis, K., Marzoni, G., Matthews, D., Mohr, C., Moomaw, E., Morse, C., Oatley, S., Ogden, R., Reddy, M., Reich, S., Schoettlin, W., Smith, W., Varney, M., Villafranca, J., Ward, R., Webber, S., Webber, S., Welsh, K. and White, J., J. Med. Chem., 34(1991)1925.

13. Varney, M., Marzoni, G., Palmer, C., Deal, J., Webber, S., Welsh, K., Bacquet, R., Bartlett, C., Morse, C., Booth, C., Herrmann, S., Howland, E., Ward, R. and White, J., J. Med. Chem., 35(1992)663.
14. Reich, S., Fuhry, M., Nguyen, D., Pino, M., Welsh, K., Webber, S., Janson, C., Jordan, S., Matthews, D., Smith, W., Bartlett, C., Booth, C., Herrmann, S., Howland, E., Morse, C., Ward, R. and White, J., J. Med. Chem., 35(1992)847.
15. Shoichet, B., Stroud, R., Santi, D., Kuntz, I. and Perry, K., Science, 259(1993)1445.
16. Webber, S., Bleckman, T., Attard, J., Deal, J., Kathardekar, V., Welsh, K., Webber, S., Janson, C., Matthews, D., Smith, W., Freer, S., Jordan, S., Bacquet, R., Howland, E., Booth, C., Ward, R., Hermann, S., White, J., Morse, C., Hilliard, J. and Bartlett, C., J. Med. Chem., 36(1993)733.
17. Varney, M., Palmer, C., Deal, J., Webber, S., Welsh, K., Bartlett, C., Morse, C., Smith, W. and Janson, C., J. Med. Chem., 38(1995)1892.
18. Jones, T., Varney, M., Webber, S., Lewis, K., Marzoni, G., Palmer, C., Kathardekar, V., Welsh, K., Webber, S., Matthews, D., Appelt, K., Smith, W., Janson, C., Villafranca, J., Bacquet, R., Howland, E., Booth, C., Herrmann, S., Ward, R., White, J., Moomaw, E., Bartlett, C. and Morse, C., J. Med. Chem., 39(1996)904.
19. Cunningham, D., Zalcberg, J., Smith, I., Gore, M., Pazdur, R., Burris, H., Meropol, N., Kennealey, G. and Seymour, L., Ann. Oncol., 7(1996)179.
20. Hakansson, K. and Liljas, A., FEBS Lett., 350(1994)319.
21. Gao, J., Qiao, S. and Whitesides, G., J. Med. Chem., 38(1995)2292.
22. Gowravaram, M., Tomczuk, B., Johnson, J., Delecki, D., Cook, E., Ghose, A., Mathiowetz, A., Spurlino, J., Rubin, B., Smith, D., Pulvino, T. and Wahl, R., J. Med. Chem., 38(1995)2570.
23. Beckett, R., Davidson, A., Drummond, A., Huxley, P. and Whittaker, M., Drug Discov. Today, 1(1996)16.
24. Rockwell, A., Melden, M., Copeland, R.A., Hardman, K., Decicco, C.P. and DeGrado, W.F., J. Am. Chem. Soc., 118(1996)10337.
25. Hauske, J., Dorff, P., Julin, S., Dibrino, J., Spencer, R. and Williams, R., J. Med. Chem., 35(1992)4284.
26. Andrus, M. and Schreiber, S., J. Am. Chem. Soc., 115(1993)10420.
27. Babine, R., Bleckman, T., Kissinger, C., Showalter, R., Pelletier, L., Lewis, C., Tucker, K., Moomaw, E., Parge, H. and Villafranca, J., Bioorg. Med. Chem. Lett., 5(1995)1719.
28. Babine, R., Bleckman, T., Littlefield, E., Parge, H., Pelletier, L., Lewis, C., French, J., Imbacuan, M., Katoh, S., Tatlock, J., Showalter, R. and Ernest, J., Bioorg. Med. Chem. Lett., 6(1996)385.
29. Egner, U., Hoyer, G. and Schleuning, W., J. Comput.-Aided Mol. Design, 8(1994)479.
30. Obst, U., Gramlich, V., Diederich, F., Weber, L. and Banner, D., Angew. Chem., Int. Ed. Engl., 34(1995)1739.
31. Grootenhuis, P., Westerduin, P., Meuleman, D., Petitou, M. and Van Boeckel, C., Nat. Struct. Biol., 2(1995)736.
32. Jones, D., Atrash, B., Tegernilsson, A., Gyzander, E., Deinum, J. and Szelke, M., Lett. Pept. Sci., 2(1995)147.
33. Fevig, J., Abelman, M., Brittelli, D., Kettner, C., Knabb, R. and Weber, P., Bioorg. Med. Chem. Lett., 6(1996)295.
34. Lombardi, A., Nastri, F., Dellamorte, R., Rossi, A., Derosa, A., Staiano, N., Pedone, C. and Pavone, V., J. Med. Chem., 39(1996)2008.
35. Fethiere, J., Tsuda, Y., Coulombe, R., Konishi, Y. and Cygler, M., Protein Sci., 5(1996)1174.

36. Fischer, B., Schlokat, U., Mitterer, A., Savidisdacho, H., Grillberger, L., Reiter, M., Mundt, W., Dorner, F. and Eibl, J., J. Biol. Chem., 271(1996)23737.
37. Vonitzstein, M., Dyason, J., Oliver, S., White, H., Wu, W., Kok, G. and Pegg, M., J. Med. Chem., 39(1996)388.
38. Vonitzstein, M., Wu, W., Kok, G., Pegg, M., Dyason, J., Jin, B., Phan, T., Smythe, M., White, H., Oliver, S., Colman, P., Varghese, J., Ryan, D., Woods, J., Bethell, R., Hotham, V., Cameron, J. and Penn, C., Nature, 363(1993)418.
39. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112(1977)535.
40. Peters, K.P., Fauck, J. and Frömmel, C., J. Mol. Biol., 256(1996)201.
41. Bayada, D.M. and Johnson, A.P., personal communication.
42. Goodford, P.J., J. Med. Chem., 28(1985)849.
43. Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C., J. Med. Chem., 32(1989)1083.
44. Wade, R.C., Clark, K.J. and Goodford, P.J., J. Med. Chem., 36(1993)140.
45. Wade, R.C. and Goodford, P.J., J. Med. Chem., 36(1993)148.
46. Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 8(1994)347.
47. Gehlhaar, D.K., Moerder, K.E., Zichi, D., Sherman, C.J., Ogden, R.C. and Freer, S.T., J. Med. Chem., 38(1995)466.
48. Rotstein, S.H. and Murcko, M.A., J. Comput.-Aided Mol. Design, 7(1993)23.
49. Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36(1993)1700.
50. Miranker, A. and Karplus, M., Proteins, 11(1991)29.
51. Pearlman, D.A. and Murcko, M.A., J. Comput. Chem., 14(1993)1184.
52. Bohacek, R.S. and McMartin, C., J. Am. Chem. Soc., 116(1994)5560.
53. Pearlman, D.A. and Murcko, M.A., J. Med. Chem., 39(1996)1651.
54. Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, Ser. B, 236(1989)101.
55. Böhm, H.-J., J. Comput.-Aided Mol. Design, 6(1992)61.
56. Böhm, H.-J., J. Comput.-Aided Mol. Design, 6(1992)593.
57. Clark, D.E., Frenkel, D., Levy, S.A., Li, J., Murray, C.W., Robson, B., Waszkowycz, B. and Westhead, D.R., J. Comput.-Aided Mol. Design, 9(1995)13.
58. Gillet, V.J., Myatt, G.M., Zsoldos, Z. and Johnson, A.P., Perspect. Drug Discov. Design, 3(1995)34.
59. Allen, F.H. and Kennard, O., Chem. Design Automat. News, 8(1993)130.
60. Klebe, G., J. Mol. Biol., 237(1994)212.
61. Laskowski, R.A., Thornton, J.M., Humblet, C. and Singh, J., J. Mol. Biol., 259(1996)175.
62. Marshall, G.R., Barry, D., Bosshard, H.E., Dammkoehler, R.A. and Dunn, D.A., In Olsen, E.C. and Christoffersen, R.E. (Eds.) Computer-Assisted Drug Design, ACS Symposium Series, Vol. 112, American Chemical Society, Washington, DC, 1979, pp. 205–226.
63. Mayer, D., Naylor, C.B., Motoc, I. and Marshall, G.R., J. Comput.-Aided Mol. Design, 1(1987)3.
64. Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B. and Marshall, G.R., J. Comput.-Aided Mol. Design, 9(1995)491.
65. Sheridan, R.P., Nilakantan, R., Dixon, J.S. and Venkataraghavan, R., J. Med. Chem., 29(1986)899.
66. Ghose, A.K., Logan, M.E., Treasurywala, A.M., Wang, H., Wahl, R.C., Tomczuk, B.E., Gowravaram, R., Jaeger, E.P. and Wendoloski, J. J., J. Am. Chem. Soc., 117(1995)4671.
67. Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7(1993)83.

68. Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9(1995)532.
69. Perkins, T.DJ. and Dean, P.M., J. Comput.-Aided Mol. Design, 7(1993)155.
70. Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 9(1995)479.
71. Martin, Y.C., J. Med. Chem., 35(1992)2145.
72. Willett, P., J. Chemometrics, 6(1992)289.
73. Hurst, T., J. Chem. Inf. Comput. Sci., 34(1994)190.
74. Clark, D.E., Jones, G., Willett, P., Kenny, P.W. and Glen, R.C., J. Chem. Inf. Comput. Sci., 34(1994)197.
75. Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S., In Roberts, S.M. (Ed.) Molecular Recognition: Chemical and Biological Problems, Royal Society of Chemistry, London, 1989, pp. 182–196.
76. Lauri, G. and Bartlett, P.A., J. Comput.-Aided Mol. Design, 8(1994)51.
77. Ho, C.M.W. and Marshall, G.R., J. Comput.-Aided Mol. Design, 7(1993)3.
78. Lawrence, M.C. and Davis, P.C., Proteins, 12(1992)31.
79. Moon, J.B. and Howe, W.J., Proteins, 11(1991)314.
80. Itai, A. and Nishibata, Y., Tetrahedron, 47(1991)8985.
81. Nishibata, Y. and Itai, A., J. Med. Chem., 36(1993)2921.
82. Tschinke, V. and Cohen, N.C., J. Med. Chem., 36(1993)3863.
83. Ho, C.M.W. and Marshall, G.R., J. Comput.-Aided Mol. Design, 7(1993)623.
84. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins, 19(1994)199.
85. Gillet, V., Johnson, A.P., Mata, P., Sike, S. and Williams, P., J. Comput.-Aided Mol. Design, 7(1993)127.
86. Gillet, V.J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z. and Johnson, A.P., J. Chem. Inf. Comput. Sci., 34(1994)207.
87. Mata, P., Gillet, V.J., Johnson, P., Lampreia, J., Myatt, G.J., Sike, S. and Stebbings, A.L., J. Chem. Inf. Comput. Sci., 35(1995)479.
88. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8(1994)513.
89. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8(1994)527.
90. Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8(1994)545.
91. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9(1995)341.
92. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9(1995)351.
93. Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9(1995)359.
94. Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R. and Kuntz, I.D., J. Mol. Graph., 10(1992)66.
95. Roe, D.C. and Kuntz, I.D., J. Comput.-Aided Mol. Design, 9(1995)269.
96. Ryckaert, J.P., Cicotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23(1977)327.
97. Miranker, A. and Karplus, M., Proteins, 23(1995)472.
98. Johnson, A.P., unpublished results.
99. Clark, D.E. and Murray, C.W., J. Chem. Inf. Comput. Sci., 35(1995)914.
100. Ajay and Murcko, M.A., J. Med. Chem., 38(1995)4953.
101. Green, S.M. and Marshall, G.R., Trends Pharmacol. Sci., 16(1995)285.
102. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110(1988)5959.
103. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8(1994)243.
104. Clark, A.J. and Johnson, A.P., personal communication.
105. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118(1996)3959.
106. Jain, A.J., J. Comput.-Aided Mol. Design, 10(1996)427.

New trends in computational structure prediction of ligand-protein complexes for receptor-based drug design

Paul A. Rejto, Gennady M. Verkhivker, Daniel K. Gehlhaar and Stephan T. Freer

Agouron Pharmaceuticals Inc., 3565 General Atomics Court,
San Diego, CA 92121-1121, U.S.A.

The molecular docking problem: Thermodynamic and kinetic aspects

A number of challenging computational problems arise in the field of structure-based drug design, including the estimation of ligand binding affinity and the *de novo* design of novel ligands. An important step toward solutions of these problems is the consistent and rapid prediction of the thermodynamically most favorable structure of a ligand–protein complex from the three-dimensional structures of its unbound ligand and protein components. This fundamental problem in molecular recognition is commonly known as the docking problem [1–3]. To solve this problem, two distinct conditions must be satisfied. The first is a thermodynamic requirement: the energy function used to describe ligand–protein binding must have the crystal structure of ligand–protein complexes as its global energy minimum. The second is a kinetic requirement: it must be possible to locate consistently and rapidly the global energy minimum on the ligand–protein binding energy landscape. While the first condition is necessary for successful structure prediction, it is by no means sufficient. Without kinetic accessibility, the global minimum cannot be reached during docking simulations, and computational structure prediction will fail. Here we review approaches to address both the kinetic and thermodynamic aspects of the docking problem.

Most docking algorithms fall into two general categories: surface complementarity methods [4–15] that match specific ligand–protein interactions and sample a relatively limited number of relevant conformational states, and more detailed methods that couple atomic representations of the intermolecular interactions with stochastic searching techniques designed to explore significant portions of the available configurational space [16–23]. Surface complementarity has been shown to be an important component of molecular recognition functions, but it alone does not make possible identification of the correct structure when alternative ligand–protein complexes have energies similar to those of native complexes [3,4,9–11]. Such ambiguous structural predictions may result from an incomplete description of the energy function, although simple hydrophobic energy functions often perform as well as more detailed representations of ligand–protein interactions [15].

The problem of docking flexible ligands, even with a protein held fixed in its bound conformation, requires determination of the global energy minimum from an enormous number of available conformations. This problem belongs to the class of nondeterministic polynomial time (NP)-hard tasks [24], and represents another manifestation of the famous Levinthal paradox also encountered in protein folding [25]. Analyses of the energy landscapes of protein folding [26–34] have revealed that uniqueness, stability and kinetic accessibility of proteins depend critically on the extent of complexity and frustration in the energy landscape. These studies demonstrate that proteins with robust folding kinetics and thermodynamic stability must obey the principle of minimum frustration [26,27,33].

The inherent complexity of a force field that can completely describe the binding process results in a frustrated energy landscape with many energetically similar, yet structurally different, local minima that are separated by large energy barriers. A binding energy landscape characterized by both of these features is defined to be frustrated. This frustration may result, first, from the inability to satisfy simultaneously both intermolecular ligand–protein interactions and intramolecular ligand constraints, and, second, when misdocked structures are not connected to the global energy minimum by broad networks of configurations with low barriers [30,34]. Energy barriers arise primarily from steric constraints, and also when attractive but nonnative ligand–protein contacts are made during the docking process. If the global energy minimum is not kinetically accessible, the docking process can be trapped in local minima even when the native complex is thermodynamically favorable. Hence, even if an accurate energy function suitable for rigorous predictions of binding affinity were available, it would not necessarily resolve the kinetic bottleneck of the docking problem.

The quality of a molecular recognition energy function is generally evaluated on its thermodynamic properties, i.e. the location of the global energy minimum and its energy relative to the energies of alternative binding modes [3–9], whereas efforts to satisfy the kinetic criteria for rapid structure prediction have concentrated on developing search algorithms [16–23]. An alternative approach to satisfy the kinetic criteria for robust structure prediction is to modify the shape of the energy landscape so as to reduce the complexity of the search problem. Such landscapes, which have a reduced level of frustration, are characterized by ‘funnels’ that connect conformational states to the global energy minimum [31–33]. We have investigated both approaches, and here review the application of two search techniques based on the concept of natural selection to docking simulations, genetic algorithms and evolutionary programming, and then discuss efforts to develop molecular recognition energy functions with reduced frustration.

Search techniques

A variety of stochastic optimization techniques have been used in docking simulations including Monte Carlo methods [16–20], molecular dynamics [21,22], and dead-end elimination [23]. These studies focused primarily on the flexible docking of ligands into proteins held in their bound conformation, in which the internal degrees of freedom of the ligand and its rigid-body variables are optimized. Genetic algorithms [35] have been used to solve a variety of optimization tasks, including molecular docking [36–41], conformational search of small molecules [42,43], and protein folding [44–47]. Recently, evolutionary programming [48] has emerged as another promising search technique for protein folding [49] and molecular docking applications [50–52].

Genetic algorithms

In the application of genetic algorithms to ligand–protein docking [41], each chromosome represents a member in a population of ligand conformations, and is encoded as a vector comprised of the six rigid-body orientational and positional coordinates along with the dihedral angles of all rotatable bonds. Initial ligand conformations are generated by randomizing the encoded vector while requiring that the center of mass of the ligand reside within the rectangular parallelepiped defining the binding site. During the search, the fitness of each chromosome in the population is evaluated and the chromosomes are ranked. An elitist mechanism, whereby survival of the most fit member is guaranteed, preserves the quality of the genetic material. Replication of all the other members is done by a selection process based on a 'roulette wheel' mechanism whereby the probability of selecting a particular chromosome for mating is proportional to its fitness. The selected chromosomes are duplicated in the next generation. They are then replicated, subject to mutation and crossover, to obtain a population of a defined size. The mutation operator, in which each bit of the chromosome is subject to a defined flip probability, imparts diversity to the population, while the crossover procedure exchanges pieces of the parents' chromosomes at randomly selected crossover points.

Evolutionary programming

Evolutionary programming is another stochastic search technique based on the ideas of natural selection. During the course of the simulated evolution, a population of candidate ligand conformers competes for survival against a fixed number of opponents randomly selected from the remainder of the population (Fig. 1). A win is assigned to the member of this set with the lowest energy, and the number of competitions a member wins determines its survival into the next generation [51]. All surviving members produce offspring by Gaussian mutation so as to maintain a constant population size. In the population of ligand conformers, each member represents an encoded vector comprised of its six rigid-body coordinates and the dihedral angles about its rotatable bonds. The initial ligand conformations are generated by randomizing the encoded vector, given that the ligand center of mass must lie within the rectangular parallelepiped that defines the active site. Rigid rotation and rotatable dihedral angles are uniformly randomized between 0° and 360°.

In the studies reviewed here, the evolutionary search was performed for a total of 120 generations with a population size of 1200. In order to maintain a diverse

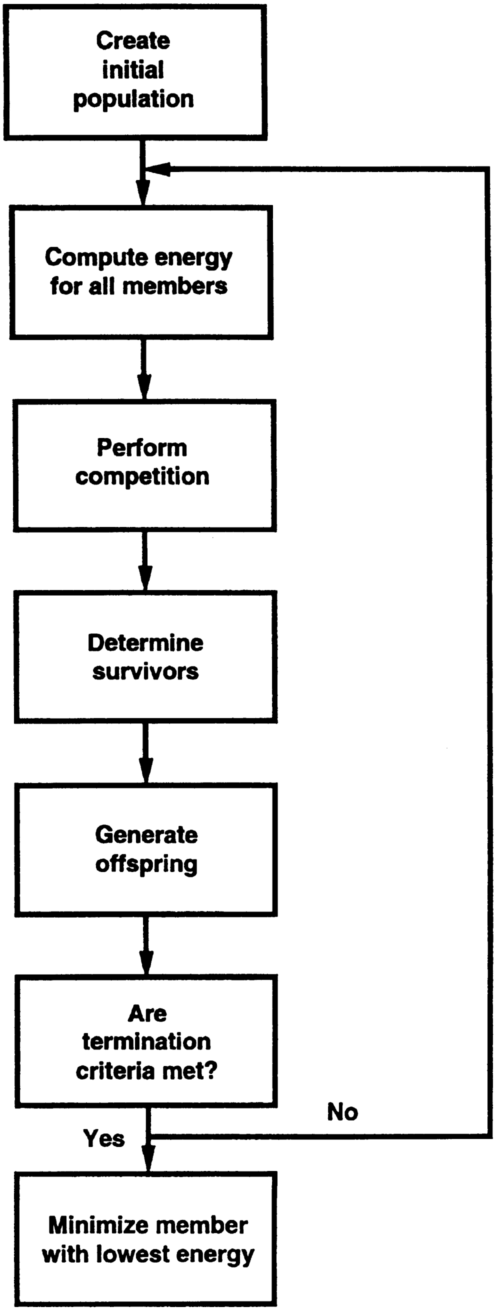

Create initial population
Compute energy for all members
Perform competition
Determine survivors
Generate offspring
Are termination criteria met?
No
Yes
Minimize member with lowest energy

Fig. 1. Flow chart for the evolutionary programming strategy protocol.

population, each member competes against only three opponents at each generation [51,52]. The size of the standard deviation of the Gaussian mutation that generates offspring from a parent must be determined with care. If the mutations are too small, the system does not explore the search space efficiently, while if they are too large, offspring bear little or no resemblance to the parent, and the search becomes undirected. Consequently, in a process resembling simulated annealing, large mutations are allowed early in the simulation to facilitate rapid searching, analogous to high temperature, while small mutations are made late in the simulation to refine solutions near the global optimum, analogous to low temperature. Because it is difficult to predict the most appropriate scaling scheme for the mutations, the standard deviation of the Gaussian mutation is varied adaptively throughout the simulation by using selection pressure [53,54]. The member of the final generation with the lowest energy is minimized [55].

Low-resolution molecular recognition energy models

While studies have shown that simple energy functions are robust in protein structure prediction [56], sophisticated search procedures outperform standard techniques only when the underlying energy surface is relatively unfrustrated [57]. We recently evaluated a family of simple molecular recognition energy functions for use in docking simulations and analyzed the ability of the energy function itself to reduce frustration in the binding energy landscape [41]. We found that the 'hardness' of the ligand–protein interaction energy was critical for satisfying both the thermodynamic and kinetic requirements for docking, and there is an optimal range of values for the repulsive hardness of the ligand–protein interaction. If the repulsion is too soft, the crystal structure is no longer the global energy minimum, while if it is too hard, the crystal structure is not kinetically accessible during the docking simulation.

The molecular recognition model used for the ligand–protein interaction energy includes steric and hydrogen bond contributions calculated from a piecewise linear potential together with a simple angular dependence for the hydrogen bond interaction (Fig. 2). This model is not intended to be a complete description of ligand–protein interactions, but rather attempts to capture the minimal requirements of the energy function that are necessary for robust structural assessments during docking simulations [41,51,52]. The parameters of the pairwise potential depend only on four different atom types: hydrogen-bond donor, hydrogen-bond acceptor, both donor and acceptor, and nonpolar. Primary and secondary amines are defined to be donors while oxygen and nitrogen atoms with no bound hydrogens are defined to be acceptors. Crystallographic water molecules and hydroxyl groups are defined to be both donor and acceptor. Carbon atoms are defined to be nonpolar.

The intermolecular interactions between ligand and protein atoms are represented by steric and hydrogen-bond-like potentials (Table 1), both of which have the same functional form. The parameters (Table 2) were extracted from a potential used for the *de novo* design of enzyme inhibitors [58], and subsequently refined to yield the

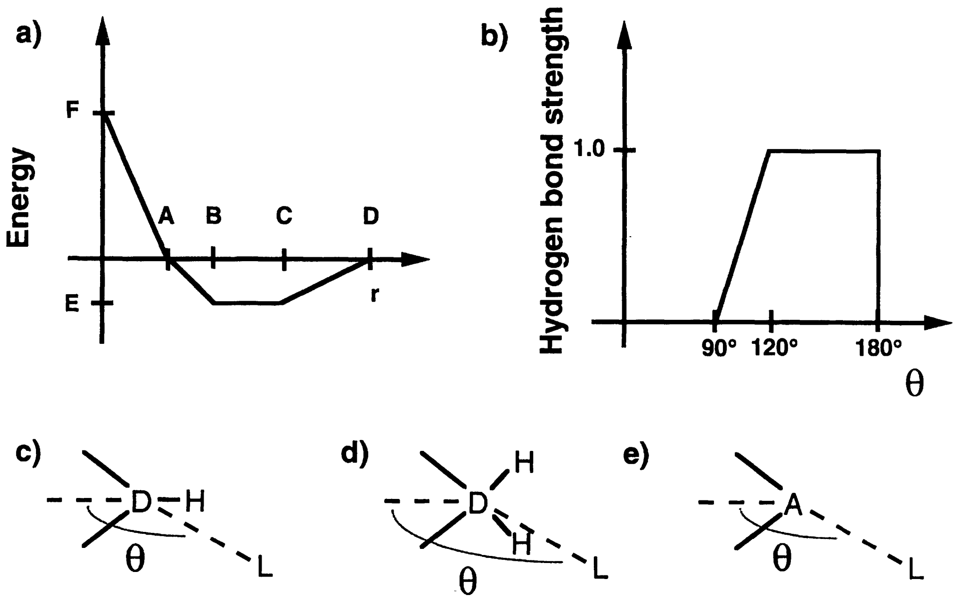


Fig. 2. (a) The potential energy function used to compute the pairwise ligand–protein interaction energy. Hydrogen atoms are not included in the calculation. The parameter values are given in Table 2, where the units of energy are arbitrary. (b) The hydrogen bond interaction energy is multiplied by the hydrogen bond strength term, which is a function of the angle θ determined by the relative orientation of the protein and ligand atoms. The angle θ is shown in panels (c)–(e). (c) A protein donor atom D bound to one hydrogen atom H makes an angle θ with the ligand atom ***L**. (d) A protein donor atom D bound to two hydrogen atoms H makes an angle θ with the ligand atom **L**. (e) A protein acceptor atom **A** makes an angle θ with the ligand atom **L**. In all cases, the range of the angle θ is between 0° and 180°.*

experimental crystallographic structure of a set of ligand–protein complexes as the global energy minimum [51]. They assume that a single hydrogen bond has more interaction energy than a single steric interaction [59]. An advantage of this representation is that the ligand–protein interaction function is well behaved even when

Table 1 *Pairwise atomic interaction types for the molecular recognition model*

Ligand atom type	Protein atom type			
	Donor	Acceptor	Both	Nonpolar
Donor	Steric	Hydrogen bond	Hydrogen bond	Steric
Acceptor	Hydrogen bond	Steric	Hydrogen bond	Steric
Both	Hydrogen bond	Hydrogen bond	Hydrogen bond	Steric
Nonpolar	Steric	Steric	Steric	Steric

Table 2 *Parameters of the atomic pairwise ligand–protein potentials*[a]

Interaction type	A	B	C	D	E	F
Steric	3.4	3.6	4.5	5.5	– 0.4	20.0
Hydrogen bond	2.3	2.6	3.1	3.4	– 2.0	20.0

[a] A, B, C, and D are in Å. E and F are in arbitrary energy units.

there are severe close contacts between protein and ligand atoms, which occur during the initial stages of the docking simulation when the ligand conformations are largely in random orientations or when the crystal structure itself has close contacts. To permit the ligands to penetrate the protein core during these early stages, the repulsive parameter of the intermolecular ligand–protein interaction potential is linearly scaled from zero to its final value throughout the simulation.

The intramolecular potential also contributes to the conformations adopted by the ligand in the protein active site. These ligand conformations may differ substantially from the conformation of the same ligand in the gas phase or in aqueous solution. A natural question then arises: is the conformation adopted by the ligand dominated by the interaction with the protein or by its internal constraints? In the course of developing potential energy functions for use in docking simulations, additional questions arise: what is the appropriate level of detail to use for a ligand intramolecular potential, and what effect does this have on the thermodynamic and kinetic requirements for docking? In the next section, we present new results for three different levels of approximation for the ligand intramolecular potential: one with no intramolecular potential, one with a simple hard-sphere interaction for atoms connected by at least three intervening atoms, where every atom has a radius of 2.35 Å and a repulsive barrier height of 2000.0, and one with a complete force field model [60]. In all three cases, a limited number of bonds are allowed to rotate, namely bonds linking sp^3 atoms to either sp^3 or sp^2 atoms, as well as nonconjugated single bonds linking two sp^2 atoms. The ligand bond distances and bond angles, as well as the torsional angles of nonrotatable bonds, were obtained from the crystal structure of the bound ligand–protein complex and were held fixed during the docking simulations. No *a priori* assumptions regarding either favorable ligand conformations or any specific ligand–protein interactions are made. The ligand conformations and orientations are searched in a rectangular box that encompasses the binding site obtained from the structure of the crystallographic ligand–protein complex with a 2 Å cushion added to every side of this box, and the intermolecular potential is precalculated on a 0.2 Å grid that covers the protein binding site. A constant energy penalty of 200.0 is added to every ligand atom outside the box. All crystallographic water molecules are included in the simulations as part of the protein structure.

Conformationally flexible docking of HIV-1 protease complexes

HIV-1 protease is a symmetric homodimer, although the symmetry may be broken upon ligand binding, resulting in two, nearly degenerate binding modes. The presence of these two symmetry-related binding modes complicates the docking even of a completely rigid ligand into the bound conformer of the protease. For flexible ligands, the number of possible alternative solutions increases dramatically, and consequently an accurate molecular recognition model must be able to distinguish between these modes. An additional problem is that the binding site is highly constrained because of the protease 'flaps' that enclose the bound ligand. This confined site leads to large energy barriers that separate the alternative binding modes. The combination of these factors makes the docking of ligands into HIV-1 protease a particularly demanding problem.

A set of 25 HIV-1 protease complexes [61] was obtained from the Brookhaven Protein Data Bank (Table 3). These structures were supplemented by an additional 36 HIV-1 protease complexes that were solved at Agouron Pharmaceuticals [62,63]. No energy minimization or additional processing of either the ligand or the protein structures was performed. These structures represent a rather diverse set of ligands that bind to HIV-1 protease, containing from 31 to 61 heavy atoms, and between 6 and 29 rotatable bonds. For each ligand–protein complex, 100 docking simulations have been performed with both rigid and conformationally flexible ligands, and the structure with the lowest energy is taken to be the computational prediction of the ligand–protein complex.

First, a set of docking simulations was performed while keeping the conformation of the ligand rigidly fixed in its bound state. Two binding modes were predicted, corresponding to the observed crystallographic mode and a symmetry-related mode (Fig. 3). For all the ligands studied, the energy of the symmetry-related mode was correctly predicted to be higher than the energy of the crystallographic mode, except in those cases where the bound conformation of the HIV-1 protease retains its twofold symmetry and both conformations are isoenergetic.

Table 3 *HIV-1 protease complexes obtained from the Brookhaven Protein Data Bank*

Complex	Ligand	PDB entry	Complex	Ligand	PDB entry
1		aaq	14	A78791	1hvj
2	SB203238	1hbv	15	A76928	1hvk
3	SKF108738	1hef	16	A76889	1hvl
4	SKF107457	1heg	17	XK263	1hvr
5	CPG53820	1hih	18	A77003	1hvs
6	U75875	1hiv	19	SB203386	1sbg
7	SB204144	1hos	20	MVT-101	4hvp
8	SB206343	1hps	21	L-700,417	4phv
9	VX478	1hpv	22	Acetylpepstatin	5hvp
10	GR123976	1hte	23	JG-365	7hvp
11	GR126045	1htf	24	U-85548E	8hvp
12	GR137615	1htg	25	A-74704	9hvp
13	A77003	1hvi			

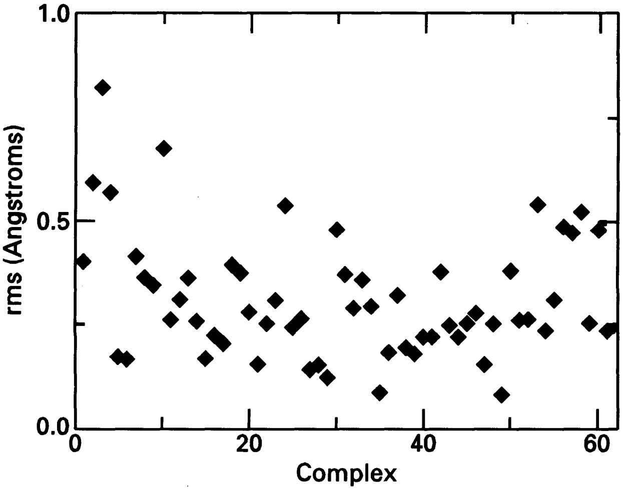


Fig. 3. Root-mean-square (rms) deviation of the lowest energy structure relative to the crystallographic conformation obtained from 100 individual rigid docking simulations for each of 61 different HIV-1 protease–ligand complexes. The structure with the lowest energy is the computational prediction for each ligand–protein complex. The first 25 complexes correspond to those listed in Table 3. The remaining complexes were solved by Agouron Pharmaceuticals [62].

Consistent structural prediction of the crystallographic conformation of the protein–ligand complex requires that the crystallographic conformation be the global energy minimum of the binding energy landscape and that this conformation be kinetically accessible. Energy landscapes that satisfy the principle of minimum frustration [26] have a relatively large separation in energy between the crystallographic conformation and other misdocked conformations. Such landscapes may also increase the kinetic accessibility of the crystallographic conformation [29], and therefore satisfy both the thermodynamic and kinetic requirements for robust docking. Consistent with this principle, the probability of successfully predicting the crystallographic conformation for XK263 (1hvr, Table 3) and AG-1343 [51], with eight and nine rotatable bonds, respectively, increases dramatically for the hard sphere or complete force field models compared to using no internal interactions at all (Fig. 4). The success rate of predicting the crystallographic conformation improves because when no internal potential is used there are many incorrectly positioned structures, whereas there are fewer incorrect structures for either the hard-sphere internal energy or the complete ligand intramolecular potential. Associated with this increase in the probability of successful structure prediction is an increase in the stability gap [27,33], the energy difference between the lowest energy structure and the energies of misdocked structures [52]. The presence of internal energy increases the magnitude of this stability gap by destabilizing incorrectly docked ligands, thereby increasing the probability of predicting the crystal structure in the course of a docking simulation

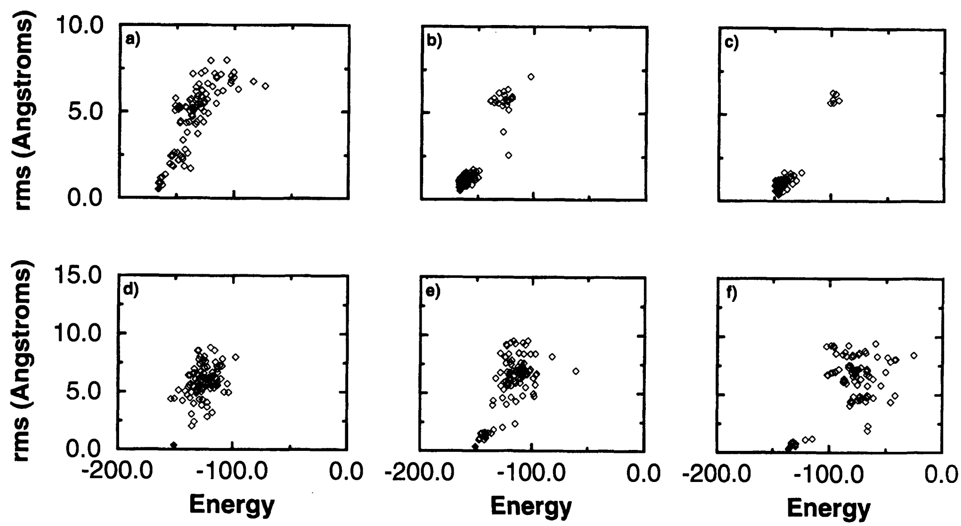


Fig. 4. Rms deviation relative to the crystallographic conformation versus energy for 100 individual docking simulations (open diamonds) for XK263, which contains eight rotatable bonds. The structure with the lowest energy is the prediction of the simulation: (a) no internal energy; (b) hard sphere; and (c) Dreiding intramolecular potential. The results for AG-1343, which contains nine rotatable bonds: (d) no internal energy; (e) hard sphere; and (f) Dreiding intramolecular potential. The structure obtained by minimization of the crystallographic conformation is represented by a filled diamond.

(Fig. 4). For these ligands, the more detailed ligand intramolecular energy models provide the best discrimination between the crystallographic structure and alternative binding modes. Interestingly, the hard-sphere internal potential is sufficient to increase the stability gap and promote kinetic accessibility for these ligands with a moderate number of rotatable bonds.

Computational structure prediction clearly becomes more difficult as the number of degrees of freedom increases. We found that the bound structures for the inhibitors AG-1002 [62] and JG-365 [64], which contain 25 rotatable bonds, were correctly predicted (Fig. 5) only when using the complete ligand intramolecular force field. Compared to using no internal potential, the hard sphere potential does destabilize some of the misdocked structures (Fig. 5), although it is not sufficient to make the crystallographic structure both thermodynamically stable and kinetically accessible during the docking simulation. Using the complete ligand intramolecular energy function, the crystal structures of complexes for a majority of conformationally flexible ligands have been assessed correctly (Fig. 6a). Due to symmetry, there is a second binding mode that is nearly isoenergetic with the crystallographic structure, and most of the misdocked structures represent these symmetry-related binding modes (Fig. 6b).

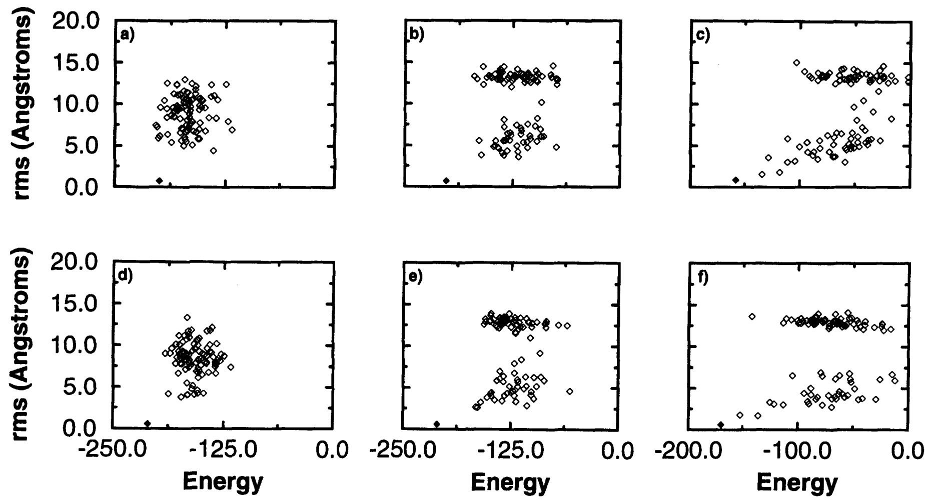


Fig. 5. Rms deviation relative to the crystallographic conformation versus energy for 100 individual docking simulations (open diamonds) for JG-365, which contains 25 rotatable bonds during the docking simulation. The structure with the lowest energy is the prediction of the simulation: (a) no internal energy; (b) hard sphere; and (c) Dreiding intramolecular potential. The results for AG-1002, which contains 25 rotatable bonds, during the docking simulation: (d) no internal energy; (e) hard sphere; and (f) Dreiding intramolecular potential. The structure obtained by minimization of the crystallographic conformation is represented by a filled diamond.

Discussion and Conclusions

A model of molecular recognition that combines a simple description of ligand–protein steric interactions and ligand–protein hydrogen bonding with a complete ligand intramolecular potential has been shown to be suitable for the structure prediction of HIV-1 protease complexes across a diverse set of ligands with a large number of degrees of freedom. While a far more detailed model that includes solvation and entropic effects is required to predict binding affinity, a model that produces a simple and relatively smooth energy landscape with few local minima is preferable for structure prediction, given that the global energy minimum is retained at the crystallographic conformation of the complex. In this way, both the thermodynamic stability and the kinetic accessibility of the ligand–protein complex are realized while the minimum frustration principle is satisfied.

Conventionally, scoring functions that are too simple to account rigorously for the binding affinity of enzyme–ligand complexes are believed to represent a stumbling block in achieving more accurate structure predictions in docking simulations. However, misdocked low-energy structures of ligand–protein complexes are caused by the rugged energy landscape of molecular recognition, which is an inherent feature of current models used to represent ligand–protein interactions. Consequently, the most

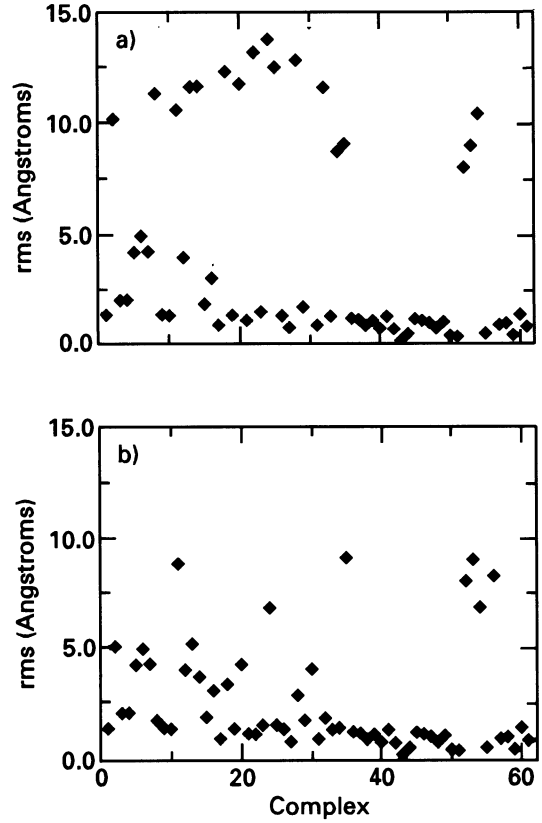


Fig. 6. (a) Rms deviation of the lowest energy structure relative to the crystallographic conformation obtained from 100 individual flexible docking simulations for each of 61 different HIV-1 protease–ligand complexes, using the Dreiding intramolecular force field. (b) The predicted structure is compared both to the correct binding mode and the symmetry-related binding mode, and the lowest rms difference is shown.

appropriate energy function for the computational structure prediction of ligand–protein complexes may differ from standard molecular mechanics force fields.

The results reported here show that a complete description of the ligand intramolecular interaction improves both the thermodynamic and kinetic requirements for robust structure prediction. Although the ligand intramolecular energy landscape also has large energy barriers, which arise from the steric interactions, these barriers appear not to be a source of additional frustration. The kinetics of the docking simulation is apparently dominated by intermolecular interactions, which favor relatively extended conformations of the HIV-1 protease ligands, thereby reducing the formation of strong nonnative intramolecular contacts that otherwise may have led to additional kinetic bottlenecks. Although a hard-sphere potential is adequate for ligands with a modest number of rotatable bonds, a more complete internal potential is necessary for ligands with a large number of rotatable bonds. These more complete internal potentials increase the stability gap between the crystal structure and misdocked structures, which increases the probability of predicting the crystal structure. Hence, complete force fields may be suitable for describing the ligand intramolecular interactions during docking simulations, while simplified representations of the intermolecular interactions may be necessary to satisfy kinetic requirements. For robust ligand–protein docking, then, the binding energy landscape must be smooth to minimize the likelihood of the simulation becoming trapped in local minima, and the landscape must have funnels leading to the global free-energy minimum that corresponds to the crystallographic conformation of the ligand–protein complex. The precise choice of force field may not be critical provided that the resulting energy landscape has reduced frustration.

References

1. Wodak, S.J. and Janin, J., J. Mol. Biol., 124(1978)323.
2. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161(1982)269.
3. Cherfils, J. and Janin, J., Curr. Opin. Struct. Biol., 3(1993)265.
4. Shoichet, B.K. and Kuntz, I.D., J. Mol. Biol., 221(1991)327.
5. Wang, H.J., J. Comput. Chem., 12(1991)746.
6. Jiang, F. and Kim, S.H., J. Mol. Biol., 219(1991)79.
7. Desjarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 29(1986)2149.
8. Desjarlais, R.L. and Dixon, J.S., J. Comput.-Aided Mol. Design, 8(1994)231.
9. Shoichet, B.K. and Kuntz, I.D., Protein Eng., 6(1993)723.
10. Walls, P.H. and Sternberg, M.J.E., J. Mol. Biol., 228(1992)277.
11. Jackson, R.M. and Sternberg, M.J.E., J. Mol. Biol., 250(1995)258.
12. Stoddard, B.L. and Koshland, D.E., Proc. Natl. Acad. Sci. USA, 90(1993)1146.
13. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A., Proc. Natl. Acad. Sci. USA, 89(1992)2195.
14. Fisher, D., Lin, S.L., Wolfson, H.J. and Nussinov, R., J. Mol. Biol., 248(1995)459.
15. Vakser, I.A. and Aflalo, C., Proteins Struct. Funct. Genet., 20(1994)320.
16. Goodsell, D.S. and Olson, A.J., Proteins Struct. Funct. Genet., 8(1990)195.
17. Yue, S.Y., Proteins, 4 (1990) 177.
18. Caflisch, A., Niederer, P. and Anliker, M., Proteins Struct. Funct. Genet., 13(1992)223.
19. Hart, T.N. and Read, R.J., Proteins Struct. Funct. Genet., 13(1992)206.
20. Totrov, M. and Abagyan, R., Nat. Struct. Biol., 1(1994)259.
21. DiNola, A., Roccatano, D. and Berendsen, H.J.C., Proteins Struct. Funct. Genet., 19(1994)174.
22. Zacharias, M., Luty, B.A., Davis, M.E. and McCammon, J.A., J. Mol. Biol., 238 (1994)455.
23. Leach, A.R., J. Mol. Biol., 235(1994)345.
24. Kuhl, F.S., Crippen, G.M. and Friesen, D.K., J. Comput. Chem., 5(1984)24.

25. Levinthal, C., In DeBrunner, P., Tsibris, J. and Munck, E. (Eds.) Mossbauer Spectroscopy in Biological Systems, Proceedings of a meeting held at Allerton House, Monticello, Urbana, IL, University of Illinois Press, Champaign, IL, 1969, pp. 22–24.
26. Bryngelson, J.D. and Wolynes, P.G., Proc. Natl. Acad. Sci. USA, 84(1987)7524.
27. Goldstein, R.A., Luthey-Schulten, Z.A. and Wolynes, P.G., Proc. Natl. Acad. Sci. USA, 89(1992)9029.
28. Shakhnovich, E.I. and Gutin, A.M., Proc. Natl. Acad. Sci. USA, 90(1993)7195.
29. Sali, A., Shakhnovich, E.I. and Karplus, M., J. Mol. Biol., 235(1994)1614.
30. Chan, H.S. and Dill, K.A., J. Chem. Phys., 100(1994)9238.
31. Leopold, P.E., Montal, M. and Onuchic, J.N., Proc. Natl. Acad. Sci. USA, 89(1992)8721.
32. Socci, N.D. and Onuchic, J.N., J. Chem. Phys., 101(1994)1519.
33. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. and Wolynes, P.G., Proteins Struct. Funct. Genet., 21(1995)167.
34. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S., Protein Sci., 4(1995)561.
35. Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
36. Xiao, Y.L. and Williams, D.E., J. Phys. Chem., 98(1994)7191.
37. Oshiro, C.M., Kuntz, I.D. and Dixon, J.S., J. Comput.-Aided Mol. Design, 9(1995)113.
38. Judson, R.S., Tan, Y.T., Mori, E., Melius, C., Jaeger, E.P., Treasurywala, A.M. and Mathiowetz, A., J. Comput. Chem., 16(1995)1405.
39. Clark, K.P. and Ajay, J. Comput. Chem., 16(1995)1210.
40. Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245(1995)43.
41. Verkhivker, G.M., Rejto, P.A., Gehlhaar, D.K. and Freer, S.T., Proteins Struct. Funct. Genet., 25(1996)342.
42. McGarrah, D.B. and Judson, R.S., J. Comput. Chem., 14(1993)1385.
43. Judson, R.S., Jaeger, E.P., Treasurywala, A.M. and Peterson, M.L., J. Comput. Chem., 14(1993)1407.
44. Unger, R. and Moult, J., J. Mol. Biol., 231(1993)75.
45. Sun, S., Protein Sci., 2(1993)762.
46. Dandekar, T. and Argos, P., Protein Eng., 5(1992)637.
47. Dandekar, T. and Argos, P., J. Mol. Biol., 236(1994)844.
48. Fogel, D.B., Evolutionary Computation: Toward a New Philosophy of Machine Intelligence, IEEE Press, Piscataway, NJ, 1995.
49. Bowie, J.U. and Eisenberg, D., Proc. Natl. Acad. Sci. USA, 91(1994)4436.
50. Gehlhaar, D.K., Verkhivker, G., Rejto, P.A., Fogel, D.B., Fogel, L.J. and Freer, S.T., In McDonnell, J.R., Reynolds, R.G. and Fogel, D.B. (Eds.) Proceedings of the 4th Annual Conference on Evolutionary Programming, MIT Press, Cambridge, MA, 1995, pp. 615–627.
51. Gehlhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.B., Fogel, L.J. and Freer, S.T., Chem. Biol., 2(1995)317.
52. Verkhivker, G.M. and Rejto, P.A., Proc. Natl. Acad. Sci. USA, 93(1996)60.
53. Schwefel, H.-P., Numerical Optimization of Computer Models, Wiley, Chichester, 1981.
54. Standard deviations of the Gaussian mutations S for each variable were generated from parent values s as follows: $S_i = s_i \exp(T N(0,1) + t N_i(0,1))$ $(T = 1/\sqrt{2n}$, $t = 1/\sqrt{2\sqrt{(n)}}$, where N(0,1) represents a zero-mean, unit variance Gaussian random number, and n is the number of variables in the optimization. $N_i(0,1)$ indicates that a different random number

is chosen for each component of the individual. The learning rate T influences the movement of the individual with respect to the parent, while the learning rate t influences-variations between components of the individual. This formula was obtained from Ref. 53.

55. Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in C. The Art of Scientific Computing, Cambridge University Press, Cambridge, 1992.
56. Yue, K. and Dill, K.A., Protein Sci., 5(1996)254.
57. Elofsson, A., Le Grand, S.M. and Eisenberg, D., Proteins Struct. Funct. Genet., 23(1995)73.
58. Gehlhaar, D.K., Moerder, K.E., Zichi, D., Sherman, C.J., Ogden, R.C. and Freer, S.T., J. Med. Chem., 38(1995)466.
59. Knegtel, R.M.A., Antoon, J., Rullmann, C., Boelens, R. and Kaptein, R., J. Mol. Biol., 235(1994)318.
60. Mayo, S.L., Olafson, B.D. and Goddard III, W.A., J. Phys. Chem., 94(1990)8897.
61. Wlodawer, A. and Erickson, J.W., Annu. Rev. Biochem., 62(1993)543.
62. Appelt, K., Perspect. Drug Discov. Design, 1(1993)23.
63. Reich, S.H., Melnick, M., Davies II, J.F., Appelt, K., Lewis, K.K., Fuhry, M.A., Pino, M., Trippe, A.J., Nguyen, D., Dawson, H., Wu, B.-W., Musick, L., Kosa, M., Kahil, D., Webber, S., Gehlhaar, D.K., Andrada, D. and Shetty, B., Proc. Natl. Acad. Sci. USA, 92(1995)3298.
64. Swain, A.L., Miller, M.M., Green, J., Rich, D.H., Schneider, J., Kent, S.B.H. and Wlodawer, A., Proc. Natl. Acad. Sci. USA, 87(1990)8805.

Estimation of binding affinity in structure-based design

Dave Timms and Anthony J. Wilkinson
Zeneca Pharmaceuticals, Alderley Park, Macclesfield, Cheshire SK10 4TG, U.K.

1. Introduction

With the increasing availability of protein structures, the potential for their use in the design of novel, pharmaceutically relevant inhibitors becomes an increasing reality. Over the last few years a number of programs have become available which aim to enable inhibitors to be designed *de novo* against specific protein active sites. Verlinde and Hol [1] classified these methods into three types: docking, linking and growing. The first of these encompasses database searching methods like DOCK [2]. The second includes vector database methods such as CAVEAT [3], EMPTOR [4] and other approaches including BOOM [5], HOOK [6] and CCLD [7] and carbocyclic linker methods such as SPROUT [8]. The final group includes LUDI [9], GROWMOL [10] and GROW [11]. These have been used with varying degrees of success in model systems and in 'real' applications, for inhibitor design.

All of these methods produce large numbers of potential inhibitors of interest. These hit lists need to be filtered and ranked in some way. Clearly, some filtering can be achieved by discarding chemically unattractive molecules and those that clash with the protein or which can only bind in energetically unfavourable conformations. However, all these design methods need to make some quantitative or semiquantitative assessment as to the quality of the binding interaction between putative inhibitor and protein. Some approaches may be more dependent on the accuracy of the binding affinity prediction than others. For instance, those methods which involve growing the ligand in the active site are usually dependent on some fitness function to guide the progress of the evolving design. If the approach attempts to be discriminating during this process of ligand growth, then the fitness function needs to provide a reasonable reflection of the strength of the protein–ligand interaction, otherwise 'good' molecules will not be designed. Other methods will usually employ a less discriminating function in the design stage and so all molecules designed are 'seen' by the user.

A range of methods have been explored in the literature in recent years to estimate binding affinity. These range from the computationally intensive free energy perturbation (FEP) methodology to empirical methods based, for example, on the number of hydrogen bonds and the degree of hydrophobic surface area that is buried. In this paper we review critically a selection of the methods that have been explored and consider their applicability in inhibitor design problems, their accuracy and their limitations. We report work carried out in our own laboratory and consider it alongside work carried out elsewhere. We make no attempt to comprehensively review the area. An excellent review on this topic has been published by Ajay and Murcko [12] and readers are directed to this as a complementary reading to this paper.

2. The problem

Figure 1 captures the problem schematically. We are interested in the bound and the unbound states of the protein and ligand and the differences in energy between these. It is essential to remember that in both states the protein and ligand are interacting with their water environment and that this changes between the two states. In analysing these states we build upon the work of Vajda et al. [13] and use their terminology where appropriate. The energy of the free state can be defined by

$$E_f = E_f^i + E_f^e + E_f^w + E_f^{i\text{-}w} + E_f^{e\text{-}w} \qquad (1)$$

where superscripts i, e and w refer to inhibitor, enzyme and water, respectively. The third term refers to the self-energy of water and the last two terms refer to the energy of interaction between the water environment and the inhibitor and enzyme, respectively. The energy of the bound state can be similarly defined as

$$E_b = E_b^i + E_b^e + E_b^w + E_b^{e\text{-}i} + E_b^{ei\text{-}w} \qquad (2)$$

where the fourth term represents the interaction between enzyme and inhibitor and the last term represents the interaction between the enzyme–inhibitor complex and

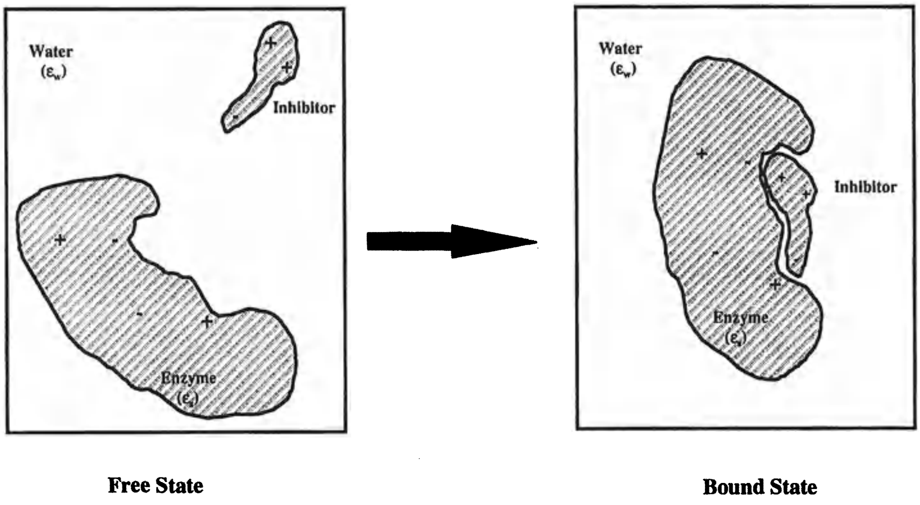


Fig. 1. Schematic representation of enzyme–inhibitor binding in an aqueous environment. ε_s and ε_w represent the dielectric of the solute and water respectively.

Table 1 *Values for the cratic (dilutional) term plus the loss in rigid-body translational and rotational entropy on complexation* ($-T\Delta S_{rigid}$)

Authors	$-T\Delta S_{rigid}$ (kcal mol^{-1})	Reference
Andrews et al.	14.0	[67]
Williams et al.	13–18	[70]
(also gives the enthalpy term)	(− 1.7)	
Krystek et al.	11.0	[42]
Vajda et al.	9.0	[13]
Böhm	1.3	[68]
Bohacek and McMartin	4.4	[10]
Wallqvist et al.	2.4	[73]
Consensus	12.0	

the water environment. Thus, the difference in potential energy between the states can be defined as

$$\Delta E = \Delta E^{i} + \Delta E^{e} + \Delta E^{w} + \Delta E^{e\text{-}i} + \Delta E^{(ei\text{-}e\text{-}i)\text{-}w} \tag{3}$$

where the last term represents the difference in the interaction of the water environment with the different enzyme and inhibitor species. In rigid systems the conformational strain terms, ΔE^{i} and ΔE^{e}, disappear.

Although this analysis assumes a single configuration for each of the three species, each term should in fact involve a Boltzmann ensemble. Allowing for this the potential energy terms described above can be equated to the enthalpic contribution to binding.

To obtain free energies of binding, we need to supplement the above with entropic contributions. The primary entropic contributions associated with binding are usually partitioned into (i) the loss of cratic (dilutional), translational and rotational entropy, ΔS_{rigid}, (ii) the change in vibrational entropy, ΔS_{vib}, (iii) the loss of conformational entropy, ΔS_{conf}, and (iv) the change in entropy associated with solvation, ΔS_{solv}.

There are wide variations for ΔS_{rigid} used in various applications in the literature and a number are given in Table 1. It is likely that in many instances the change in vibrational entropy compensates for the loss of rotational and translational entropy and thus figures at the upper end of this range are likely to be an overestimate of the combined effect. In any case, in structure-based design applications one is almost always interested in relative ranking of a number of ligands against a single protein target. In this instance the differences in ΔS_{rigid} and ΔS_{vib} between different ligands are likely to be small and so this term is often ignored.

Contributions to ΔS_{conf} potentially arise from both the ligand and the receptor. This term is often ignored or confined solely to the ligand. This may be a justifiable approach in some instances or when one is evaluating relatively similar molecules and binding modes, but as has been shown by Vadja et al. [13] this term can contribute significantly to differences in binding between ligands.

Possibly the most difficult component to quantify rigorously is ΔS_{solv}, or more generally ΔG_{solv}. The latter comprises a number of factors including the changes in self-energy of the solvent (ΔE^{w}), the changes in the energy of interaction between solute and solvent ($\Delta E^{(ei-e-i)-w}$) and the work required to create a cavity in solution. Almost all treatments involve some measure of molecular or accessible surface area or volume. How these are used depends on whether some aspects of the solute–solvent interaction have been catered for elsewhere in the method and this has led to some confusion between different authors as to what their 'solvation' term actually represents. This will be discussed in more detail later.

Thus when pulling together the components of relative binding energies, for which ΔS_{rigid} and ΔS_{vib} are ignored, we obtain:

$$\Delta\Delta G_{binding} = \Delta\Delta E_{strain} + \Delta\Delta E_{int} + \Delta\Delta G_{solv} - T\Delta\Delta S_{conf} \tag{4}$$

where

$$\Delta E_{strain} = \Delta E^{i} + \Delta E^{e} \tag{5}$$

$$\Delta G_{solv} = \Delta E^{(ei-e-i)-w} + \Delta E^{w} - T\Delta S_{solv} \tag{6}$$

Thus, Eq. 4 represents a version of what Ajay and Murcko [12] referred to as the 'master equation'. As they point out there is no rigorous partitioning and the one shown above is just one version which we have found useful in our thinking. As indicated earlier, all the terms in the equations above refer to ensemble averages. For example, the solvation energy for a flexible ligand should be derived from an ensemble of conformations rather than a single low-energy conformation. In practice, this is often ignored despite work that has shown its importance in certain cases [13].

3. Approaches

In this section we examine a range of approaches to tackle the problem outlined above. These range from explicit simulations which are time-consuming and impractical for large numbers of diverse compounds, through to methods which partition the free energy as described above and then attempt to estimate each component using theoretically based methods. At the other end of the spectrum are regression or statistically based models of binding affinity which are extremely fast and which can handle a very large number of molecules. The approach adopted will depend on the nature of the problem, the accuracy required and the number of molecules which need assessing. A summary of the range of methods used is given in Tables 2 and 3. The classification of methods into explicit, partitioning and statistical/regression-based is somewhat arbitrary, but is useful for reviewing the different approaches.

3.1. Explicit simulation methods

These are nonpartitioning methods that do not use Eq. 4 explicitly and that rely on molecular dynamics or Monte Carlo methodology to generate a thermodynamic

Table 2 *Partitioning approaches to the estimation of the free energy of binding*

Authors	ΔG_{int}	ΔG_{solv}	ΔE_{strain}	ΔS_{conf}	Comments	Reference
Zhang and Koshland	Electrostatic interaction energy from FDPB ($e_s = 2$)	Polar term from FDPB. Nonpolar using ASA and a coefficient of 20 cal mol^{-1} Å^{-2}	None	None	Limited data set of nine mutants of isocitrate dehydrogenase each with seven substrates, but correlation with observed relative binding energies was very good (average SD = 0.4 kcal mol^{-1})	[36]
Hecht et al.	Electrostatic interaction energy from FDPB ($e_s = 4$)	Polar term from FDPB. Nonpolar using ASA and a coefficient of 59.8 cal mol^{-1} Å^{2} and curvature corrected	None	None	Study of drug–DNA binding	Personal communication
Shen and Quiocho	Electrostatic interaction energy from FDPB ($e_s = 3$)	Polar term from FDPB. Assumes differences in the non polar term are small within series	None	None	Arabinose binding protein and sulphate binding protein	[35]
Engels et al.	Electrostatic interaction energy from FDPB ($e_s = 20$)	Nonpolar using ASA and a coefficient of 9 cal mol^{-1} Å^{-2}	None	None	Thrombin– ,and trypsin–inhibitor complexes. Dielectric and nonpolar scaling terms derived by goodness of prediction of binding energy	[37]
Timms and Wilkinson	Electrostatic interaction energy from FDPB ($e_s = 4$) plus vdW energy	Nonpolar using MSA and a coefficient of 9 cal mol^{-1} Å^{-2}	None	Uses 0.7 kcal mol^{-1} per rotatable bond	Thrombin–inhibitor complexes	This work

Table 2 (*continued*)

Authors	ΔG_{int}	ΔG_{solv}	ΔE_{strain}	ΔS_{conf}	Comments	Reference
Janin	vdW and electrostatic interaction energy ($\varepsilon = 3$)	ASP approach – derived from vacuum/water hydration energies	None	Uses Pickett and Sternberg empirical scale for side-chain conformational entropy	Barnases–barstar, lysozyme–antibody and subtilisin–eglin complexes. Also includes changes in rotational, translational and vibrational entropy in an attempt to estimate absolute affinities	[55]
Vadja et al.	Electrostatic interaction energy ($\varepsilon = r$)	ASP (based on ensemble average) – derived from octanol/ water partitioning	Non-vdW energy only included (based on ensemble average derived for free state)	Side-chain and backbone entropy included. P_i derived from X-ray structures	Demonstrates the importance of the polar contribution to solvation and the need to deal with ligand flexibility. Studies with streptavidin, MHC and serine proteases	[13]
Krystek et al.	Electrostatic interaction energy only ($\varepsilon = 4r$)	Nonpolar using ASA and a coefficient of 25 cal mol^{-1} $Å^{-2}$	None	Uses 0.6 kcal mol^{-1} per rotatable bond	Serine protease–protein inhibitor and antibody–antigen complexes	[42]
Grootenhuis and van Galen	vdW and electrostatic interaction energy ($\varepsilon = r$)	None	None	None	Thrombin–inhibitor complexes – average error for predicted pK_i was 1.45, although with three compounds removed this fell to 0.97	[64]
Perakyla and Pakkanen	QM (6-31G) calculation plus classical long-range electrostatics ($\varepsilon = r$)	AM1 or SM2 solvation model	None	None	Arabinose binding protein–sugar complexes	[65]
Kroeger-Smith et al.	vdW and electrostatic interaction energy	Enthalpic term calculated with a 6Å layer of water molecules	Full molecular mechanics calculation of strain eenrgy	None	HIV reverse transcriptase–inhibitor complexes. All calculations were carried out on a single system configuration	[66]

Table 2 (*continued*)

Authors	ΔG_{int}	ΔG_{solv}	ΔE_{strain}	ΔS_{conf}	Comments	Reference
Kurinov and Harrison	vdW and electrostatic interaction energy	None	Full molecular mechanics calculation of strain energy	None	Trypsin–inhibitor complexes. Incorporates fast orientational search. Flexibility handled by interspersing MD calculations	[41]
Wang et al.	vdW and electrostatic interaction energy ($\varepsilon = 80$)	Method of Privalov and Makhatadze used	Full molecular mechanics calculation of strain energy	Configurational entropy calculated directly from partition function. Obtains 0.48 kcal mol^{-1} per rotatable bond	Calculation of relative binding energies of thrombin–hirudin analogues using effective harmonic well method. Partition function of bound and free states calculated using systematic search techniques	[57]
Luty et al.	vdW and electrostatic interaction energy ($\varepsilon = r$)	Occupancy desolvation method (related to the ASP approach)	Not relevant in example used	None	Trypsin–benzamidine complex. Pre–calculated grid to enable the method to be used in docking calculations. Used for ranking hits rather than predicting free energy of binding	[58]
Viswanadhan et al.	vdW and electrostatic interaction energy ($\varepsilon = 1$)	Enthalpic term calculated by interaction with bath of water molecules. Nonpolar contribution using an atom-based hydrophobic interaction energy	Full molecular mechanics calculation of strain energy	None	HIV protease–inhibitor complexes. Regression analysis required to obtain realistic $\Delta\Delta G$ values	[43]

Table 2 (*continued*)

Authors	ΔG_{int}	ΔG_{solv}	ΔE_{strain}	ΔS_{conf}	Comments	Reference
Holloway et al.	vdW and electrostatic interaction energy ($\varepsilon = 1.5$)	None	None	None	HIV protease–inhibitor complexes. Regression-based. Thirty-four complexes used to produce correlation, 16 for prediction. Cross-validated correlation around 0.76. Average error on prediction of pK_i was 1.01 (0.79 with one compound omitted)	[40]
Wilson et al.	vdW and electrostatic interaction energy ($\varepsilon = r$)	ASP method – derived from octanol/water partitioning	None	None	α-Lytic protease–substrate binding. Relative weights for nonbonded and solvation terms derived empirically ($w_{NB} = 0.031$, $w_{sol} = 1.98$). All accessible side-chain rotamers contribute to free energy calculation	[56]

ensemble from which relative free energies of binding can be derived. The two most frequently used methods are free energy perturbation (FEP) and thermodynamic integration (TI). In both these approaches, a series of nonphysical perturbations to the free and bound inhibitor are carried out and relative free energies are obtained using the relevant thermodynamic cycle. These approaches have attracted considerable attention over recent years and have achieved some success both in predicting relative binding energies and in providing insight into the nature of molecular recognition [14,15]. However, the methodology has a number of limitations which restrict their value in structure-based design [16]. In particular, to achieve convergence in the simulations, only relatively small differences between ligands can be investigated and, even in these cases, they are extremely computer intensive. Although computers will get faster and the methodology continues to improve, it is difficult to believe that these methods will be of central value to the structure-based design process where one usually needs to investigate the potential of a wide variety of ligands and where one is looking for the results from calculations to be obtained at least 1–2 orders of magnitude faster than the compound can be synthesised.

Two approaches that attempt to overcome the limitations of FEP/TI approaches are worthy of mention. Van Gunsteren and co-workers [17–19] have explored the possibility of predicting free energy differences between a manifold of molecular states from a single simulation representing one reference state using extrapolation methods. Acqvist and co-workers [20–23] have used a linear approximation procedure (LIE) to estimate absolute binding free energies from two aqueous simulations – one involving the inhibitor alone and the other involving the protein–inhibitor complex.

The underlying principle of extrapolation methods is that the effect of a range of changes to a reference ligand can be estimated from a single simulation of that reference ligand. Van Gunsteren and co-workers [17–19] have carried out some key studies to understand and extend the applicability of this approach. In their most recent paper in this area, they showed that, using the free energy perturbation formula together with an appropriate reference state, it was possible to reproduce, with a reasonable degree of accuracy, the 'exact' free energies calculated using normal thermodynamic integration methodology. Where changes in the ligands involved the deletion or creation of atoms, it was necessary to employ a biased, nonphysical reference state generated using a reference ligand which included soft core dummy interaction sites. Although no data have yet been published on protein–ligand applications using this latest variation of the methodology, the approach holds some promise.

In the LIE procedure, the binding energy is evaluated from the difference in ligand–environment interaction energy (both electrostatic and van der Waals energy) in the bound and free states. The environment in the simulation of the enzyme-inhibitor complex includes the protein as well as the solvent. The absolute free energy of binding is given by

$$\Delta G_{\mathrm{bind}} = 1/2 <(V^{\mathrm{el}}_{\mathrm{bound}} > - < V^{\mathrm{el}}_{\mathrm{free}} >) + \alpha(< V^{\mathrm{vdW}}_{\mathrm{bound}} > - < V^{\mathrm{vdW}}_{\mathrm{free}} >) \qquad (7)$$

where α is an empirical constant, the optimal value of which has been determined by Aquist to be 0.161. It is not clear why the procedure should be able to give absolute

Table 3 *Regression-based approaches to the estimation of the free energy of binding*

Authors	ΔG_{int}	ΔG_{solv}	ΔE_{strain}	ΔS_{conf}	Comments	Reference
Horton and Lewis	Polar and vdW interactions subsumed into solvation terms	Hydrophobic and polar solvation terms based on accessible surface area and parameterised coefficients	None	None	Uses 6.2 kcal mol^{-1} to correct for rigid-body entropy losses on binding	[71]
Böhm	Ionic and neutral hydrogen bonds	Lipophilic contribution based on coarse-grid-based accessible area	None	Uses 0.3 kcal mol^{-1} per rotatable bond	Accurate to ca. 1.4 log units for affinity. Incorporates an intercept/rigid-body entropy contribution of 1.3 kcal mol^{-1}	[68]
Ortiz et al.	Coulombic and vdW terms	None explicit	Intramolecular terms both geometric and non-bonded	Torsional terms implicit in intramolecular energy terms	Uses ligand–receptor interaction energies as variables in a PLS generation of a system-specific regression model	[79]
Head et al.	Electrostatic and vdW interactions augmented by 'steric fit' term	Partition coefficient plus lipophilic and hydrophilic contact surface areas	Difference between energy of bound conformation and nearest local minimum in generalised Born/surface area solvent model	Uses 0.6 kcal mol^{-1} per rotatable bond including (n – 4) bonds for alicyclic rings of size n	Uses the various terms in a PLS or a neural network generation of a system specific model	[80]
Bohacek and McMartin	Hydrogen bond term	Hydrophobic term subsumes dispersion interaction energy, etc.	None	None	Linear regression approach for log K_i. Large intercept of 3.16	[10]

Table 3 (*continued*)

Authors	ΔG_{int}	ΔG_{solv}	ΔE_{strain}	ΔS_{conf}	Comments	Reference
Verkhivker et al.	Knowledge-based mean-field potentials including interactions with bound water	Polar and nonpolar accessible surface area	None	Protein side-chain entropies + 0.6 kcal mol^{-1} per non-peptide rotatable bond	Uses 15 kcal mol^{-1} for the rigid-body entropic contribution	[74]
Wallqvist et al.	Uses vdW-type steric weighting of contact preferences between surface patches	Term subsumed into contact preferences	None	None	Incorporates a cratic term of 2.39 kcal mol^{-1}	[73]
Sobolev et al.	Uses a complementarity function based on 'legitimate' contact surface area + repulsion and non-ideal hydrogen bond penalty terms	Incorporated into the complementarity function	None	None	Normalises the complementarity function by dividing the surface area by that of the isolated ligand	[72]
Jain	Weighted sum of Gaussian and sigmoid functions for hydrophobic and polar terms + charge components	Hydrophobic function + any loss of hydrogen bond capacity on binding	None	Uses 0.2137*2.303RT = 0.3 kcal mol^{-1} per rotatable bond	Rigid-body entropic contribution for binding the ligand is given as 1.4 log(ligand MW) kcal mol^{-1}	[78]

binding energies as claimed by the authors, when the interactions of the free protein and solvent are ignored, or why α appears to be transferable to different proteins. One could classify this method as a partitioning approach (*vide infra*). The value of α would then equate to the scaling factor used in other methods which correlates the surface area or volume to the dispersion-repulsion and cavity terms. Acqvist has reported a number of protein systems where reliable estimates of binding energies have been obtained, including calculations on HIV protease [20], glucose/galactose binding protein [21], endothiapepsin [22] and trypsin [23]. This method does have the advantage of working from an ensemble of molecular conformations and thus molecular flexibility can be properly accounted for. In addition, it appears applicable to a wide range of inhibitors for any given protein. More recently, Paulsen and Ornstein have applied the method with some success to P450cam. In this work the optimal value of α was 1.043 [24]. It is not clear at this stage whether this difference reflects the different force fields used by the two groups or the nature of the protein–ligand complexes investigated. Jorgensen has applied the method to the estimation of solvation free energies and found the optimal value of α to range from 0.3–0.6 [25]. Despite the reservations about the transferability of α, the method holds promise and deserves further investigation.

3.2. Partitioning approaches

There has been much discussion in the literature about the validity of partitioning the free energy calculated by FEP/TI methods [26,27]. The argument is that, although free energy is a state function and therefore the pathway one chooses to calculate it is irrelevant, this does not apply to the components of the free energy as these are not state functions. The same arguments can be made against the 'partitioning' approach to free energy calculation. This is implicit in the acceptance of the arbitrary nature of the 'master equation'. However, analysis of binding free energy in this way provides valuable insight into the principles of the binding process and the limitations and applicability of the method. So although the approach may not be rigorous, it has proven of enormous utility.

In this section we consider some of the key contributors to binding free energy which were captured in Eq. 4. In particular, we consider the polar and nonpolar contributions to the changes in solvation energy and in the conformational entropy on binding. Some of the applications of partitioning methods are summarised in Table 2.

3.2.1. Nonpolar contributions to binding

There has been considerable discussion in the literature as to the most appropriate method of calculating the nonpolar contribution to binding. Most methods use relationships derived using solvent accessible surface area (SAS), although Jackson and Sternberg [28] and Pitarch et al. [29] have recently suggested that the use of molecular surface area (MSA), unlike SAS, leads to a potential of mean force that corresponds to that observed in simulations and therefore is more appropriate.

a)

$$\begin{tikzcd}
E_g \arrow[d, "\Delta G^{E}_{solv}"] & + & I_g \arrow[r, "\Delta G_{int}"] \arrow[d, "\Delta G^{I}_{solv}"] & E\text{-}I_g \arrow[d, "\Delta G^{EI}_{solv}"] \\
E_w & + & I_w \arrow[r, "\Delta G_{bind}"'] & E\text{-}I_w
\end{tikzcd}$$

b)

$$\begin{tikzcd}
E_g \arrow[d, "\Delta G^{E}_{solv,o}"] & + & I_g \arrow[r, "\Delta G_{int}"] \arrow[d, "\Delta G^{I}_{solv,o}"] & E\text{-}I_g \arrow[d, "\Delta G^{EI}_{solv,o}"] \\
E_o \arrow[d, "\Delta G^{E}_{part}"] & + & I_o \arrow[r, "\Delta G_{int,o}"] \arrow[d, "\Delta G^{I}_{part}"] & E\text{-}I_o \arrow[d, "\Delta G^{EI}_{part}"] \\
E_w & + & I_w \arrow[r, "\Delta G_{bind}"'] & E\text{-}I_w
\end{tikzcd}$$

Fig. 2. (a) A thermodynamic cycle describing enzyme–inhibitor binding based on gas to aqueous phase transfer from which the nonpolar contribution to binding can be extracted using relative hydration free energies. (b) A thermodynamic cycle based on oil to aqueous phase transfer from which a nonpolar contribution to binding can be extracted using relative partition free energies. The relationship to the gas-phase cycle is also shown.

There are two thermodynamic cycles which are commonly used to define the nonpolar contribution to binding. The first involves the use of gas to aqueous phase transfer and the second involves nonpolar liquid to aqueous phase transfer. The former is summarised by the thermodynamic cycle shown in Fig. 2a. The binding free energy in water is given by the gas-phase interaction energy plus the difference in hydration energy (ΔG_{solv}) between the bound and free species.

$$\Delta G_{bind} = \Delta G_{int} + \Delta G^{EI}_{solv} - \Delta G^{E}_{solv} - \Delta G^{I}_{solv} \tag{8}$$

The hydration energy contribution can be considered to comprise a term involving the cost of cavity formation and two terms, a dispersion-repulsion and an electrostatic term, describing the favourable interaction of the solutes with water.

$$\Delta G_{solv} = \Delta G_{cav} + \Delta G_{d\text{-}r} + \Delta G_{ele} \tag{9}$$

The gas-phase interaction term involves dispersion-repulsion and Coulombic interaction terms between the interacting molecules. The subscripts vdW and coul are used to distinguish these from the interactions between solute and solvent.

$$\Delta G_{int} = \Delta G^{EI}_{coul} + \Delta G^{EI}_{vdW} \tag{10}$$

Thus,

$$\Delta G_{\text{bind}} = \Delta G^{\text{EI}}_{\text{coul}} + \Delta G^{\text{EI}}_{\text{vdW}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{cav}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{d-r}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{ele}} \tag{11}$$

The superscript EI-E-I indicates the difference for each energy contribution between the three species. Some authors make the assumption that, in a well-packed system, the dispersion-repulsion energy will be conserved and thus the $\Delta\Delta G_{\text{d-r}}$ and ΔG_{vdW} terms will cancel, giving

$$\Delta G_{\text{bind}} = \Delta G^{\text{EI}}_{\text{coul}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{cav}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{ele}} \tag{12}$$

If the dispersion energy is included explicitly, then Eq. 11 can be rewritten as

$$\Delta G_{\text{bind}} = \Delta G^{\text{EI}}_{\text{coul}} + \Delta G^{\text{EI}}_{\text{vdW}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{np}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{ele}} \tag{13}$$

where

$$\Delta G_{\text{np}} = \Delta\Delta G^{\text{EI-E-I}}_{\text{cav}} + \Delta\Delta G^{\text{EI-E-I}}_{\text{d-r}} \tag{14}$$

This is convenient representation because, for nonpolar molecules with no dipolar nature (i.e. $\Delta G_{\text{ele}} = 0$), differences in hydration free energy solely reflect differences in cavitation and dispersion-repulsion energies and an empirical relationship with solvent accessible surface area can be derived.

$$\Delta G_{\text{solv}} = \Delta G_{\text{np}} = \gamma_{\text{solv}}\,\text{MSA} + C_{\text{solv}} \tag{15}$$

Hydration free energies for alkanes are given in Table 4. These data indicate that the value of γ_{solv} should be 5.8 cal mol^{-1} Å^{-2}. This corresponds to the relationship used

Table 4 *Solvation energies and transfer free energies for alkanes – relationship with surface area*

Alkane	MSA[a]	SAS[a]	ΔG_{part}	$\Delta G^{\text{a}}_{\text{part}}$	$\Delta G^{\text{b}}_{\text{solv}}$	$\Delta G^{\text{c}}_{\text{solv}}$
	(Å^2)	(Å^2)	(octanol/water)	(alkane/water)	(gas/alkane)	(gas/water)
			(kcal mol^{-1})	(kcal mol^{-1})	(kcal mol^{-1})	(kcal mol^{-1})
Ethane	69	180	2.47		– 1.04	1.77 (3.34)
Propane	89	211	3.22	4.05	– 1.80	1.98 (4.07)
Butane	109	242	3.92	4.92	– 2.55	2.15 (4.89)
Pentane	129	274	4.62	5.82	– 3.30	2.34 (5.57)
Hexane	149	306	5.32		– 4.06	2.55 (6.29)
fCH$_2$	20	31	0.72	0.89	– 0.75	0.18 (0.74)
γ (MSA)[d] (cal mol^{-1} Å^{-2})			36.1	44.5	– 37.5	9.0 (37)
γ (SAS)[d] (cal mol^{-1} Å^{-2})			23.2	28.7	– 24.2	5.8 (24)

[a] Data presented are for gas/n-hexane transfer free energy. These data together with the surface area data are from Jackson and Sternberg [28].

[b] Data presented are for cyclohexane/water transfer free energy.

[c] Figures in parentheses are values of corrected hydration energy. Data from Sharp et al. [31].

[d] The γ value is derived from the relationship between solvation and transfer free energy and molecular and solvent accessible surface area. Thus, from octanol and hexane transfer free energies γ_{part} is obtained, from hexane solvation energies $\gamma_{\text{solv,o}}$ is obtained, and from the hydration energies γ_{solv} is obtained.

by Sitkoff et al. [30] in their work on the estimation of hydration energies using continuum methods. Sharp et al. [31] have argued that one needs to correct the experimental values of transfer and solvation free energies for changes in volume entropy. For hydration free energy, they obtained a corrected value for γ_{solv} of 24 cal mol^{-1} Å^{-2}. It is not absolutely clear whether the corrections derived by Sharp et al. are theoretically sound for small molecules. Jackson and Sternberg [28] have argued that they are not. It is likely that the 'true' value for γ_{solv} lies somewhere between 5.8 and 24 cal mol^{-1} Å^{-2}.

The second thermodynamic cycle that is commonly used to derive estimates for the nonpolar contribution is that involving transfer between water and a nonpolar solvent. This is based on the belief that the core of a protein is more like a hydrophobic solvent than the gas phase. The relevant thermodynamic cycle and its relationship to the gas-phase cycle can be seen in Fig. 2b.

Similarly to the gas-phase model described above, the binding energy in the oil phase can be defined by

$$\Delta G_{\mathrm{bind}} = \Delta G^{\mathrm{EI}}_{\mathrm{int,o}} + \Delta\Delta G^{\mathrm{EI\text{-}E\text{-}I}}_{\mathrm{part}} \tag{16}$$

As described above the solvation energy comprises a cavity energy term and dispersion-repulsion and electrostatic solvation energy terms. The transfer free energy, described by the subscript part, is the difference between these contributions in the two solvents.

$$\Delta G_{\mathrm{part}} = \Delta G_{\mathrm{cav,part}} + \Delta G_{\mathrm{d\text{-}r,part}} + \Delta G_{\mathrm{ele,part}} \tag{17}$$

The dispersion-repulsion energy for a molecule in any solvent is normally considered to be the same and is thus ignored (although this need not be the case). Thus,

$$\Delta G_{\mathrm{bind}} = \Delta G^{\mathrm{EI}}_{\mathrm{coul}} + \Delta G^{\mathrm{EI}}_{\mathrm{vdW}} + \Delta\Delta G^{\mathrm{EI\text{-}E\text{-}I}}_{\mathrm{cav,part}} + \Delta\Delta G^{\mathrm{EI\text{-}E\text{-}I}}_{\mathrm{ele,part}} \tag{18}$$

The nonpolar term (which now is solely a cavitation term) can be estimated through the relationship between surface area and water/hydrocarbon transfer free energy for alkanes.

$$\Delta G_{\mathrm{np,part}} = \gamma_{\mathrm{part}}\mathrm{MSA} + C_{\mathrm{part}} \tag{19}$$

It is worth considering the relationship between models produced from the two cycles. Both can be represented by the general equation

$$\Delta G_{\mathrm{bind}} = \Delta G^{\mathrm{EI}}_{\mathrm{coul}} + \Delta G^{\mathrm{EI}}_{\mathrm{vdW}} + \Delta\Delta G^{\mathrm{EI\text{-}E\text{-}I}}_{\mathrm{ele}} + \Delta\Delta G^{\mathrm{EI\text{-}E\text{-}I}}_{\mathrm{np}} \tag{20}$$

The Coulombic and electrostatic terms in the two models will vary in a manner defined by the solute dielectric. The nonpolar contribution to binding will be more favourable using the hydrocarbon model, because no allowance for the loss of favourable solute–water dispersion interactions is required. The transfer free energy values for alkanes are given in Table 4. The frequently quoted values for γ_{part} used by several groups are derived from experimental transfer free energy values in octanol. This is understandable given the extensive data that exist for octanol partition coefficients. The values of 23.2 and 28.7 cal mol^{-1} Å^{-2} obtained for octanol and hexane, respectively, are in agreement with other work [32,33] and are significantly different from the uncorrected value obtained from hydration free energies (5.8 cal mol^{-1} Å^{-2}). As can be seen from the table, the difference between these terms derives from the solvation energy of the solute in the hydrocarbon solvent.

As to which model for nonpolar binding to use, the issue becomes to what extent the method used to evaluate the intrinsic energy of interaction mimics the interaction in the hydrocarbon phase versus the gas phase. Where the intrinsic interaction energy is estimated using standard molecular mechanics force fields with gas-phase charges and a dielectric of 1, then the gas-phase cycle is more relevant. Alternatively, using continuum methods with a solute dielectric of 4 to determine interaction energy and electrostatic desolvation, it may be that the hydrocarbon cycle is more appropriate. It is interesting to note that using 'corrected' hydration energy, the value for γ_{np} obtained is very close to that obtained from the transfer free energy data. In the above we have referenced the values of γ corresponding to the SAS; however, equivalent values can be derived based on MSA and these are also given in Table 4.

This assumes that the complete analysis involves the explicit calculation of the van der Waals interaction energy between the species in the bound state. If this is not the case, the dispersion-repulsion term within the solvation energy contribution should also be dropped. In this case we are left with a 'nonpolar' contribution comprising only a cavity term and the surface area scaling factor should be somewhat larger.

In summary, there appears to be no universally accepted rigorously defined value for γ. Values of between 20 and 30 cal mol^{-1} Å^{-2} seem to be most commonly used. Given the uncertainties, authors are justified in using an empirically defined value that best suits the system under investigation.

3.2.2. Electrostatic contributions

Continuum methods: The use of continuum approaches to the estimation of solvation, conformational and binding free energies has become increasingly popular and well validated in recent years. Numerical approaches to the solution of the Poisson–Boltzmann equation have been the cornerstone of this approach, and a seminal paper by Gilson and Honig in 1988 [34] has formed the basis for much of the more recent work. The approach consists, in essence, of three finite difference Poisson–Boltzmann (FDPB) calculations (two if only relative binding energies to the same rigid protein are required), on the inhibitor and protein when free in solution and the protein–ligand complex. Conceptually, the overall binding energy can be seen as comprising the Coulombic interaction energy as a result of binding, the difference in reaction field energy (or electrostatic self-energy) between the free and bound states and the nonpolar contribution. The latter is calculated as described in the preceding section, based on molecular or solvent accessible surface area. As indicated previously, unless the surface area term is scaled to represent the cavity only (rather than the complete nonpolar contribution), the van der Waals energy of interaction should also be included. In fact, this has rarely been the case in examples published to date. In our experience the short-range nature of this energy and its sensitivity to the precise position of the atoms in the protein–ligand interface make it difficult to include

Table 5 *Prediction of relative binding affinity of thrombin/inhibitor complexes*

Inhibitor	$\Delta\Delta G_{coul}$	$\Delta\Delta G_{ele}$	$\Delta\Delta G_{np}$	$-T\Delta\Delta S_{conf}$	$\Delta\Delta G_{calc}$	$\Delta\Delta G_{obs}$	Error
1(NAPAP)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	− 0.3	− 0.7	1.8	0.0	0.8	0.7	0.1
3	0.1	0.7	− 0.1	0.0	0.7	1.7	− 1.0
4	24.3	− 21.6	1.5	0.0	4.2	5.6	− 1.4
5	11.5	− 9.1	− 1.8	0.7	1.3	0.4	0.9
6	10.1	− 10.5	1.6	0.0	1.2	0.9	0.3
7	10.2	− 10.6	2.5	0.0	2.1	1.7	0.4
8	11.4	− 11.9	3.1	0.0	2.6	2.0	0.6
9	11.3	− 8.9	0.5	0.7	3.8	4.7	− 0.9

All energies (in kcal mol^{-1}) are given relative to the thrombin inhibitor NAPAP. $\Delta\Delta G_{coul}$ is the relative coulombic interaction energy. $\Delta\Delta G_{ele}$ is the relative electrostatic desolvation energy. Electrostatic calculations carried out with DELPHI [34]. $\Delta\Delta G_{np}$ is the relative non-polar contribution to binding based on accessible surface area and a scaling factor of 40 cal mol^{-1} Å^{-2}. $\Delta\Delta S_{conf}$ is the relative contribution due to loss in conformational entropy. $\Delta\Delta G_{obs}$, the experimental values, are taken from Grootenhuis and van Galen [64].

explicitly. For this reason we tend to adjust the scaling factor for the surface area term to allow for this contribution.

In work in our laboratory, this approach has been used successfully against a range of protein–ligand systems including trypsin, thermolysin, barnase and thrombin. The results for a series of thrombin inhibitors, based on NAPAP and argatroban, are given in Table 5. We have used a dielectric of 4 in the FDPB calculation together with solvent accessible surface area and a scaling parameter γ of 40 cal mol^{-1} Å^{-2} for the nonpolar contribution. The relative binding energies are predicted to a mean error of 0.7 kcal mol^{-1}, which is generally the level of accuracy we have seen with the other protein systems. These results emphasise the importance of the electrostatic desolvation term. The overall electrostatic contribution (i.e. Coulombic plus electrostatic desolvation term) tends to be finely balanced and, in our experience can, often be an unfavourable contribution to binding. This demonstrates the need for polar groups to achieve a compensating favourable interaction on burial from solvent, in order to achieve a reasonable binding affinity.

Shen and Quiocho [35] obtained a similar level of accuracy with their work on sulphate binding protein and arabinose binding protein, although they did not take into account the nonpolar/cavity term in their analysis. Zhang and Koshland [36], in their calculations on seven different substituted R-malate substrates bound to nine different mutants of isocitrate dehydrogenase, obtained a standard deviation between predicted and observed relative binding energies of 0.4 kcal mol^{-1}. Engels et al. [37] obtained an accuracy of 1.1 kcal mol^{-1} in their work on thrombin and trypsin inhibitors, although the results were very dependent on the choice of solute dielectric and the nonpolar scaling parameter γ. They chose a solute dielectric of 20 and a γ value of 9 cal mol^{-1} Å^{-2}. Interestingly, in this work the electrostatics were only responsible for around 10–20% of the variance in data. As part of his computational combinatorial ligand design (CCLD) program, Caflisch [7] has developed this approach further using a full molecular mechanics description to obtain a gas-phase interaction energy, together with FDPB calculations to estimate the solvent-shielded intermolecular association energy and the electrostatic desolvation energy. He used the changes in solvent accessible surface area and a scaling factor of 25 cal mol^{-1} $\mathrm{Å}^{-2}$ to estimate the nonpolar contribution.

We have found the results of this approach to be relatively insensitive to the precise charges used for ligand and protein. We have used PARSE charges [30] for the protein and CHARMM-based [38] charges for the ligand and PARSE vdW radii for all atoms. The results, however, are sensitive to the precise positioning of the ligand. We always relax the complex through 500 steps or so of energy minimisation prior to carrying out the FDPB calculation. The protein is usually kept fixed during this process, although hydroxyl hydrogens are allowed to move. One of the drawbacks of this approach is that the geometry, binding conformation and orientation of ligands need to be generated using one theoretical model (i.e. standard molecular mechanics) and then the binding energy calculation carried out using another. This can lead to some inconsistencies which need to be taken into account. The treatment of crystallographic water that appears to be playing a structural role in the protein or protein–ligand complex is another unresolved problem. We have found that it is usually necessary to include these explicitly in the calculation as part of the protein. Numerical accuracy of the FDPB algorithm can also be a problem for large proteins.

These calculations are almost always carried out on a single conformation of protein and ligand, as it is currently not possible to be able to carry out calculations on a realistic ensemble. Each calculation usually takes around 10 min CPU on SGI R4400. Thus, the method is not well suited to ranking large hit lists of molecules generated by structure-based design programs, but is probably useful for problems involving a moderate number of ligands (up to 50). A recent publication by Schaeffer and Karplus [39] describes a new analytical treatment of continuum electrostatics (ACE) which, if it proves to be generally applicable, is reported to be sufficiently fast to use alongside the most prolific structure-based design methodology.

Atomic solvation parameter approaches: The most common alternative approach to the use of the Poisson–Boltzmann methodology is the use of standard molecular mechanics potentials, or more occasionally quantum mechanical methods, to estimate relative binding affinities. There are examples where a reasonable relationship between gas-phase binding energies, incorporating dielectric screening usually through the use of a distance-related dielectric, has been obtained [40,41,64]. The slopes of the lines fitting experimental and calculated binding affinities are always significantly greater than 1 in these instances. Given the omission of electrostatic solvation terms and therefore the vast overestimation of the electrostatic contribution to binding, these relationships must be considered exceptional and unlikely to be transferable to a wide range of ligands or proteins.

It is not appropriate to use the nonpolar solvation contribution alone in conjunction with an estimation of gas-phase interaction energies, even when the electrostatic contribution is reduced using an arbitrary choice for dielectric, although some workers have taken this approach [42,43]. Some account must be made for the electrostatic aspects of desolvation. Other than the use of continuum methods, the most common approach is the use of the atomic solvation parameter (ASP) method [31, 44–52]. Here the solvation contribution is estimated using the relationship

$$\Delta G_{solv} = \sum_{i} \sigma_i A_i \tag{21}$$

where A_i is the solvent accessible surface area of each atom while σ_i is the atomic solvation parameter for that atom. Unlike in the approaches where the nonpolar solvation contribution to binding is estimated separately from the polar solvation contribution, each atom type is associated with a different σ_i. These parameters have been derived from transfer free energy data, including water/octanol and gas/water partitioning. The nonpolar (and primarily entropic) contribution results from the positive values of σ_i for carbon atom types, while the polar (and primarily enthalpic) solvation contribution results from negative values of σ_i for polar atom types.

It is worth comparing the two approaches to the calculation of overall solvation contribution. For the two methods to give equivalent results, then

$$\sum \sigma_i A_i \sim \sum \gamma_{np} A_i + \Delta G_{es(FDPB)} \tag{22}$$

For carbon one would anticipate that σ_i will be close to γ_{np}. As indicated by Smith and Honig [53] however, the FDPB/γ approach is more physically realistic. It cannot be strictly valid to use surface area to estimate the electrostatic contribution to solvation. For example, using the ASP approach the electrostatic free energy does not change once an atom is buried, yet in reality it is dependent on the depth beneath the surface and the shape of the interface. Because the ASP approach cannot account for dielectric screening, it is usually used in combination with an effective dielectric (e.g. $\varepsilon = r$), whose value is arbitrary. As a result, there can be difficulties in scaling ASP solvation energies to the rest of the force field. Despite these shortcomings, this approach has found great utility in a number of areas [13,54–59]. Perhaps the greatest practical limitation is the need for additional atom types to reproduce accurately the solvation energy of the range of molecules encountered in a drug discovery project and the limited solvation energy data that are available to derive parameters for these new atom types.

3.2.3. Conformational entropy and strain energy

Many workers make the assumption that both ligand and protein are rigid and thus ignore contributions arising from conformational strain energy and conformational entropy. While this assumption can be justified on occasion either because both protein and ligand are rigid or because within a series the relative binding energy is not affected by these contributions, in general these aspects must be considered. Usually any strain energy associated with the protein is ignored, the assumption being made that all accessible conformations in both the free and complexed states are equi-energetic. For the ligand, the strain energy must be estimated based on the ensemble of conformations observed in the free state.

$$\Delta E_{strain} = E_b - < E_f > \tag{23}$$

It is worth emphasising at this point that when a ligand or protein displays flexibility, this should also be incorporated into the solvation energy contributions, which should be calculated from a representative conformational ensemble. Vajda et al. [13] demonstrate the importance of this in their studies on MHC–peptide complexes. The generation of the conformational ensemble can be achieved by a systematic conformational analysis or from simulation. In either case the ensemble may be biased by the 'gas-phase' energy function that is usually used and, if so, this needs to be addressed.

If one assumes that all conformational flexibility is lost on complexation, the loss of conformational entropy on binding is given by

$$\Delta S_{conf} = -R\sum p_i \ln p_i \qquad (24)$$

where p_i represents the fractional population of each conformation of both protein and ligand in the free state. p_i can be determined from an ensemble generated as described above or for particular classes of molecules, such as proteins or peptides from experimental data [60–62].

Making the assumption that all conformational flexibility is lost on binding is simplistic and some workers determine the residual conformational entropy in bound protein–ligand complexes by assuming that all conformational entropy is lost when 60% or more of a side chain is buried. Between 0 and 60% buried the 'intrinsic' entropy loss is scaled with the degree of surface area buried [13]. This approach can also be used to determine the conformational entropy associated with the free protein. It is not appropriate to assume that side chains in an uncomplexed protein active site have conformational degrees of freedom equivalent to those available to a linear peptide.

Many studies on peptide ligands focus upon the conformational entropy loss associated with side-chain flexibility; however, for linear peptides, the contribution from backbone flexibility in the unbound state can be equally important. This can be addressed in a similar way. For example, Vadja [63] has developed a method in which 16 regions of the (ϕ, ψ) map are defined as conformational states. The transition probability associated with each state is derived from high-resolution X-ray structures.

Such entropy scales are not appropriate for the more general set of ligands that are encountered as part of a drug discovery project. For these, the entropy contribution must be calculated explicitly as described by the equation above or, alternatively, a more rapid but approximate approach can be used where it is assumed that all accessible states are equi-energetic and that conformational entropy inherent in the unbound state is given by

$$\Delta S_{conf} = -R\sum \ln W \qquad (25)$$

where W is the number of different accessible conformations in the unbound state. This can then be approximated based on the number of single bonds in the ligand and by assuming that each sp^3-sp^3 single bond can adopt three states. This leads to an entropy contribution of $-R\ln 3$ (i.e. 2.2 cal K^{-1} mol^{-1}) per rotatable bond. Despite the approximations inherent in this approach, the value obtained correlates remarkably well with the more sophisticated analyses described above [61].

3.3. *Regression-based approaches*

Where large numbers (>500) of hits from high throughput screening are to be evaluated via ligand docking to a target protein, the speed with which an acceptably accurate estimate of binding affinity can be determined is of overriding importance. These two elements, speed and accuracy, currently are difficult to achieve at the same time. A number of empirical approaches have been developed in which available structural and binding data have been used to calibrate the contribution to affinity of chosen factors in the context of particular functional formats. Some of these are summarised in Table 3.

The most basic of these approaches is that of Andrews et al. [67], where binding data alone were analysed to derive functional group contributions to affinity. A regression equation involving 10 functional groups plus terms for the loss of rigid-body and conformational entropies on binding predicts the binding free energies for 200 complexes to an accuracy of 4.6 kcal mol^{-1}. Although this work was intended to provide insight as to the likely effect of incorporating particular functional groups in drug design, it also provides a useful baseline for binding affinity prediction.

Combining known structures with known binding data, Böhm [68] has generated a regression equation which employs the relative contribution of the number and nature of hydrogen bonds and hydrophobic effects in addition to conformational and rotation/translation entropic terms. For 45 complexes the standard deviation between calculated and observed binding energies was 1.9 kcal mol^{-1}. A similar approach has been taken by Bohacek and McMartin [10], who map the environment of the protein to a cubic grid and evaluate the correspondence between the ligand atoms and the grid point nearest to them to determine the number of hydrogen bonds and lipophilic contacts. The accuracy of binding energy prediction for nine thermolysin complexes was 0.5 kcal mol^{-1}; however, we have extended this approach using the program SCORER, developed at Zeneca, to 102 complexes covering a wider variety of ligands and proteins. This resulted in a prediction accuracy of 2.3 kcal mol^{-1} (see Fig. 3). Another way of evaluating the value of this approach in examining a hit list from a ligand design program might be to ask how many molecules are correctly predicted to be better than 10^{-6} M as these are likely to be the molecules of real interest. SCORER predicts 77% of the molecules correctly with 5% false positives and 18% false negatives.

As further structural and binding data are generated, these regression equations can be rederived, although there is no reason to believe that the accuracy can be improved beyond 1.5–2 kcal mol^{-1}. This results from limitations in the structural data, from inaccuracies and a lack of uniformity in the binding data, in addition to the simplistic approximations of the calculation. An additional limitation of this approach is that unfavourable binding interactions (e.g. a steric clash or a buried polar group) rarely

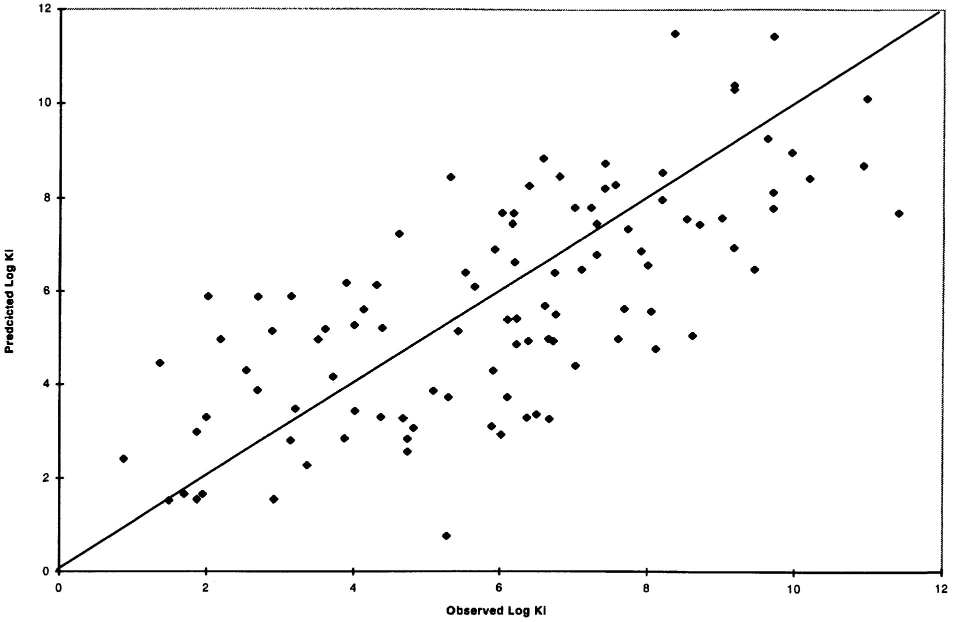


Fig. 3. SCORER predictions versus observed binding affinity for 102 protein–ligand complexes.

occur in known protein–ligand complexes and are thus difficult to include in these methods.

The magnitudes of the contributions of the various factors determined by these regression approaches show an approximate correspondence to values determined by experiment [69] or by theoretical methods [70] (see Table 6). The regression coefficients obtained are, to some extent, dependent on the data set used and, as indicated earlier, these data are probably inadequate for sophisticated statistical analysis. If a set of coefficients based on experimental/theoretical values are adopted, one might generate an equation of similar predictive capability but without any dependence on the calibration data set.

An alternative approach to the calibration of predefined physical contributions to the binding energy relates the interaction energy to the size and nature of the contact surface. Horton and Lewis [71] used the ASPs referred to earlier from liquid partitioning data, but with further calibration for the binding energies of protein–protein complexes. A similar approach using contact surface complementarity along with a repulsion term has been developed by Sobolev et al. [72] to guide docking calculations. This idea has been extended by Wallqvist et al. [73], who have calibrated the contact areas of 10 HIV protease ligand complexes in terms of pairwise atom–atom preferences for 21 atom types. The accuracy of the binding energy predictions for this limited data set was ± 1.5 kcal mol^{-1}; again extension to a broader range of complexes may lead to poorer prediction.

Table 6 *Comparison between regression-based and experimental/theoretical values for contributions to ligand binding energetics*

Contribution	Regression (Böhm)[a] (kcal mol^{-1})	Experimental/ theoretical (kcal mol^{-1})	E/T[b]
Cratic and rigid-body translation and rotational entropy term	1.3	13–18	T
Conformational entropy term per rotatable bond	0.3	1.3	T
Neutral hydrogen bond – exposed	− 0.4	0.0 to − 0.5	E
Neutral hydrogen bond – buried	− 1.1	− 1.0 to − 2.0	E
Ionic hydrogen bond – exposed	− 0.7		
Ionic hydrogen bond – buried	− 1.9	− 2.9 to − 4.8	E
Hydrophobic contribution accessible area coefficient (cal Å^{-2})	26	15–53	E
Hydrophobic contribution per CH_2–CH_2 contact; assumes change in area = − 66 Å^2	− 1.6	− 1.0 to − 3.5	E

[a] From Böhm [68].
[b] E: experimental [69]; T: theoretical [70].

The idea of calibrating pairwise atom–atom preferences has been developed by Verkhivker's group [74] as mean-field distance potentials for 12 atom-type pairs. These short-range potentials were derived from 30 HIV protease complexes following the knowledge-based approach of Sippl [75]. The potentials are augmented by desolvation and rigid-body/conformational entropy terms along with allowances for bound water and conformational changes from the free species. In a subsequent paper from the Agouron group [76], a pairwise linear distance potential has been adopted to describe hydrogen bonding and steric contributions to binding in a mean-field characterisation [77] of ligand–protein complexes. Jain [78] has likewise adopted a functional form to describe hydrophobic and polar complementarity. This function is a combination of Gaussian and sigmoidal terms and, again, there are associated repulsive, solvation and entropic contributions. When calibrated using 34 ligand–protein complexes, the accuracy of binding energy prediction was ± 1.2 kcal mol^{-1}. The method predicted the notoriously difficult streptavidin–biotin affinity quite well (log K_i = − 12.5 cf. −13.4) and it will be of interest to see if this accuracy is maintained on extension to complexes beyond the training set.

The determination of the contributions to binding energy by fitting to an experimental data set can be formulated in an analogous fashion to classical quantitative structure–activity relationship (QSAR) methods used to correlate drug potency and physicochemical factors in the pharmaceutical industry. The approach adopted by

Ortiz et al. [79] generates a partial least-squares (PLS) regression for a series of inhibitors binding to a given protein. This uses, as variables, force-field-derived components of the interaction energy between ligand and protein and also their intramolecular geometric and nonbonded terms. A similar PLS approach using 12 calculated properties, including steric and electrostatic energies along with various contact surface area terms, etc., has been calibrated by Head et al. [80] using 51 complexes. The accuracy of prediction in the training set was 1.4 kcal mol^{-1} and, on further test sets, ranged from 1.0 to 2.6 kcal mol^{-1} for HIV protease and thermolysin complexes, respectively.

In summary, a variety of fast empirical methods based on the calibration of the contribution of various components in the context of a number of functional forms can predict binding energies to ± 1.5 kcal mol^{-1} in favourable circumstances. There is evidence of training set bias in a number of cases and the adoption of values for contributions based on experiment or simple theory may be more generally applicable but of no greater accuracy.

4. Insights from calorimetry

The most obvious components of the free energy of binding are the changes in enthalpy and entropy on complex formation. Many of the contributions considered earlier have been associated with a particular thermodynamic component, e.g. hydrophobic terms are regarded as largely entropic and hydrogen bonding as mainly enthalpic.

Following calibration of an accessible area approach to calculating the thermodynamics of protein unfolding, Freire and co-workers [81] have applied the same methodology to the calculation of the entropic and enthalpic components of ligand–protein complex formation. The advent of an increasing volume of directly determined calorimetric binding data [82] has allowed these approaches to be more widely applied and appraised.

There is also a third thermodynamic parameter, namely ΔC_p, the change in heat capacity on binding, which has been the subject of much interpretation. For protein unfolding, ΔC_p is generally large and positive and has been directly related to the increase in accessible hydrophobic surface area in the denatured state. In the case of ligand–protein binding, large negative ΔC_p values are often observed [83] and, by analogy, these have been interpreted as indicating the burial of hydrophobic surface on complex formation. For a number of systems, employing the calibration obtained from protein unfolding leads to estimates of buried hydrophobic surface far in excess of that identified from the structures of the complexes [84].

One explanation has been suggested [85] to involve a contribution from water molecules sequestered into the binding interface. Indeed, the majority of known experimental structures for ligand–protein complexes contain such sequestered water [86]. Depending on the particular environment at the interface, a sequestered water molecule can contribute of the order of $-60\,\text{cal}\,\text{K}^{-1}\,\text{mol}^{-1}$ to the ΔC_p [87], and such waters tend to link a hydrophilic group in the ligand to one in the protein.

Ben-Naim has suggested the term *hydrophilic interaction* for this effect [88] and postulated a free energy contribution of $-(2.5–3.0)$ kcal mol^{-1}. However, based on perturbation free energy calculations, Sun and Kollman [89] indicate that this may be an overestimate. Analysis of calorimetric data suggests [86] that the hydrophilic effect contributes -1 to -3 kcal mol^{-1} to the binding enthalpy and that this is largely compensated by a $-T\Delta S$ contribution of similar magnitude but opposite sign. The resultant free energy for the hydrophilic contribution is of the order of -0.5 kcal mol^{-1} per water.

It has been suggested [88,84] that such hydrophilic interactions be incorporated in ligand design. Conversely, the replacement of such a water may allow the retention of the enthalpic polar contribution without invoking the associated entropy loss on immobilising the water, hence leading to improved affinity [90]. Thermodynamic data for the binding of a series of analogues to a protein often show enthalpy–entropy compensation [91]. If the ΔC_p values are large and negative and are more negative for analogues which exhibit more negative binding enthalpies, and if these enthalpies are associated with compensating changes in entropic contribution, one might infer a role for sequestered water.

In conclusion, calorimetric data can give insights as to the role of water in ligand–protein complexes. In addition, the burial of surface can be related to entropic hydrophobic and enthalpic polar contributions. The fact that water molecules may be involved in the binding interface and may impact on the binding energetics means that they need to be taken into account when calculating binding energies. This may be straightforward when indicated in an experimental structure, but is problematical when the structure of the complex is the result of modelling or docking calculations.

5. Summary

A number of new approaches to the prediction of binding affinity have appeared over the last 3–4 years. In particular, the exploitation of finite difference Poisson–Boltzmann methodology has provided much encouragement. Many problems, however, still need to be addressed, including the speed and numerical accuracy of the methodology, protein and ligand flexibility, the role of structural water, the development of a general atomic charge set and the treatment of the nonpolar contribution to binding.

The extensive structural information now available on protein–ligand complexes has enabled a number of authors to develop regression-based approaches to binding energy prediction. These have been used with some success but current methods tend to be limited to molecules within or close to the training set. As one expands the number of protein–ligand complexes examined, the accuracy of the models tends to decline.

As the quantity of experimental structural and binding data on protein–ligand complexes expands rapidly, our understanding of the contributions to the binding process increases also. Isothermal microcalorimetry is providing important insights which should enable a new generation of methods to be developed over the next few years. The prediction of binding affinity remains one of the greatest challenges to the wider application and acceptance of structure-based design. Much progress has been made in recent years, but much remains to be done.

Acknowledgements

Valuable discussions and support from Andy Grant, Charlotte Adcock and Anthony Nichols on the use of Poisson–Boltzmann methodology for free energy calculations and programming support from Dannyal Conn in the development of SCORER are gratefully acknowledged.

References

1. Verlinde, C.L.M.J. and Hol, W.G.J., Structure, 2(1994)577.
2. Desjarlais, R.I., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31(1988)722.
3. Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S., In Roberts, S.M., Ley, S.V. and Campbell, M.M. (Eds.) Molecular Recognition: Chemical and Biological Problems, Special Publication, Vol. 78, The Royal Society of Chemistry, London, 1989, pp. 182–196.
4. Cosgrove, D.A., EMPTOR, Zeneca vector database searching software.
5. Cosgrove, D.A. and Kenny, P.W., J. Mol. Graph., 14(1996)1.
6. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins Struct. Funct. Genet., 19(1994)199.
7. Caflisch, A., J. Comput.-Aided Mol. Design, 10(1996)372.
8. Gillet, V., Johnson, P., Mata, P., Sike, S. and Williams, P. J., J. Comput.-Aided Mol. Design, 7(1993)127.
9. Böhm, H.-J., J. Comput.-Aided Mol. Design, 6(1992)61.
10. Bohacek, R.S. and McMartin, C., J. Am. Chem. Soc., 116(1994)5560.
11. Moon, J.J. and Howe, W.J., Proteins Struct. Funct. Genet., 11(1991)314.
12. Ajay and Murcko, M.A., J. Med. Chem., 38(1995)4953.
13. Vadja, S., Weng, Z., Rosenfeld, R. and DeLisi, C., Biochemistry, 33(1994)13977.
14. Staratsma, T.P., Zacharias, M. and McCammon, J.A., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 349–367.
15. Reddy, M.R., Varney, M.D., Kalish, V., Viswanadhan, V.N. and Appelt, K., J. Med. Chem., 37(1994)1145.
16. Van Gunsteren, W.F., Beutler, T.C., Fraternali, F., King, P.M., Mark, A.E. and Smith, P.E., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.) Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications, Vol. 2, ESCOM, Leiden, 1993, pp. 315–348
17. Liu, H., Mark, A.E. and van Gunsteren, W.F., J. Phys. Chem., 100(1996)9485.
18. Smith, P.E. and van Gunsteren, W.F., J. Chem. Phys., 100(1994)577.
19. Gerber, P.R., Mark, A.E. and van Gunsteren, W.F., J. Comput.-Aided Mol. Design, 7(1993)305.

20. Hansson, T. and Aquist, J., Protein Eng., 8(1995)1137.
21. Aquist, J. and Mowbray, S.L., J. Biol. Chem., 270(1995)9978.
22. Aquist, J., Medina, C. and Samuelsson, J.E., Protein Eng., 7(1994)385.
23. Aquist, J., J. Comput. Chem., 17(1996)1587.
24. Paulsen, M.D. and Ornstein, R.L., Protein Eng., 9(1996)567.
25. Carlson, H.A. and Jorgensen, W.L., J. Phys. Chem., 99(1995)10667.
26. Mark, A.E. and van Gunsteren, W.F., J. Mol. Biol., 240(1994)167.
27. Boresch, S. and Karplus, M., J. Mol. Biol., 254(1995)801.
28. Jackson, R.M. and Sternberg, M.J.E., Protein Eng., 7(1994)371.
29. Pitarch, J., Moliner, V., Pascaul-Ahuir, P., Estanislao, S. and Tunon, I., J. Phys. Chem., 100(1996)9955.
30. Sitkoff, D., Sharp, K.A. and Honig, B., J. Phys. Chem., 98(1994)1978.
31. Sharp, K.A., Nicholls, A., Friedman, R. and Honig, B., Biochemistry, 30(1991)9686.
32. Reynolds, J.A., Gilbert, D.B. and Tanford, C., Proc. Natl. Acad. Sci. USA, 71(1974)2925.
33. Herman, R.B., Proc. Natl. Acad. Sci. USA, 74(1977)4144.
34. Gilson, M.K. and Honig, B., Proteins Struct. Funct. Genet., 4(1988)7.
35. Shen, J. and Quiocho, F.A., J. Comput. Chem., 16(1995)445.
36. Zhang, T. and Koshland, D.E., Protein Sci., 5(1996)348.
37. Engels, M., Schaeffer, M., Karplus, M. and Grootenhuis, P., Molecular Interaction, 15th International Conference, Molecular Graphics and Modelling Society, 1996.
38. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4(1983)187.
39. Schaeffer, M. and Karplus, M., J. Phys. Chem., 100(1996)1578.
40. Holloway, M.K., et al., J. Med. Chem., 38(1995)305.
41. Kurinov, I.V. and Harrison, R.W., Nat. Struct. Biol., 1(1994)735.
42. Krystek, S., Stouch, T. and Novotny, J., J. Mol. Biol., 234(1993)661.
43. Viswanadhan, V.N., Reddy, M.R., Wlodawer, A., Varney, M.D. and Weinstein, J.N., J. Med. Chem., 39(1996)705.
44. Eisenberg, D. and McLaclan, A.D., Nature, 319(1986)199.
45. Ooi, T., Oobatake, M., Nemethy, G. and Scheraga, H.A., Proc. Natl. Acad. Sci. USA, 84(1987)3086.
46. Eisenberg, D., Wesson, L. and Yamashita, M., Chem. Scrip., 29A(1989)217.
47. Vila, J., Williams, R.L., Vasquez, M. and Scheraga, H.A., Proteins Struct. Funct. Genet., 10(1991)199.
48. Wesson, L. and Eisenberg, D., Protein Sci., 1(1992)227.
49. Schiffer, C.A., Caldwell, J.W., Kollman, P.A. and Stroud, R.M., Mol. Sim., 10(1993)121.
50. Stouten, P.F.W., Frommel, C., Nakamura, H. and Sander, C., Mol. Sim., 10(1993)97.
51. Juffer, A.H., Eisenhauer, F., Hubbard, S.J., Walther, D. and Argos, P., Protein Sci., 4(1995)2499.
52. Privalov, P.L. and Makhatadze, G.I., J. Mol. Biol., 232(1993)660.
53. Smith, K.C. and Honig, B., Proteins Struct. Funct. Genet., 18(1994)119.
54. Abagyan, R. and Totrov, M.M., J. Mol. Biol., 235(1994)983.
55. Janin, J., Proteins Struct. Funct. Genet., 21(1995)30.
56. Wilson, C., Marc, J.E. and Agard, D.A., J. Mol. Biol., 220(1991)495.
57. Wang, J., Szewczuk, Z., Yue, S., Tsuda, Y., Konishi, Y. and Purisima, E.O., J. Mol. Biol., 253(1995)473.
58. Luty, B.A., Wasserman, Z.R., Stouten, F.W., Hodge, N., Zacharias, M. and McCammon, J.A., J. Comput. Chem., 16 (1995)454.

59. Cummings, M.D., Hart, T.N. and Read, R.J., Protein Sci., 4(1995)2087.
60. Pickett, S.D. and Sternberg, M.J.E., J. Mol. Biol., 231(1993)825.
61. Doig, A.J. and Sternberg, M.J.E., Protein Sci., 4(1995)2247.
62. Koehl, P. and Delarue, M., J. Mol. Biol., 239(1994)249.
63. Vadja, S., J. Mol. Biol., 229(1993)125.
64. Grootenhuis, P.D.J. and van Galen, P.J.M., Acta Crystallogr., Sect. D, 51(1995)560.
65. Perakyla, M. and Pakkanen, T.A., Proteins Struct. Funct. Genet., 20(1994)367
66. Kroeger-Smith, M.B., Rouzer, C.A., Taneyhill, L.A., Smith, N.A., Hughes, S.H., Boyer, P.L., Janssen, P.A.J., Moereels, H., Koymans, L., Arnold, E., Ding, J., Das, K., Zhang, W., Michejda, C.J. and Smith Jr., R.H., Protein Sci., 4(1995)2203.
67. Andrews, P.R., Craik, D.J. and Martin, J.L., J. Med. Chem., 27(1984)1648.
68. Böhm, H.-J., J. Comput.-Aided Mol. Design, 8(1994)243.
69. Fersht, A.R., Jackson, S.E. and Serrano, L., Phil. Trans. R. Soc. London, Ser. A, 345(1993)141.
70. Williams, D.H., Cox, J.P.L., Doig, A.J., Gardner, M., Gerhard, U., Kaye, P.T., Lal, A.R., Nicholls, I.A., Salter, C.J. and Mitchell, R.C., J. Am. Chem. Soc., 113(1991)7020.
71. Horton, N. and Lewis, M., Protein Sci., 1(1992)169.
72. Sobolev, V., Wade, R.C., Vriend, G. and Edelman, M., Proteins Struct. Funct. Genet., 25(1996)120.
73. Wallquist, A., Jernigan, R.L. and Covell, D.G., Protein Sci., 4(1995)1881.
74. Verkhivker, G., Appelt, K., Freer, S.T. and Villafranca, J.E., Protein Eng., 8(1995)677.
75. Sippl, M.J., J. Comput.-Aided Mol. Design, 7(1993)473.
76. Verkhivker, G. and Rejto, P.A., Proc. Natl. Acad. Sci. USA, 93(1996)60.
77. Finkelstein, A.V. and Reva, B.A., Nature, 351(1991)497.
78. Jain, A.N., J. Comput.-Aided Mol. Design, 10(1996)427.
79. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., J. Med. Chem., 38(1995)2681.
80. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., J. Am. Chem. Soc., 118(1996)3959.
81. Gomez, J. and Freire, E., J. Mol. Biol., 252(1995)337.
82. Ladbury, J.E. and Chowdhry, B.Z., Chem. Biol., 3(1996)791.
83. Connelly, P.R. and Thomson, J.A., Proc. Natl. Acad. Sci. USA, 89(1992)4781.
84. Morton, C.J. and Ladbury, J. E., Protein Sci., 5(1996)2115.
85. Ladbury, J.E., Wright, J.G., Sturtevant, J.M. and Sigler, P.B., J. Mol. Biol., 238(1994)669.
86. Holdgate, G.A., Tunnicliffe, A., Ward, W.H.J., Weston, S.A., Rosenbrock, G., Barth, P.T., Taylor, I.W.F., Pauptit, R.A. and Timms, D. (1997) submitted.
87. Habermann, S.M. and Murphy, K.P., Protein Sci., 5(1996)1229.
88. Wang, H. and Ben-Naim, A., J. Med. Chem., 39(1996)1531.
89. Sun, Y. and Kollman, P., J. Phys. Chem., 100(1996)6760.
90. Lam, P.Y.S., Jadhav, P.K., Evermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, M.L., Wong, N.Y., Chang, C.H., Weber, P.C., Jackson, D.A., Sharpe, T.R. and Erickson-Viitanen, S., Science, 263(1994)380.
91. Grunwald, E. and Steel, C., J. Am. Chem. Soc., 117(1995)5687.

Computer languages in pharmaceutical design

Barry Robson*
The Dirac Foundation, The Royal Veterinary College, University of London,
Royal College Street, London NW1 0TU, U.K.
Current address: Principal Scientist, MDL Information Systems Inc., 14600 Catalina Street,
San Leandro, CA 94577, U.S.A.

Synopsis

Features of computer languages relevant to drug design, as actually used (or at least described) in the public domain, are reviewed. Ideally, the current molecular designer needs to know about eight different languages. There is a 'Babel' of forms used in conjunction, including Unix, C, C++, HTML, Java, VRML, SQL, PERL, editing languages and specific CAD software command/control languages. There is a need for a single *lingua franca.* No one language yet meets all requirements, but the way forward is becoming clearer. . Difficcties anand possssible limitits of compguages, and by extension the computational approach, are also discussed.

1. Introduction

All chemical agents affecting the body are 'drugs'; however, it is the beneficial subset, useful in prevention, diagnosis and treatment of disease, to which the word 'drug' relates in terms such as *drug discovery* [1]. Successful drug discoverers are reasonably considered important benefactors of mankind. If so, then they deserve to be equipped with the most appropriate and powerful tools.

What tools are these? The popular view considers drug discovery in terms of the ever-vigilant bacteriologist and the serendipity of bacteriological petri dishes with contaminating mould, or the chance discovery of the chemical genius collecting samples in a rain forest. In the latter decade of the 20th century, such events and individuals are rare if not apocryphal. Drug development is too expensive, and the needs too pressing, for chance to be the primary tool. Moreover, when such examples do occur, it is rare that the original agent 'discovered' represents the final commercial agent: a degree of tailoring and redesign is usually required (for example, the original penicillin molecule was too susceptible to the acid of the gut). It is true that

*Visiting researcher and teacher in bioinformatics at Stanford University: Stanford Bioinformatics Resource, Department of Biochemistry, Beckman Center, Stanford University School of Medicine, Stanford, CA 94305-5323, U.S.A.

biotechnology seems at present an exception to all this: the original agent often represents the final product, but this reflects our level of ignorance (about how to design *de novo* or even routinely implement useful changes) rather than aspiration. The situation is very simply that we need to develop more successful arts of discovery, tailoring and redesign [2,3].

The information to be handled in the practice of these arts of tailoring and redesign may be subtle, based on deep chemical, physical and mathematical principles. The information is also often substantial: even if some molecules seem promising straightaway, they are not immediately rushed into the clinic but must be placed in line and compared with others for extensive testing. The huge amount of data which might need to be sifted in order to discover the drug as it may be buried in that data, requires a very high degree of well-organised data management. More generally, then, *rational drug discovery is the process of obtaining and manipulating information such that one can develop an acceptable, novel drug product. The tools required are those of computational informatics.*

The thesis of the present discussion is that *amongst the most potent of the 'hands-on' tools of computational informatics are the computer languages.* Computer languages play two roles in computer-aided drug discovery: (i) They are required for the development of drug design systems: they underlie whatever means of man–machine interaction is used. (ii) They are represented by the control languages used by the drug design software. There is every good reason why two such languages should be the same. The modern approach of moving toward modular ways of building systems and high-level object-oriented approaches justifies a specialised chemistry and biology language in any event. System developers themselves would certainly appreciate the value of working with the end-user scientific experts to identify the essential component ideas in drug development, so as to avoid the writing of new code every time the system is extended or a new system is developed.

In 1996, however, the emphasis is still not on the use of powerful languages at the user end: "Rather than performing time-consuming and costly laboratory tests, the drug designer can use software to *visualize* molecules, determine chemical properties and revise the proposed chemical structure", said Thomas Raechle, Director of Applications at Cray Research Inc. [2]. That is, the current emphasis on computer graphics. Nonetheless, while three-dimensional molecular graphics plays a powerful role, it must not be forgotten that one-dimensional human language serves as a means of communication which, in addressing a broader range of issues than just visual matters, is of unparalleled power. A picture may speak a thousand words, but there may be countless millions of concepts in human endeavour which a picture cannot portray, at least not without becoming a language (e.g. Bliss Symbolics). The competitive position of graphics in so much as it currently challenges language-based graphics is simply because the current technology of available high-level drug design languages is inferior. There are as yet no widely used public domain tools of sufficient power to rise to the graphics challenge.

To be sure, graphics will always play a role. There are increasingly powerful virtual reality helmets to manipulate molecular models, and there will emerge a sophisticated world of artificial intelligence and 'smart rooms' which will facilitate drug design using only human language. However, it remains the case that, to allow efficient human–machine interaction, the visual models represent the molecules as if they were entities of the macroscopic world of familiar experience. Conversely, if the user were to be immersed in a more realistically simulated version of the real world of the very small, it would be an alien one. Some of the differences are the themes of the books of Ref. 4.

There are difficulties in meaningfully visualising the population and time average physical properties of single molecules, quantum mechanical effects, the effects of thermal agitation, and systems away from equilibrium. According to several authors, including Penrose [4], it is possible that the quantum state vector does not collapse to distinguish different macroscopically perceived outcomes of different events except for masses and momenta much larger than those associated with single protein molecules, for example. We cannot escape the need to phrase our questions, commands and data in the logic of the molecular world since the molecular world does not behave in ways altogether similar to our experiences in everyday life. We will need to instruct the systems with which we communicate as our ambassadors to the alien molecular world.

Relevant features of computer languages as used or previously described in the public domain are reviewed here. At the present time there are no languages which well address *all* the fundamental issues raised in this review. Although the author and his colleagues believe they have made a useful attempt in the commercial pharmaceutical sector, it would indeed be arrogant to assume that all the above matters of the molecular world are yet adequately understood, or even that we can pre-empt future needs and preferences of drug designers. It is possible, however, to lay a sounder basis for the future. What is needed is language forms which will evolve and naturally lead in turn to more sophisticated forms so as to make the transition, to a satisfactory global language solution, as seamless as possible. To do this we must still think ahead as best we can about the forms of language which humans and machines must master alike, in order to reach a common understanding. *A major argument of the present review is that the ideal languages for drug design are likely to be structural in character, emphasising the inherent hierarchic structure of programs, computing systems, molecules, molecular data, and the design process. Amongst other improved features, they will be more fundamentally, more powerfully, structural paradigm languages than exist today.*

Irrespective of the choice we make regarding paradigm, there are important justifications for starting afresh with sound engineering principles.

1. Firstly, to evolve efficiently and avoid 'white elephant' or 'legacy systems' which are hard to change, the ability to evolve smoothly and naturally needs to be designed into the system created. Efficient evolution does not come for free.

2. Secondly, unless standard high-level environments are created, we will be stuck with a huge variety of disparate systems sharing numerous features which are redundantly repeated, and many novel features which are impossible to integrate with other standard and novel approaches.

3. Thirdly, unless the work performed by different approaches can be inte grated sufficiently so as to be automated, drug development will be slow, the expert will be

doomed to repeat similar protocols of design for rather similar problems, and we will never have true reproducibility, which is the cornerstone of science.

Underlying the third point in the above is a need to have *a uniformity of feel* through the system – effectively, as few paradigms and styles as possible. This criterion is not yet fully met by general computer systems. The operating system, file editors, programming languages and expert systems, for example, are normally distinct languages. A system which has a degree of uniformity of feel with uniform access across all relevant systems at the same level is sometimes, perhaps misleadingly, called a *polymorphic programming environment.*

An underlying theme throughout this review is the need to consider the *structure* of the design problem, or of each of its aspects. This is considered as much as the structure of any programming language (or polymorphic programming environment) which is required. Indeed, in an ideal world there will be a very close relationship between these structures. Unfortunately, there has been relatively little work looking at the problem 'top down' which will help us design a more appropriate language. Such a language would have structures which map directly to the design problem. The best this review can do is to point out some directions.

1.1. Drug discovery – Relation to modern experimental methods

Not all aspects of drug discovery are drug design. Drug discovery can take several forms with somewhat different information management needs. *Drug screening* does not require information management in the same sense as *drug design,* at least not until refinement of the leader is required. The present review relates primarily to drug design, but two aspects of screening deserve to be noted. In drug screening a track has to be kept of combinatorial libraries and the products of screening operations. This can require substantial data management. Further, the closer the screening approach comes to having the flexibility and power to deliver a drug product directly, the more closely it approaches a kind of molecular-scale analogue computer, in which the information is sorted by carefully constructed possibilities for molecular interactions. One can reasonably speculate about an exciting technological future for deep integration between screening and design methods, within the same hardware complex. Such integration will have to use a language common to screening chip, digital chip and human brain alike, a language which relates to concentrations and probabilities.

1.2. Drug discovery – Theoretical chemistry

In comparison to the drug screening approach mentioned above, in *drug design* an important feature of information management is that it includes prediction of the properties of a molecule which does not yet exist. Hence, drug design is by *definition* primarily a problem in *theoretical chemistry.* This has several important implications for computer-based drug discovery methodology, which the author has considered extensively elsewhere [3]. The implication considered in the present review is that it determines the character of the language to be used as the tool. It is interesting to compare drug screening in another sense: seen as an analogue computation, the singular advantage of the experimental screening method is that it always gets its parameters (e.g. potential functions, quantum mechanical basis sets) and simulations right, albeit right by definition: reality is taken as the gold standard. How can we assure that our theoretical considerations will come as close as possible to that gold standard – in other words, how can we assure its predictive power? The possibility that such a theoretical chemistry does in fact exist, at least in principle, was perhaps first fully appreciated by Dirac [4].

"*The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to handle*".

In principle, then, it should be possible to obtain the properties of molecules by pure calculation, *without recourse to experimental data*, and hence also the properties of molecules not yet synthesised. Dirac, however, also highlighted the practical limitation that the equations rapidly became too complex to be directly soluble. This relates to the restriction represented in the dimension of time (i.e. there is rarely enough). More recently, recognition of the importance of *ab initio* methods (i.e. methods based on a limited number of fundamental first principles with little input from empirical data) has increased dramatically.

Whether at the quantum mechanical or Newtonian level, these are time-consuming calculations and require computer methods. Hence they are said to be problems in *computational chemistry*. We find by 1950 Boy's paraphrase "*It has thus been established that the only factor limiting the calculation of the wave function of any molecule ... is the amount of computing necessary.*" The rapid advances in computation leading to the supercomputers of the early late 1970s and early 1980s further improved matters, as Clementi was quick to appreciate in the 1970s: "*We can calculate everything*" (for a historical review and discussion of these aspects, see Robson and Garnier [3]). Much of the language of theoretical chemistry is thus the language of quantum mechanics. Subsequently, computers have gained in power. In the summer of 1995, the U.S. Department of Energy and Intel Corporation announced the agreement to build the world's first teraflop computer (1 trillion floating point operations per second) consisting of 4096 32 GHz optical-wire linked, combined processor-plus-memory chips, bathed in liquid helium.

However, computer time is always finite. Clementi's view, somewhat unfairly quoted out of context above, also encompassed the need to develop good approximate methods based on quantum mechanical study, in order to make the most of the available computer power. Even then, computers can handle only the simplest systems by this method. For example, we do not yet know how to predict the structure of a protein molecule from its chemical formula (amino acid sequence) and most workers believe that teraflop machines will *not* solve 'the folding problem'. Dirac's cautionary note still holds true, and empirical data must be introduced into the problem, to help guide to a conclusion in reasonable time. This also affects the language form of choice: it must be, in part, not only a simulation language, but also a data acquisition and assimilation language.

2. Computational languages

Here are outlined the basic principles of computation with particular emphasis on the modern world of the multiprocessor, multiuser systems and the basic principles of communication and data gathering [5–7].

2.1. The roots of computer language – The minimal requirements

For completeness, this section is included for the reader who has not been formally trained in programming.

The simplest conceivable useful computing device which is in principle capable of being suitable for our purposes (as a digital agent – see below) is an *automaton* [6]. This is a minimal hypothetical *or simulated* or real device (if it is a real device, then we neglect here hardware considerations). Since it is a programmable machine and it can be simulated, its properties are also the minimal requirement for a computer program. The Turing machine is a particular hypothetical automaton which can be considered as a simple program moving along a tape containing information which it can process and transform. The data on this tape could itself represent a language, and in computational theory frequently does [6]. Specifically, an automaton may be considered as the smallest possible process capable of processing information in a meaningful manner, consisting of at least one function capable of transforming states, the input, and states in which to hold the results of the transformations (including those which will provide output). The most basic computer language is the 'intra-automata' language that will provide the facilities for a procedure with at least five components which make up a 'finite automaton'. These are

S a set of states

I a set of input symbols

F a transition process (e.g. as defined by a function, look-up table, etc.)

S(I) an initial state

S* a subset of states to be used as an accepting state, some of which will be used for output

The simplest interesting thing that an automaton will do is transform each input symbol S to an output symbol *predetermined* by function F. If the output symbol is not found as one of the accepting states, the input is *accepted*; otherwise it is rejected. This can achieve a surprising number of useful applications. If we allow these into the class of digital agents, they will be able to perform useful data transformations, for example. However, the type of transformation cannot be changed. This is still very limited and not very useful for studying the general properties of computer programs, which was the original reason for invoking these devices. Broader activity is permitted to a 'nondeterministic' machine. Then, F is not a deterministic function whose returned value is 'cast in stone' once its argument X is specified, but may be a relation which is dynamically alterable (a user-defined or otherwise modifiable table).

In the hypothetical Turing machine, input I includes a blank symbol and the transforming process involving F or a look-up table respectively has a range (Si, I, {L,R,H}), where symbols L,R,H control the movements of the head reading the input symbols on the linear tape. When in the nondeterministic mode (by definition a deterministic form of the Turing machine is also allowed) the table returns more than one state, the machine is considered as being in all those states (the complete nondeterministic Turing machine acts as if it had infinitely many processors available).

A program asked to solve a tough program may never stop. A digital agent (see below) sent off to treat an excessively complex problem, and told not to return until the task is completed, may never return. Automata are the classic minimalist concepts used for exploration of the notion of knowing when to stop, and related to this they are used to develop the notion of complexity of languages and computer programs. There are important implications for the theory of computation in the matter of 'acceptance', or computation, of formula(X). Of particular interest is whether a Turing machine can perform its computation in a time t.

Whether the time required increases as a polynomial function of some size of the task, or as a nonpolynomial (i.e. exponential) function, will basically determine whether the problem is, or is not, tractable. Languages are also possible input to the machine and represent problems of differing complexity. Class P languages are those languages which can be recognised in polynomial time on a deterministic Turing machine, and class NP is that class of languages which can be recognised in polynomial time on a nondeterministic machine.

It would be inappropriate to describe the Turing machine in detail here, but the important thing is that Church's thesis, which can formally be shown to be equivalent to the Turing representation in its ability to generally describe computational processes, expresses the minimum kind of statements you need for a language (and does so in a more standard programming type of format).

```
LABEL1: X = X + 1, go to LABEL2

LABEL3: IF X NOT = 0 THEN X = X-1 ELSE go to instruction labelled
  LABEL4

LABEL5: Halt and display X
```

where each instruction may occur more than once with a different label, no label is duplicated and each LABEL pointer corresponds to one unique physical LABEL. Note that for generality and comparison with what follows, a label is associated with each instruction. A typical program command normally on completion initiates its successor (e.g. 'on the next line' of a program). It is merely a special case of an implied label.

Church also developed a calculus – the Lambda calculus – which cleverly described the minimum elements of a language in a functional format. Aspects of this notation are relevant to styles being considered in potential language development. Amongst the practical deficiencies of Church's representation as a high-level language, numbers (e.g. 6) are expressed as a repeating cycle of + 1 operations (+ 1 + 1 + 1 + 1 + 1 + 1). This is hardly the epitome of 'user friendliness'. The minimum commands for practical purposes are

LABEL1:X = formula1(X):go to LABEL2

LABEL3:IF formula2(X) NOT = 0 THEN X = formula3(X) ELSE go to instruction LABEL4

LABEL5:Halt and display X

where, again, each instruction may occur more than once with a different label, no label is duplicated and each LABEL pointer corresponds to one unique physical LABEL. With facility for more than one variable X and a few write/write utilities, the ability to write permutations of (essentially) these three instructions (essentially) represents machine code.

Machine code is the deepest level of programming, neglecting, for present purposes, microcode which may be hardwired or software features of fundamental chip operation. The commands of machine code are ultimately in binary numbers and have the form

Command, Address

where the address is the location in the machine where the number representing the command is to be placed. In essence, the number identifies the particular mechanism of the processor which is to be activated. At a slightly higher level, assembler code can be used, which exploits programs called assemblers to help more convenient forms of machine code to be entered and displayed. Assembler input, or 'assembly language', has the format

Label, Mnemonic, Operand, Comment

The label is a symbolic reference to the memory location or register where the next instruction is located, typically used as the destination of a jump (or subroutine call). In essence, there are two types of location, those which simply store and which play the role of variables (named boxes in which to place specific values), and those which store and carry out some action. Moving some information to a processor register such as the addition unit to a memory location which simply stores information, or the converse, or movement between two memory locations, is a typical process, and it is customary to speak of the source and the destination of such information transfer. Slightly higher forms of Assembler have appearances such as

begin: MOV AH2,02;MOV contents of hex 02 to register AH.

The minimal basic set of machine code instructions which would meet the requirements of the Church-like expression above are

MOV destination to source

ADD add value in destination to source

SUB subtract value in destination from source

JMP jump to target

CMP compare destination to source

JNZ jump if not zero

JMP jump if zero

INT halt

With some ingenuity this set can be reduced, and on some machines one or two more instructions would be required in practice, but for all practical purposes the above exemplified the absolute minimum components of any language.

At a somewhat more convenient level, the more practical Church-like options described above also resemble FORTRAN in its original form. FORTRAN (FORmula TRANslator) in effect facilitated the writing of the formulae in the above expression.

```
100 X = A*COS(THETA) + B*SIN(THETA)-C

200 IF (X) 500,600,700
```

which branches to statements labelled 500,600,700 according to whether X is less than zero, zero, or greater than zero. Such a conditional jump as represented by the IF statement can implement any form of conditional flow, just by the way in which we calculate X. We note the advantage of IF statements of more readable form

```
IF (X.GT.5) GOTO 700
```

Later languages had not only numeric but also logical types of expression, and the extended convenient set of operations:

numerical variable = numerical formula

logical expression = logical formula

numerical expression = logical formula (type conversion)

logical expression = numerical expression (type conversion)

IF logical expression then go to LABEL

Halt

We note that in real languages the formulae on the right hand side above may be special simple cases of a constant or other variable, that *coercion* can be defined within a language to convert type within a formula (e.g. the system can be programmed to know that TRUE = FALSE + 1 in expressions such as y = x + y where x is logical and y numeric), and that the GOTO statement can in general be replaced by any statement to be conditionally executed, though there must be a go to statement, or an implicit 'go to' action, within the set of instructions allowed.

In higher languages the 'go to' instruction is discouraged and even disappears. Sole use of 'Go to' discourages portraying the program in a structured way, and leads to 'spaghetti' codes in which the program flow weaves backward and forward in a tortuous, confusing manner. It is replaced by a block structure of BEGIN and END brackets or equivalent which actually state "If logical statement is true then do this whole block of commands, then return to continue with next statement.", viz.

```
IF logical statement

BEGIN

many commands..

END

next command to be executed IF logical statement was untrue.
```

It is obvious that this implies 'go to' commands in relation to the first command within BEGIN and END and the first command after END. The implied 'go to' actions can however be complex because many BEGIN END brackets can be nested. In such cases, however, it is clear that the use of structure imposed by BEGIN END makes the program much easier to develop and read. It is possible to demonstrate rare cases where, even if such a block structure is available, a GOTO statement is still required. However, in some systems the practice has, at least until 1995, been to allow the statement only to system developers, not end-users.

We also note the use of indexed arrays such as X(I) (meaning that X is a list of numbers and the (I)th is the one being referred to) and further convenient instructions which can act on commands or blocks of commands, the most important being the loop control instruction

```
DO I = 1,200
```

This repeats the following statement or block for different values of I from 1 to 200 (unless otherwise specified, in steps of 1) and allows, amongst other things, more direct treatment of a mathematician's indexed counting as in

$$\sum_{I = 1,200} x(i)$$

which would in FORTRAN be

```
DO I = 1,200

XSUM = XSUM + X(I)
```

For practical purposes the above plus commands for read and writing data to and from the program are the basic elements with which programmers must work. Even so, in some paradigms such as the functional programming paradigm (see below) the above simple structure can be hidden.

2.2. *Environments and agents*

In a well-developed computational system or network it is possible to take a 'top-down' view which ignores the fine details. At the level of resolution of interest here, one can consider the computer language with computer language statements as the smallest 'atomic' items. Like atoms in molecules, we can, if we wish, consider that only a limited number of types of statement exist, and that elaborate forms are really 'molecular' statements. Church allowed only some three statements as absolute primitives (see above). The colloquial term 'code' is often used below for a collection of statements. 'Statement' is synonymous with 'commands', 'instructions', if we assume that the language is of the classical imperative form (see below).

For human convenience, and possibly also as a consequence of the principles of complexity and of systems emergent by selection, it is not efficient that any system is a kind of uniform sprawling chaos in which we cannot distinguish structure and modular character. In computer systems including networks, we can in practice distinguish (i) the internal environment or 'sea' and (ii) 'digital entities' or '*digital agents*' which function in that 'sea'. We can also consider higher hierarchic structures of these agents, much as in the hierarchy biomolecules < organelles < cells < tissues < organs < organisms < societies. All of these, including the background 'sea', can still be considered as made of these computer language components.

First it is useful to indicate what digital agents are not. There is a level of organisation intermediary to statements and agents, the level of *objects* (strictly, 'data objects' to distinguish them from functional forms). A definition of an object, like the systems discussed below, contains both data and a degree of code concerned with some distinct kind of information, but some current definitions do not allow them to have sufficient code to lead an independent existence within the background sea, and some computer language purists, in order to enhance that distinction, argue that there should be no additional code in an object definition at all. Examples of objects are 'a chart, table or a short movie' [5] expressed in code or data, and, for present purposes, typically a representation of a molecule, up to and including its molecular dynamics or simulation history, and optionally the style of representation. Objects are essentially manifestations of concepts on which agents act. Hence, digital agents are more complex than objects.

A computer virus is an example of a digital agent. A more constructive example is represented by the imminent rise of the *applet*. An applet or 'miniature application' is an entity which '*may live partly on the Net. The client might send the server a little program that initiates a customised database search, for example. This environment of full two way interaction looks more like a lively exchange of digital agents than the static world of the Web today*' (see Wolf [5]). We can, beyond this, consider more sophisticated entities, possessed of a high degree of intelligence, roving and exploring different parts of a network to gather fresh input and the data from lengthy calculations, automatically updating user sites.

2.2.1. Principles of protocol exchange

Digital agents can be considered as composed of, represented by, capable of moving in, and responding to and processing, the statements ('code'). Simple programs residing at only one location are allowable as examples of digital agents but only as static, primitive ones. Mobility includes both movement within locations in a single processor, but most particularly between processors. True mobility implies compatibility and specifically comprehension of, and by, the agent at both sites A and B. A sufficient degree of sophistication allows true *asynchronous communication* throughout a computer network. For most purposes, this simply means that an agent can be dispatched at any time which carries the information for how its transmission should be interpreted, and in particular it defines the communication protocol. This ability depends on matters of language. The definition of an agent allows the existence of agents sufficiently sophisticated to be able to interpret, act on, and to define languages. The statements which make up our digital agents and their environment only have meaning as perceived by the agents which read and manipulate those statements. Indeed, it is possible that a set of statements would have different meaning to a different agent. There is the minimal requirement in asynchronous communication that either (i) the same language environment must be supported, at least as an executable option, on A and B, or (ii) that the transmitted agent can interact with the receiving site sufficiently to generate the required language environment. To this effect, Babel, or at least unnecessary effort in recognition and translation, is avoided by use of an agreed global language.

Computational theory already uses the notion of placing Turing machines and other automata in series or in parallel, in order to analyse basic problems in computer science. Though a Turing machine can perform any kind of calculation by laborious programming, this assembly is a way to consider less fundamental kinds of complexity which aid in programming these devices. In particular, we may want to add convenient language features to control the relationship of the automata to each other in a network. In terms of the automata notation, there are transforming functions T(S*, A) which will read as input I the states (S*) of automaton A. If these transforming functions are deterministic, with fixed input–output relations, the network defined by the set of T(S*, A) is static. By analogy to the above discussion regarding the transforming function F, a language which allows T to change its action by the choice of A has greater complexity rather than simply a deterministic function. In such a case the network is dynamic and evolving.

```
LABEL1: X = T(S*, A), go to LABEL2

LABEL3: X = formula1(X)

LABEL4: IF formula2(X) NOT = 0 THEN X = formula3(X) ELSE go to
  instruction labelled LABEL5

LABEL6: A = T1(X)

LABEL7: IF T2(A) NOT = 0 THEN A = T3(A) ELSE go to instruction labelled
  LABEL8

LABEL9: Halt and display X
```

where again each instruction may occur more than once with a different label, no label is duplicated and each LABEL pointer corresponds to at least one unique physical LABEL.

The practical complexity is further increased by adding, to such sets of instructions, further instructions

LABEL: COPY(LABEL1,A1;LABEL2,A2)

which copies the line labelled LABEL1 in automaton A1 to the line LABEL2 in A2. This allows the case A1 = A2 and the case LABEL1 = LABEL2. A set of such instructions would allow new automata to be created and old automata to be modified. If the arguments are again computable, e.g. can be replaced by formula(A), this adds an ultimate level of practical complexity. A set of such automata and capable of moving around in the larger computational system or computational network in which they are embedded performing a useful function can be identified with an Applet (mini-application). In this case, A1, A2 might be considered as the specified processors or the site address of the remote computer, and LABEL1 would be the address within those processors or computing systems.

In practice, the above is incomplete and further steps are required for its use. For example, it is not sufficient to force information onto another device without ensuring that it is in a recipient state. Several operations might need to be followed to allow fluid communication:

1. attract attention of other machine to receive above application and 'handshake'
2. verify environment and adapt remote-install procedure to accommodate
3. remote-install the application
4. leave tidy – instruct remote application as how to respond to future incoming instructions
5. attract attention of the installed remote application and 'handshake'
6. check local relevant status before updating information
7. transmit information
8. check information has transmitted, verify detailed content if necessary
9. leave tidy – instruct remote application as how to react to future incoming instructions
10. disconnect

2.2.2. Practical aspects of protocol exchange: Examples of the Web, HTML and JAVA

The Web is distinct from the network of computers known as the Internet. The Web is a set of software components, i.e. an application system. The earliest interest in software of this type was first developed by the nuclear physics community for communication purposes between their computers, and more recently it has been greatly exploited by the bioinformatics, and hence pharmaceutical, community. In 1996, there is increasing emphasis on software components which are digital agents (see above) and which live on computer networks [5].

Being a set of software components which organise use of information, application of the Web is not specifically confined to the Internet, nor necessarily to any network system in the physical sense. Admittedly, it is best known in the Internet context. Rather, it is a 'web' in the sense of software handling a network of concepts and information irrespective of the hardware and communication implementation.

Wherever it is implemented in the computer network mode, the Web can be viewed as a communication system based on the following:

1. The paradigm of the client–server model. A client is a software running on the end-user's local computer, while a server is a software running on the information provider's host machine. A 'browser' is typically a program by which the Web may be observed, and may be taken as synonymous with 'client' in the server–client paradigm. In 1993 one of the first sophisticated browsers, 'Mosaic', was developed by Marc Andreseen and colleagues. Netscape and Microsoft's Internet Explorer are the current dominant commercial browser systems [5].

2. The paradigm of the knowledge net, as a set of links or 'threads' which connect each feature on a page of information to another page of information, allowing the user to work his way along a chain of references. The notion of associatively linking information in this way was developed in 1945 by Vanevar Bush.

3. In protocol, a type of communication system called a hypertext communication system (see below). The term 'hypertext' was coined by Ted Nelson in 1965 to describe the chaining of text in a computer, on the Bush model. Tim Berners-Lee in 1989 proposed the specific hypertext format which would provide an interface across many platforms, access a variety of document types and information exchange protocols, and all universal access to any user at any site (see Ref. 5 for a review of the historical development). By 1990 the first modern versions were implemented on the NeXT computer. This allowed the now-familiar implementation of the linkage mechanism and its navigation by a user. For example, 'clicking' on protein might lead to a page defining proteins and naming different basic types. Clicking on the type globular protein would then lead to a page specifying this class in greater detail. Note that a logical 'knowledge structure' is implied. Clicking on globular proteins and obtaining a page about the mating habits of wombats would be regarded as a linkage error. In the current world of the Web, the downloaded content is not confined to text but extends to multimedia defined in terms of MIME (Multimedia Internet Mail Extensions) embodying a variety of multimedia document and communication protocol formats.

Finally, in subject matter, it is the set of the following components which are the *objects* of discourse in describing and using the Web:

pages (or 'nodes') – are blocks of information displayed

links – the connections between items of information

anchors – the items themselves (which are source or destination of links)

the host computers

the users

the servers

the executable programs

the information (e.g. database, image, sound)

user interactions

data communications

network connections

What specialised languages are used to facilitate the interplay of all the above? Most notably, the Web has a hypertext markup language HTML [5] used to create the Web page documents (strictly, HTML is not a complete layout design language, but a notation which provides 'tags' which make the appearance dependent on the browser or search engine while preserving the integrity of the information and its original content and relationships). With the available language system up to about 1995, the user could choose and observe content, rather as with the U.K. Teletext system, and as with that system the user could not generally interact. The approach also necessitates the use of helper applications. To view a movie in a multimedia system, the user had to first install software found in a helper applications set. The user must also have a graphical system to display the result (e.g. X-windows, Macintosh Operating System or Microsoft Windows). Note that the movie display software, installation method, and local graphics system are of a variety of types.

To overcome these limitations, one needs a common command content for the transmitting source, the entity transmitted, and the receiving environment. Java, developed by Sun Microsystems, is a programming language specially suited for overcoming the above limitation (see December [5]). Originally, the name Java, properly applied, stood for a suite of tools used to create and implement 'executable content' using the Java programming language, but it is frequently employed for the language itself. Like the Web, Java is not specifically an Internet matter. Indeed, Java interactive technology can potentially be implemented in 'embedded systems' such as telephones and TV and VCR controllers.

HotJava [5] is the browser or search engine built to 'show off' the advanced capabilities of the Java programming language.

Java is suited for the construction of the first generations of true *Applets*, as well as potentially more powerful self-intelligent net-roving constructs. A *Java Applet* is a program that can execute with the Java interpreter inside the HotJava browser or a browser that supports *Java* code. A *Java application* is a program that can execute outside of the HotJava browser.

A typical Java application is a program which commences with a specification such as

class ApplicName {public safe void FunctionName (String args[ ]) {*body-of-code*}}

The *body-of-code* is a set of Java commands which are rather reminiscent of C and fairly typical in having assignment statements of the general type

```
variable = (constant, variable or expression)
```

and IF, WHILE and other flow control statements.

The Java Applet differs principally in not having a component such as Function Name(String args[ ]), since arguments are not passed.

The Java Applet is compiled with the Java compiler ('javac') producing a 'class file', and is included in an HTML file through the APP element, i.e. a Hypertext item called < code > app < /code > which flags that the code describes a Java Applet. A Java-enabled browser or search engine not only sees into HTML and displays MIME specifications, but also automatically copies to the user's local processor a file containing Java 'bytecodes', i.e. codes in format which can be interpreted and executed by the user's local Java implementation. An example HTML file including the Applet 'MySearch' would appear as

```
< HTML >

< HEAD >

< TITLE >  My search applet  < /TITLE >

< /HEAD >

< BODY >

  "Here it is"  < APPLET Class = "MySearch" >

< /BODY >

< /HTML >
```

(Note the HTML style which uses < section > to indicate "Begin section" and < /section > to indicate "End section".)

The rather complex opening of the Java application/applet definitions (with the keyword 'class') requires some explanation. Although not the principal consideration in the present section, it is notable and relevant to the overall theme that the developers have chosen to make the Java programming language as an example of an object-oriented language. Applications could in fact be objects such as animals which interact with each other and with an object describing the hero in an adventure game. The FunctionName is a specific instance of the ApplicName class. A keyword 'extends' is available to relate an Applet or Application to a broader class of which all the principal features and methods are *inherited.* Thus one might develop an Application with all the properties of *vertebrate*s and more specific Application with the properties of *mammals* and even more specific Applications which have the properties of *cats.*

The progress in these areas is spectacular. Since writing the first draft of this Chapter, Java has gained enormous popularity, such that in Northern California it appears to be almost as widely distributed amongst the general population as that other beautiful language, Spanish. On one roadside a sign reads 'Artichokes $1 a bag. Also Java programmed'. It may emerge as the best known programming language outside classical programming circles.

The needs of drug design develop at least as rapidly, however. The author feels that for drug design purposes, Java is rather artificially inserted into HTML, at least in the original manner, as shown above. He has very recently developed specifications and pilot forms of TRIDENT© (Text-Rich Data Engineering and Networking Tool). This language embeds HTML commands as a subset, introduces many new commands in HTML < Command ... > format, allows nesting of < ... > commands, and has block structure < block > .. < /block > structures with conditional and loop control. The language is more naturally and fully HTML-miscible. Generally, like HTML, the language sees the commands as relatively sparse features embedded in extensive text (e.g. the sequence of the Human Genome), and as concerned with the transformation of that text, on display. One key feature is the extension of the < FORM ... > ... < /FORM > command to assign values to local variables (read using e.g. < INPUT ... >) as well as to pass on information to a URL or file. These variable values are then used to interact with and control the display. Such features allow HTML-type philosophy to lend itself naturally to bioinformatic computing. Notably, the < DATA ... > *... embedded text data ...* < /DATA > statement is introduced. This has the HTML-like and extended parameters type = , display = , which may be set (by assigning variable or constant) to transform on screen (or to a specified file) the embedded text. For example, type = what_user_sets, type 'DNA', display = 'protein'. The embedded text is usually a DNA sequence which is transformed to a protein sequence in one of the six (three per strand) reading frames. A DNA or protein sequence can be converted to a predicted glycosylated sequence, a secondary structure prediction as a string of characters H (helix), B (sheet) or L (loop), and some 20 other various useful representations (which are computable transformations of the initial DNA or subsequently derived protein sequences). Clearly, some transformations, such as display = 'tertiary structure', could be computationally intensive.

2.2.3. Present (1996) structure of 'the Net'; bioinformatics sources as examples

Apart from Java (see above) and the Internet exchange formats themselves, protocols not universal and not very intelligent, but at least there is a sensibly limited standard set such as *http* ('Hypertext transfer protocol'). The principal others at present include *ftp*, *file*, *gopher*, *mailto*, *news*. Computer addresses are identified by universally recognised codes known as uniform resource locators or URLs, such as

http://gnn.com/

To the above may also be appended a chain of subsystems and directories which will direct the searcher to a specific process or file.

http://gnn.com/gnn/meta/edu/index/html

(AOL advice page 1995)

To illustrate the opportunities at this time in 1996, the following are of interest. For http the WWW World Wide Web affiliation http://www. followed by further specifiers is one of the more popular uses. Efficient searching, whether by humans or by automated procedures, is an issue. General success depends on the power of the software responsible for the details of the search – *the search engine.* R. Finn's Lycos search engine for the WWW can be found at http://www.lycos.com and useful search facilities are obtained from the World Area Information Service (WAIS). Other search and display tools for drug designers and biotechnologists are found in Pedro Coutino's Biomolecular Research tools at http://www.public.iastate.edu/-pedro/research-tools.html. For molecular display of the results one may use Roger Sayle's celebrated molecular display package RASMOL at ftp://ftp.dcs.ed.ac.uk/pub/rasmol or at ftp://src.doc.ic.ac.uk/packages/rasmol (for general image processing one may note ftp://zippy.nimh.nih.gov/pub/nih-image).

At several such addresses there are a number of gene and protein sequence databases and other information. One of the most important is 'Entrez' to genome data at NIH ftp://ncbi.nim.nih.gov.entrez. This can access some 300 000 protein sequence and 500 000 DNA nucleotide sequence records, and well over a million MEDLINE (Biomedical citation) records. Other examples are

European Bioinformatics Institute
ftp://ftp.ebi.ac.uk/
gopher://ftp.ebi.ac.uk/
http://www.ebi.ac.uk/biocat/biocat.html
Bio archive biology software and data
ftp://iubio.bio.indiana.edu/
gopher://iubio.indiana.edu/
http://iubio.bio.indiana.edu
National Institute of Genetics gene data
ftp://fto.nig.ac.jp/
gopher://gopher.nig.indiana.edu
http://www.nig.jp/
Houston gene database
ftp://ftp.bchs.uh.edu/pub/gene-server/
gopher://ftp.bchs.uh.edu/
http://www.bchs.uh.edu
Houston biomedical archive
ftp://dean.med.uth.tmc.edu
gopher://dean.med.uth.tmc.edu/
http://dean.med.uth.tmc.edu/

It is notable that addresses of this type have a branching structure, the specific names represent the path taken at each multiple fork in the road, qualifying further the target location. With the increased importance of the Internet and multiple processor single machines, tree structures and network structures in general have gained increased importance, and the ideal language should recognise that. In addition, the language itself should have flowing, branching character (see above) and last, but not least, so should the structure of molecules and the organisation of the data we associate with them. This is considered in the form of a simple language in Sec. 2.3 below and, following the consideration of some top-down requirements in Sec. 2.4, is readdressed in Sec. 2.5.

2.3. A bottom-up design of a general structured language – A simple example

Here it is demonstrated that quite reasonable language forms can be written to emphasise the branching or network structure of the programs both in terms of the program flow and the way arguments are passed. These forms can be made internally consistent with structured data forms considered below.

By 'bottom up' is simply meant that one thinks of the details of the programming, while making it as generally powerful as possible (e.g. in handling mathematics), and then thinks of the application areas where special cases are developed for molecular work. The first generation of the Prometheus$^{\text{TM}}$ system [8] at Proteus was largely developed that way. An example processor network language based on these fundamental concepts can readily be constructed. A simple example avoiding the details of the internal workings of procedures or processes is as follows. Except in its notation to implement parallelism, it resembles the language FORTH and, like FORTH, it concentrates on the aspect of program flow, rather than trying to emphasise details. We can consider that the minimal requirements inside each named visible statement such as JOHN, which would make each component work, are those FORTRAN-like primitive statements discussed above.

```
JOHN MARY
```

where JOHN and MARY may be considered as procedures or applications, at the very least containing code made up from use of the minimal 'Church-like' function evaluation and IF test statements. Note that the label pointers to the next application or routine are implicit and tucked away in the details inside each statement: when JOHN is complete, MARY is initiated. Information can be passed from the first executed to the second. We can consider all global variables (those not declared just to be private to each application or procedure) to be accessible to the subsequent command. In practice we can define one such procedure as an input procedure (e.g. READ) and one as an output procedure (e.g. WRITE) to display data. For our practices we would assign specific molecular manipulation functions to such procedures or processes, e.g.

```
BUILD_PROTEIN   FOLD_PROTEIN   DRAW_RESULT
```

The fact that we can have named procedures, commands and processes arranged sequentially is straightforward and typical. Where one may begin to have specialised customisation of a language is at the level of considering whether separators between the procedures etc. might have alternative forms with different significance for data transfer. For example, we can also construct such a language so that information to the screen is sometimes automatically output when each is executing. In one development of this experimental language (which the author implemented) a convenient notation comparable with UNIX and DOS is to include a symbol which also channels any screen output as input to the procedure following.

```
JOHN > MARY
```

and other transfer notations can also be defined which specify which type of data we want transferred, e.g. > X,Y > might be taken to mean that only information about X and Y are transferred (it will be apparent from what follows that > X,Y > and > X Y > could be used to distinguish parallel and sequential transmission). In this mode we might use some applications or procedures to represent variables, e.g. X%, from which data can be recovered later, viz.

```
JOHN > X > Y% MARY JILL Y% > HARRY
```

where the value of X returned from JOHN is stored in Y% to be picked up later by HARRY. One might also allow some inputs to be constants, such that we can write

```
"LYSOZYME" > BUILD_PROTEIN   FOLD_PROTEIN
  DRAW_PROTEIN >  STORE
```

A consistent notation could also be developed such that Y% > X > HARRY meant pass the value of Y% and assign this to Y on entry to X, in the manner of passing function arguments. In practice, there is no reason why a more classic style of function brackets as in HARRY(Y%) should not be permissible and effective, but both can coexist, with Y% in the above example playing more the role of a global variable.

There are quite a few languages like this. The principal reason for an informal definition of this simple language is however to emphasise its role in constructing networks of computation. Languages for parallelism are of course known; the following are principally illustrative of convenient notation. As in normal algebra for identical numeric operations between symbols, simply inserting brackets in a string is redundant and has no effect.

```
JOHN MARY [BILL JOHN JACK] SANDRA
```

This means do John then Mary then, on completion of MARY, do BILL JOHN JACK also sequentially, then SANDRA. The difference arises when the comma operator is introduced, meaning do in parallel:

```
JOHN MARY [BILL, [JOHN FREDA], JACK] SANDRA
  FRED | DONE(MARY) TOM | ~ DONE(MARY).
```

Here do BILL, JOHN and JACK in parallel after completing MARY. When JOHN is complete, initiate FREDA. When all are complete do SANDRA. Then do FRED on condition MARY was complete and FRED on condition MARY was not complete.

LAURA(JOHN, MARY)
Use JOHN and MARY as arguments for LAURA

FRED(JOHN MARY)
Use JOHN as an argument for FRED then repeat using MARY.

The usefulness of this notation is a matter of taste. Traditionally it might suggest "do JOHN then MARY" and use any result as the argument. In the present case JOHN and MARY are *still* done sequentially, but are *used to activate* FRED in some way, as each is done.

LAURA(JOHN,BILL MARY,JILL)
Use John and Bill as arguments for Laura, then repeat using MARY,JILL as arguments

FRED(SMITH({1 ... 10}))
Repeat for SMITH(1), SMITH(2), ...

FRED(SMITH({1 ... 10})|EQUAL(X,Y))
Start to repeat for SMITH(1), SMITH(2) ... only if X equals Y.

FRED(SMITH{1 ... STOP|EQUAL(X,Y)})
Repeat while X equals Y.

FRED(SMITH({1 ... 10}|EQUAL(SMITH(X),Y)))
Repeat only for cases where SMITH(X) = Y.

JIM(X,Y):[MARY(X) TOM(Y)]
Define Jim as MARY followed by TOM

COPY(JIM:{1 ... END}, PROCESS5, 100)
Copy all procedure JIM to process P5, line label 100)

2.4. Top-down design of languages

The broadest choice that we make at top level in designing a language is to choose the paradigm. Language types are generally classified into paradigms, largely reflecting the way in which the human user prefers to think in order to get a particular job done. Some paradigms are not mutually exclusive. Where two paradigms can coexist with various possible combinations of partners, the most popular marriage is taken to be the current most popular combination of relevance to molecular computation (somewhat reflecting the author's personal view), included here as follows.

Group 1: Imperative languages

1. Formula translation paradigm – FORTRAN.

2. Algorithmic and block structure languages – using BEGIN ... END or equivalent block structures to emphasise the structure of the algorithm and to facilitate algorithm development.

3. Mathematical operator languages – such as APL, emphasising the conversion of mathematical operations as a mathematician's 'sketch pad' ideas into executable code.

4. Data interrogation and recovery languages, such as relational database systems, Awk, and particularly PERL (pattern extraction and report language).

5. Networking languages.

Group 2: Nonimperative languages

1. Some list processing languages, and potential languages based on Church 'Lambda' calculus – languages which to various extents see the language grammar and/or the data on which it acts as a series of operations.

2. Functional programming systems (FPS) – which emphasise the program structure from the viewpoint of information flow through functions (input innermost, output outermost), and in the purest form disallowing any 'hidden actions'.

3. Predicate calculus languages such as PROLOG, emphasising the relations between data relevant to logical reasoning.

4. Some expert systems, such as EMYCIN, emphasising capture of human expertise in a program.

Group 3: Object and related concept languages

(These may, in principle, be combined with imperative or nonimperative forms, but are typically primarily of imperative form.)

1. Object-oriented languages – which display encapsulation, modularization, and polymorphism. That is, they act on information in unified blocks with special properties, such that all operations and consequences of dealing with a single conceptual structure can be done at once, without the need to process element by element of the constituent information. They are typically polymorphic in the sense that identical commands will have equivalent effects in different application areas of the software. The approach has a certain affinity with FPS systems (since functions are, to some thinkers, a natural kind of object), or can be considered as a sophisticated extension of type definition in the simpler languages.

2. Relational languages, which can be also typically of some object-oriented character. In such languages, concepts akin to +,-,*, (. .) in ordinary mathematics act instead on *sets of data* describing entities such as molecules. The sets of data are typically considered as objects.

3. Systems resembling human languages, with objects and verb forms.

4. Graphics languages, emphasising visual interaction and pictures as objects.

Group 4: Structural languages

1. Related to object languages, these are to be distinguished from 'structured languages' in the sense of the latter having the BEGIN ··· END block structure or equivalent. It is argued below that they are well suited to be the paradigm of drug design languages.

Of course, in molecular work as any other, there are times in which almost all of these paradigms have some merit, but of principal interest is the object-oriented approach since we will wish to manipulate conveniently (i) molecules and (ii) modelling procedures as single concepts, and the enrichment of that as the structural paradigm. Of enormous importance, however, is also database interrogation and information gathering procedures, and expert systems (since drug design is a complex, problem-dependent process requiring great expertise).

Having said this, there is a need to consider carefully the broader extent to which a language addresses the more general programming issues, since programming, or the construction of some kind of high-level protocol, will be something which the user still frequently has to do. Moreover, such issues control the 'look and feel', which will mould the philosophy with which we address problems, and ensure a uniform 'style', not a patchwork 'Frankenstein's monster' composed of disparate last-generation pieces. These latter aspects will not in general be matters in the universe of chemistry, but there will be some choices which will more naturally lend themselves to the way in which chemists think at this moment in history.

2.5. Object-oriented and structural languages

Almost all concepts have structure, and hence so does human language. Where structure is missing we could choose to impose it: even the concept of the empty set 'nothing' might be further resolved into more specific types of nothing in terms of types of nothing for poetic, philosophical, or even mathematical reasons. In the phrase 'Black Cat', the adjective qualifies the world of Cats to point to 'black cats' as opposed to cats which are white, tortoise shell, etc. In the phrase 'Move up', the preposition qualifies the action of the verb to distinguish it from 'move down', 'move in', 'move out', etc. In the phrase 'very quickly', the adverb is 'hedge' to distinguish it from 'slightly' or 'to an average degree'.

Objects are entities which can be distinguished from other entities by special attributes which they carry with them. The notion of structure comes in because objects may also share some attributes in common, and, in particular, one type of object can have a subset of the attributes of another, making that object a special instance, or case, of the other object, as a cat is a special case of a mammal. These are issues of *class*, and descriptions of classes and subclasses are used to define objects in many object-oriented languages.

Object technology and much technology which makes use of structural considerations has roots in the earliest (FORTRAN) notions which distinguish variables and constants of different types. Variables of type integer would not be stored, or generally treated, in exactly the same way as variables of type real. For example, in the assignment z = x/y, different results would be obtained according to which variables are of type real as opposed to type integer (notably, in relation to 'round-off'). For this to occur, 'type' is clearly a property which must be 'carried with' each variable, in addition to the variable name and the value stored in it. It would affect the interpretation of that value, and the nature of the operations applied to it. In most languages, if there is an ambiguity in relation to the action of an operator, the consideration of type is dominant over the operator. Hence in BASIC, the operation ' + ' will mean addition if the arguments are of type integer or real, and will mean string and concatenation if of type string. Note also the common use in imperative general languages of type array, which implies an ordered list of data accessed by the order number on that list (e.g. XARRAY(6) is the sixth item in the list).

PERL [7] is an example (also discussed later below as a data-oriented language) of a language which extends this notion to other types of object, though the extension is not as pure, complete, or general as in the full objects oriented languages (OOLs). It has proven a tool of choice for bioinformatics and related disciplines however. The dialect BIO-PERL has been established at the University of London. Its variables are classified as scalars, arrays-of-scalars, and associative arrays of scalars. The latter is accessed by a string, as opposed to a number (e.g. XARRAY('MONDAY') is the data item relating to the Monday case). Since the language specialises in manipulating data on files, further objects are file handles, directory handles, and formats.

Modern object-oriented (OO) systems properly reflect, however, more than one paradigm [7].

1. The types are not fixed as in non-OO languages but are 'user-definable'. User-definable types are sometimes referred to as abstract data types (ADTs) [7].

2. They show modularity/identity. As with fixed types in non-OO languages such as FORTRAN, the implementation detail (e.g. how z = x/y is interpreted) does not flow over into other code. The only exceptions to this in more advanced languages are in the use of *coercion operators* which force reinterpretation of the type of a following constant, variable or expression. Otherwise, it is invisible inside the module which is treated as a 'black box'. The contents of such a black box may be much greater than in the above FORTRAN integer/real example, and may be of the kind of complexity one sees in a FORTRAN subroutine or function. In this sense, and in respect to FORTRAN which first exploited the user-defined subroutine/function idea, it does maintain the FORTRAN notion of the distinct subroutine or function code which one was *supposed* to treat as a distinct entity. One only has to worry about attaching the visible inputs and outputs of the box into the surrounding environment. Also this means that even if code is common, this makes the module distinct from any other module.

3. They show abstraction. The visible inputs and outputs are confined only to the characteristics relevant to the current purpose. This is not fundamentally different to the argument list in a FORTRAN subroutine or function, but the equivalent of such a list is hidden (by being hidden, it has some analogy with FORTRAN's use of named 'common blocks').

4. They show classification. The current purpose (and hence the relevant characteristics) relates to the level or scale at which we wish to examine the problem. A fine level of consideration will focus on special cases, say on cats in mammals. There is no obvious relation to the FORTRAN case here, save that functions (and subroutines) could be called by other functions (and subroutines). Partly for this reason and the above, there is generally held to be a degree of affinity between functions and objects.

5. They show inheritance. As some consolation to the 'black box' structure prohibiting code reuse, one can reuse previously defined classes as the basis for new objects. If the OO system allows only single inheritance, then a subclass can only inherit from one parent class. If this prohibition is relaxed, it can be misused and lead to nonsensical contraintuitive relations, and increased complexity.

6. They show specification inheritance (essential inheritance). It is in true OO systems using this that the 'is-a-kind-of' aspect inherent in human speech truly appears. Unlike inheritance as above, it is not code-reuse but encoding semantic relations which is intended [7].

7. They show polymorphism. In its simplest sense, this simply means that, like in FORTRAN, the operation depends on the type. When user-defined types provide a multiplicity of types, the complexity of an OO system is reduced by the provision of consistent semantics, which are chosen by the programmer. An important point in polymorphism is that such choices should make operations with the same name perform the same generic function. For example, objects could represent physical objects such as furniture, which can be moved, but not eaten or walked-through, for example. Since it is up to the programmer, choices could be irrational, or at the very least the complexity greatly increased instead of diminished.

8. They show structure by arranging the objects into classes and in some OO systems the structural relations are particularly emphasised with provision of operations to manipulate and borrow from structures.

In structural languages which are not otherwise particularly object-oriented, it is the above last aspect which is emphasised.

The advantages of OO systems over structural systems lacking some or all other OO qualities are in the recursive or iterative life cycle of development. This includes an OO efficient reuse and avoidance of 'white elephant' legacy systems. They also encourage peer-to-peer message levels and a sharing of knowledge [7]. However, a good structural system will at least retain some of the advantages in indicating distinctness and relationship, in inheritance, and in facilitating development of abstraction levels.

2.5.1. Molecules as objects

One of the most important features a chemistry language needs is the facility to define chemical structure. The concept in a simple tree format exists for files in UNIX and DOS, where each directory can be considered as branching into further subdirectories and, finally, files. This branching structure is also inherent in the specification of address on the Internet. The program itself has a branching structure, as emphasised by the language of Sec. 2.3.

A structure applied to data is a form of object in which we can address the hierarchic levels of data within the object. We can extend this to almost any kind of data.

The structure of a molecule, for example, benefits from being able to specifically define entities in a manner such as

alanine = NH.CH(CH3).CO

To do this in practice will typically require a definition, implicit or explicit of type STRUCTURE akin to INTEGER, ARRAY, COMPLEX.

(Comparable representations for molecules, often referred to as SMILES code, were also much used in the 1970s by the Weizmann Institute group, particularly by Michael Levitt.)

Structure representations in powerful OO languages can be used much more generally, however. The assignment-like statement where ': = ' means 'of' (i.e. CONF: = X means TAKE (AND KEEP AS CONF) THE CONF OF X)

```
CONF: =
   MOLECULE.LYSOZYME(XYZ,PHIPSI).HYDRATION_WATERS(XYZ)
```

means copy into structure variable CONF data of the type of CONF, the molecule lysozyme with variables XYZ (Cartesian coordinates) and PHIPSI backbone dihedral angles, and its associated water molecules with Cartesian coordinate data only.

The statement

```
CONF: = MOLECULE.LYSOZYME.HYDRATION_WATERS
```

would be equivalent if there were only types of coordinate XYZ and PHIPSI, but otherwise it would copy all data of the type of CONF, including perhaps nonbonding parameters for the molecule, atom by atom. *What is not specified in this notation means assume the whole class.* Hence the command

```
CONF: = MOLECULE
```

would copy all data of the type of CONF starting with the name MOLECULE, including data for LYSOZYME above and also for all other molecules specified, such as MOLECULE.MYOGLOBIN (XYZ,PHIPSI).WATER(XYZ)

The power of such notations can be greatly extended by structure class operators, such as 'not'

```
CONF: = MOLECULE. ~ (LYSOZYME, TRYPSIN)
```

which transfers data of the type of CONF for everything but the LYSOZYME and TRYPSIN data.

This is an exclusion operator. Useful operators which act between variables of type structure include the relational operators of union and conjunction. Closely related are operators related to the relational database approach, since we can regard structures as databases:

1. + add structures at a specified node and take out duplicating redundant branches;

2. − subtract out structures at a specified node (take everything out of the first structure which is common to the second);

3. form more complex trees as the formal product of two simpler trees (in one formulation, every branch of one tree is replaced by an image of the whole tree of the other, and then inconsistencies and redundancies are removed).

Functions such as COMPOUND could be defined to search a database for the molecule corresponding to the given molecular structure.

CONF: = MOLECULE. ~ (COMPOUND(CH3.CO.*.CO.CH3))

which assigns to CONF the conformational data for all molecules in the set MOLECULE which are not N-acetyl N′-methylamide derivatives.

It should be possible to define a language with a uniform feel by combining the structured language flow of Sec. 2.3 with the structure of the above data item, which are molecules, directories, and so on. Within the structured language, the structure object might appear as an argument in the following example format:

ENERGY_OF(CH3.CO.NH.CH(CH3).CO.NH.CH3, CONF)

which calculates the energy of the molecule in a conformation defined by CONF,

ENERGY_OF(CH3.CO.NH.CH(CH3).CO.NH.CH3,

({ − 180 ... + 180}, { − 180 ... + 180},{ − 180 ... + 180}))

which repeatedly calculates the energy of the molecule 360 × 360 × 360 times, for each increment by 1 degree of each of three variables.

2.6. Language content – Structure in the levels of discourse in molecular design

It is demonstrated here that the modelling and design studies can themselves be ordered in a hierarchic way which could be exploited in a structural drug design language.

There is a singular advantage of the theoretical drug design: the 'trawl space' is very large. That is, more molecules exist in principle, waiting to be discovered in some kind of theoretical chemical 'virtual reality', than can ever exist in practice in a petri dish, a rain forest, a laboratory full of test tubes, a biological broth, or a combinatorial chemistry chip. We can speak of a very large *possibility space*. This is particularly easy to see in the case of possible protein structures. The number of possible amino acid sequences of a protein that is N amino acid residues long is 20 raised to the Nth power. For each such chemistry there is a conformation of some number of distinguishable conformational possibilities, say C, of each residue, this C being also raised to the Nth power. For the medium to larger proteins, there are not enough fundamental particles in the universe to make even just one conformation of every possible variant. Indeed, as far as we know the capacity of the real universe could not even be a faint scratch on the menu of possibilities from which one might choose novel molecules. In various related senses described below, however, the dimension of time rather than of matter and space provides an important restriction in design. The universe will not last long enough to allow us to generate and explore the many molecular possibilities. Indeed, there are many who have questioned even the possibility of predicting the folded structure of a single reasonably large globular protein, from first principles alone.

What kind of languages will ultimately be required to explore this possibility space? To understand that, it helps to appreciate that the full possibility space is arranged as a hierarchy, such that each space is embedded in a space of further dimensions. These dimensions and the points and volumes defining specific regions of the space are formally and ultimately the objects of the ideal computer language for drug design. Each dimensional level represents the 'level of discourse'. To this level, to define the character of language required, we must add the operations which combine and separate spaces, or which transform one point or region to another. In principle, all other matters are derivable from these by simple fundamental rules. 'The rest is silence.' In practice, nonetheless, a language has to be comprehensible to humans, and we require operations that allow the operations to be managed, and the consequences to be understood, in human terms. Each level will have its own vocabulary, appropriate to that level. Generally speaking, it would also seem that we want to maintain the same grammar at all levels of consideration. However, while it is desirable that a larger universal grammar is available, it does not follow that the component features selected to handle information at one level will be the same as those at another. For example, some language features are well suited to consider molecular formulas as connected graphs, and others for addressing levels which involve forming operations on regions of (effectively) continuous space using relational algebra.

2.6.1. Atom-set space

How is the possibility space comprised? The simplest relevant space is perhaps *the atom-set space*, describing the types and number of atoms which are available to us for considering molecules. The key feature which will determine the way we consider the higher spaces is which atoms are considered as distinguishable and which are not. For example, this partly determines the value of the grand partition function, a recipe which gives the probability distribution of different molecules in different aspects of behaviour in the higher spaces considered below. Here we consider that we select N atoms from this space to build a molecule of interest (the implications of the full set including those atoms not selected are considered below).

2.6.2. Connectivity space

The next level is *connectivity space*, the set or field of all the used $N \times (N-1)/2$ possible modes of connections between the N atoms, such that one point in that space defines the organic chemist's specific structural formula for the molecule. Each such point has its own further dimensions so that the connectivity space can be regarded as a subspace of a higher-dimensional fuller description. Specifically, there is the description of the molecule represented by each specified structural formula in terms of energy as a function of the positions of its atoms, and hence its chemistry and conformation.

2.6.3. Conformational space

This is a 3N-6 dimensional *conformational space* of many maxima, minima and saddle points. Workers discuss this energy in terms of quantum mechanical language with terms such as 'basis set' 'variation principle' in the quantum limit, and in terms of e.g. 'potential functions', 'force field', 'minimisation' for the construction of the empirical counterpart of the quantum mechanical energy in the classical limit.

2.6.4. *The higher spaces*

2.6.4.1. *Phase space*

This above description is embedded in turn in the higher 6N-12 dimensional description which includes the conjugate momenta, namely the *phase space.* Here we treat not just the potential energy as a function of the positions of the atoms, but also the kinetic energy of all the atoms. At this level, we most typically use the approximation of the Newtonian world of forces, velocities and accelerations. This is the realm of the molecular dynamics simulation and the rich language which has developed for that discipline. In providing a dynamical description, it also enters the realm of thought and language of dynamic systems theory, and one may speak of 'periodic behaviour', 'quasi-periodic behaviour', and, over longer timescales, 'chaotic behaviour' and 'attractors'.

2.6.4.2. *Design game-state space*

In turn, this representation is embedded in a higher *design game-state space* which is the realm of all possibilities for the particular approach taken. The analogy is with all possible legal board layouts of a chess game representing the state space for chess. This space can also be considered as an extension of the conformational or phase space for design purposes. Assembling and trying to build new molecules by taking atoms in and out of a pot can be considered a broader case of moving atoms around in space to change the conformation. Indeed, if in considering the above atom-set space we consider the larger set available to build any kind of molecule, then the set which is put aside when we select a subset to build our specific molecule will partly define the higher dimensions of this space.

2.6.4.3. *Complexity of the higher spaces*

Topological and related methods exist for formally (if abstractly) constructing the higher state spaces to match the (molecular design) algorithms used, working outward from the descriptions familiar to theoretical chemists. The resulting spaces are however highly topologically complex.

The phase space has surfaces and manifolds (higher than two-dimensional 'surfaces') representing the regions of constant total potential plus kinetic energy (other surfaces exist for the constant temperature case). It is likely that the language of topology will come increasingly into play as this deep relationship is increasingly appreciated. To some extent, however, a deeper understanding depends on the ability of mathematicians to continue to develop the discipline of topology. For example, there is a deep relationship between the network which represents the structural formula of the molecule (i.e. a point in connectivity space) and the topology of the isoenergy surfaces in phase space, but the nature of the relationship is not well understood.

Whether the drug designer thinks of it that way or not, however, and even more disturbingly whether he knows it or not, in many aspects of his approach he is subject to topological considerations. Simply running a simulation forever will not necessarily help, for example. Most often, he is bound for practical purposes to some kind of function surface, a discontinuous manifold of great topological complexity like a system of strong tides in an ocean of higher dimensions. If not simply trapped in the doldrums, the simulation procedure used for design purposes is doomed to wander for all eternity like a digital Flying Dutchman. These aspects are currently expressed in the language used by search-methodologists, such as 'global minimisation/optimisation', 'simulated annealing', 'force bias methods'. These are heuristic tricks and devices for helping searches by more logical strategies. Nonetheless, if we try to apply a little creative scientific technology and try and jump like the Starship Enterprise in hyperspace, but do so only on the local information near the starting point, then we may be completely lost and disoriented.

There is also need of external empirical data and of specific requirements as a guide. The language of external data and of external requirements is currently rich in terms like 'penalty functions', 'constraints', 'biases', 'umbrella sampling', 'targets', 'target functions'. At this level too the commercial goals become a consideration. What type of drug is the target? With a little effort one can, in the grander scheme of things, imagine a procedure which seeks to choose the molecule so as to optimise the commercial profit with the human benefit. The language of economics and of the ethical committee and the FDA comes into play. At present, we believe that no method without guiding data will guarantee a solution to these problems at any of the higher levels, in reasonable time. If, however, we *could* explore this space instantaneously, then we could reach and assess all possibilities in the full possibility space. A portion of this space, governed by the practical, ethical and commercial considerations, is the only portion of interest to us. Nevertheless, this is in the sense that those portions are the goals. The possibility space is still the landscape which must be searched in order to locate those goals, and the ultimate language must address both this and all the underlying levels.

Experimental data, however creditable, are not in a convenient form for many computational experiments and either the computational results or the experimental results must be processed to bring them into correspondence. Reality is glimpsed 'through a glass darkly'. Whereas reproducible effects of the real world may be deemed true, they do not always carry the information we would like. We require the tools of *statistics*. Statistical aspects are inherent and will need to be part of the language system. Data are intrinsically noisy, because of experimental error, and in other aspects (which we describe as intrinsic error) it is meaningful to describe this in terms of a 'chaotic attractor'. Further, the results of experiment are averages over the behaviour of large populations (roughly Avogadro's numbers) of molecules over a long period of time, and only a crude average picture is seen. To analyse this picture and to bring our computations into line with it, we need the tools of *statistical mechanics*. Since computers carry much less information than the real world, and run much less efficiently, we are confined to looking at one molecule or relatively small populations of them over relatively short periods of time. By virtue of this limitation, however, we can obtain great insight from a picture which is not a gross population and time average. We can see more meaningful short-timescale 'action replays' or simple 'snapshots' in time, of the molecule, as if enlarged by a supermicroscope.

While both simulation and reality have defects in providing the quality of information we would like, it is important for computational prediction that the imperfections are of different types. Thus the theoretical and experimental approaches complement each other: on this basis Pauling and Corey put theoretical models together with ambiguous X-ray diffraction data to obtain the helical and pleated sheet structures of proteins, and similarly Crick and Watson obtained the double helical structure of DNA.

2.6.5. The higher space of external information

One potential solution to the complexity of searching the above complex spaces lies in the simple fact that even in the realm of theoretical calculation, *we do not have to deny ourselves recourse to experimental and other empirical data.* In computational pharmaceutical science, much greater emphasis is now given to the *integration* of *ab initio* information tools as with other kinds of information as follows, and the languages must be appropriate to handle them. For example, this problem has been partly addressed, and is being increasingly addressed, in regard to gene and protein structure by the discipline of bioinformatics. The control languages of bioinformatics software can be regarded as languages of this class. The more general language PERL (pattern extraction and report language) is also gaining wide support.

The kind of data which must be manipulated includes:

1. Direct experimental data obtained from direct studies on the molecule of interest, either freshly obtained or recovered from database. This includes sequence data, X-ray crystallographic data, nuclear Overhauser distances from nuclear magnetic resonance spectroscopy, circular dichroism data, difference spectroscopy data, hydrodynamic and viscometric data, light scattering data, immunological data, solubility and partitioning data, and pharmaceutical data.

2. Indirect experimental data in databases where such data are not obtained from experimental studies on the molecule of interest, but from other members of the class to which that molecule belongs. The assumption that some information peculiar to that class will also apply to the molecule of interest. This includes, for example, conformational data about proteins with sequences homologous to the sequence of a protein of interest, for which a conformation may not be known. With the advent of genetic engineering approaches, it is extremely common that the chemistry of a protein is known only through the characterisation of its gene.

3. Information in human expertise. The fact that the design problem must primarily be one of theoretical chemistry remains true even if theoretical considerations take place in the mind of the expert pharmaceutical chemist (even one who may be cynical about the contribution which computers can make). The theoretical aspect need not represent just the application physicochemical principles at the quantum mechanical or Newtonian level, but less tangible information held in the computer though drawn from expert human sources, including the way the expert utilises experimental data. It was precisely the existence of such information which gave pharmaceutical chemists a competitive edge in the 'computer chemistry versus chemist debate'.

2.6.6. Examples of code features addressing the above issues

2.6.6.1. Early forms

Operating system aspects including flexible file management is a feature of several molecular calculation languages, a feature currently expressed most elegantly in the general programming language PERL (see below). By way of example to show the merits of a high-level approach, the first example is the simple but interesting control language of LUCIFER of the University of Manchester developed in the 1970s. This was academic code which would be an insult to the modern structured programmer, but its unique command language is of interest in regard to its form, if only to show the direction for improvement in look-and-feel aspects of languages. LUCIFER used a specially written batch editor language BRED ('B.R.'s Editor'!), a code which allowed procedures to be written for fairly intelligent accommodation to new formats of incoming data. This provision was not well met by existing file editors available at Manchester at that early time, and none available as imports could be easily integrated with the modelling software. There were some examples of routines using BRED which could hunt out and adapt the format of a large variety of foreign files. The BRED command language was based on a one-verb one-argument format, but had a limited block structure with the provision to change the action within the block. Even earlier forms written by the author in the early 1970s had, for practical purposes, indefinite nesting of BEGIN and END, and the facility to write and call named procedures. With only slight modifications for readability, a typical input for most recent forms might be as follows.

```
ASSIGN FILES FIND FILE ATOMS
TAKE FILE
FROM LINE 20 UP TO LINE 300 VIEW FROM COL 2 VIEW TO COL 60
VIEWTIMES 3
  BEGIN
  EDIT LINE /ATOM/ EDIT LINE /CH/C1/ ON ERROR EDIT LINE /CA/C2/
  END
MAKE OK
MAKE FILE NEW_ATOMS

SEQUENCE  + (A.V.G.G.K.L.L.M.N.G.S.S.G.P.Q.Q)-
BUILD
IF ERROR THEN DUMP

TAKE FILE OUTTREE
THROUGHOUT EDIT VIEW TO COL 2 EDIT LINE /N/
NEXT 6 EDIT LINE /H/ IF OK SCRAP LINE IF ERROR THEN LISTFILE
WARNINGS
INPUT FILE GLYCOSYLATION
IF ERROR THEN DUMP

MAKE FILE INTREE
```

```
FROM FILE OUTTREE TO FILE INTREE COPY FILE

SET DIELECTRIC 10.0 BUILD

MINIMISE 2000 CALL_MD 50000.
GO TO FILE REPORTS GO TO MARK LAST ENTRY
IF OK FROM MARK TO END LISTFILE
IF ERROR THEN DUMP
```

This hunts out files and attaches them to standard data streams ready for use. It takes the file with label atoms and, from lines 20 to 300 and between columns 2 and 60, examines only lines containing the characters ATOM, and up to three occurrences of atom type name CH are changed to C1. If CH is not found in the above range, CA is located and changed instead, if present. The output tree structure (the chemical formula expressed as linkages) is then modified to make a new input file by adding glycosylation to the appropriate site on the asparagine. The molecule is then rebuilt with a dielectric set by rescaling the charges. The molecule is built ready for up to 1000 energy minimisation steps followed by 50 000 dynamics iterations. The last entry on the output file is scanned. If no label called 'LAST ENTRY' was written, an error condition is indicated and a print-out is made for debugging. An important feature of such simple languages is the ERROR COUNT where a counter is incremented by one every time an error is found. An error is deemed to have occurred when a subroutine reports a difficulty, and, most importantly, an error is incremented when something is not found. Note the use of MAKE OK to reset the error count to zero.

Clearly, this is clumsy: it requires detailed attention of the programs about where to look and what to do. It is moderately difficult to use and time-consuming to write. The better language might say

```
BUILD PROTEIN  + (A.V.G.G.K.L.L.M.N.G.S.S.G.P.Q.Q)-
ADD MOLECULE GLYCOSYLATION AT FIRST N
SET DIELECTRIC 10.0
MINIMISE, DO DYNAMICS TILL ENTHALPY CONVERGENT
```

A 20–30-fold reduction in written instructions is typical for languages which go from the former explicit form to the latter more 'intelligent' form. One aspect of 'intelligence' at a basic level is sensible defaults. Extended lavishly, providing sensible defaults suited to circumstances is an example of an expert system.

2.6.6.2. Current forms

The languages used for molecular modelling are, with the principal exception described below, not true languages at all. They are *command languages* which must be mixed with operating systems, other languages and database searching tools in order to function.

Following Java, there has been a second relevant and dramatic language development since this article was first written. This is VRML or Virtual Reality Modelling Language, developed by Gavin Bell, Anthony Parisi and Mark Pesce at Silicon

Graphics Inc. It could well have been discussed alongside Java, as VRML is also becoming the standard for delivering three-dimensional images across the Internet. The information is in three-dimensional coordinates and so these images can be manipulated by the local VRML server; moreover, objects within the three-dimensional field can be anchors (the hyperlinks) on which the user can click to access information located at the same or at remote Internet sites. The author had little difficulty in locating and downloading a VRML interface to his laptop, and manipulating downloaded files of molecular display (but while on-line he had problems in getting the anchors to remote sites to perform). Although VRML has no special features for molecular calculations, it has become popular amongst a small but growing group of virtuosos for displaying molecules including proteins and, importantly, to display proteins in motion. Some, notably Hardy and Robinson at Oxford, have begun to build specific VRML applications. There is also already a well-established protein motions database. Display modes for proteins are lines or 'wires' (traditionally this was the faster mode in molecular graphics, and still is), tubes, and ellipsoids of thermal motion. Examples of displays which are available at this time include Alan Robinson's simulation of a bilayer of 166 lipid molecules with a Gramicidin channel running through it, texture mapping of molecules to illustrate electrostatic surfaces, and even a whole bacteriophage. It is not hard to see how the hyperlink action connecting other locations can be used to introduce not only a query capability, but dynamics and energy calculations as well. VRML is well worth noting here as a 'current form' because of this kind of imminently realizable potential, and because of its popularity for protein work. Indeed, Robinson and others have recently made a strong case for VRML in chemistry, but at present it is, as a molecular programming language, incomplete. It remains a graphics language whose objects are general three-dimensional images and styles, rather than a molecular modelling and design language *per se*. Tomorrow may however be different!

Extending the power of C + +, Java, and VRML, is a new *architecture* which is, more precisely, an adaption of existing language systems (especially by C + + subset extensions). CORBA (The Common Object Request Broker Architecture) is an emerging standard which has gained particular popularity in the field of bioinformatics. Its application in more general drug design arenas may therefore follow, but what CORBA is, and also what it fails to be, more clearly defines the future directions required. CORBA was first defined by the Object Management Group (OMG) in late 1990. CORBA may be considered a more security-minded form of Java. Like Java, it is object-oriented. This is in contrast to its nearest procedure-based contender, DCE (Distributed Computing Environment) which was developed by the Open Software Foundation (OSF). It is important to appreciate that both these languages (CORBA and DCE) are also characterised by being primarily for use as 'middleware'. Middleware such as CORBA and DCE glues existing programs together by acting as an intermediate environment, which overlays the operating system. In this way, these are important 'glue' media which therefore must be mentioned. However, at the time of writing, this binding of utilities is static. The CORBA bindings require definition prior to run time as an intermediary specification known as an OMG IDL file. This restriction does not seem to exist for interactive glue languages such as GLOBAL. Such differences and the ‘middleware’ philosophy as a whole more clearly define, to the author’s mind at least, the notion of ‘everyware’. Middleware increases the multiplicity of languages, rather than decreasing it. *It is ‘everyware’ which is required.* Broader power and more dynamic capabilities may still be required for true ‘everyware’ in drug design applications.

Today, UNIX and C are the commonest support tools and allow the manipulation of text of files *and* communication between separate UNIX processes. However, these jobs are still implemented only with difficulty. PERL was conceived as a data reduction language which is becoming popular as a background support for conformational chemists. BIO-PERL for bioinformatics is a dialect developed at the University of London. It is a language to navigate amongst files in a somewhat arbitrary but efficient fashion, to invoke from this searching commands which obtain dynamic data, and to output the findings as easily formatted reports. However, this role soon expanded to encompass the roles of the operating shell itself. Files can be easily manipulated as a whole, and processes can be created or destroyed, the information flow between them controlled, processed and formatted. It is now above all things a networking language, with the ability to unite tasks and activities on different machines [7]. It makes use of the Regular Expression pattern matching facility of UNIX, a powerful generalisation of the ‘wildcard’ ‘*’ and ‘?’ symbols in DOS filenames. Its overall structure is block-like, like the ALGOL/Pascal group. However, it could permit a functional programming (‘FPS’) approach with little or no modification.

Compared with BRED above, PERL code can appear very like human speech, even poetry:

```
BEFOREHAND: close door, each window & exit; wait until time.
   Open spellbook, study, read (scan, select, tellus);
write it, print the hex while each watches,
   reverse its length, write again;
      kill spiders, pop them, chop, split, kill them.
   Unlink arms, shift, wait & listen (listening, wait),
sort the flock (then, warn the “goats” & kill the “sheep”);
 kill them, dump qualms, shift moralities, ...
```

(Anonymous [7]) and so on in that vein. This actually parses and it gives some feel for the operations encountered in file data manipulation, but such examples are highly contrived and do not do much that is useful. As a more typical example of useful PERL code, the following example prints out the history file pointers on a B-news system [7]:

```
≠ print out history file offsets
dbopen(%HIST, ‘/usr/lib/new/history’,0666);
while (($key,$val) = each %HIST) {print $key, ‘ = ‘,unpack(‘L’, $val,”\n”;}
dbclose(HIST);
```

When used in this way it may have much greater power but it was not developed by molecular modellers, as was BRED. It is noteworthy that, with some allowance for taste, PERL is often less readable to the uninitiated than BRED. It is excessively concise. It should be noted that there are many even far less readable examples of PERL, doing some relatively simple operations, which could be given as examples. PERL has however earned a powerful following of adherents and this is well deserved in view of the power of the language.

A detailed language comparison between the leading command languages in chemical design would be tedious and not particularly profound, but it is possible to construct some form of consensus. The following represent the collective language features which most commonly occur across the above software or which give it particular power. They exemplify or indicate (i) specifications, (ii) arguments and (iii) commands which are selected from the menu or keyboard commands for a variety of drug design and protein modelling software. They are to some extent rationalised into a convenient common framework to avoid inconsistency. They represent the minimal set required for a powerful language with a graphics orientation which will also allow a reasonable degree of automation, so justifying software development of a new package in the late 1990s, and may be helpful in the selection of software for purchase.

File types Text, commands, spreadsheet, parameters, formula, DNA_sequence, RNA_sequence, Protein_sequence, protein_secondary structure, alignments, conformation, dynamics "playback".

Filename (arguments) Preferably of structured type directory.subdirectory.subdirectory.name. Up to 256 characters per component.

Remote filenames (arguments) E.g. Remote//http://....

Virtual filenames (arguments) Input, Screen, Printer, Memory, Menu, Trash-can, Shredder.

File operations Open, Read, Read–remote, Reopen–recent, Close, Write, Write–remote, Print, Save As, Append, Move, Scrap, Rename, Obey(take as command input).

File Edit operations Move, Copy (or Cut, Paste), Find, Change.

Data types Integer, Real, Logical, Character, String, Array (for each the previous, e.g. Real Array), File, Spreadsheet, List, Menu, Structure, Object.

Data restrictions Local, Global, Default, Allowed Range, Allowed size (List, Array).

General: Program Flow: IO Define types, Define Procedure/function, Return, Call procedure, IF, Go to, Obey file as input stream, Obey String as command, Do from ... until, Do while, BEGIN ... END, Assignment (X = constant, variable, expression), Read, Write, Format.

Modelling operations – (principally calls to simulation procedures) Define bond geometry, Build (dihedral angles, bond lengths, valence angles to Cartesian coordinates), De-build (coordinates to dihedral angles, bond lengths, valence angles), Convert to secondary structure, Edit secondary structure, Convert secondary structure to built molecule, Convert built structure to planar motif representation, Edit planar motif representation, Convert planar motif representation to built molecule, Define energy parameters, minimise in Cartesian, minimise in rigid geometry (dihedral angles change only), Monte Carlo in Cartesian, Monte Carlo in rigid geometry, Dihedral angle dynamics (pseudo-molecular dynamics using rigid geometry, rigid body dynamics), Molecular Dynamics, Hybrid/Stochastic Dynamics, Add Water, Build Unit Cell, Assign Molecule as Object, Fit Object to Object, Combine Objects as Object, Take part of Molecule as Object, Convert Molecule Object to Abstract Molecule Object (calculate vectors and points to represent hydrogen bonds, charges, nonpolar points, van der Waals surface, etc.), Calculate abstract molecule to fit specified site, Calculate virtual molecule Object to fit quantitative structure activity data Object, Convert abstract molecule Object to molecule Object (grow real molecule to best represent abstract molecule), Define part of virtual molecule Object as an Object, Make Object from combined Objects.

Graphics/Menu Manipulation Examples The examples below are placed in a uniform language style to show the scope:

Obey_active_Menu_item	
Atom_xyz(6)	-Cartesian coordinates of atom 6
Cursor_xyz	-Cartesian coordinates of cursor (if x,y view, z is either zero, or deduced from cursor use in last x,z or y,z view)
Atom(66)	-Atom number 66
Type(Atom(50))	-Type of atom 50
Atom(cursor)	-Atom closest to mouse cursor
Near_Type(10, Type(Atom(60),Atom(cursor)))	-List of all atoms of same type as atom 60, which are within 10 Å of the atom nearest the cursor
Atom(Near(10,Atom(cursor)))	-All atoms within 10 Å of the atom nearest the cursor
Chain(Atom(100))	-List of all atoms in chain containing atom 100
Residue_number(Atom(66))	-Residue number containing atom 66
Residue_Type (Atom(cursor))	-Residue type closets to cursor
Atom(Adjacent_Res(1,-1,Atom(100)))	-List of all atoms in the residue containing atom 100 and all atoms in the adjacent residues before and after in the sequence
Atom(Contact(Atom(cursor)))	-List of all atoms in residues in contact with closest residue to cursor
Secondary_number(Atom(cursor))	-Secondary structure feature (e.g. helix) and number containing atom 66
Menu(cursor)	-Menu item closest to cursor
Point_Atom (30)	-Move cursor to atom number 30
Point_Menu (20)	-Move cursor to menu item number 20
Step_Menu_Cursor 7	-Point to menu item number 7 higher than current menu number
Step_Atom_Cursor 37	-Point to atom 37 higher than current atom number
Minimise_Chain(Cursor_xyz)	-minimize energy of chain including force to cursor

Graphics style arguments Stereo, superimposed, Max-fit superimposed, Perspective = , Colours = , Backbone, Ribbon, Wire, Stick, CPK, Change CPK size, Nicholson, Show hydrogen bonds, Show distances < x, Show buried residues, Show by colour progression of chain (e.g. MacroModel), Show by colour energy of interaction with all other atoms, Show mobility, Show number of references about each segment (Michael Levitt's LOOK), Grow atoms to touch, Add labels, Apply to backbone, Apply to side chains, Apply to atom range, Apply to volume radius R round point.
Graphic operations Select, Xtranslate, Ytranslate, Ztranslate, Xrotate, Yrotate, Zrotate, Centroid = , Viewpoint, Zoom.
DNA–protein sequence operations Show, Hide, Align, Unalign, Edit, Mutate, Glycosylate, > DNA (convert to best guess DNA sequence and similarly...), > RNA, > Retro, > Contra_strand (sequence as if translation of the opposite strand of DNA), > best_homolog (replace sequence by best homolog), > Next_best_homolog, > predicted_secondary_structure, > predicted_phi_psi_angles, Highlight_consensi, Highlight motifs, Model mutant from known parent conformation, add alignment constraints, remove constraints, show % similarity, show % identity.
Information Help, Search sequence database on sequence, Search sequence database on name, Search sequence database on entry/acquisition number, Search literature on name, Search for homology, Search for cryptic homology, search for homologous conformation, Use knowledge base to find all references to subject X.

The above languages are not integrated with the background operating systems though some make a significant effort. For example, Michael Levitt's LOOK is integrated with some more general data file manipulation aspects, especially Net searching tools. Generally speaking, you could not write a respectable *general* program of any value using these command languages, and in that sense they do not satisfy the test of a Turing machine for the ability to perform, in principle, any function. Nor does one address many operating system functionalities by means of the command language code. There are a few efforts, inside pharmaceutical or biopharmaceutical companies, to move towards this. They are not in the full public domain however, except to collaborators (Glaxo has traditionally claimed an open policy, and while making many innovative contributions it generally seems to rely heavily on external and academic tools). In the late 1980s and early 1990s, the author's team sought, at Proteus, to develop an integrated computational system (Prometheus^{TM}) to explore and exploit the use of as much information as possible. The system was one intended to optimise the way in which both application of quantum mechanical and Newtonian principles and stored human expertise could be manipulated together, both by utilising experimental data and by expressing the operations which an expert would form in a high-level chemistry computer language specially developed for the purpose [8]. This system was based on a deeper analysis of the design problem and how various types of experimental data [9] might be utilised (especially human molecular modelling and design expertise, and generics data) to generate proposals for drugs automatically [10–14].

Prometheus^{TM} was largely constructed in a proprietary language GLOBAL © as described in greater detail in the above references (especially Refs. 8 and 9; note: the present proprietary language may have since changed substantially). This language had several PERL-like features. Indeed it deliberately included a number of features from other languages. The examples given thus also describe features of other powerful languages. Two principal points dictate the power of GLOBAL:

As in APL, a function can be defined to be evoked in several ways. Also, as in PERL and a few other languages, a procedure once defined can appear in several guises. A procedure

```
BEGIN PLUS (x,y);
z = x + y;
RETURN z;
END;
```

could be called as a nonadic, monadic or diadic operator

```
3 [PLUS] 6
```

as a function

```
PLUS(3,6)
```

or as a command (with some flexibility allowed in appearance, e.g.)

```
PLUS 3,6;
```

The second is that such a command can not only follow another command sequentially, but be embedded within it. In such a case the full statement brackets must be made explicit $ ⋯ ; or {⋯} as an option for embedded forms. Such an embedded statement, which is evaluated before the embedding statement, leaves its result as a *trace*, a string of ASCII characters which replace the occurrence of the embedded in the embedding statement. It is, in effect, a form of macrosubstitution as in UNIX, DOS and other operating systems. Note that a statement which is a simple constant, variable or expression leaves a trace which is the (string) value of that constant, variable expression. Typical simple use of this is

```
SPECIAL_DAY = “BIRTHDAY”;
DAY_TO_REMEMBER = {ANSWER WHAT IS YOUR {SPECIAL_DAY}
PLEASE?};
```

which asks WHAT IS YOUR SPECIAL DAY? and stores the user’s answer in the variable DAY_TO_REMEMBER. It is noteworthy that functional form ANSWER(...) could validly have been used in the context, so this approach would not have great advantage over a functional substitution. However, the second inner statement {SPECIAL_DAY} could not be passed in the functional manner. The brackets { ... } (or $ ... ;) are like BEGIN; END; special operations controlling program structure and flow. The quotes “ ... ” are the inverse operation of { ... }, for example, so that X, “{X}”, {“X”} are equivalent.

A consequence of the difference between a trace left by an embedded statement and a function passing a value is that the trace can be placed anywhere, even as part of a variable name. For example, one form of the expert system approach in GLOBAL© is that variables can be variable by name as well as content and so represent general statements which may be used in probability, truth and predicate calculus computations. The syllogisms can be encoded, as in

```
ALL_{X}_ARE_{Z} = ALL_{X}_ARE_{Y} [AND] ALL_{Y}_ARE_{Z};
```

and supported by extrinsic procedures which compute and transfer logical values or probability, as in

```
SURE = 1.0;
LESS_SURE = 0.7;
WHEN IT_RAINS, THE_STREETS_ARE_ WET, SURE;
WHEN THE_STREETS_ARE_WET, IT_IS_RAINING, LESS_SURE;
IF IT_IS_RAINING; SAY TAKE UMBRELLA;
```

Another bioinformatics example is

```
PREDICT_2RY_STRUCTURE{ALIGN{MY_SEQUENCE}
{MOST_HOMOLOGOUS{MY_SEQUENCE}}}
```

which is representative rather than actual since the current names of procedures may currently be different within the Prometheus system (the fact that new statements and functions are readily defined and hence renamed nonetheless makes this an attainable example!).

Finally, as noted with PERL, GLOBAL© can make extensive use of Regular Expression notation as a fundamental feature of the language. One could readily address all entries of an array where the search string matched the contents, or a list in which the name of the location matched the search string, for example. The definition and use of the Regular Expression is exactly that as found in UNIX, probably with some extensions after 1995.

3. Automatic language-based approaches – Difficulties and limitations

3.1. Language difficulties

A user does not work well with a language mode which is unnatural to him, and not every user enjoys the same paradigm in thinking about computer-aided drug design. Functional programming systems are notoriously unsettling for users who prefer the imperative paradigm of FORTRAN, PASCAL, BASIC and UNIX. Difficulty in reading initially is an issue, but not a main one. C and PERL are difficult to read when approaching them from a classical FORTRAN/ALGOL/PASCAL background, but they have essentially the same imperative and block character.

Those with a tendency to mathematics, or who enjoy the esoteric, often tend to represent one extreme. They tend to like languages which have very few symbols and rules, but which, by an ingenious choice of grammar, appear elegant. In contrast, many of a more engineering inclination often prefer bare bones imperative forms which are practical mnemonics without too much unnecessary sophistication in human grammar.

For example, the standard system editor available at Manchester University in the 1970s had essentially only three instructions: mark this spot, move to the next spot copying the material crossed over to another specified file, and move to the next spot not copying. This was sufficient to perform all basic editing operations but was irksome to some users who thought it 'too clever for the average user'. In consequence, there was a spate of writing editors from scratch, with instructions like COPY LINES 20–30 TO 50. One such led to the command system for the LUCIFER modelling suite described above.

Another example of a language structure which has elegance but which is confusing to the uninitiated is that it would be perfectly possible to allow within a language a type FILE so that operations resembling mathematical operations, and formally consistent with the relational calculus, do all the file and data manipulations. For example, a command form such as

```
COPY B TO A
```

is adequately represented by the assignment

```
A:= B
```

when A and B are type file. Similarly

```
A:= B + C
```

would concatenate C after B and copy result in A. In the relational database application, the action would also remove redundant duplicated information which had arisen as a result of the operation. With this important feature in mind, it may be noted that the operators + − * with specific relational algebraic meanings would, taken in various combinations such as in

```
A:= (B + C)*(D − E)
```

perform all the operations of the relation database approach. (For completeness it may be noted that D − E removes all information from D that is found in E, and * forms the formal product such that every information item in the result of (B + C) is associated with all the information items in the result of (D + E), with inconsistencies and degeneracies removed.) This is powerful, but the important point for present purposes is that these new concepts do not come naturally to the average user. Nonetheless, fairly fluid access to relational databases such as Oracle and Ingres is found in ISIS software from MDL. In this case, a less concise approach comes more naturally.

An elegant choice can become acceptable with familiarity, but there is energy needed to acquire understanding and familiarity. There is one paradigm which we can be reasonably sure will be preferred by those who have a job to do (as opposed to exploring elegant new features for their own sake). This can be described as ‘the paradigm of the banal’. The need for familiarity was the most critical factor for the average user in the development and application of the PROMETHEUS^{TM} system, and the trick was to bury the sought-for enhanced power and sophistication within that familiarity. Not to get on with design would seem frustrating. To press on with the job, users would prefer to press on with the language features with which they are familiar. There was a strong requirement by management for users to automate their expertise in GLOBAL©. Yet the users would sometimes ask ‘Why do I have to learn another language?’ even when the language familiar to them was entirely unsuited to efficient expertise capture, e.g. it would not be at a sufficiently high, chemistry-oriented level. Not only are end-users important to appease, the smooth extraction of expertise from the expert user, into re-executable code, depends on an affinity between the expert user and the language; if this is lacking, expertise will not be captured even if the language is much better suited to the task in principle. The problem was that users had different language backgrounds.

Thus to overcome this, GLOBAL © had many language features which could be mixed and matched from other languages, and although essentially of a functional programming system form, this would normally appear as facilities in the more familiar imperative language form. There were FORTRAN-like read/write facilities, ALGOL/PASCAL-like block structure, a PERL and C content. The language also allowed a smooth blend with host operating system commands which, for example, allowed UNIX to be used within programs.

Such an approach may be referred to as *banalisation*, meaning ‘to make ordinary’. The further trick, nonetheless, is to make a smooth whole without appearing as a discordant mix. On the whole I believed that GLOBAL © achieved this. The biggest difficulty was the need to mix familiar FORTRAN-like forms in which the string constant part is specifically indicated by quotes “ ... ”, with the natural functional/macro-editing character of GLOBAL where it is the variable part which is specifically indicated by curly brackets { ... }. Hence there would be two types of solution to reading and writing allowed, e.g. coding for input and output, for example:

1. FORTRAN-like method

```
S = “name”;
PRINT (*,“What is your ”, A10, “?”) S;
READ (*,20A4) Your_answer;
```

2. ‘Canonical GLOBAL’ method

```
S = “name”;
Your_answer = {ASK What is your {S}?};
```

It is interesting to recall that in the GLOBAL language generally the inverse relation between “ ... ” meaning a constant and brackets meaning a variable part

{...} had nonetheless a degree of consistency, since they were in fact inverse operations with respect to each other such that, for example, "{X}" = {"X"} = X (see above). Nonetheless, in the above choices of command line form remain inconsistencies. They might well irritate the mathematical style thinker who prefers limitation of choice and elegance, though the aims of the above were, of course, commercial, not academic.

In allowing mixed forms in a language, the form most intrinsic to the paradigm and style of the language and inherent in its underlying structure, is termed 'canonical'.

The approach of also allowing a language to resemble several possible languages is effective in getting the job done in the commercial sector. The primary difficulty is that one can tell from the code the background of the writer. In the extreme, a user's code could look like UNIX, FORTRAN or PERL. Thus editing a routine was not always trivial to an expert coming from a different computer language background. What justifies this approach, however, is that it is much easier to read a language in another paradigm than it is to write it, as it is easier for an Englishman with some smattering of understanding to read French than it is for him to write it. Bearing in mind that in updating the expertise in a block of code the user is not required to confine himself to the style of the surrounding code since all GLOBAL © styles are compatible, this is a moderately satisfactory state of affairs. However, it can lead to code which is efficient but not aesthetic. It can also tend to lead to users confining their expertise updates, when extensive, to separate functional modules invoked from the older text. (This is sometimes deemed a desirable thing to do in any event.)

What is the solution? The difficulty resides not in any one language, but in the fact that there is a multiplicity of them. Since familiarisation overcomes many difficulties, there is a need to reduce the energy in acquiring familiarity. Amongst other things, this is an argument for standardisation to a single, global *lingua franca.*

3.2. *Prediction is not design?*

It has been possible to make computer-based predictions of molecular behaviour, for many years, particularly since the 1960s. One of the criticisms was that doing a calculation, and thereby making a prediction, was not the same as actually designing something. This is certainly true, though there are comprehensible relations between prediction and design. This matter is discussed in detail in Sec. 4.

3.3. *Need for human creativity?*

In such criticisms as voiced in Sec. 3.2, there is buried the further argument that design demands human creativity, and specifically that computer approaches, synthetic chemistry and testing are necessary but not sufficient to drug development. There are three levels of objection and counter objection: (i) There is the argument that only human beings armed with consciousness can add the essential spark of creativity. However, even if it is argued that consciousness is a special human quality which can never be implemented in a computer, it does not automatically follow that human consciousness is required for drug design. (ii) One can nonetheless believe that in practice, until truly intelligent and creative machines come along, drug design must involve human beings. Nobel Laureate Peter Mitchel believed so [15]. He nonetheless appreciated the scientific merit of automation by seeking to enforce reproducibility, which is, after all, the cornerstone of science [16].

Automation of creativity may not be seen as an intractable problem, but it can still be seen as a hard problem. In principle, an *expert system*, combined with *artificial intelligence*, might capture human creativity. If we allow that this is the case (as does this author), then there are still deep practical problems. Notably, the *Feigenbaum bottleneck* refers to the enormous difficulty of capturing human expertise. This is largely because much expertise is subconscious, but there may be other issues. The expert might be resistant to the process because of a fear of being shown to be unable to defend his views on logical grounds, for example, or of losing an elite position or even his job (the 'Luddite response'). Debriefing an expert is a skilled task performed by a 'knowledge engineer'. If we provide the appropriate computational interface, and teach the expert how to express his expertise *himself* in computer terms, and at the same time the expert has also a job to do in terms of meeting deadlines in molecular design, we run into further difficulties. It takes time to make the first pass in automating expertise, and the expert may even believe that an interactive graphics method *feels* more like real work is being achieved. (Similar observations have been made for the 1970s when interactive program editors and user operating systems replaced punched cards and careful reading of lineprinter output.)

3.4. Difficulty of predictability due to complexity of biological systems

Effective design is also often distinguished from prediction because the complexity and the unknown factors in complex biological systems are a bar to full understanding and predictive power. Unexpected, strong and undesirable side effects are usually evocative of the thalidomide tragedy, though it is clear that this was less a design matter and, in part, due to the need to pay attention to the alternative chiral forms of the molecule which were present in pharmaceutical production. More illustrative of the problem in its purest form is the case of IL-12, which seemed so promising in research, but which caused severe toxic effects and two unexpected deaths in clinical trials. "The episode was all the more shocking because the patients, who were suffering from kidney cancer, were given doses that had previously proved tolerable... The drug apparently had a unique property which couldn't have been foreseen" [17]. Inasmuch as binding to other receptors may be an important origin of toxicity, theoretical design could be applicable here too, once the receptors are known, by reducing the binding to known undesirable sites. This is for the longer term but, equivalently, steps are already increasingly being made to check that epitopes selected for peptide vaccines have reduced likelihood to cause autoimmune effects by resemblance to segments of other gene products. These are further strong cases for the value of the human genome project; however, many toxic effects are perhaps too complex in character for such data to be useful in the near future. They might involve the disruption of complex systems in a less specific, more diffuse manner. Further, difficulties like those experienced above [17] might often arise because of complex control processes modifying the response of 'magic bullet' targets and so possibly shifting the efficacy and toxicity of a drug. In a study which seemed to have real potential for developing oral substitutes for insulin, we noted a promising highly specific 'magic bullet' action of a plant extract. It was one of the first known specific inhibitors of mitochondria fatty acid oxidation, with consequent potent useful hypoglycaemic effects [18]. However, it had a much more complex toxic effect of potently uncoupling oxidative phosphorylation and, further, it emerged that the degree of fatty acid inhibition was unexpectedly extremely sensitive to diet.

In addition to such familiar difficulties, the precise mechanism of action of a drug may be misconceived. The intensive "... efforts to develop antisense compounds as therapies for cancer, AIDS, and other diseases have encountered some unexpected questions about how the drugs really work" [19]. In encountering such difficulties, however, there is usually the assumption that the barriers are not fundamental, and that even if pharmaceutical researchers "don't fully understand an insidious effect" there will usually be a happy conclusion that they "know how to avoid it" [17].

3.5. *Fundamental difficulties for design involving biological systems*

In rarer cases, however, criticisms of prediction in relation to final pharmaceutical action relate to the notion that there are *fundamental* differences in living systems. In pharmaceutical discovery we are ultimately concerned with the repair or amelioration of complex biological systems. In this worthy task, it is highly desirable that special, e.g. vitalist, considerations do *not* apply, else we are restricted. Criticism of computer-aided design on the above grounds may not be extremely rare. In the present author's personal experience, not all scientists, biotechnological industrialists and investment bankers really share a pure nonvitalist view of biological molecules. For example, a vitalist consideration does seem occasionally to arise in the commercial biotechnology sector, at least in some 'fuzzy' guise, as to whether a protein derived by chemical synthesis is likely to be satisfactory for some purpose even if its chemical constitution is correct. However, the 'vitalist objection' is often tangled up with a more valid objection. There is recognition in the biotechnology industry that following cloning, expression, there is also the need for correct folding of the product proteins [20]. The vitalist objection has valid overtones in some criticisms regarding the difficulty of *chemically* synthesising and folding complex molecules with the same precision as can be achieved in biological systems. However, such syntheses are possible, and the molecules can be folded to functional forms [21–24]. This need for precision on a large molecular scale is the same problem that must be addressed in the nascent nanotechnology industry, and is discussed in more detail in Sec. 5.

3.6. *Limits to computational reproducibility*

Reproducibility implies reproduction at different times, and the dimension of time introduces a special complication, of which one must be aware. Dynamical systems theory studies how the phenomena in reality (or manifest in molecular mechanical computations) change with real or simulated time. In the theoretical computer realm, with the dimension of time introduced, the calculations are referred to as *simulations*, typically as molecular dynamics simulations (generally taken to mean at the Newtonian level of molecular mechanics, but they may contain a quantum mechanical element). The popular branch of dynamical systems theory known as Chaos theory emphasises how some processes, in the real world and in simulation, can diverge exponentially after minute perturbation. In effect, the consequence of a minute incalculable effect leads to progressive loss of ability to predict the history of a particle and, conversely, what its future will be. Hence in the real world, and in the computer, not all results of interest can be obtained reproducibly. In reality this might be due to perturbation by something as weak and remote as an unaccounted electron in a distant star, and in the computer to changes such as running the simulation on a different computer with a different numerical precision. It is also possible that some results can appear reproducible within certain limits, because of the presence of an *attractor* in the underlying mathematical space. Both (i) simulations of weather and (ii) the development of weather in the real world are subject to chaos and hence limited predictability. But it can still be said that the temperatures of summer and winter will be largely confined, or, better, 'attracted', to a distinct range. Note that two or more variables may change chaotically, but some mathematical function of them, such as their sum, might represent a conservation law.

4. How computers can design automatically

There is the concern expressed above that, though computers can predict, prediction is not the same as design. Further, there was the concern that the design component was necessarily a matter for humans and could not be automated.

4.1. Prediction

The general features of the molecular design process can be considered from a purely 'mathematical' viewpoint. We can think of the resulting simple and general mathematical descriptions as strategies which must be 'fleshed in' with details to develop the protocol for a drug design study. Consider first a molecule with chemistry C and with properties P. Tools such as quantum mechanics and molecular mechanics provide the basis for simulation algorithms which *predict* the properties from the chemistry given, and play the role of a transforming operator T which transforms the C to P.

$$\mathrm{P} = \mathrm{T(C)} \qquad (1)$$

The molecule with chemistry C need not actually yet exist. When the molecule with chemistry C already exists, there is a certain analogy with performing an experiment on a real molecule. To explore the consequences of such an analogy, it is often useful to consider the real world as a kind of sophisticated computer program, and it is by common consensus that we will favour the calculations resulting from that program as being *inevitably* 'true', provided they satisfy the condition of *reproducibility*. Clearly, if experiment yields a result P′ in one case and a different result, say generally $\sim$ P′ (meaning *not* P′), in another case, we cannot say which of the outputs is true or false. In contrast, the predictions of a theoretical calculation are held to be true only if they match the results of the real world.

4.2. Design

Prediction is not the same as *design*. This objection is correct. Equation 1 begs the question of how one chooses the C to generate the P of interest, say P*. Two approaches can be conceived.

In the first case, we consider prediction as a subcomponent of the overall design process. One adjusts the chemistry C to minimise the discrepancy between the predicted property P and the required property P*, as a function of chemistry C.

$$C^* = \min\{\mathrm{abs}(T(C)\text{-}P^*)\,|\,C\} \tag{2}$$

This is akin to aiming a gun, or let us say optimising the strength of the explosive powder C, to hit a target P*. The predictive component in the heart of this feedback cycle will typically contain at least some molecular dynamics simulation. Since Newtonian dynamics can be nonlinear, and since in any event the above equation implies a negative feedback, drug design by this route is potentially Chaotic. Chaotic, that is with a capital 'C', does not in itself mean a bad result. What it does mean is that for long computations we might obtain two different good results on two different computers. This approach sets a fundamental limit to the argument that automation implies reproducibility.

In the second case, we consider prediction as the inverse of design

$$C^* = T^{-1}(P^*) \tag{3}$$

In general, computer programs cannot be written such that one can predict P from C and also C from P, by working backwards. This would be akin to feeding the target into the mouth of our gun to magically produce the required powder at the other end. It would seem to imply time reversal or the breach of the entropy principle. Specifically, with the above considerations of dynamical system theory in mind, it requires a nondegenerate flow of information, without losses, in each direction. However there are, in principle, specific procedures, such as the use of neural nets with bidirectional associative memories, which could be programmed. Such nets could be trained to predict properties by feeding them a set of known properties with known chemistries, and then effectively driven backwards with a required property as input. For fundamentally related reasons, reversible logic systems are also of considerable interest.

A special case of Eq. 3 arises when we consider C and P (actually or mathematically) as two surfaces which are complementary. Then we may think of each as a template for the other. In particular, we can imagine the details of the binding site of a receptor as driving the assembly of a molecule from chemical components to form a final molecule which will fit that site (hence, 'induced design'). Solutions are degenerate and a receptor binding site might define many chemistries which will fit it. If we rank these in some way such that the fit of one molecule can be said to be better by so much than the fit of another, then by exhaustive examination of all the fits, or more realistically by screening a combinatorial selection of generated hits, or by a more intelligent search procedure, one may obtain a best fit.

4.3. Some basic operations with peptides as objects

The following are amongst those operations of potential value both in developing computation-based *and* experimental protocols. In precise use many will obviously require arguments to qualify further their action, e.g. to set ranges of effect in an amino acid sequence. They may indeed be quite lavish protocols involving several computational procedures. In cases like 'make a dimer' or 'Pegylate' (add polyethylene glycol), either when implemented by a human or by a computer, they will themselves contain optimisation features and feedback cycles ('wheels within wheels'!).

( )	bracketed form implies reference to chemistry
[ ]	refers to sequence of polymeric units (e.g. amino acid units, but not to components within units)
{ }	applies to electrostatic and van der Waals density in 3D space
{ }vw	applies to van der Waals density in space only
$^{-1}$	do inverse of operation (postfix superscript)
t	do transpose of operation (where meaningful)
c	form complement (see above)
+	add new specified modular component
—	delete specified modular component
x	swap specified modular components (diadic)
∪	use all groups/sequence segments from both specified entities (diadic)
∩	use only groups/sequence segments common to both specified entities (diadic)
~	use only groups/sequence segments from first specified entity which do not occur in second (diadic operator)
> <	remove all redundant (e.g. second) occurrences of substructures such as sequence segments (but not redundancies within polymeric units)
i	make mirror image (enantiomer)
q	swap hydrogen donor (basic) for acceptor (acidic) residues
r	synthesise a polymer sequence with units in reverse order
p	'pegylate' (add polyethylene glycol)
< 2 >	make dimer
a	swap i − 3,i − 4 and i + 3,i + 4 residues over specified sequence range
b	swap polar and nonpolar amino acid residues over specified sequence range

g	replace specified glycine residues by D-alanine
h	replace hydrophobic group by the next most hydrophobic group

and so on ('diadic' operators normally have two arguments, e.g. A x B).

For example, the operation

$$C^* = i(r((C)))$$

represents the retroinverso approach of synthesising the sequence backwards using D-amino acids (see Sec. 5.4) and

$$C^* = i(r(q(C)))$$

may be regarded as an improvement on retroinverso when applied to regions where there are strong backbone-to-sidechain dipole interactions (e.g. hydrogen bonds).

In computational approaches this range is greatly extended because changes can be described and implemented in intimate molecular detail not routinely available to the laboratory bench chemist with pencil and paper.

However, the lack of full invertability of procedures (e.g. due to degeneracy) in the case of Eq. 3 leads to the generally preferred use of Eq. 2, involving a minimisation procedure with various degrees of intelligence and sophistication.

4.4. Heuristically aided design

As noted above, the need for heuristic information follows from the arguments of Paul Dirac. An *ab initio* solution from fundamental theoretical physicochemical principles alone was, according to Dirac, possible in principle (at least in matters of terrestrial chemistry). In principle, no higher intelligence or clever programming is required to manipulate the state space: the chess game should play out itself. It is a game or puzzle which contains the seeds of its own optimal solution (the minimisation of the free energy of the system). A computer program applying this notion may be described as a *Dirac engine*. In practice, as noted above, the equations for any but the most trivial problems are too complex to form and solve in reasonable time. The reason is that the mathematical surfaces of the processes described, which with a given starting point determine the dynamical behaviour of the system, are topologically extremely complex. They are filled with many local attractors as well as the strongest global one, even assuming that a unique strongly global one exists. In considering part of the problem, the conformational energy surface as a function of conformational variables of a molecule, one speaks of the 'multiple minima problem' which underlies the 'protein folding problem'. As discussed earlier we can consider this level of description of molecular behaviour as embedded, in turn, in the tougher description which includes the conjugate momenta. This gives the phase space. In turn, this representation is embedded in a higher state space which is the state space of the approach taken (see above). The notion of conformational space, being an energy function of the relative positions of different atoms in space, can be extended to taking atoms in and out of a pot to assemble new molecules. In any event, the resulting spaces are highly complex. In consequence, they cannot be searched so as to guarantee a solution in reasonable time.

In the author's view, heuristic approaches including expert systems are best seen in conjunction with extensive uses of fundamental physicochemical principles and implemented at a higher level ('metalevel') than those principles, in effect so as to manage their use and search their consequences. We may compare a chess-playing program where the rules of chess are the fundamental principles, and the intelligent chess-playing program as analysing, manipulating and predicting based on those principles. Heuristic approaches must then, in the above 'metalevel' philosophy, have some kind of *ab initio* calculation, or some kind of simulation, somewhere, on which to act. This does not seem unreasonable: a chemical computer-based chemistry approach which is of heuristic character and yet does not address in any way at least *some* of the theoretical insight gained in the 19th and 20th centuries would indeed be hard to imagine.

More rigorous approaches lead to rather similar conclusions. For example, a powerful argument due to the statistical mechanician Jaynes (see Ref. 8 for a discussion) was that all such theoretical studies contained 'loose probabilities' which might be set by the prejudices of the researcher. The Jaynes recipe for reducing subjectivity was to choose all the values of these probabilities so as to maximise the entropy. However, isolation of those probabilities which may be assigned values according to well-founded prejudices, including those from experimental data, is a plausible method for combining the *ab initio* and heuristic information. In certain routes to achieving this, such as sampling based on biases introduced by Monte Carlo methods, one may remove the biases retrospectively so as not to interfere with the statistical mechanical averages. In finite time, such an approach becomes a particular form of guided search.

What will heuristic information typically be? Experimental data about the molecule and about similar molecules are the most widely used. However, it is not helpful to assign the term too broadly and thus it is instructive to note what kind of information is *not* heuristic. Information directly related to the understanding of the biological action at the molecular level does not itself necessarily constitute heuristic information. This is because the receptor target should properly and formally be treated as part of the system to be simulated (see, however, the next paragraph). The design approach is then based on assembling molecules to fit cell-surface protein receptor (or other) targets such as saccharides, internal receptors, enzymes or DNA [10]. This is generally called *direct drug design.* Of course, where receptor data are incomplete, information of heuristic character may be brought in to help deduce the receptor structure. Prior to and during the development of the Prometheus^{TM} system there were ongoing studies on how experimental data and expertise in the analysis and modelling of protein structures might be analysed [8] and captured [9]. Such studies provided a basis primarily for the analysis or modelling of protein structures, such as receptor structures, such that organic drugs or peptides might be designed to bind to them and bring about activation or inhibition as required.

Receptor information may not be available. It may also not be reliable since it is also inevitably incomplete. We note that information about the receptor, even X-ray crystallographic data, is not sufficient to define the 'switched-on' state of a receptor [1–4] and hence to design an agonist or antagonist. This depends on interactions of the receptor with the rest of the cell. Even the same receptor molecule in different cells may have different switched-on states so that an agonist in one case is an antagonist in another. Thus, the physicochemical system really means much more than just the ligand plus receptor. Use of 'circumstantial evidence' about the structure and activity relationships of molecules related to the ligand of interest is then required. Instead of automatically developing molecules to fit the binding site of the target, one develops them to fit quantitative structure–activity (QSAR) data from the probe compounds. This is called *indirect drug design.*

Indirect drug design does use data which should be regarded as heuristic. A variety of methods have been explored and implemented, primarily reflecting the choice of target function and minimisation procedure in Eq. 2. By the analysis of many related molecules, three-dimensional QSAR data representing a molecular field analysis (MFA) represent a kind of van der Waals electrostatic complementary image to the receptor site which is deduced indirectly [10]. One may also attempt to assign to components of the chemistry (regarded as a formula or treated as a three-dimensional object) additive elements related to the activity [10]. These can be regarded as representing a molecule-family profile, which isolates the salient features explaining the efficacy of the family. Fast assembly of drugs to fit such data on a trial-and-error basis can use various optimisation techniques. In particular, a genetic algorithm [13] was found effective. Combinatorial methods which more efficiently generate large numbers of drugs from a finite set of components, in order to fit a receptor site, have also recently proven effective.

5. The objects of prediction and design

5.1. 'Organic' drugs

The cunning of the synthetic chemist has for many years allowed the synthesis of a variety of novel forms of organic molecules of molecular weight typically less than, say, 800 Da. A good indication of size limit is that earlier ISIS/Host software from MDL Information Systems could not handle more than 256 atoms per organic molecule. This has now been increased, but most molecules of interest to pharmaceutical companies and considered as organics (other than polymers) are still within that size. They have proven extraordinarily valuable as orally administerable compounds. With the advent of biotechnology it became important to understand that 'organic' is a historical term and now simply means 'containing carbon atoms'. The accepted implication in pharmaceutical jargon is that they are low-molecular-weight compounds which are *not* typically derived from biological sources. The word is usually taken to mean the very opposite of 'from an organic source'. For example, 'organic molecules which bind DNA' would actually imply 'molecules binding DNA which are not proteins or other nucleic acids'. Corporate lawyers often get confused on that point! Nature provides an abundance of examples of low-molecular-weight organic compounds, such as penicillin, which have potent effects on other organisms. It is based on these examples that the organic chemists of the 19th and early 20th centuries learned to study the principles of carbon-containing chemistries. The components of organic drugs which are commonly used nonetheless form a finite set of groups such as benzene, carbonyl, amine, etc., and dictionaries of these are assembled in automated drug design programs. In the Prometheus system, a high weighting is given to cyclic structures which have maximum connectivity and minimal conformational entropy on binding to a molecular surface, as this is a common feature of many organic drugs, partly for similar entropic reasons.

5.2. Peptides

These are entities in the 1–35 residue range, generally lacking a cohesive hydrophobic core. At least about 40 residues are typically required to form such a core. Avian pancreatic polypeptide of 36 residues only forms a well-defined core by making a dimer. Oligopeptides of some 12–46 residues are typically roughly linear. In the range 10–15 a protein can fold back on itself, but the chain is likely to be otherwise β-like, with a length of 3.8 Å/residue, to allow enough contact surface, which is mainly by hydrogen bonds. These define 'β-loops' or 'hairpin loops'. Lengths of 20 do not in general have tight well-organised structures, but fragments of partially synthesised C-terminal 20-residue fragment immunoglobulin folds in aqueous medium seem able to form a complex compact structure (Gryphon Sciences, personal communication). A high degree of twist is seen in the larger structures of this class, often such that the loop itself can be considered as bent back on itself in a higher-order loop. In some cases this approaches 'Greek Key' motifs. Lengths of 30 or more are really required to guarantee that at least one arm of the hairpin is an α-helix, which has a length of only 1.5 Å/residue.

5.3. Proteins

Molecules in excess of 40–50 residues may be classified as fibrous or globular. If globular, it implies that they are compactly folded in a specific way with a well-defined hydrophobic core with a degree of hydrogen bonding skeleton (if absent, it is likely that the compact form would be a 'molten globule'; however, some large polypeptides such as parathyroid hormone are relatively polar, unfolded, and do not form a molten globule). Functional proteins can be made chemosynthetically, with no possible help from biological machinery [21–24]. As reviewed and analysed elsewhere [25–30, 35, 36], the industrial importance of the view of Anfinsen that protein molecules carry their own information (in the amino acid sequence) for folding up correctly was increasingly appreciated in the early 1970s. No special machinery is required to achieve the spatial structure. However, proteins such as chaperonins may catalyse the process and help avoid aggregation difficulties due to intermolecular interactions between proteins.

Computationally, the linear sequence of amino acid residues can be considered as a linear code coding for the three-dimensional structure. The year 1974 was a watershed year for recognition [25–28] of the importance of understanding this code, which also relates directly to the 'folding problem' as discussed above [29]. Then, the late 1970s saw fairly intensive activity to understand the code, and optimism ran high with occasional unjustified excitement that the code had been broken. The author also critically reviewed this later phase in a series of News-and-Views in Nature (for a compilation see Ref. 30) and Amino Acids Peptides and Proteins (for a compilation see Ref. 31) throughout this period. Despite Anfinsen's thesis that the native structure of a protein is the accessible state of least free energy (see Ref. 29 for a discussion), there is a justifiable further concern.

In implementing new designs, caution should be exercised. Since synthetic proteins [21–24] and cloned proteins start from different spectra of configurations, they could have different folding paths (e.g. Ref. 28). Different solvents can affect the folding history [32]. In practice, the biggest problem is that one can end up with wrongly connected disulphide bonds. Mass spectroscopy is a powerful tool [33], and there are improved methods for using it to verify the required connectivity [34]. The reason for optimism is precisely that proteins do carry such information for their own folding and can be carefully folded and refolded correctly (to recover their functions), provided they have not been extensively modified since biosynthesis (see, for example, Refs. 35 and 36). These latter studies also provide a sound rationale on how to identify, study and characterise protein folding intermediates, which may well be directly related for aggregation. Provided that disulphides are shown to be joined correctly, one can have reasonable confidence in the final structures produced by these precision chemistries.

The 1980s to 1990s have shown a progressive interest in actually implementing the design of novel proteins, even though the stereochemical code has not been broken. Four approaches have been taken, exemplified as follows.

A. *Editing secondary structures.* This relates particularly to β-barrels. A well-known example is the Richardson 'β-bellin' [37], and other pleated sheet structures were tried in the 1990s [38–40]. A β-sheet has the advantage of being a large entity poised somewhere between a secondary structure feature and a supersecondary structure feature, and β-sheet barrels naturally give a globular overall form. This should be distinguished from editing whole domains such as helix bundles, which relates more to C below. Editing α-helices alone is technically interesting (kinetic and thermodynamic studies on nucleation have been performed in Baldwin's laboratory for example), but a helix would generally be considered a peptide rather than a protein.

B. *Editing known folding domains.* It is possible to adapt or edit sections of proteins for which the amino acid sequence and conformation is known experimentally. For example, there has been a recent design [41] of a metal-binding protein with a novel fold, but the fold is adapted from the immunoglobulin folding motif.

C. *Assembling known segments.* The feeling is also that by assembling segments from nature for which the folding pattern is known, one could build up more complex proteins. This is reminiscent of the modular approach described in regard to synthetic and bionic nanoscopic structures described above, and both may make use of secondary and supersecondary (subdomain) components drawn from known protein structures. Bryson et al. [42] provide a concise review which captures the optimistic mood which is currently prevailing.

D. *Replacing natural by unnatural amino acids and novel linkages.* This might also be described as 'editing the genetic code' [43–48]. The above novel proteins are novel only in regard to changes in amino acid sequence using the 20 naturally occurring amino acids. However, the ability to insert more than the natural 20 amino acids into expression systems, albeit still a very restricted set of chemistries, also allows a detailed exploration of novel proteins with novel chemical features and even adds understanding to the properties of natural protein features [43]. There are much fewer restrictions if the manufacture is totally by chemical synthesis. By 1994 Keith Rose held a record for the largest precision made artificial protein at 20 kDa [24]. A superoxide dismutase [44] analogue of 15 kDa has been made chemosynthetically by Gryphon Sciences [45]. Other comparable large structures made entirely synthetically include proteins which, however, contain features which could not be incorporated by cloning and expression (such as unusual linkages or N-terminal to N-terminal chain connections). These include a leucine zipper heterodimer [47] and a solubilised receptor [48]. The implications for 'breaking the shackles of the genetic code' have been reviewed elsewhere [48]. It may be presumed that in some cases certain physicochemical properties of natural amino acids will need to be conserved by their unnatural substituents [49–51].

E. *Total de novo design of proteins.* In trying to generate entirely novel sequences with correct fold and function, significant and reproducible success has not been achieved. Even when *de novo* design is the stated aim, this approach has been largely confined to developing the requisite computational tools and calibrating them on known systems. Purely statistical tools such as the GOR method [52,53] and related methods [54] have been extensively tested [55] and developed [56–59], but can only act as starting points for modelling, despite high expectations by some early authors. A set of references reflecting the efforts in our own laboratory and which provided the background for the LUCIFER modelling suite at Manchester and the Prometheus^{TM} suite at Proteus are given in Refs. 60–66. By way of example, these studies and those underlying the MacroModel suite developed primarily at Columbia University are discussed in more detail below. LUCIFER, Prometheus and MacroModel pay particular attention to the problem of searching conformational space (see e.g. Refs. 67–69). These methods have also been used to study problems as diverse as enzyme activity and flexibility [70–74] and problems in chemical synthesis [73].

5.4. D-proteins

D-proteins are made entirely of D-amino acids. They fold up like globular proteins, but in mirror image form [74–75]. These molecules should not be confused with the widespread use of peptides with just one or a few D-amino acid insertions. These systems are in purest form 'all D' and clearly represent a special and unique case. Only once can the 'mirror be flipped' on biological systems. There are no other true reflections of the L-proteins than the D-proteins, i.e. the proteins made of D-amino acids. What are the principal advantages? Despite the relation to protein sequence in genomic databases, all these D-peptides and proteins also share many properties with organic molecules, and represent a kind of unexplored continent between biotechnology products and classical organic drugs. For example, the data so far suggest that they will be highly resistant to proteolysis and relatively invisible to the immune system (see, for example, Ref. 76). These 'stealth' properties will be also particularly important when the next generation of high molecular weight precision therapeutics is constructed, since such nanoscopic machinery is susceptible to proteolysis and, particularly, immune response. The important components of such a system can be encoded in D-amino acid format, and yet can still be assigned function (see below). There is a further bonus, and hope, that like organic drugs they may, in some cases, be subject to oral administration [77].

Although for design purposes we can in many cases start with biological knowledge and L-protein sequence data, the D-peptides and D-proteins are incapable at this time of being synthesised in biological systems. At the same time, D-peptides and D-proteins are not true xenobiotics. D-amino acids appear in the ageing of natural proteins, by post-translational modification (via an amino acid residue epimetase), and occur naturally in products of bacterial infection. D-peptidases and D-amino acid oxidases also exist in the body. It is also worth noting that a number of peptide analogies containing one or a few amino acid residues are also available as approved drugs.

When the substrate is chiral, D-enzymes will only act on the full mirror image enantiomer of the substrate [74]. This seemingly restricts the range of application to achiral or approximately achiral target systems when protein sequence data can be directly used. Compared with an original native L-protein hormone, the mirror image of the required van der Waals surface would be as diametrically opposed as one can get, to that which would fit the original receptor, for example. To assign function to ('program function into') these molecules, one could use standard discovery methods of trial-and-error screening. One could also use sophisticated design methods as for organic molecules, and indeed the techniques developed for the *de novo* design of proteins (CAPE – computer-aided protein engineering). The latter does allow access to some genomics by potentially using structural data from homologous proteins, for example. However, in the case of screening and when design is based purely on basic physicochemical principles, these routes certainly do *not* 'map us directly' to genomic data.

However, there are special design shortcuts. The retroinverso approach [78] has proven promising. In this approach, the sequence is synthesised not only with D-amino acids, but also backwards (C-terminal-most residue at the N-terminus and *vice versa*). The principal consequence of these operations is that it is similar to taking the normal L-sequence, but with the amine and carbonyl backbone groups exchanged. This does not in principle affect the internal hydrogen bonds (NH ··· OC becomes CO ··· HN with equivalent geometry) but the carbonyl–β-carbon interactions in particular change, affecting the handedness of helices and sheets (which inevitably have a degree of twist). This information is still 'genomic' in origin: we apply simple operations to protein sequence data from biological systems. Gryphon Sciences has developed further 'quasi-retroinverso' approaches based on further 'rules of thumb' and powerful proprietary *experimental* methods for the programming function, based on genomic data, into these molecules, so that they may interact with their biological targets. Finally, we note that D-peptide structures may also contain a number of L-amino acids without being subjected to proteolysis.

5.5. Ribozymes

These can be produced by DNA-dependent RNA polymerases, as proteins can be produced biologically on ribosomes by cloning and expression, even when the producing organism is not the natural origin (e.g. as in the use of bacteria and yeast in biotechnology). The question of the difficulty of precision chemical synthesis in the laboratory, outside cells or ribosomal systems, does not usually arise.

5.6. Nanoscopic structures

Making large complex molecules with precision has been described by several different terms: precision macromolecular chemistry; ultrastructural chemistry; macrostructural chemistry; supramolecular chemistry; nanoscopic chemistry; nanochemistry; chemical nanotechnology. They all relate to compounds approximately in the 1000 Da range and higher, and typically to structures 20–2000 nm (0.000000001 m) across. However, the term would cover the case of assembling these modular fashions to form larger entities, and to some extent to the production and linkage of the lower-molecular-weight building blocks to produce the essential basic structure.

Natural proteins may be modified by chemical methods to include unnatural sections [79]. Nanoscopic structures may also contain protein-like components chemically synthesised and cloned and expressed. When such components are mixed, one can speak of the molecules being 'bionic' (having synthetic parts as well as natural parts). An advantage of mimicking nature as closely as possible is that it helps design. Designing complex structures can be difficult. Although we do not yet know how proteins achieve their three-dimensional structure (see the discussion of the folding problem above), we can at least borrow from the structures generated by the design process of nature, which is some 3–4 billion years of mutation and natural selection. It is rather as if nanomachinery from an extraterrestrial race has fallen into our hands, and we can to some extent reverse-engineer it.

Nanoscopic systems are conveniently assembled from modular components and are well suited to carry out multiple functionalities, rather like a Swiss Army Knife^{TM}. Gryphon Sciences also uses the colloquial description of 'Molecular Battlewagon' for more sophisticated constructions. An advantage of this modular aspect means also that the parts can be preplaced (at least in subsequent resynthesis of the whole entity) cassette fashion. By keeping the molecular compounds down to the minimum without extraneous complexity, this allows a more rigorous exploration of how to optimise the functionality of the component for pharmaceutical applications.

Synthetic systems such as synthetic peptide vaccines and the even more sophisticated gene therapy vehicles require a variety of functionalities. These are amongst the earliest of the entities that may be deemed 'nanoscopic'. Good modern synthetic peptide vaccines require a molecular frame and functionalities including B-epitopes, T-epitopes, CTL-epitopes, molecular adjuvant, immunostimulation, targeting and delivery.

DNA vaccine vectors require a molecular frame and functionalities for (i) holding the DNA plasmid, (ii) targeting, cell entry, and (iii) endosome escape.

Gene therapy devices require a molecular frame and functionalities for (i) holding the DNA plasmid, (ii) targeting, cell entry, (iii) endosome escape, (iv) nuclear entry and (v) incorporation into chromosome.

It is easy to see that the molecular weights of such structures can easily exceed 150 kDa. This quite naturally brings such therapeutics into the range of known nanoscopic chemistry and, at the very least, borders on meeting the long-awaited chemical requirements for nanotechnology. In other words, the nanoscopic character is need-driven and emerges from what we wish to make, rather than being a technology looking for an application.

A constant feature of this nanoscopic chemistry is the need for 'molecular due diligence'. The majority of steps should be followed by high-resolution atomic mass spectroscopy, high-performance liquid chromatography and nuclear magnetic resonance spectroscopy (see, for example, Ref. 33). In this verification, one must be prepared to accept a reduction in yield. This helps ensure the precision of the chemistry, i.e. that, chemically speaking, each atom is in its place, and that the molecule is folded up correctly, so that each atom is in its right spatial position. The criteria for success may be more demanding than the normal analytical criteria, so the term 'precision' as opposed to 'purity' is often employed. This problem of placing every atom in its correct spatial position applies equally to folding.

6. Available molecular modelling software

Conformational energy and related calculations relate the chemical formula to physicochemical properties as a function of conformation. Classical organic drugs are not as rich in conformational possibilities as their larger counterparts, so conformation, while remaining an issue, is much less of a concern. Biomolecules in general are a more difficult task, and proteins remain the supreme problem.

6.1. Software using statistical methods

These methods are wholly empirical; they make predictions and build models on the basis of what has been seen before. Energy is only implicit, in the form of probabilities or 'equilibrium constants' for distributions of conformational studies in databases of sequence structure relationships. Only the method developed by the present author and colleagues is mentioned here, despite *many* important contributions by other workers. Variations of this method, at least in the earliest research phases, had language-like forms to facilitate exploring the effect of addition and neglect of different contributions (e.g. from pairs of residues, hydrophobic patterns, different definitions of secondary structure state). Early published forms even had a simple language-like 'driver' with a facility for redefining details of automatic runs (see Appendix III of Robson and Garnier (1984,1986)) [54]. Such language-like features at early stages of development, perhaps just as much as its sound theoretical basis of broad power, was a major factor leading to its widely applauded automatic use and reproducibility (for a discussion see Robson and Garnier (1993) [54]).

Statistical methods are well illustrated by references to studies on peptide and protein systems, and the stereochemical code. To have a greater mastery of the stereochemical code would however allow more than simple 'adapting'. Indeed, it would not confine us to protein systems, but extend the scope to nanoscopic pharmaceuticals. Earlier the author has argued that the stereochemical code should be seen as a code having the form

$$\{S\} = T\{R\}$$

where $\{R\}$ is the string of amino acid residues, $\{S\}$ is the string of residue conformational states, and T is the transformation operator, the elucidation of which by any means would represent breaking the code. Statistical analyses have been applied in the author's laboratory to a variety of other possible correlates, such as the way in which amino acid residues substitute in the course of evolution [50], and to spatial distributions of amino acid residues [51]. However, simply considering the amino acid sequence $\{R\}$ as an input message with a linear output $\{S\}$ has been the most revealing.

In the realm of secondary structure prediction and engineering, this leads to the widely used GOR method [52], the algorithm for which has been reported as a computer program [5,53] and widely reproduced in most commercial and academic bioinformatics software. In fact, the method has a long history commencing in 1970, and some of the statistical reasoning described in these earlier studies has more general application (see Ref. 54 for a compilation). Notably, the method has been formally expressed as a theory of expected information [49]. Precisely, it represented the use of Bayes theory of probability as degrees of belief to obtain expected values of information from finite data. The Bayes approach, poorly exploited at that time save at the Department of Statistics at Cambridge and by Simon French at Oxford, is now widely used and the similar use of Bayes factors is recognised as a powerful approach to quantifying evidence in favour of a scientific theory [55]. These observations are relevant to combining in a proper, fairly weighted, manner the contributions of information from *ab initio* data, experimental data, databases, and human expertise sources for drug design as discussed above. The ability of Bayes reasoning to link human belief and confidence in a theory is important here. Similarly, the chemical results of information theory may also be related to the statistical weights of thermodynamics and statistical mechanics. A simple result from helix-coil transition theory illustrates the point [52]. Though the 1978 GOR method is still widely used, it has been continuously refined over the years (e.g. Refs. 56–58) and the widely distributed software based on the earlier method is not really representative of the full power of the approach.

This view of protein structure remains essentially linear. A full three-dimensional appreciation requires molecular modelling by energy (or force) calculations.

6.2. Conformational modelling software

A great deal of background work tends to go into a modelling suite. Whereas the front end language may be prominent to the routine user, the matters behind the scenes are the development of the force field, the development of methods of simulation and the searching of conformational space, and the calibration and testing of all of these.

A variety of available software is shown in Table 1. The language aspect is fairly well developed in Tripos and BioSym codes, and though Polygen is now amalgamated the original Polygen codes were fairly sophisticated. BioSym programs Insight and Discover represent some of the earliest and still best established software in the field. BioSym was scientifically founded by Dr. Arnie Hagler. BioSym's original language-like instructions did have some nice human language-like features, but are basically keyword based in the manner of keyboard-operated adventure games. At MAG (Molecular Application Group – a commercial enterprise), Michael Levitt's LOOK emphasises use as a front end to the Internet, and with the principal exception of his search language SOOP, it is a 'Web-user' Menu feel which is implemented. For the most part in all the above, the underlying codes are not generally written in a specialised proprietary language, though some, like the software of Chemical Design, use nice underlying customised language to control screen display. Otherwise, apart from some batch-type capability using the control language, they are for the most part written in standard programming languages, especially C, and FORTRAN for the older forms.

Rightly or wrongly these aspects have so far been all intertwined and the force field and search methods are often unique features of systems, 'coming with them' as an inherent part of their culture. Amber uses its own force field and the BioSym force field has its origin in the consistent force fields of S. Lifson's laboratory at the Weizmann Institute, which were conceptually also the starting point for LUCIFER (see below) parameters, the parameters of Michael Levitt, and the 'universal force field' found in the graphics interface Pimms of Oxford Molecular. Many of the other programs use MM1, MM2 or MM3 force fields, as does Macromodel from Columbia University, which is one of the best vehicles for these widely used sets of potentials. Although protein calculation parameters tend to be of similar end result quality in most systems, Macromodel has tended to pay attention to sugar parameters [77],

Table 1 *Available software*

Analex Laboratories for Molecular Design
3550 General Atomics Ct.
San Diego, CA 92121
Phone: 619-455-3200
Fax: 619-455-3201

Aldrich Chemical Co.
P.O. Box 2060
Milwaukee, WI 53201
Phone: 414-273-3850
Fax: 414-287-4079

APOCOM
1020 Commerce Park Drive
Oak Ridge, TN 37830
Phone: 423-482-2500
Fax: 423-220-2030

Biosoft U.K.
49 Bateman St.
Cambridge CB2 1LR, U.K.
Phone: 1223-68622
Fax: 1223-312873

Biosoft U.S.
P.O. Box 10938
Ferguson, MO 63135
Phone: 314-524-8029
Fax: 314-524-8129

Chemical Design Inc.
Roundway House
Cromwell Park
Chipping Norton
Oxfordshire OX7 5SR, U.K.
Phone: 1608-644000
Fax: 1608-642244

Chemsoft Inc.
892 Main St.
Wilmington, MA 01887
Phone: 508-567-8881
Fax: 508-657-8228

Cherwell Scientific Publishing Lpd.
Magdalen Centre
Oxford Science Park
Oxford OX4 4GA, U.K.
Phone: 0865-784800
Fax: 0865-784801

Cray Research Inc.
644-A Lone Oak Dr.
Egan, MN 55121
Phone: 612-683-3538
Fax: 612-683-7198

DNASTAR
1228 S. Park St.
Madison, WI 53715
Phone: 608-258-7420
Fax: 608-258-7439

Genomic SA
B.P. 43, F-74160 Collonges sans Saleve
France
Phone: 5043-6765
Fax: 5043-6870

Hypercube, Inc.
419 Phillip St.
Waterloo, ON
Canada N2L 3X2
Phone: 519-725-4040
Fax: 519-725-5193

Molecular Applications Group
445 Sherman Ave.
Palo Alto, CA 94306
Phone: 415-473-3030
Fax: 415-473-1795

Molecular Arts Corp.
1532 East Katella Ave.
Anaheim, CA 92805
Phone: 714-634-8100
Fax: 714-634-1999

MDL Information Systems Inc.
14600 Catalina St.
San Leandro, CA 94577
Phone: 510-895-1313
Fax: 510-483-4738

Molecular Simulations Inc.
16 New England Executive Park
Burlington, MA 01803
Phone: 617-229-9800
Fax: 617-229-9899

National Biosciences Inc.
3650 Annapolis Lane #140
Plymouth
MN 55447
Phone: 612-550-2012
Fax: 612-550-9625

Oxford Molecular
The Magdalen Centre
Oxford Science Park
Stanford on Thames
Oxon
OX4 4CA, U.K.
Phone: 1865-7846000
Fax: 1865-784601

Oxford Molecular U.S. Office
700 E. El Camino Real
Mountain View
CA 94040
Phone: 415-952-7300
Fax: 415-962-7302

Softshell International
715 Horizon Dr. #390
Grand Junction
CO 81506
Phone: 303-242-7502
Fax: 303-242-6469

Terrapin Technologies Inc.
750-H Gateway Blvd.
South San Francisco
CA 94080
Phone: 415-244-9303
Fax: 415-244-9388

Tripos Inc.
1699 S. Hanley, Suite 303
St. Louis, MO 63144
Phone: 314-647-1099
Fax: 314-647-9241

WindowChem Software
1955 West Texas St.
#7-288, Fairfield,
CA 95433-4462
Phone: 707-864-0845
Fax: 707-864-2815

which have in the past been somewhat neglected in commercial software, and its force fields overall fare well in comparative tests [78]. Though LUCIFER is now largely obsolete, its force fields are in the public domain and remain of value. Two semiapproximate force fields for peptides and proteins are of interest in LUCIFER. A method suitable for the rigid backbone approach is increasingly popular [61], and even cruder models, suitable for early stages of modelling, have been parametrized [62]. At the other end of the scale of resolution, LUCIFER OFF (orbital force fields) force fields separately represented non-core orbitals such as lone pair orbitals and hybrid forms, first identified by high grade quantum mechanical calculation. They have been calibrated not only for modelling of proteins [63], and biologically active peptides [64], but also for nucleic acids and particularly analysis of sugar systems (for a compilation see Ref. 65). There were also special approximate methods for vaccine design. For a compilation of calculations using LUCIFER and related software, see Ref. 66.

Most of these are strong on graphics, compared with Prometheus^{TM} described below which is (or was in 1995) more than acceptable on graphics but exceptionally strong on the language element. Graphics systems sometimes approach a language element if they have a powerful menu system. In the extreme case, the selection of a limited vocabulary from a Menu can be equivalent to writing a language input. Unichem of Cray Research Inc. is essentially Menu and graphics driven to make supercomputer chemistry more approachable. The commands relate to running programs in a batch mode at a remote site, particularly of quantum mechanics (most recently, of density functional theory type) and molecular dynamics simulations. SCULPT illustrates particularly well some of the strengths of a graphics approach, in which modelling proceeds by pulling the chain round with a cursor, with the chain responding in a natural way. HyperChem produced by Hypercube also does an excellent job of blending dynamics simulations with graphic manipulation.

Some systems emphasise data management in regard to chemical structures and properties. MDL information systems are well known in the area and use the concept of spreadsheets (they allow integration with Excel spreadsheet software). The language is well suited to this: a single command, for example, places chemical structures and data onto the spreadsheet. Chemical Design can also produce excellent structure searching software and has a neat little facility for generating combinatorial chemistries by computer and searching on them. SoftShell and also WindowChem Software emphasise not only the drawing of molecules and the management of structure databases, but also the preparation of documents and interaction with the Web. SoftShell claims a powerful use of both formatted and free-formatted chemical data. Aldrich Chemical Co. also produces software which shows the spectra of over 10 000 substances and facilitates their drawing. TRAP^{TM} from Terrapin Technologies Inc. predicts relationships between chemicals and their binding targets; by computationally screening compound libraries, TRAP matches chemical fingerprints against those in the database.

Some systems emphasise searching conformational space both to locate the lowest free energy state of molecules, and to calculate the properties of flexible systems.

Searching conformational space was the major point of LUCIFER academic software at the University of Manchester, developed early in the 1970s to 1980s. Amongst the 'physics-flavoured' *ab initio* approaches independent of any experimental data or expert system style input, the more interesting search options include, in the system Prometheus^{TM} at Proteus (see below), a number of methods which modify the laws of physics (for a compilation see Ref. 67). The ability to separate harmonic or rigid parts of the calculation from Newtonian laws calculated by reiteration has recently attracted interest [68,69]. MacroModel reflects an interest in searching conformational space, and many excellent routines which emphasise this aspect have been implemented in MacroModel [80–87]. The ability to study motion and broad conformational changes (which relates to the matter of entropy) is important in designs where one may not wish to have a compact structure. It has been argued that simple artificial enzymes, based on peptides, benefit from a high degree of flexibility [70]. In some cases a better approach to design might be based on statistical coil theory [5,71]. One may also wish to select sequences which are to be deliberately floppy and so avoid difficulties of compaction ('premature folding') in the conditions of expression or chemical synthesis [72]. At the other extreme, relatively rigid sections might be linked by synthetic bonds [79,88]. Ultimately, all these matters are to do with a correct and full treatment of the entropy, which is enormously difficult. Future languages might well emphasise this aspect.

Pangea is a company in Northern California that has emphasised integration tools. Integration is already an important feature of the large BioSym and Tripos codes. Tripos also excels in the manipulation of chemical structure data, and blends this in a fairly powerful language construct with simulations and quantitative structure–activity analysis. This allows a high degree of integration to combine conformational field analysis, distance constraints, and constrained conformational searching in modelling studies. The system is fairly large to accommodate all this. It includes model building, generation of spectra and other data, including structure–activity relationship data, in an elegant manner. It can also perform property calculations, protein structure analysis, evaluate binding energies for drug design, generate proposal drugs, etc.

The need for a system to *integrate* molecular work is increasingly recognised. Oxford Molecular, in comparison, had a huge variety of software from academic contributions, and from corporate acquisitions, and this was in consequence poorly integrated. It is clear that this integration is part of the battle plan with a major deal struck between them and Glaxo Wellcome. "The integrated and expandable environment will combine the proprietary features of Oxford's molecular management tools with additional advanced analysis and computational chemistry methods from Glaxo Wellcome – explains A.F. Marchington, CEO, of Oxford Molecular – Our goal is to streamline the drug discovery process by developing a computational environment that is rich in functionality, familiar in look and feel across platforms and integrated so that results can be shared by scientists working at each stage of the drug discovery" [2]. Integration at this kind of level was already a key design goal in the development of the system Prometheus™, and most of the aspirations are already implemented and being explored (in 1995), albeit on a currently small network of some 32 workstations and 2–3 mini-supercomputers. The concept of the uniform environment as a 'polymorphic programming environment' [8] was also inherent in the Prometheus system. The GLOBAL © language also allowed true asynchronous communication and fluid flow of data and routines across different hardware platforms.

GLOBAL © and the Prometheus system also sought to exploit one aspect of power of the kind of network described by Marchington, namely to capture expertise. Of course, as discussed above, expertise capture is difficult. Sophisticated expertise embodiment was however achieved in the Prometheus system by 1994–1995 in some important areas, most successfully in regard to protocols for the automatic generation of organic molecules and peptides to fit binding sites, structure–activity data, or both, and, in the case of proteins, in elaborate forms of improved modelling by homology. More powerful metasystems, expert in extracting general expertise, are probably required, but restrictions in the power of computer languages as currently perceived may be a limiting factor in this.

7. Conclusions

The vast majority of software available does not have a truly powerful, flexible and general command language facility in comparison to that in general programming languages, operating systems and expert systems, though graphics are often well developed. The best one can do at any time is build in the capability for development and growth. Integration in the commercially available systems is often good internally, but the intersystem communication capabilities are still primitive.

Generally, the biggest difficulty is perhaps that more intelligent systems are required to facilitate the capture of expertise into a language. Complex protocols in modelling and design cannot easily be automated to a level required for expertise capture. It is likely that relating the structure of the problem better to the structure of the language will facilitate the problem-to-solution mapping. In any event, there is currently insufficient means to trap high levels of expertise which will allow the expert in future to be free to concentrate on other tasks, or tackle the design task at a higher level without needing to concentrate on fine details. Modelling work of a complex nature is often not exactly reproducible. The system Prometheus™ and its language GLOBAL © sought to overcome these difficulties. The capture of human expertise still proves difficult, so clearly language improvements are required to facilitate this. Ultimately a system expert in catching expertise is required, and this should hold for all systems who wish to develop with use and not be set in stone as legacy systems.

The considerations for powerful, integrated, automatic drug design languages are complex, but at least appear to be definable. What is clear is that great care must be taken in the definition and development of the languages required, since a plethora of different nonstandard forms would retard the field. It is timely to consider cautiously the design of a single unified *lingua franca* in computational drug design.

References

1. Goodman, A.G., Rall, T.W., Nies, A.S. and Taylor, P. (Eds.) Goodman and Gilman's The Pharmacological Basis of Therapeutics, McGraw Hill, New York, NY, 1993; Issebacher, K.J., Braunwald, E., Wilson, J.D., Martin, J.B., Fauci, A.S. and Kasper, D.L. (Eds.) Harrison's Principles of Internal Medicine, McGraw-Hill, New York, NY, 1992, see especially pp. 401–412, 2425–2429.
2. Newspaper report: interview with Thomas Raechle, Genet. Eng. News, October 8(1995)11.
3. Robson, B., Trends Biochem. Sci., 2(1980)240; Robson, B., CRC Crit. Rev. Biochem., 14(1984)273; Robson, B. and Garnier, J., Introduction to Proteins and Protein Engineering, Elsevier, Amsterdam, New York, NY, Oxford, 1986, 1988, see especially Chapter 12.
4. Dirac, P.A.M., Foundations of Quantum Mechanics, Clarendon Press, Oxford, 1947; Rae, A., Quantum Physics: Illusion or Reality?, 1986; Hawing, S. and Penrose, R., The Nature of Space and Time, Princeton University Press, Princetone, NJ, 1992; Drexler, K.E., Nanosystems. Molecular Machinery, Manufacturing and Computing, Wiley, New York, NY, 1992.
5. Wolf, G., Wired, February (1996)107; Decembers, J., Presenting Java, Sams.net Publishing, Indianapolis, 1995.
6. Witala, S.A., Discrete Mathematics: A Unified Approach, McGraw-Hill, New York, NY, 1987.
7. Wall, L. and Swartz, R.L., Programming Perl, O'Reilly and Associates Inc., Sabestepol, CA, 1991; Henderson-Sellers, B., A Book of Object-Oriented Knowledge, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1995; Coad, P. and Yourdon, E., Object Oriented Design, Prentice-Hall/Yourdon Press, 1991; Sakkinen, M., ECOOP '89: Proceedings of the European Conference on Object Oriented Programming, Cambridge University Press, Cambridge 1989; Borgia, A., Mylopoulos, J. and Wong, H.K.T., On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases, and Programming Languages, Springer, Berlin, 1984; Rumbaugh, J., Balha, M., Premerlani, W., Eddy, F. and Lorensen, W., Object-Oriented Modelling and Design, Prentice-Hall, Englewood Cliffs, NJ, 1991; Armstrong, J.M. and Mitchell, R.J., Software Eng. J., January (1994)2; Henderson-Sellers, B. and Edwards, J.M., BOOK TWO of Object-Oriented Knowledge: The Working Object, Prentice-Hall, Englewood Cliffs, NJ, 1994; Henderson-Sellers, B., Report on Object Analysis and Design, 1995, pp. 48–51; Graham, I.M., Migrating to Object Technology, Addison-Wesley, Reading, MA, 1995.
8. Ball, J., Fishleigh, R.V., Greaney, P.J., Marsden, A., Platt, E., Pool, J.L. and Robson, B., In Bawden, D. and Mitchell, E.M. (Eds.) Chemical Information Systems – Beyond the Structure Diagrams, Ellis Horwood, Chichester, 1990, pp. 107–123; Robson, B., Platt, E. and Li, J., In Beveridge, D.L. and Lavery, R. (Eds.) Theoretical Biochemistry and Molecular Biophysics 2 Proteins, Adenine Press, Guilderland, NY, 1992, pp. 207–222; Fishleigh, R.V., Robson, B., Garnier, J. and Finn, P.W., FEBS Lett., 2(1987)219.
9. Robson, B., Ball, J., Fishleigh, R.V., Greaney, P.J., Li, J., Marsden, A., Platt, E. and Pool, J.L., Biochem. Soc. Symp., 57(1991)91.
10. Clark, D.E., Frenkel, D., Levy, S.A., Li, J., Murray, C.W., Robson, B., Waszkowycz, B. and Westhead, D.R., J. Comput.-Aided Mol. Design, 9(1995)13.

11. Waszkowycz, B., Clark, D,E., Frenkel, D., Li, J., Murray, C.W., Robson, B. and Westhead, D.R., J. Med. Chem., 37(1994)3994.
12. Westhead, D.R., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B. and Waszkowycz, B., J. Comput.-Aided Mol. Design, 9(1995)139.
13. Frenkel, D., Clark, D.E., Li, J., Murray, C.W., Robson, B., Waszkowycz, B. and Westhead, D.R., J. Comput.-Aided Mol. Design, 9(1995)213.
14. Murray, C.W., Clark, D.E., Frenkel, A.D., Li, J., Robson, B., Waszkowycz, B. and Westhead, D.R., Proceedings of the 1st European Conference on Computational Chemistry, Circulated Report.
15. Robson, B., The Biochemist, October/November (1994)46.
16. Robson, B. and Garnier, J., Nature, 361(1993)506.
17. Cohen, J., Science, 270(1995)908.
18. Senior, A.E., Robson, B. and Sherratt, H.S.A., Biochem. J., 110(1968)511.
19. Gura, T., Science, 270(1995)575.
20. D'Amico, E., Chem. Week, October 25(1995).
21. Muir, T.W. and Kent, S.B., Curr. Opin. Biotechnol., 4(1993)420.
22. Dawson, P.E., Muir, T.W., Clark-Lewis, I. and Kent, S.B., Science, 266(1994)776.
23. Offord, R.E., In Hook, J.B. and Poste, G. (Eds.) Plenum, New York, NY, 1990, pp. 253–282; Rose, K., Zeng, W., Brown, L.E. and Jackson, D.C., Mol. Immunol., 32(1996)1031; Mikolajczyk, S.D., Meyer, D.L., Starling, J.S., Law, K.L., Rose, K., DuFour, B. and Offord, R.E., Bioconjugate Chem., 5(1994)636.
24. Rose, K., Chem. Soc., 116(1994)30.
25. Robson, B., Nature, 248(1974)636.
26. Robson, B., Nature, 249(1974)409.
27. Robson, B., Nature, 250(1974)707.
28. Robson, B., Nature, 262(1974)447.
29. Robson, B., Trends Biochem. Sci., 3(1976)49; Robson, B. and Garnier, J., Introduction to Proteins and Protein Engineering, Elsevier, Amsterdam, New York, NY, Oxford, 1986, 1988, see Chapters 3, 5 and 7.
30. Robson, B., Nature, 254(1975)386; 256(1975)89; 267(1977)577; 283(1980)622; Robson, B. and Jones, M.N., Nature, 272(1978)206.
31. Robson, B., Amino Acids, Peptides and Proteins, The Stonebridge Press – John Wright and Sons Ltd. for The Chemical Society, 1971; 4(1972)224; Amino Acids peptides and Proteins, 5(1974)180; Theoretical Aspects of Protein Folding, Hagler, B.A.T. and Robson, B., Amino Acids Peptides and Proteins, 6(1975)206; Robson, B. and Osguthorpe, D.J., Amino Acids Peptides and Proteins, 7(1976)176; Robson, B., Asher, M. and Osguthorpe, D.J., Amino Acids Peptides and Proteins, 8(1977)181; Robson, B. and Osguthorpe, D.J., Amino Acids Peptides and Proteins, 9(1978)196; 10(1979)208; see also Robson, B., BioEssays, 8(1988)93.
32. Robson, B., Douglas, G.M., Metcalf, A., Woolley, K. and Thompson, J.S., In Franks, F. and Mathias, S.F. (Eds.) Biophysics of Water, John Wiley, New York, NY, 1982, pp. 21–23.
33. Loo, J.A., Bioconjugate Chem., 6(1995)644.
34. Gray, W.R., Protein Sci., 2(1993)1732; see also 2(1993)1749.
35. Robson, B. and Pain R.H., Biochem. J., 155(1976)325.
36. Robson, B. and Pain, R.H., Biochem. J., 155(1976)331.
37. Richardson, J.S. and Richardson, D.C., Trends. Biochem. Sci., 14(1989)303.

38. Hill, C.P., Anderson, D.H., Esson, L., Delgrado, W.F. and Eisenberg, D., Science, 249(1990)543.
39. Hecht, M.H., Richardson, J.S., Richardson, D.C. and Ogden, R.C., Science, 249(1990)884.
40. Federov, A.N., J. Mol. Biol., 225(1992)927.
41. Pessi, A., Bianchi, E., Crameri, A., Venturini, S., Tramantano, A. and Sallazzo, M., Nature, 362(1993)367.
42. Bryson, J.W., Betz, S.F., Lu, H.S., Suich, D.J., Zhou, H.X., O'Neil, K.T. and Delgrado, W.F., Science, 270(1995)935.
43. Thorson, J.S., Chapman, E. and Schultz, P.G., J. Am. Chem. Soc., 117(1995)9361.
44. Page, H.E., Hallewell, R.A. and Tainer, J.A., Proc. Natl. Acad. Sci. USA, 89(1992)6109.
45. Borman, S., Scientists Refine Understanding of Protein Folding and Design, pp. 29–35, Chem. Eng. News, May (1996).
46. Strader, C., Fong, T.M., Garaziano, M.P. and Tota, M.R., FASEB J., 9(1995)745.
47. Canne, L.E., Ferre-D'Amare, A.R., Burley, S.K. and Kent, S.B.H., J. Am. Chem. Soc., in press.
48. Muir, T., Williams, M.J., Ginnsberg, M.H. and Kent, S.B.H., Science, 33(1994)7701.
49. Robson, B., Biochem. J., 141(1974)853.
50. French, S. and Robson, B., J. Mol. Evol., 19(1983)171.
51. Crampin, J., Nicholson, B.H. and Robson, B., Nature, 272(1978)558.
52. Garnier, J., Osguthorpe, D.J. and Robson, B., J. Mol. Biol., 120(1978)97.
53. Robson, B., Douglas, G.M. and Garnier, J., In Geisow, M.J. and Barrett, A.N. (Eds.) Computing in the Biomedical Sciences, Elsevier, Amsterdam, 1983, pp. 132–142.
54. Robson, B. and Garnier, J., Introduction to Proteins and Protein Engineering, Elsevier, Amsterdam, 1984, 1986 (see the Appendices); Pain, R.H. and Robson, B., Nature, 227(1970)62; Robson, B. and Pain, R.H., J. Mol. Biol., 58(1971)237; Pain, R.H. and Robson, B., Proceedings of the lst Biophys. Congress, Vol. 1, 1971, pp. 33–37; Robson, B. and Pain, R.H., Nature New Biol., 238(1972)107; Robson, B. and Pain, R.H., Conformation of Biological Molecules and Polymers, Academic Press 5(1973)161; Robson, B. and Pain, R.H., Biochem. J., 141(1974)869; Robson, B. and Pain, R.H., Biochem. J., 141(1974)883 ; Robson, B. and Pain, R.H., Biochem. J., 141(1974)899; Schulz, G.E., Barry, C.D., Griedman, J., Chou, P.Y., Fasman, G.D., Finkelstein, A.W., Lim, V.I., Ptitsyn, O.B., Kabat, E.A., Wu, T., Levitt, M., Robson, B. and Nagano, K., Nature, 250(1974)140; Robson, B. and Suzuki, E., J. Mol. Biol. 107(1976)327; Garnier, J. and Robson, B., Rap. d'Activite Scientific du CECAM, 1979, pp. 147–149; Garnier, J., Gaye, P., Mercier, J.C. and Robson, B., Biochimie, 62(1980)231; Levin, J., Robson, B. and Garnier, J., FEBS Lett., 205(1986)303; Robson, B. and Garnier, J., Nature, 361(1993)506.
55. Kass, R.E. and Raferty, A.E., J. Am. Statist. Assoc., 90(1995)773.
56. Suzuki, E. and Robson, B., J. Mol. Biol., 107(1976)357.
57. Gibrat, J.F., Garnier, J. and Robson, B., J. Mol. Biol., 198(1988)425.
58. Gibrat, J.F., Robson, B. and Garnier, J., Biochemistry, 30(1991)1578.
59. Garnier, J. and Robson, B., In Fasman, G.D. (Ed.) Prediction of Protein Structure and the Principles of Protein Conformation, Plenum, New York, NY, 1989, pp. 417–465.
60. Robson, B., Hillier, I.H. and Guest, M., J. Chem. Soc., Faraday Trans. II, 74(1978)1311; Hillier, I.H. and Robson, B., J. Theor. Biol., 76(1978)83; Robson, B., CRMC2 Reports, A Circulated Report, Marseille, 1978, pp. 44–45; Robson, B., Stern, P., Hillier, I.H., Osguthorpe, D.J. and Hagler, A.T., J. Chim. Phys., 76(1979)831; Hagler, A.T., Osguthorpe, D.J. and Robson, B., Science, 208(1980)599; Finn, P., Morffew, A., Robson, B., Glasel, J.A.,

Freer, R.J. and Day, A.R., Daresbury Meeting on Simulation of Biomolecules. A Circulated Report, 1981; Platt, E., Robson, B. and Hillier, I.H., J. Theor. Biol., 88(1981)333; Robson, B., In Franks, F. and Mathias, F.S. (Eds.) Wiley, New York, NY, 1982, pp. 66–70; Platt, E. and Robson, B., J. Theor. Biol., 96(1982)381; British Biophys. Soc. Abstracts, Robson, B., A Circulated Report, Leeds University, 1982, p. 1; Platt, E. and Robson, B., In Geisow, M.J. and Barrett, A.N. (Eds.) Computing in the Biomedical Sciences, Elsevier, Amsterdam, 1982, pp. 91–131; Robson, B., Biochem. Soc. Trans., 10(1982)297.

61. Robson, B. and Platt, E., J. Mol. Biol., 188(1986)259; Robson, B., Platt, E., Fishleigh, R.V., Marsden, A. and Millard, P., J. Mol. Graph., 5(1987)8.

62. Robson, B. and Osguthorpe, D.J., J. Mol. Biol., 132(1979)19; Robson, B., Rap. d'Activite Scientifique du CECAM, 1979, pp. 117–120; FEBS Lett., 120(1980)295; Robson, B. and Platt, E., J. Comput.-Aided Mol. Design, 1(1987)17; Robson, B. and Platt, E., J. Comput.-Aided Mol Design, 4(1990)369; Platt, E. and Robson, B., Proc. R. Soc. Edinburgh, 99B (1992)123; Robson, B., Collura, V.P. and Greaney, P.J., Protein Eng., 7(1994)221.

63. Finn, P.W., Robson, B. and Griffiths, E.C., Regul. Pept., 7(1983)286; Finn, P.W., Robson, B. and Griffiths, E.C., Int. J. Pept. Protein Res., 24(1985)407; Ward, D.J., Griffiths, E.C., Robertson, R.G. and Robson, B., Regul. Pept., 13(1985)73; Fishleigh, R.V., Ward, D.J., Griffiths, E.C. and Robson, B., Biol. Chem. Hoppe-Seyler, 367(1986)1112; Robson, B., Ward, D.J., Marsden, A., Fishleigh, R.V., Griffiths, E.C. and Platt, E., J. Mol. Graph., 4(1986)235; Fishleigh, R.V., Ward, D.J., Griffiths, E.C. and Robson, B., Biol. Chem. Hoppe-Seyler (Suppl.), 367(1986)266; Prediction of Preferred Solution Conformers of Analogues and Fragments of Neurotensin, Ward, D.J., Fishleigh, R.V., Platt, E., Griffiths, E.C. and Robson, B., Regul. Pept., 15(1986)197; Griffiths, E.C., Robson, B. and Ward, D.J., Br. J. Pharmacol., 87(1986)177S; Baris, C., Griffiths, E.C., Robson, B., Szirtes, T., Starkie, D. and Ward, D.J., Br. J. Pharmacol., 87(1986)173S; Ward, D.J., Griffiths, E.C. and Robson, B., Int. J. Pept. Protein Res., 27(1986)461; Fishleigh, R.V., Ward, D.J., Griffiths, E.C. and Robson, B., Biochem. Soc. Trans., 14(1986)1259; Griffiths, E.C., Robson, B. and Ward, D.J., Br. J. Pharmacol., 88(1986)361S; Ward, D.J., Finn, P.W., Griffiths, E.C. and Robson, B., Int. J. Pept. Protein Res., 30(1987)263; Griffiths, E.C., Millard, P. and Robson, B., In Metcalf, G. and Jackson, I.M.D., (Eds.) Biomedical Significance, Ann. New York Acad. Sci., 553(1989)487; Griffiths, E.C., Kelly, J.A., Ashcroft, A., Ward, D.J. and Robson, B., In Metcalf, G. and Jackson, I.M.D. (Eds.) Thyrotrophin-Releasing Hormone: Biomedical Significance, Ann. New York Acad. Sci., 553(1989)17; Ward, D.J., Chen, Y., Platt, E. and Robson, B., J. Theor. Biol., 148(1991)193.

64. Wynn, C.H., Marsden, A. and Robson, B., J. Theor. Biol., 119(1986)81; Wynn, C.H., Marsden, A. and Robson, B., Biochem. Soc. Trans., 14(1986)707; Marsden, A., Robson, B. and Thompson, J.S., Biochem. Soc. Trans., 14(1986)530; Marsden, A., Robson, B. and Thompson, J.S., Biochem. Soc. Trans., 14(1986)629; Wynn, C.H. and Robson, B., J. Theor. Biol., 123(1986)221.

65. Robson, B., Platt, E., Finn, P.W., Millard, P., Gibrat, J.G. and Garnier, J., Int. J. Pept. Protein Res., 25(1985)1; Morrison, C.A., Fishleigh, R.V., Ward, D.J. and Robson, B., FEBS Lett., 214(1987)65; Robson, B., Fishleigh, R.V. and Morrison, C.A., Nature, 325(1987)395; Bomford, R., Garnier, J. and Robson, B., In Vassarotti, A. and Magnien, E. (Eds.) Biotechnology R&D in the EC (BAP) 1985–1989, Vols. I and II, Elsevier, Paris, 1990, p. 105 (Vol. I), pp. 59–64 (Vol. II).

66. Robson, B., In Darbre, A. (Ed.) Practical Protein Chemistry – A Handbook, Wiley, London, 1986, pp. 567–607; Robson, B., Stat. Mech. and Thermodynam. Group and SERC/Collaborative Computer Projects, 5(1981)36; Robson, B., Douglas, G.M. and Platt, E., Biochem. Soc. Trans., 10(1982)388; Robson, B., Applied Biotechnology, Proceedings of Biotech '86 Europe, held in London, May 1986, Vol. 1, B9–B14; Robson, B., Ward, D.J. and Marsden, A., Chemical Design Automation News, 1(7)(1986)9; Robson, B., Appl. Biotechnol., 1(1986)B9; Cyber 205 Newsletter, No. 5, January 1986; Royal Society Reports, Royal Society – Sandoz Symposium, 1986; Robson, B., In Hadzi, D. and Jerman-Blazic, B. (Eds.) QSAR in Drug Design and Toxicology, Elsevier, Amsterdam, 1987; Robson, B., Transcript of lecture at CDC seminar in Heidelberg, 1988; Garnier, J. and Robson, B., In Silverman, P.D. (Ed.) Computer Simulation of Carcinogenic Processes, CRC Press, Boca Raton, FL, 1988, pp. 67–90; Robson, B., published transcripts, (1989); Robson, B., In Fauchere, J.L. (Ed.) QSAR: Quantitative Structure–Activity Relationships in Drug Design, Alan R. Liss, New York, NY, 1989, pp. 227–231; Li, J., Brass, A., Ward, D.J. and Robson, B., Parallel Comput., 14(1990)211; Ward, D.J., Brass, A.M., Li, J., Platt, E., Chen, Y. and Robson, B., In Ward, D.J. (Ed.) Peptide Pharmaceuticals, Open University Press, Milton Keynes, 1991, pp. 83–129; Ward, D.J., Brass, A.M., Li, J., Platt, E., Chen, Y. and Robson, B., In Ward, D.J. (Ed.) Conference Proceedings, Peptide Pharmaceuticals, Elsevier, New York, 1991, pp. 83–134.
67. Robson, B., Brass, A., Chen, Y. and Pendleton, B.J., Biopolymers, 33(1993)1307; Li, J., Platt, E., Waszkowycz, B., Cotterill, R. and Robson, B., Biophys. Chem., 43(1992)221; Byrne, D., Li, J., Platt, E., Robson, B. and Weiner, P., J. Comput.-Aided. Mol. Design 8(1994)67.
68. Janezic, D. and Merzel, F., J. Chem. Inf. Comput. Sci., 35(1995)321.
69. Turner, J., Weiner, P.K., Robson, B., Venugopal, R., Schubele III, W.H. and Singh, R., J. Comput. Chem., 16(1995)1271.
70. Robson, B. and Marsden, A., Biochem. Soc. Trans., 15(1987)1191.
71. Robson, B., J. R. Stat. Soc. B, 44(1982)136.
72. Robson, B., Trends Biochem. Sci., 6(1981)XIII; Robson, B. and Millard, P., Circulated document, Biochem. Soc. Symp., Oxford, 1982.
73. Baris, C., Brass, A., Robson, B. and Tomalin, G., In Epton, R. (Ed.) Innovations and Perspectives in Solid Phase Synthesis, 1st International Symposium SPCC, Birmingham, 1990, pp. 441–445.
74. deLisle Milton, R.C., Milton, S.C.F., Scholzer, M. and Kent, S.B.H., Science, 256(1992) 1445.
75. Zawadzke, L.E. and Berg, J.M., Protein Struct. Funct. Genet., 16(1993)301
76. Dintzis, H.M., Symer, D.E., Dintzis, R.Z., Zawadzke, L.E. and Berg, J.M., Proteins Struct. Funct. Genet., 16(1993)306.
77. Senderowitz, H., Guarnieri, F. and Clark Still, W., J. Am. Chem. Soc., 117(1995)8211.
78. Gundertofte, K., Lileforjs, T., Norrby, P. and Pettersson, I., J. Comput. Chem., 17(1996)429.
79. Kent, S.B.H., Baca, M., Elder, J., Miller, M., Milton, R., Milton, S., Rao, J.K.M. and Schnolzer, M., In Takahashi, K. (Ed.) Aspartic Proteinases, Plenum, New York, NY, 1994.
80. Senderowitz, H., Parish, C. and Clark Still, W., J. Am. Chem. Soc., 118(1996)2078.
81. Pappenheimer, J.R., Dahl, C.E., Karnovsky, M.L. and Maggio, J.E., Proc. Natl. Acad. Sci. USA, 91(1994)1942.

82. Chorev, M. and Goodman, J.M., TIBTECH, 13(1995)438.
83. Lipton, M. and Clark Still, W., J. Comput. Chem., 9(1988)343.
84. Chang, G., Guida, W.C. and Clark Still, W., J. Am. Chem. Soc., 111(1989)4379.
85. Saunders, M., Houk, K.N., Wu, Y., Clark Still, W., Lipton, M., Chang, G. and Guida, W.C., J. Am. Chem. Soc., 112(1990)1419.
86. Goodman, J.M. and Clark Still, W., J. Comput. Chem., 12(1991)1110.
87. Guarnieri, F. and Clark Still, W., J. Comput. Chem., 15(1994)1302.
88. Gaertner, H.G., Offord, R.E., Cotton, R., Timms, D., Camble, R. and Rose, K., J. Biol. Chem., 269(1994)7224.

Characterization of the effect of functional groups substitution at the 2-position of adenine on the stability of a duplex dodecamer d(CGCGAATTCGCG)$_2$ by molecular mechanics and free energy perturbation method

Hiroshi Kuramochi[a] **and U. Chandra Singh**[b]

[a] *Nippon Kayaku Co., Ltd., 1-12, Shimo 3-Chome, Kita-ku, Tokyo 115, Japan*
[b] *AM Technologies Inc., 14815 Omicron Drive, Texas Research Park,*
San Antonio, TX 78218, U.S.A.

Introduction

The dynamical structure and stability of duplex DNA is closely related to biological phenomena such as transcription, replication and regulation of gene expression. The modification of bases in DNA is known to have an important influence on the biological function of DNA. Modified bases cause a change in the structure and stability of DNA and, as a result, the interaction of enzymes and regulatory proteins with DNA is modified. There are several reports [1–6] in which the structure change of DNA by base modification is theoretically treated using molecular modeling, molecular mechanics and quantum chemistry, but few studies have adopted a dynamical approach such as molecular dynamics and free energy calculation. Thus, we tried such a dynamical approach to understand the effect of base modification on the stability of the DNA duplex.

The EcoRI endonuclease–DNA system is an attractive system for theoretically studying the effect of base modification on DNA–enzyme interaction, because the crystal structure of the enzyme–DNA cocrystal has been elucidated [7–9] and, based on the structure, the effect of base modification on cleavage reaction has been examined in some detail [10–13]. The crystal structure of the endonuclease–DNA(TCGCGAATTGCGC) complex shows the following characteristics. The DNA is kinked at three sites upon binding of the endonuclease. One kink is located at the center of the recognition site (neo-1) and the other two are located at both termini of the recognition site (neo-2). The kinks appear to be necessary for the enzyme to be accommodated into the major groove and contact functional groups located on the nucleobases. The endonuclease interacts with the recognition sequence only in the major groove of the DNA and the minor groove is open to the solvent. The N7 and N6 of the adjacent adenine and the N7 and O6 of guanine interact with the amino acid side chains of the endonuclease.

Brennan et al. [10] and McLaughlin et al. [11] have reported the effect of functional group change in the recognition site on the cleavage reaction catalyzed by the endonuclease. Their studies have clarified that the change of functional groups which interact directly with the endonuclease causes a varying effect upon the cleavage reaction, as expected. In addition, the introduction of an NH_2 group into the 2-position of either adenine in the recognition site reduced the reactivity. Since this NH_2 group is in the minor groove, it does not appear to interact directly with the endonuclease. The kinetics data show that the modification mainly influences the k_{cat} value and not the K_m value; namely, the catalytic step is inhibited, while the binding step is not. Since all the functional groups of the bases which interact with the endonuclease are not changed by the modification, no alteration of the binding appears to be reasonable. On the other hand, the mechanism of inhibition against the catalytic step is not clear. It appears to come from the alteration of the intrinsic property of DNA, and not from that of the interaction mode with the endonuclease. Thus, we tried to approach this problem from the point of view of the dynamical stability of DNA. We examined the effect of the introduction of NH_2 and other groups into the 2-position of an inner adenine in the recognition site on DNA stability using the free energy perturbation method based on molecular dynamics (MD). The dodecamer duplex of sequence d(CGCGAATTCGCG) was chosen as target DNA, because not only is the dodecamer the same as that of the endonuclease–DNA cocrystal but it is also a typical example of B-form DNA, the crystal structure of which has been studied in detail by Dickerson and co-workers [14].

Molecular dynamics has been developed as a powerful theoretical method to study the dynamical properties of macromolecules, and several groups have reported the MD simulation of DNA [15–25]. It is well known that DNA shows plastic properties such as the bending of whole structure. The negative charges on the phosphate groups mutually induce strong repulsive interactions, and the presence of counterions and solvent water strongly influences the structure of DNA. Such unique properties of DNA have made its MD simulation more difficult as compared to that of most small globular proteins. Although the MD simulation of DNA including water and counterions explicitly has been carried out by a few groups [19,20,25], the conclusion obtained by them is not necessarily in agreement, since they use different types of force fields and oligonucleotide duplexes. Thus, as a first step we examined the MD simulation of DNA in detail using AMBER and a B-form dodecamer duplex, d(CGCGAATTCGCG), which contains one complete helix turn. Then, based on the result of MD simulation, we calculated the effect of base modification on the dynamical stability of DNA.

The thermodynamic perturbation method has been used in calculating the free energy change not only for the chemical alteration of small molecular systems but also for macromolecular systems such as proteins [26]. This method has succeeded in a number of studies on protein–ligand interactions and the stability of protein conformations during the past five years. In contrast, there are only a few reports [27–29] on free energy simulations for nucleic acid systems such as DNA and RNA structures. Thus, the small alteration of a base appears to be a proper model for the free energy calculation of the DNA system.

The stability of the DNA duplex is defined by the equilibration constant between the double-stranded DNA and the two single-stranded DNAs which make up the double strand, that is, the free energy change between them. Since it is extremely

$$\begin{tikzcd}
\text{ss-DNA} + \text{ss-DNA} \arrow[r, "\Delta G_1"] \arrow[d, "\Delta G_3"] & \text{ds-DNA} \arrow[d, "\Delta G_4"] \\
\text{ss-DNA}^* + \text{ss-DNA} \arrow[r, "\Delta G_2"] & \text{ds-DNA}^*
\end{tikzcd}$$

$$\Delta\Delta G = \Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$$

ss-DNA : single-stranded DNA
ss-DNA*: modified single-stranded DNA
ds-DNA : double-stranded DNA
ds-DNA*: modified double-stranded DNA

Fig. 1. Thermodynamic cycle for the formation of a double-stranded DNA from single-stranded DNAs.

difficult to directly simulate the formation and dissociation process of a double-stranded DNA even with the most powerful computer hardware, we used the thermodynamic cycle perturbation method [26] which evaluates only relative free energies. The difference in the free energy change due to the base modification can be evaluated from both the free energy change of the base transformation in a double-stranded DNA and that in a single-stranded DNA instead of the association path corresponding to the physical process as shown in Fig. 1, since free energy is a state function. This nonphysical transformation path is accompanied by less conformational changes as compared to the physical one and, therefore, is much easier to deal with. However, a single-stranded DNA is more flexible and spans a much wider conformation space than a double-stranded DNA. It is not possible to properly sample such a wide conformation space on a picosecond scale. Therefore, we used trinucleotides containing a modified adenine in the middle position instead of modified single-stranded DNAs. This treatment would be valid because the influence of distant residues is supposed to be small in single-stranded DNA. In addition, isolated 9-methyl-adenine derivatives are used to simulate free energy change when the base is completely exposed to solvent water, which is thought to correspond to an unstacked random coil state.

Free energy perturbation method

The free energy perturbation method is based on the statistical perturbation theory developed by Zwanzig [30], and the detailed implementation of the free energy perturbation method into a molecular dynamics program is discussed elsewhere [31]. The essential features of the method are briefly described here.

The free energy difference between two states of a system is computed by transforming one state into the other by changing a single coupling parameter in several steps. If the two states A and B are represented by Hamiltonians H_A and H_B, respectively, the intermediate state between them is given by a dimensionless coupling parameter, λ, as

$$H(\lambda) = (1 - \lambda) H_A + \lambda H_B, \quad 1 \geq \lambda \geq 0 \tag{1}$$

This state is a hypothetical mixture of A and B. When $\lambda = 0$, $H(\lambda) = H_A$, and when $\lambda = 1$, $H(\lambda) = H_B$. Therefore, the conversion of state A into state B can be made smoothly by changing the value of λ in small increments, dλ, such that the system is in equilibrium at all values of λ. The Gibbs free energy change due to the perturbation of the Hamiltonian from $H(\lambda)$ to $H(\lambda + d\lambda)$ is given by

$$\Delta G = - k_B T \ln < \exp(- \Delta H(\lambda)/k_B T) >_0 \tag{2}$$

where k_B is the Boltzmann constant. The average of $\exp(- \Delta H(\lambda)/k_B T)$ is computed over the unperturbed ensemble of the system. If the range of λ is divided into N windows, $\{\lambda_i, i = 1, N\}$, the solute state is perturbed to λ_{i+1} and λ_{i-1} states at each window λ_i, and the free energy difference between states A and B is computed by summation over all the windows as

$$\Delta G = \sum \Delta G(\lambda_i) \tag{3}$$

During the mutation of a molecule, it is necessary to reset its coordinates at every intermediate state, since the coordinates of the final state differ from those of the initial state. Therefore, the coordinates are reset using the technique of coordinate coupling outlined in our earlier work [32].

Computational details

The partial atomic charge assignments for modified bases were calculated by *ab initio* quantum mechanical methods using QUEST [33] with an STO-3G basis set [34], in which quantum mechanically derived electrostatic potentials are fitted to a point charge model [35]. Their force field parameters were assigned from standard values in the AMBER set [36,37], by direct comparison with analogous fragments in the AMBER parameter library, where available. The other parameters that are not in the standard AMBER set were assigned from the MM2 force field set [38] and those bond lengths and angles from the microwave spectra data or X-ray crystallographic data. The atomic charges and van der Waals parameters employed for modified bases are listed in Tables 1 and 2. The initial structure of a duplex dodecamer, d(CGCGAATTCGCG), was built using a standard B-DNA geometry, and the 2-substituted-Ade6 residue of modified duplexes was built using standard bond lengths and bond angles, if necessary, with the dihedral angles adjusted to form a hydrogen bond with the O2 of Thy19. Phosphates were neutralized with 22 Na^+ counterions by placing them 4.0 Å from the OPO bisector. The structure of a single-stranded trimer, d(ApApT), in which the middle adenine is modified, is the same as the corresponding

Table 1 *Charges on the atoms of 2-substituted adenine residues*

Atom	-H	-F	-Cl	-NH_2	-OH	-SH
N9	– 0.088	– 0.021	– 0.029	– 0.019	– 0.004	– 0.047
C8	0.266	0.225	0.239	0.211	0.210	0.243
H8	0.059	0.066	0.067	0.061	0.066	0.061
N7	– 0.539	– 0.524	– 0.526	– 0.533	– 0.525	– 0.543
C5	– 0.076	– 0.015	– 0.017	0.029	0.015	0.009
C6	0.774	0.660	0.671	0.594	0.604	0.647
N6	– 0.778	– 0.699	– 0.708	– 0.679	– 0.672	– 0.712
HN6A	0.321	0.304	0.309	0.297	0.297	0.309
HN6B	0.339	0.326	0.330	0.316	0.317	0.321
N1	– 0.770	– 0.766	– 0.731	– 0.747	– 0.767	– 0.725
C2	0.617	0.888	0.822	0.918	0.941	0.685
X2	– 0.026	– 0.217	– 0.261	– 0.794	– 0.554	– 0.210
HX2A	–	–	–	0.308	0.332	0.132
HX2B	–	–	–	0.321	–	–
N3	– 0.684	– 0.714	– 0.672	– 0.700	– 0.703	– 0.648
C4	0.530	0.432	0.451	0.362	0.388	0.423

part of a duplex dodecamer. The initial geometries of 2-substituted-9-methyl-adenines were built by using the adenine geometry with the bond lengths, bond angles and dihedral angles identical to those of duplexes. The solutes were placed in a rectangular box surrounded by repeating cubes of TIP3P water molecules [39]. Solvent molecules that were closer than 3.6 Å to any solute atom or more than 10 Å to the duplex dodecamer and the trimer and 12 Å to 9-methyl-adenine derivatives along any one of the rectangular coordinate axes were removed. The number of water molecules is approximately 3300, 980 and 740 for the duplex dodecamer, trimer and 9-methyl-adenine derivatives, respectively. The simulations were carried out with these periodic boundary conditions.

Table 2 *Nonbonded parameters of the substituents*

Atom	Atom type	R(A)	ε
H2	HC	1.540	0.010
F2	F	1.650	0.078
Cl2	Cl	2.030	0.240
N2	N2	1.750	0.160
HN2	H2	1.000	0.020
O2	OH	1.650	0.150
HO2	HO	1.000	0.020
S2	SH	2.000	0.200
HS2	HS	1.000	0.020

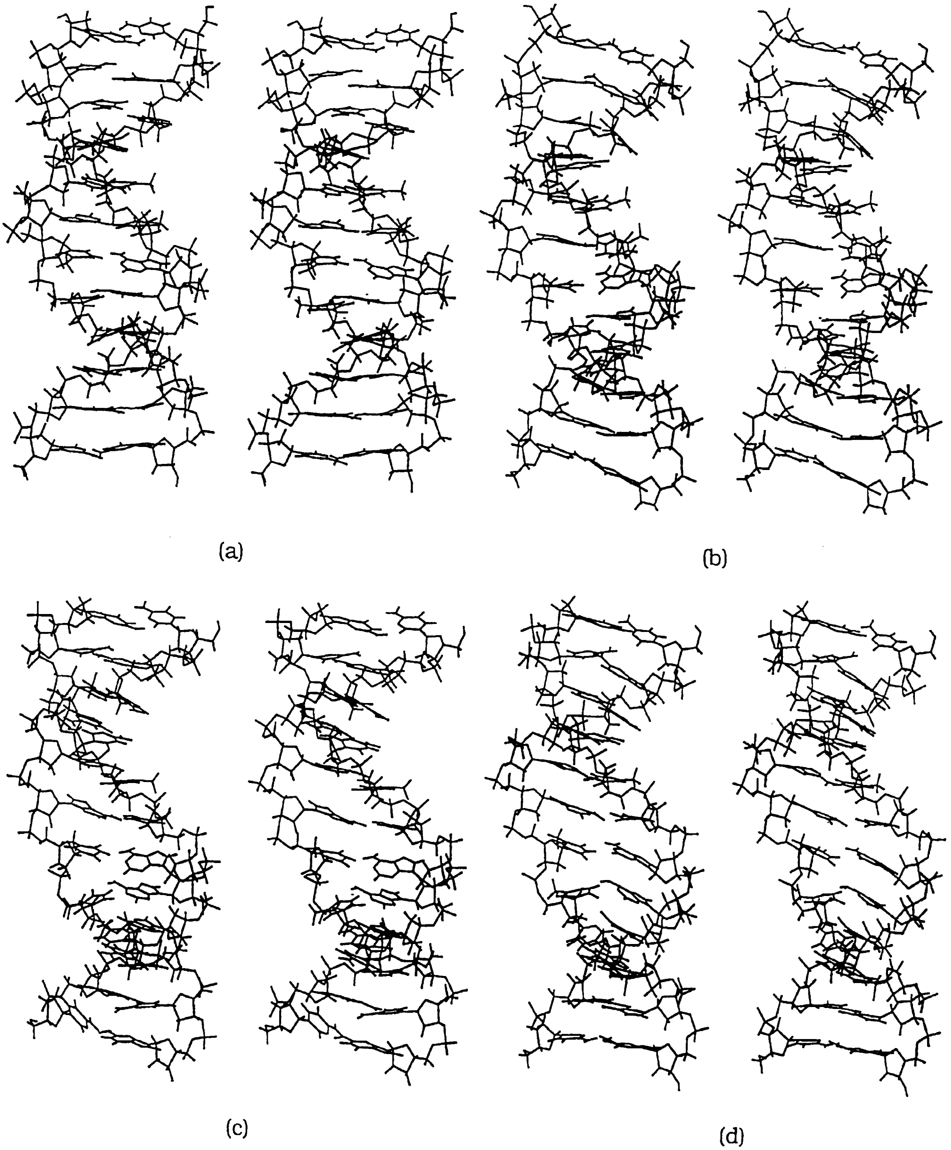


Fig. 2. Structures of d(CGCGAATTCGCG) at (a) 0, (b) 20, (c) 40, (d) 60, (e) 80 and (f) 100 ps of MD simulation. The structure at 0 ps is that after 10 ps of equilibration dynamics.

The initial molecular mechanics, molecular dynamics and free energy perturbation calculations were performed with a fully vectorized version of AMBER (version 3.3 [40a]) on a Cray Y-MP computer system at The Scripps Research Institute. Subsequent molecular mechanics, molecular dynamics and free energy calculations on

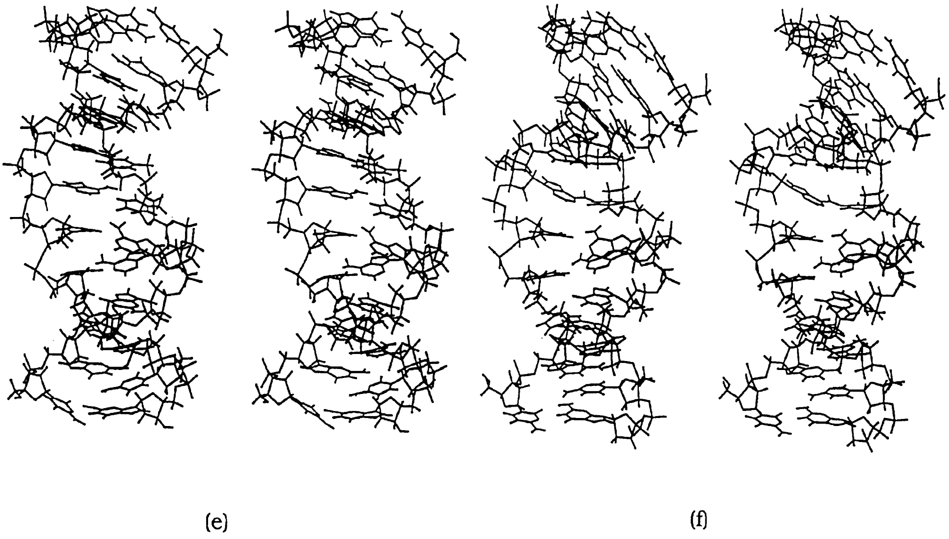


Fig. 2. (continued).

trimers in solution were performed using the program Galaxy 2.0 [40b] on IBM RS-6000/580 workstations. All calculations were made using the all-atom force field and, for the duplex dodecamer, harmonic constraints were added to hydrogen bonds involved in the Watson–Crick base pairing of the last two base pairs on one end. The force constant was 4 kcal/(mol Å^2) for the C1-G24 base pair and 1 kcal/(mol Å^2) R2 for the G2-C23 base pair. Before starting the data collection, each system was minimized by the conjugate gradient method and equilibrated for 10 ps. The equilibration step is separated into two parts. In the first 4 ps of equilibration, the solute atoms were restrained to the initial positions with a harmonic force of 1.0 kcal/mol, and then the restraints were removed and the system was equilibrated for another 6 ps. The equilibration step was started by assigning a random velocity to each atom so that the velocity distribution conformed to the Maxwellian distribution corresponding to 10 K. This is followed by heating the system to 300 K with a temperature coupling time of 0.1 ps.

Molecular dynamics simulation and free energy simulation were performed at constant temperature (300 K) and pressure (1 atm) with periodic boundaries for 100 ps and 80 ps, respectively, following an initial 10 ps of equilibration. In MD simulation, the structures were stored every 0.1 ps for data analysis. In free energy simulation, 101 windows were employed with 0.4 ps of equilibration followed by 0.4 ps of data collection at each window, unless noted. For each mutation, both forward ($A \rightarrow B$) and reverse ($B \rightarrow A$) perturbation calculations were carried out to estimate the convergence. In the case of trimers, the second run was carried out using the final coordinates from each mutation, namely, four free energy changes were calculated for one perturbation system, because single-stranded trimers are more flexible and have various conformations in solution. The SHAKE routine [41], in which all bond lengths are held constant, was used with a timestep of 0.001 ps. A constant dielectric of 1 was used for all simulations. A cutoff distance of 8 Å was used for solute–solvent and solvent–solvent nonbonded interactions. All solute–solute nonbonded interactions were included.

Computational results

Molecular dynamics: The MD simulation for an unmodified duplex dodecamer, d(CGCGAATTCGCG), was carried out for 100 ps in order to examine the dynamical behavior of the B-form DNA helix structure under the condition described above. A sequence of structure keeps the double helix intact over the course of the simulation as shown in Fig. 2, although a large bending of the duplex is observed at 100 ps. While we sometimes observed the distortion of the whole structure leading to an unstable helix without the constraint for the terminal base pairs, more stable DNA structures were obtained with the constraint. However, the calculated dynamical structures fluctuate considerably as compared to a canonical B-form. Therefore, we analyzed the dynamical structures, i.e. backbone and glycosidic torsion angles, sugar puckering and helical parameters, in detail.

Table 3 shows the conformational analysis of backbone torsional angles over a 100 ps MD simulation. ε, the C3′-O3′ angle, remains the initial trans conformation in most cases, but a $t \rightarrow g^-$ transition is observed in a significant number of cases. These transitions correlated well with those of $g^- \rightarrow t$ of ζ, the O3′-P angle. This correlation corresponds to earlier crystallographic studies, in which B_I ($\varepsilon = t$ and $\zeta = g^-$) and B_{II} ($\varepsilon = g^-$ and $\zeta = t$) forms have been identified as B-form DNA [42,43]. Another phosphodiester torsion angle, α(P-O5′), remains mainly g^- with a few g^+ or t conformations. The O5′-C5′ angles (β) are mostly in the t conformation with a few transitions of t to g^+. The C4′-C5′ angles (γ) remain in the g^+ conformations for most of the residues, with a few residues in the t conformation. The glycosidic torsion angles, χ, are all in anti-orientation in the range of 210–270° as shown in Table 4. In the middle region (AATT), relatively high anti-values are observed. These torsion angles described above are almost all in the range of B-form DNA conformation [44], except for some transitions, and the local conformational distributions are almost constant during 100 ps of molecular dynamics, that is, in the equilibrium state.

The average sugar puckerings are consistent with B-form DNA conformations [44,45] over 100 ps of molecular dynamics (Table 5). The predominant sugar conformation is C2′-endo and the conformation of O1′-endo is observed in a relatively high percentage among the other conformations. The C3′-endo conformation characteristic of an A-DNA is observed only in a few nucleotides. Although the sugar puckering is dynamically active, the proportion of conformations is almost constant during the molecular dynamics. No difference of the sugar conformation between pyrimidine and purine or between one half of the DNA structure and the other half was observed.

Table 3 *Conformational analysis of backbone torsion angles*[a]

Base	No.	I	II	III	IV	V	Base	No.	I	II	III	IV	V
ε. C4′-C3′-O3′-P													
Cyt	1	g^-	g^-	g^-	g^-	g^-	Cyt	23	t	t	t	t	t
Gua	2	t	t	t	t	t	Gua	22	t	g^-	g^-	g^-	g^-
Cyt	3	g^-	g^-	g^-	g^-	g^-	Cyt	21	t	t	t	t	t
Gua	4	t	t	t	t	t	Thy	20	t	t	t	t	t
Ade	5	t	t	t	t	t	Thy	19	t	t	t	t	t
Ade	6	t	t	t	t	t	Ade	18	t	t	t	t	t
Thy	7	t	t	g^-	t	t	Ade	17	t	t	t	g^-	g^-
Thy	8	t	t	t	t	g^-	Gua	16	t	t	t	t	t
Cyt	9	t	g^-	t	g^-	g^-	Cyt	15	t	t	t	t	t
Gua	10	t	t	t	t	t	Gua	14	t	t	t	t	t
Cyt	11	t	t	t	t	t	Cyt	13	g^-	g^-	g^-	g^-	g^-
ζ. C3′-O3′-P-O5′													
Cyt	1	t	t	t	t	t	Cyt	23	g^-	g^-	g^-	g^-	g^-
Gua	2	g^-	g^-	g^-	g^-	g^-	Gua	22	t	t	t	t	t
Cyt	3	t	t	t	t	t	Cyt	21	g^-	g^-	g^-	g^-	g^-
Gua	4	g^-	g^-	g^-	g^-	g^-	Thy	20	g^-	g^-	g^-	g^-	g^-
Ade	5	g^-	g^-	g^-	g^-	g^-	Thy	19	g^-	g^-	g^-	g^-	g^-
Ade	6	g^-	g^-	g^-	g^-	g^-	Ade	18	g^-	g^-	g^-	g^-	g^-
Thy	7	g^-	g^-	g^-	g^-	g^-	Ade	17	t	g^-	g^-	t	t
Thy	8	g^-	g^-	g^-	g^-	g^-	Gua	16	g^-	g^-	g^-	g^-	g^-
Cyt	9	g^-	t	t	t	t	Cyt	15	g^-	g^-	g^-	g^-	g^-
Gua	10	g^-	g^-	g^-	g^-	g^-	Gua	14	g^-	g^-	g^-	g^-	g^-
Cyt	11	g^-	g^-	g^-	g^-	g^-	Cyt	13	t	t	t	t	t
α. O3′-P-O5′-C5′													
Gua	2	g^-	g^-	g^-	g^-	t	Gua	24	g^-	g^-	g^-	g^-	g^-
Cyt	3	g^-	g^-	g^-	g^-	g^-	Cyt	23	g^-	g^-	g^-	g^-	g^-
Gua	4	g^+	g^+	g^+	g^+	g^+	Gua	22	g^-	g^-	g^-	g^-	g^-
Ade	5	g^-	g^-	g^-	g^-	t	Cyt	21	g^-	g^-	g^-	g^-	g^-
Ade	6	g^-	g^-	g^-	g^-	g^+	Thy	20	g^-	g^-	g^-	g^-	g^-
Thy	7	g^-	g^-	g^-	g^-	g^-	Thy	19	g^-	g^-	g^-	g^-	g^-
Thy	8	t	t	g^-	g^-	g^-	Ade	18	g^-	g^-	g^-	g^-	g^-
Cyt	9	g^-	g^-	g^-	g^-	g^-	Ade	17	g^-	g^-	g^-	g^-	g^-
Gua	10	g^-	g^-	g^-	g^-	g^-	Gua	16	g^-	g^-	g^-	g^-	g^-
Cyt	11	g^-	g^-	g^-	g^-	g^-	Cyt	15	g^-	g^-	g^-	g^-	g^-
Gua	12	g^-	t	g^+	g^+	t	Gua	14	g^+	g^+	g^+	g^+	g^+
β. P-O5′-C5′-C4′													
Gua	2	t	t	t	t	t	Gua	24	t	t	t	t	t
Cyt	3	t	t	t	t	t	Cyt	23	t	t	g^+	t	g^+
Gua	4	t	t	t	t	t	Gua	22	t	t	t	t	t
Ade	5	t	t	t	t	t	Cyt	21	t	t	t	t	t
Ade	6	t	t	t	t	t	Thy	20	t	t	t	t	t

Table 3 *(continued)*

Base	No.	I	II	III	IV	V	Base	No.	I	II	III	IV	V
β. P-O5′-C5′-C4′													
Thy	7	t	t	t	t	t	Thy	19	t	t	t	t	t
Thy	8	t	t	g^+	g^+	g^+	Ade	18	t	t	t	t	t
Cyt	9	t	t	t	t	t	Ade	17	t	t	t	t	t
Gua	10	t	t	t	t	t	Gua	16	t	t	t	t	t
Cyt	11	t	t	t	t	g^+	Cyt	15	t	t	t	t	t
Gua	12	t	t	t	t	t	Gua	14	t	t	t	t	t
γ. O5′-C5′-C4′-C3′													
Cyt	1	t	t	t	t	t	Gua	24	g^+	g^+	g^+	g^+	g^+
Gua	2	g^+	g^+	g^+	g^+	g^+	Cyt	23	g^+	g^+	g^+	g^+	g^+
Cyt	3	g^+	g^+	g^+	g^+	g^+	Gua	22	g^+	g^+	g^+	g^+	g^+
Gua	4	t	t	t	t	t	Cyt	21	g^+	g^+	g^+	g^+	g^+
Ade	5	g^+	g^+	g^+	g^+	t	Thy	20	g^+	g^+	g^+	g^+	g^+
Ade	6	g^+	g^+	g^+	g^+	g^+	Thy	19	g^+	g^+	g^+	g^+	g^+
Thy	7	g^+	g^+	g^+	g^+	g^+	Ade	18	g^+	g^+	g^+	g^+	g^+
Thy	8	t	t	t	t	t	Ade	17	g^+	g^+	g^+	g^+	g^+
Cyt	9	g^+	g^+	g^+	g^+	g^+	Gua	16	g^+	g^+	g^+	g^+	g^+
Gua	10	g^+	g^+	g^+	g^+	g^+	Cyt	15	g^+	g^+	g^+	g^+	g^+
Cyt	11	g^+	g^+	g^+	g^+	g^+	Gua	14	t	t	t	t	t
Gua	12	g^+	t	t	t	t	Cyt	13	g^+	g^+	g^+	g^+	g^+
δ. C5′-C4′-C3′-O3′													
Cyt	1	t	t	t	t	t	Gua	24	t	g^+	t	t	g^+
Gua	2	t	t	t	t	t	Cyt	23	g^+	g^+	g^+	g^+	g^+
Cyt	3	t	t	t	t	t	Gua	22	t	t	t	t	t
Gua	4	t	t	t	t	t	Cyt	21	g^+	g^+	g^+	g^+	g^+
Ade	5	g^+	g^+	g^+	g^+	g^+	Thy	20	g^+	g^+	g^+	g^+	g^+
Ade	6	g^+	g^+	g^+	g^+	g^+	Thy	19	g^+	g^+	t	t	t
Thy	7	g^+	g^+	t	g^+	g^+	Ade	18	t	g^+	g^+	g^+	g^+
Thy	8	t	t	g^+	t	t	Ade	17	t	g^+	g^+	t	t
Cyt	9	g^+	t	t	t	t	Gua	16	t	t	g^+	g^+	g^+
Gua	10	t	g^+	t	t	t	Cyt	15	g^+	g^+	g^+	g^+	t
Cyt	11	g^+	g^+	g^+	g^+	g^+	Gua	14	t	t	t	t	t
Gua	12	g^+	g^+	g^+	g^+	t	Cyt	13	t	t	t	t	t

[a] The MD simulation was divided into five time ranges as follows. I: 0–20 ps; II: 20–40 ps; III: 40–60 ps; IV: 60–80 ps; V: 80–100 ps. The average conformation was found in each time range. The ranges of the three regions g^+, t and g^- are respectively 0–120°, 120–240° and 240–360°.

The average intrastrand adjacent phosphorus atom distances every 20 ps are shown in Table 6. The distances fluctuate around the values observed in the crystal structure of the duplex [44], that is, 6.17–7.12 Å. It is noted that the distance of the two pairs P4-P5 and P16-P15 is significantly larger throughout the molecular dynamics than

Table 4 *Glycosidic torsion angles (°)*

Base	No.	Angle (rms)	Base	No.	Angle (rms)
χ. O1′-C1′-N1-C2 (pyrimidine), O1′-C1′-N9-C4 (purine)					
Cyt	1	256.4 (29.1)	Gua	24	239.7 (13.5)
Gua	2	248.7 (14.4)	Cyt	23	223.9 (14.3)
Cyt	3	264.7 (11.1)	Gua	22	251.6 (14.3)
Gua	4	229.0 (12.5)	Cyt	21	218.5 (15.9)
Ade	5	223.6 (19.5)	Thy	20	211.6 (13.4)
Ade	6	215.2 (18.0)	Thy	19	230.5 (14.6)
Thy	7	219.0 (19.4)	Ade	18	219.9 (21.5)
Thy	8	212.1 (16.8)	Ade	17	244.8 (29.9)
Cyt	9	268.2 (25.1)	Gua	16	232.8 (20.1)
Gua	10	233.4 (17.1)	Cyt	15	225.6 (16.6)
Cyt	11	217.8 (11.7)	Gua	14	236.9 (17.2)
Gua	12	223.8 (13.7)	Cyt	13	256.2 (31.2)

those of the other pairs. This reflects the transformation of backbone torsion angles from the values in the standard B-DNA, namely, the transformation of $(\varepsilon,\zeta,\alpha,\gamma)$ to (g^-,t,g^+,t) from (t,g^-,g^-,g^+). On the other hand, the two pairs P7-P8 and P19-P18 show values similar to that of a standard A-DNA, that is, 5.9 Å. This is consistent with the sugar conformation of C3′-endo in Thy7 and Ade18.

The averaged helical twist angles and the number of base pairs per unit turn are almost constant and in the range of values found in B-form DNA [44], in which the ideal helix twist angle is 36.0° and the number of base pairs is 10.0, over the dynamics trajectory. Table 7 contains the average value of the helical twist angles at each base-pair step over 100 ps of molecular dynamics. The helical twist angles of the central region of the duplex exhibit smaller values than the ideal value, indicating that

Table 5 *Predominant sugar conformation*

Base	No.	I	II	III	IV	V	Base	No.	I	II	III	IV	V
Cyt	1	2	2	2	2	2	Gua	24	3	3	2	2	2
Gua	2	2	2	2	2	2	Cyt	23	1	1	1	1	1
Cyt	3	2	2	2	2	2	Gua	22	2	2	2	2	2
Gua	4	2	2	2	2	2	Cyt	21	1	1	1	1	1
Ade	5	2	1	1	1	2	Thy	20	1	1	1	1	1
Ade	6	1	1	1	1	1	Thy	19	1	2	2	2	2
Thy	7	3	2	2	1	3	Ade	18	2	2	3	3	3
Thy	8	2	2	1	1	2	Ade	17	2	2	1	2	2
Cyt	9	3	2	2	2	2	Gua	16	2	2	1	1	1
Gua	10	2	2	2	2	2	Cyt	15	1	3	1	1	2
Cyt	11	3	1	1	1	1	Gua	14	2	2	2	2	2
Gua	12	3	3	3	3	2	Cyt	13	2	2	2	2	2

1: O1′-endo; 2: C2′-endo; 3: C3′-endo.

Table 6 *Intrastrand phosphorus atom distances (Å)*

Pair	I	II	III	IV	V	Pair	I	II	III	IV	V
P2-P3	7.11	7.18	7.13	7.03	7.11	P24-P23	6.81	6.86	6.97	6.92	6.91
P3-P4	6.72	6.49	6.71	6.65	6.84	P23-P22	6.70	6.46	6.41	6.65	6.52
P4-P5	7.61	7.64	7.64	7.65	7.60	P22-P21	6.94	6.68	6.77	6.59	6.72
P5-P6	6.63	6.62	6.81	6.69	7.17	P21-P20	6.80	6.54	6.43	6.33	6.32
P6-P7	6.64	6.66	6.23	6.38	6.49	P20-P19	6.87	6.91	7.17	7.12	7.04
P7-P8	6.64	6.72	6.56	6.07	5.98	P19-P18	7.03	6.61	6.20	6.13	5.80
P8-P9	7.61	7.20	6.65	6.19	6.07	P18-P17	6.86	6.71	6.54	6.61	6.60
P9-P10	6.54	6.64	6.85	6.89	6.79	P17-P16	7.08	6.91	6.55	6.63	6.12
P10-P11	7.20	6.90	6.96	7.07	7.04	P16-P15	6.57	6.56	6.91	6.83	7.05
P11-P12	5.80	6.57	6.66	6.63	6.93	P15-P14	7.61	7.64	7.45	7.63	7.65
Mean	6.89	6.83	6.78	6.73	6.74						
Rms	0.23	0.27	0.22	0.26	0.28						

the recognition site, AATT, tends to be unwound. Further, the result shows clearly that CpG steps have larger twist angles than GpC steps in all the cases, in contrast to Calladine's rules [46] and the crystal structure [47,48] of the dodecamer. The same conclusion is obtained from an earlier study [49], although the solvent water was not explicitly contained in the simulation.

The hydrogen bonds of the Watson–Crick paired bases remained intact throughout the simulation as shown in Table 8, and the average distances for base pairing are consistent with standard hydrogen-bond distances NH ··· O = 1.95 Å and NH ··· N = 1.99 Å reported for X-ray crystal structures [44]. In addition, the distances and

Table 7 *Average helix twist angles and number of base pairs per turn*

Base pair step	Helix twist angle (rms)					
	0–20 ps	20–40 ps	40–60 ps	60–80 ps	80–100 ps	0–100 ps
C1-G24/G2-C23	38.8 (3.9)	41.9 (3.4)	44.8 (3.3)	44.4 (3.6)	43.1 (3.3)	42.6 (4.1)
G2-C23/C3-G22	32.8 (2.4)	34.0 (2.4)	33.4 (2.4)	35.3 (2.2)	37.0 (2.6)	34.5 (2.8)
C3-G22/G4-C21	43.1 (2.8)	38.9 (3.2)	41.1 (3.2)	39.6 (3.5)	42.3 (3.4)	41.0 (2.8)
G4-C21/A5-T20	31.8 (4.5)	30.1 (3.0)	31.0 (3.0)	28.5 (2.6)	30.0 (3.6)	30.3 (3.6)
A5-T20/A6-T19	33.8 (4.2)	32.4 (3.9)	30.5 (3.9)	29.5 (2.8)	28.2 (2.9)	30.8 (4.1)
A6-T19/T7-A18	35.6 (3.0)	33.3 (4.5)	25.1 (2.9)	30.7 (4.4)	31.0 (5.2)	31.1 (5.4)
T7-A18/T8-A17	42.7 (2.6)	36.2 (3.7)	30.5 (3.8)	29.6 (4.8)	36.0 (4.5)	35.0 (6.2)
T8-A17/C9-G16	35.5 (3.7)	30.3 (3.5)	33.4 (4.1)	28.5 (4.3)	22.0 (6.0)	29.9 (6.4)
C9-G16/G10-C15	36.1 (3.9)	42.7 (4.0)	42.7 (3.8)	44.6 (3.6)	44.2 (3.1)	42.1 (4.8)
G10-C15/C11-G14	36.1 (2.9)	32.7 (3.6)	29.7 (2.7)	30.6 (2.8)	35.0 (2.9)	32.8 (3.9)
C11-G14/G12-C13	34.0 (2.7)	42.7 (3.2)	37.1 (3.8)	42.2 (3.7)	43.7 (3.9)	39.9 (5.1)
Mean	36.3 (3.4)	35.9 (3.5)	34.5 (3.4)	34.9 (3.6)	35.7 (3.9)	35.5 (4.7)
No. of base pairs per turn	9.9	10.0	10.4	10.3	10.1	10.2

Table 8 *Average base-pair hydrogen-bond distances*

Hydrogen bond	Length(Å) (rms)
Gua 06[a]-Cyt HN4A	2.01 (0.19)
Gua H1[a]-Cyt N3	1.92 (0.09)
Gua HN2A[a]-Cyt O2	1.94 (0.12)
Gua O6[b]-Cyt HN4A	2.01 (0.17)
Gua H1[b]-Cyt N3	1.93 (0.09)
Gua HN2A[b]-Cyt O2	1.93 (0.12)
Ade HN6A-Thy O4	2.09 (0.27)
Ade N1-Thy H3	1.95 (0.17)

[a] Gua-2,4,22,24.
[b] Gua-10,12,14,16.

fluctuations of four GC pairings with hydrogen-bond constraints are almost identical to those of another four GC pairings of the opposite side. The result indicates that these constraints do not significantly influence the dynamical behavior of base pairings.

All the duplexes tested remain as B-DNA structure and base pairings are conserved during the course of the dynamics simulation. Therefore, we calculated the free energy differences between dodecamer duplexes containing 2-substituted-adenine and an intact duplex.

Free energy differences: The calculated free energy changes of dodecamer duplexes, single-stranded trimers and 9-Me-adenine bases are summarized in Tables 9–11, respectively, and the variations of ΔG with λ are given in Figs. 3–6. These ΔG's

Table 9 *Free energy changes in dodecamer duplexes*

System	ΔG (kcal/mol)		Average ± SE
	(+)	(−)	
Ade → NH$_2$-Ade	− 3.70	3.74	− 4.15 ± 0.43
NH$_2$-Ade → Ade	4.59	− 4.55	
NH$_2$-Ade6, NH$_2$-Ade6, Ade18 → NH$_2$-Ade18	− 4.46	4.44	− 3.95 ± 0.50
NH$_2$-Ade6, NH$_2$-Ade6, NH$_2$-Ade18 → Ade18	3.50	− 3.39	
Ade → F-Ade	0.18	− 0.18	0.26 ± 0.08
F-Ade → Ade	− 0.31	0.36	
Ade → Cl-Ade	0.37	− 0.32	0.25 ± 0.10
Cl-Ade → Ade	− 0.18	0.12	
Ade → OH-Ade	− 2.78	2.82	− 2.86 ± 0.06
OH-Ade → Ade	2.96	− 2.88	

Table 10 *Free energy changes in trinucleotides*

System	ΔG (kcal/mol)			
	(+)	(−)	Average	Weighted average
Ade → NH_2-Ade	−5.14	5.22	−4.26 ± 0.46	−3.76 ± 0.74
	−3.29	3.43		
	−3.25	3.25		
NH_2-Ade → Ade	5.69	−5.70		
	3.28	−3.12		
	4.94	−4.84		
Ade → F-Ade	0.34	−0.32	0.16 ± 0.09	0.08 ± 0.11
	0.01	0.00		
F-Ade → Ade	0.02	0.01		
	−0.30	0.29		
Ade → Cl-Ade	0.74	−0.66	0.51 ± 0.18	0.42 ± 0.13
	−0.33	0.37		
NH_2-Ade → Ade	−0.62	0.71		
	0.30	−0.32		
Ade → OH-Ade	−4.72	4.77	−4.84 ± 0.14	−4.94 ± 0.17
	−5.10	5.22		
NH_2-Ade → Ade	4.54	−4.45		
	5.00	−4.91		

represent only the free energy contribution due to the interaction of the perturbed group with the rest of the system. The perturbed group consists of a 2-substituted-adenine base and the charge distribution and force parameters for the sugar portion are identical to that of the unperturbed group. Therefore, the only interaction of 2-substituted-adenine with the rest of the system without the sugar portion is monitored, and ΔG comes only from the electrostatic and van der Waals interaction. ΔG is not decomposed into the electrostatic contribution and the van der Waals contribution in the thermodynamic perturbation method. However, when the increment of the coupling parameter λ at each window is small enough, that is, $\Delta H(\lambda) \ll kT$, Eq. 2 can be reduced to

$$\Delta G(\lambda) = < \Delta H(\lambda) >_0 = < \Delta H_{ele}(\lambda) >_0 + < \Delta H_{vdw}(\lambda) >_0 \qquad (4)$$

In fact, ΔG(λ) obtained from Eq. 4 is almost the same as that from Eq. 2 at each window. Thus, ΔG(λ) is approximately decomposed into the electrostatic contribution and the van der Waals contribution using Eq. 4. The pattern of contribution of electrostatic and van der Waals interaction to ΔG is very similar in the forward and reverse transformations in all cases. Therefore, only the pattern of reverse transformation is shown in Figs. 3–6.

Table 11 *Free energy changes in 9-Me-adenine bases*

System	ΔG (kcal/mol)		Average ± SE
	(+)	(−)	
Ade → NH_2-Ade	− 1.63	1.65	− 1.65 ± 0.01
NH_2-Ade → Ade	1.68	− 1.63	
Ade → F-Ade	− 0.90	0.92	− 1.09 ± 0.18
F-Ade → Ade	1.22	− 1.30	
Ade → Cl-Ade	− 1.98	1.92	− 2.10 ± 0.15
Cl-Ade → Ade	2.23	− 2.27	
Ade → OH-Ade	− 3.00	3.04	− 2.91 ± 0.11
OH-Ade → Ade	2.81	− 2.80	

Simulation 1: The free energy change of the transformation of hydrogen to the NH_2 group in the 2-position of Ade6 is −4.15 kcal/mol for the duplex system, −3.76 kcal/mol for the trimer system (please refer below) and −1.65 kcal/mol for the isolated base system. The difference in the free energy change, ΔΔG, between the duplex and the trimer is −0.39 kcal/mol. This result shows that the substitution of the 2-NH_2-adenine residue for the adenine residue makes the duplex structure only slightly stable as compared to the single-stranded structure. ΔΔG between the trimer and the isolated base is −2.11 kcal/mol, indicating that 2-NH_2-adenine in the trimer strongly interacts with the neighboring bases as compared to the unmodified adenine. For the duplex, ΔG decreases linearly with λ in both the forward and reverse simulations, although there is some discrepancy between the final values of the two cases. ΔG for the isolated base decreases continuously with an initial shoulder with respect to λ and the forward and reverse simulations show almost the same variation.

In the case of the trimer system, ΔG decreases continuously with an initial shoulder as in the case of the isolated base, but the magnitude is three times that of the isolated base. This trend is observed in six simulations. The simulations are classified into two groups based on the ΔG values and the nature of the dynamical structures. In one group, ΔG is about 3 kcal/mol and the base–base interaction, including the stacking, is maintained throughout the simulation. In the other group, ΔG is about 6 kcal/mol and the stacking between the bases is disturbed and the transformed adenine residue interacts with the other residues and water molecules in a more flexible manner. The variation of ΔG_{ele} and ΔG_{vdW} for the duplex with respect to λ is shown in Fig. 3e. ΔG_{vdW} contributes negatively to ΔG mainly at lower values of λ, whereas ΔG_{ele} contributes constantly over the entire value of λ. For the isolated base transformation, ΔG_{vdW} decreases at lower values of λ, but this effect is compensated by ΔG_{ele} (Fig. 3g). ΔG_{ele} increases as a function of λ, resulting in the large solvation free energy. For the trimer system, the pattern of nonbonding contribution is different in each simulation (data not shown). In three simulations in which ΔG is about 5.0 kcal/mol, ΔG_{ele} is larger than that of the duplex and ΔG_{vdW} is much smaller. In another set of three

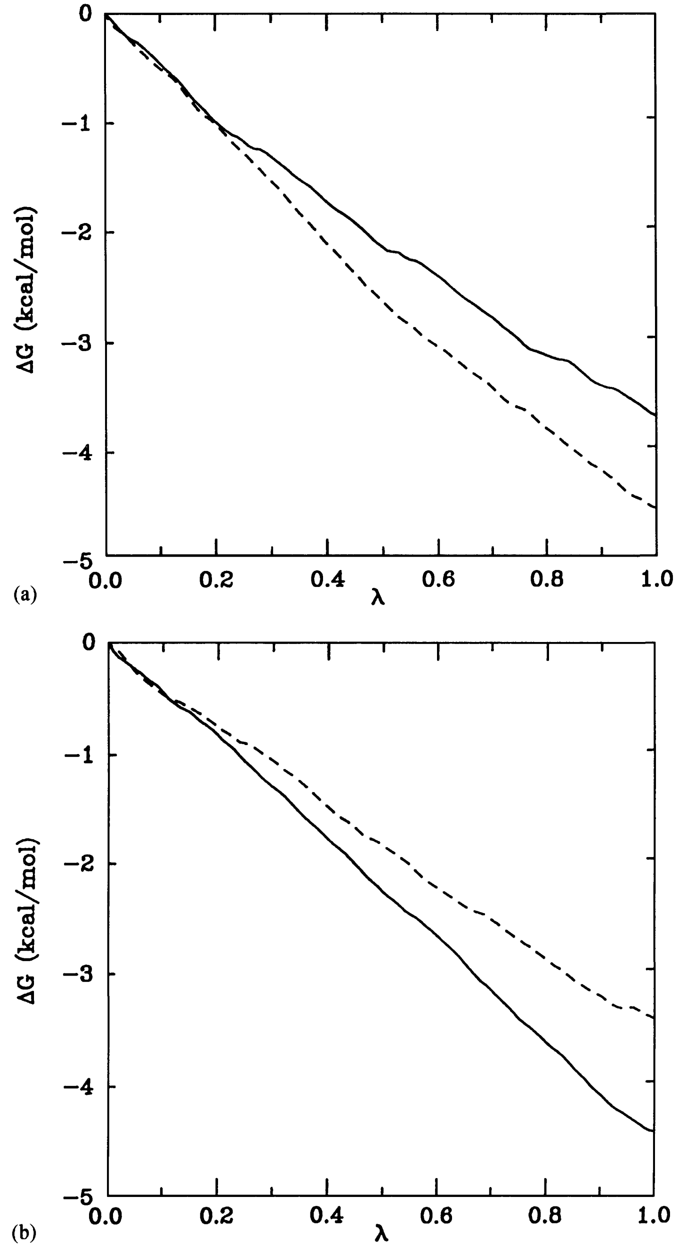
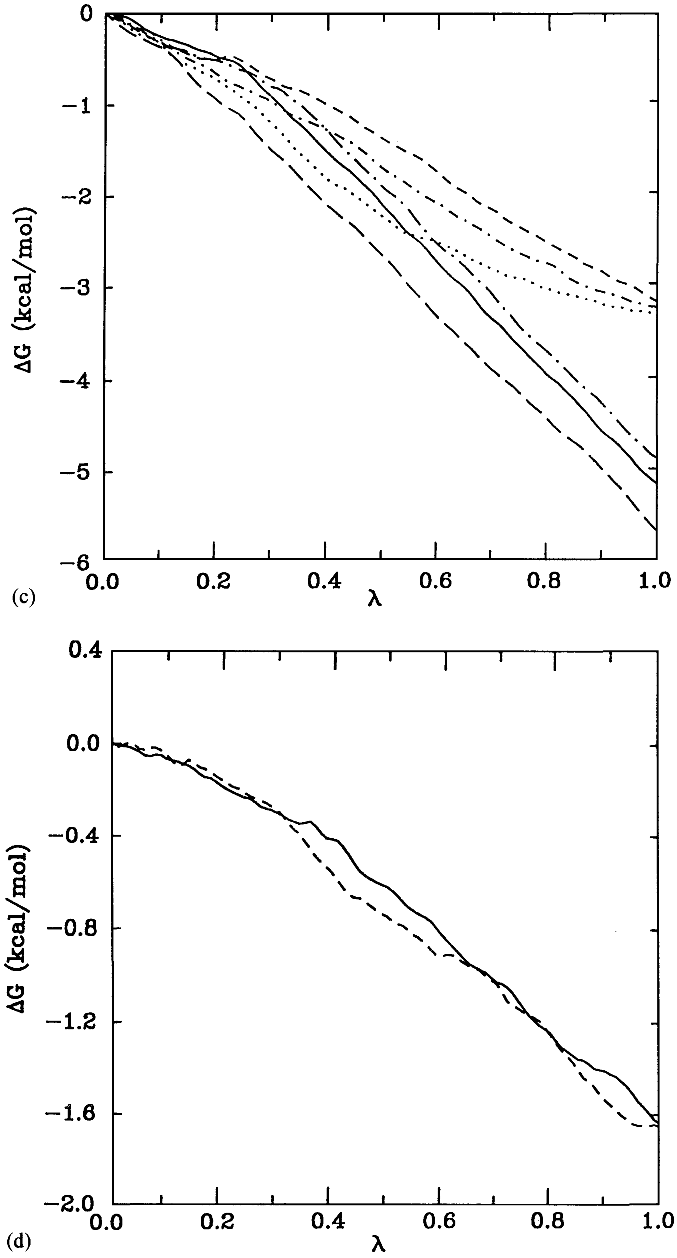


Fig. 3. Transformation between adenine ($\lambda = 0$) and 2-NH_2-adenine ($\lambda = 1$). (a)–(d): variation of ΔG with λ for the transformation. (a) Dodecamer duplex, (—) H → NH_2 and (- - -) NH_2 → H; (b) the second transformation in Ade 18 following (a), (—) H → NH_2 and (- - -) NH_2 → H; (c) single-stranded trimer, (—, ··· and - · -) NH_2 → H; (d) isolated base, (—) H → NH_2



and (- - -) $NH_2 \rightarrow H$. (e)–(g): contribution of (—) electrostatic and (- - -) van der Waals interaction to ΔG of the $NH_2 \rightarrow H$ transformation. (e) Dodecamer duplex; (f) the second transformation of the dodecamer duplex; (g) isolated base.

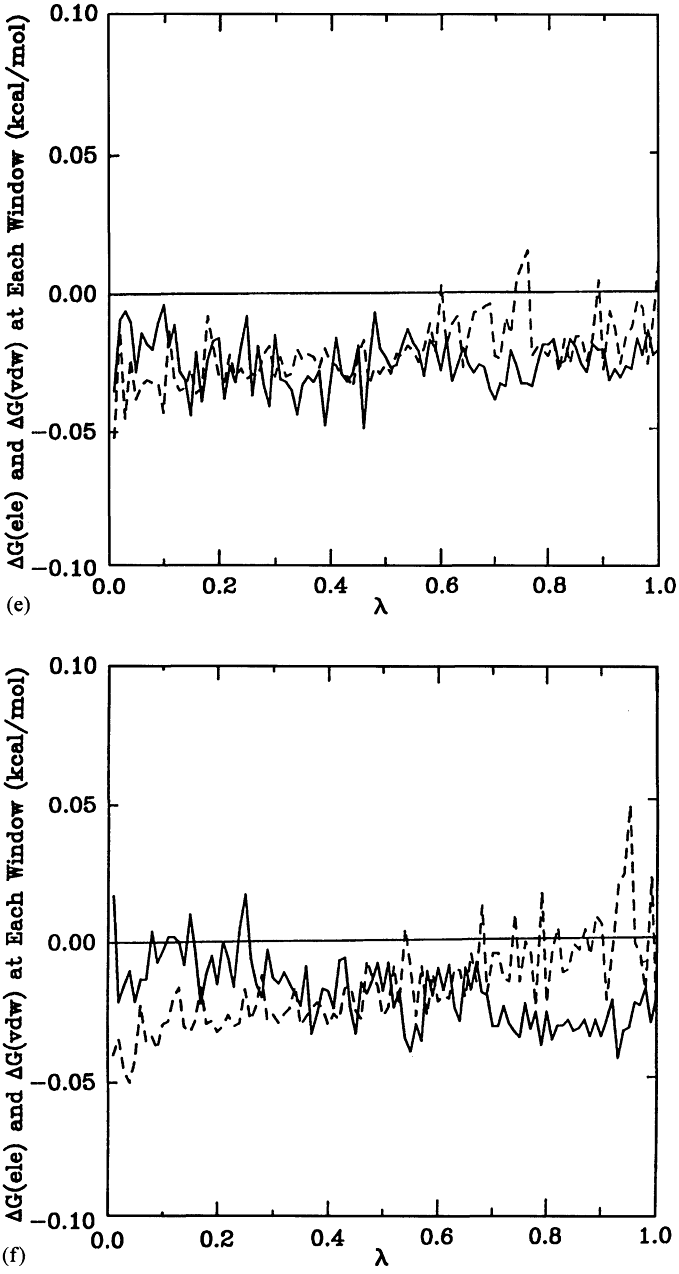


Fig. 3. (continued).

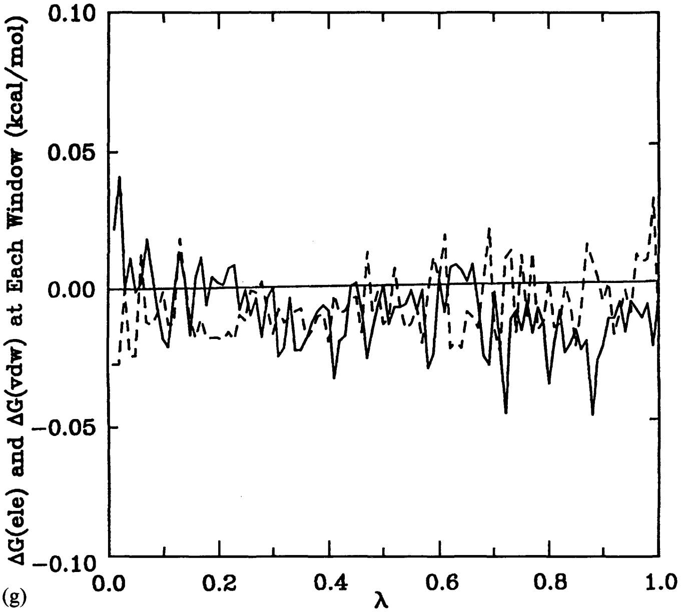


Fig. 3. (continued).

simulations, a significant contribution of ΔG_{vdW} is observed, while ΔG_{ele} is slightly larger.

It is well known that single-stranded oligonucleotides can exist in several conformational states and the conformation of single-stranded nucleic acids has been studied in many model systems [50–54]. Olsthoorn et al. [55] systematically studied the conformations of a series of oligodeoxy-adenosine nucleotides, $(dA)_n$, n = 2,3,6,9,12, in solution by means of temperature-dependent circular dichroism. Based on the thermodynamic results presented in the paper, it was estimated that a single-stranded trimer will predominantly exist in the stacked form and the percentage of the stacked form is determined to be 75.7% of the total population. The rest of the population is shown to exist in the unstacked form. These population results were used to average the free energy differences calculated by several runs for the trimer. For simulations where the stacking of the bases is maintained, a weight of 0.757 was used to compute the average free energy differences. For cases where the stacking was disrupted significantly, a weight of 0.243 was used for averaging the free energy differences.

Simulation 2: This describes the transformation of Ade18 in the other chain to a 2-NH$_2$-adenine residue. The resultant duplex is a structure in which two 2-NH$_2$-adenines are substituted for two inner adenines of the unmodified duplex. ΔG for the transformation is −3.95 kcal/mol, which is almost the same as that of simulation 1. The variation of ΔΔG and the contribution of nonbonding interaction is very similar to that of simulation 1 (Figs. 3b and f).

Simulation 3: Simulation 3 describes the transformation of hydrogen at the 2-position to fluorine. ΔG increases for the duplex and the trimer system, whereas it decreases for the isolated base system. The ΔG values for these simulations are 0.26, 0.08 and −1.09 kcal/mol, respectively. ΔΔG between the duplex and the trimer is 0.18 kcal/mol, suggesting that the introduction of fluorine has little influence on the stability of the duplex. On the other hand, ΔΔG between the trimer and the isolated base is 1.17 kcal/mol. The variations of the ΔG's with λ for the above systems are shown in Figs. 4a–c, respectively. In the case of the duplex, ΔG increases when the value of λ is above 0.5, whereas for the isolated base, it decreases almost linearly to a value of about −1.0 kcal/mol. In the trimer, ΔG tends to increase linearly in three of four simulations, although in one case it decreases rapidly in the range of λ between 0.8 and 1.0. In general, it is observed that the more the perturbed adenine is exposed to solvent water, the smaller the ΔG. ΔG_{ele} and ΔG_{vdW} to the free energy change at each window in the duplex and the isolated base are shown in Figs. 4d and e, respectively. In the case of the duplex, the ΔG_{ele} contribution is positive throughout the transformation, whereas ΔG_{vdW} is negative in the range of λ(0–0.5) and then fluctuates around a value of 0.0 in the range of λ(0.5–1.0). In contrast, ΔG_{ele} for the isolated base

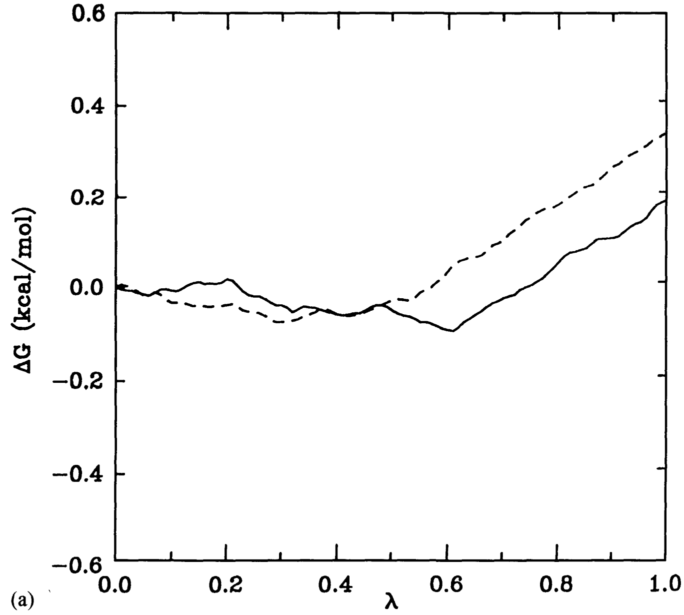


Fig. 4. Transformation between adenine (λ = 0) and 2-F-adenine (λ = 1). (a)–(c): variation of ΔG with λ for the transformation. (a) Dodecamer duplex, (—) H → F and (- - -) F → H; (b) single-stranded trimer, (— and ···) H → F and (——— and - - -) F → H; (c) isolated base, (—) H → F and (- - -) F → H. (d), (e): contribution of (—) electrostatic and (- - -) van der Waals interaction to ΔG of the F → H transformation. (d) Dodecamer duplex; (e) isolated base.

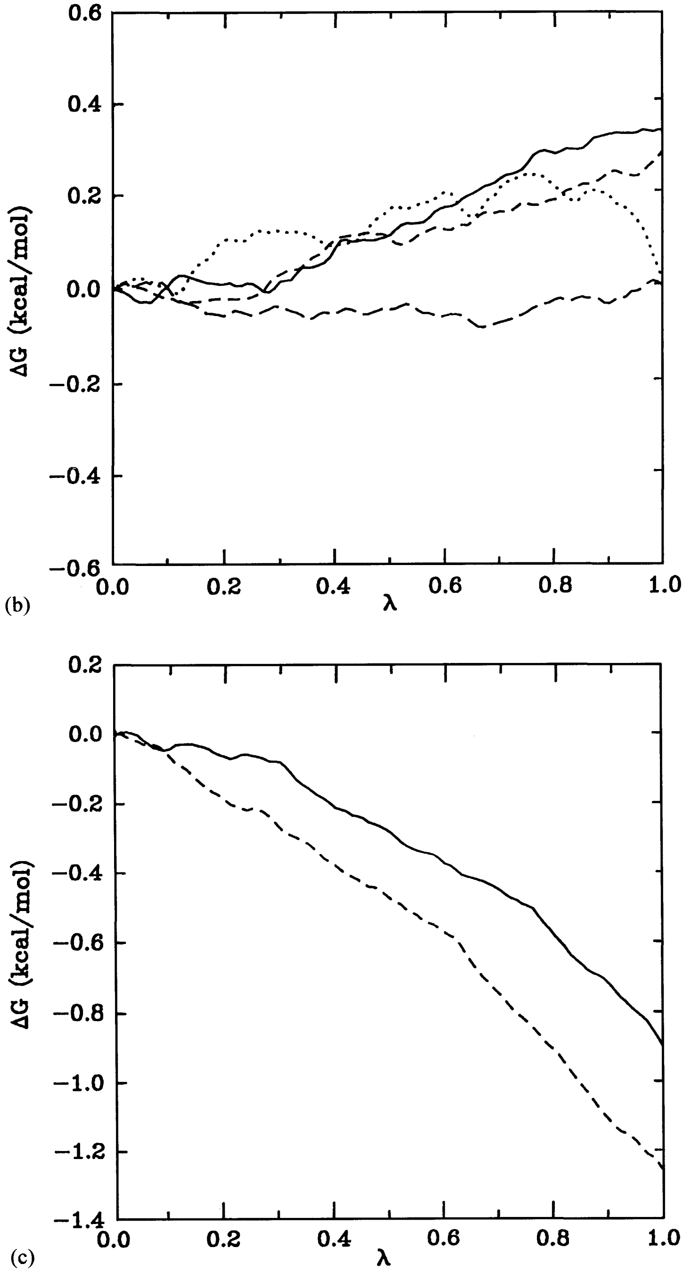


Fig. 4. (continued).

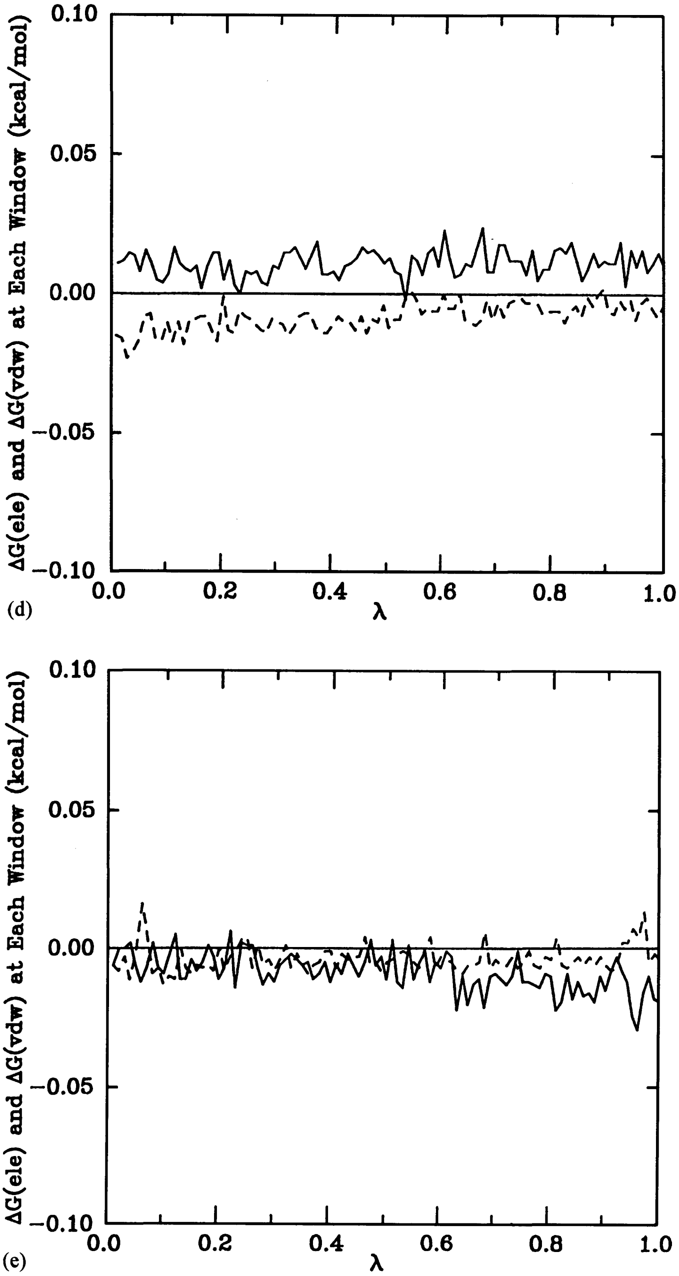
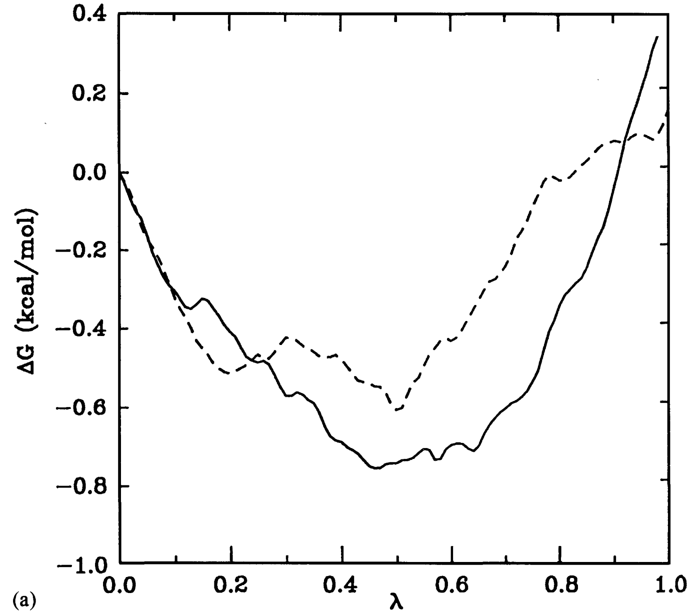


Fig. 4. (continued).

system is negative throughout the transformation, and ΔG is comparatively smaller. In the trimer, ΔG_{ele} is positive and ΔG_{vdW} is negative, although both are small in magnitude (data not shown).

Simulation 4: This describes the transformation of 2-hydrogen into chlorine. ΔG for this substitution is positive in the duplex and the trimer system, and negative for the isolated base system. The ΔG values for these simulations are 0.25, 0.42 and −2.10 kcal/mol, respectively. ΔΔG between the duplex and the trimer is −0.17 kcal/mol. This suggests that the stability of the duplex does not change by the introduction of chlorine. ΔG between the trimer and the isolated base is even larger (2.52 kcal/mol) than that of fluorine. The variations of the ΔG's for these systems with λ are shown in Figs. 5a–c, respectively. In the case of the duplex, ΔG initially decreases to −0.6 to −0.8 kcal/mol and then increases to 0.2–0.3 kcal/mol, although the curve is somewhat different between the forward and reverse simulations. On the other hand, for the isolated base case it decreases almost linearly to a value of about −2.0 kcal/mol in both the forward and reverse simulations. In the trimer, the simulations are divided into two groups. In one group, ΔG increases linearly with the λ value, and in the other it decreases. An examination of the structures reveals that in group one, the stacking between bases is preserved throughout the simulation, whereas in group two the stacking tends to be disturbed considerably. The ΔG_{ele} and ΔG_{vdW} contributions to the free energy change at each window are shown in Figs. 4b and 5b, respectively. For the duplex, ΔG_{vdW} contributes largely to the initial decrease in ΔG, whereas at higher values of λ both ΔG_{ele} and ΔG_{vdW} contribute positively to the total ΔG (Fig. 5d). In the isolated base, ΔG_{ele} contributes largely to the decrease in ΔG throughout the transformation, whereas ΔG_{vdW} is smaller except for λ values below 0.1 (Fig. 5e). In the trimer case, ΔG_{ele} shows positive values and ΔG_{vdW} shows negative values, and the pattern of the group in which ΔG is negative closely resembles the case of isolated base (data not shown).

Simulation 5: The free energy change of the transformation of 2-H into a 2-OH group is −2.86 kcal/mol for the duplex system, −4.94 kcal/mol for the trimer system and −2.91 kcal/mol for the isolated base system. ΔΔG between the duplex and the trimer is 2.08 kcal/mol, which indicates that the introduction of the OH group makes the duplex unstable compared to the single-stranded state. ΔΔG between the trimer and the isolated base is −2.03 kcal/mol. This indicates that the OH group introduced in single-stranded DNA has favorable interactions as compared to the unmodified DNA. Even in the case of trimer, the free energy change is almost the same between the simulations, although the simulations are classified into two groups from the aspect of dynamical structure as well as the other transformations. Figure 6d shows that the initial decrease in ΔG for the duplex is due to the van der Waals interaction, although it is partially compensated for by the electrostatic interaction. On the other hand, the electrostatic interaction mainly contributes to the decrease in ΔG at higher values of I. Although the electrostatic interaction is the main contributor as a whole, the van der Waals interaction also plays a significant role in the stability of OH-adenine. For the base simulation, the electrostatic contribution to the decrease in ΔG is much larger than the van der Waals one and the contribution rate increases



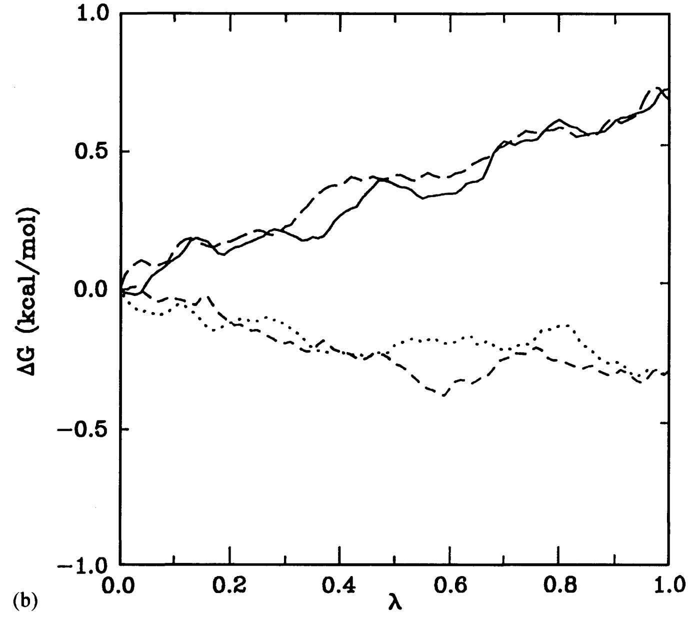
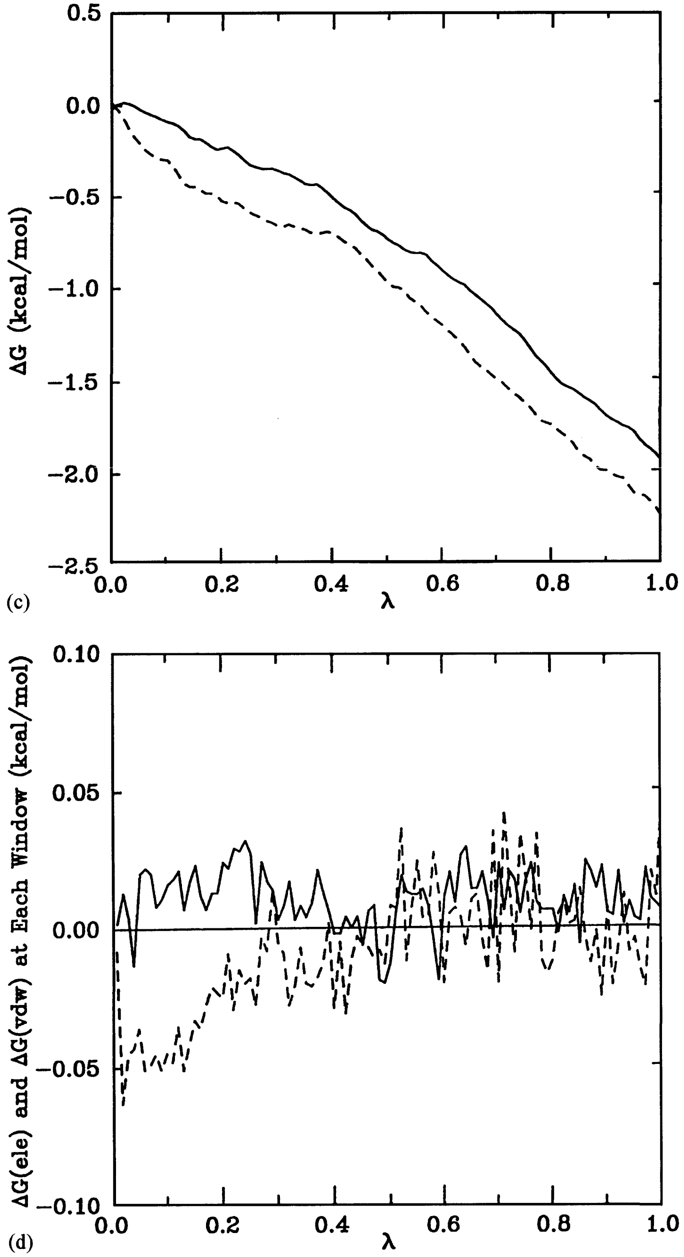


Fig. 5. Transformation between adenine ($\lambda = 0$) and 2-Cl-adenine ($\lambda = 1$). (a)–(c): variation of ΔG with λ for the transformation. (a) Dodecamer duplex, (—) H → Cl and (- - -) Cl → H; (b) single-stranded trimer, (— and ···) H → Cl and (— — — and - - -) Cl → H; (c) isolated base,



(—) H → Cl and (- - -) Cl → H. (d), (e): contribution of (—) electrostatic and (- - -) van der Waals interaction to ΔG of the Cl → H transformation. (d) Dodecamer duplex; (e) isolated base.

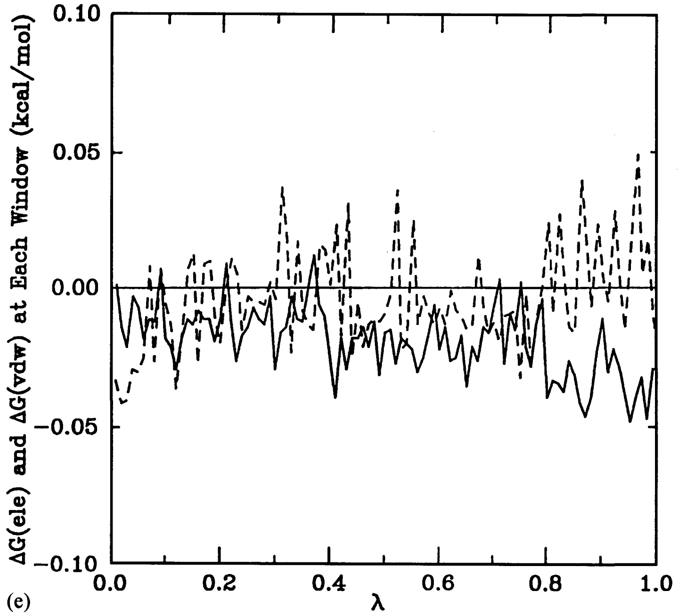


Fig. 5. (continued).

proportionally with an increase in λ (Fig. 6e). On the other hand, the contribution rate of the electrostatic interaction for the duplex is almost constant from 0.5 to 1.0 of λ. In the trimer simulation, most of the decrease in ΔG is due to the electrostatic interaction (data not shown).

Discussion

Molecular dynamics studies: The MD simulations of the DNA helix have been carried out with or without the explicit inclusion of solvent water by several groups [15–25]. Rao and Kollman [21] and Srinivasan et al. [23] reported the MD simulation on the same dodecamer duplex as in this study, that is, d(CGCGAATTCGCG), using the same AMBER force field; however, both their simulations were carried out in an *in vacuo* environment. Rao and Kollman found that the molecular dynamical structure during 84 ps of simulation stayed near a canonical B-form structure with reasonable H-bond and helical parameters. In addition, a significant bend in the DNA helix and opposite helix twist angles to that suggested by Calladine's rules were observed. Srinivasan et al. indicated that a significant deviation in the direction of the canonical A-form occurred in the base-pair orientations, although the structures generated during 100 ps of simulation remained generally in the B-family. These studies suggest that the *in vacuo* MD calculations of DNA simulate the structure

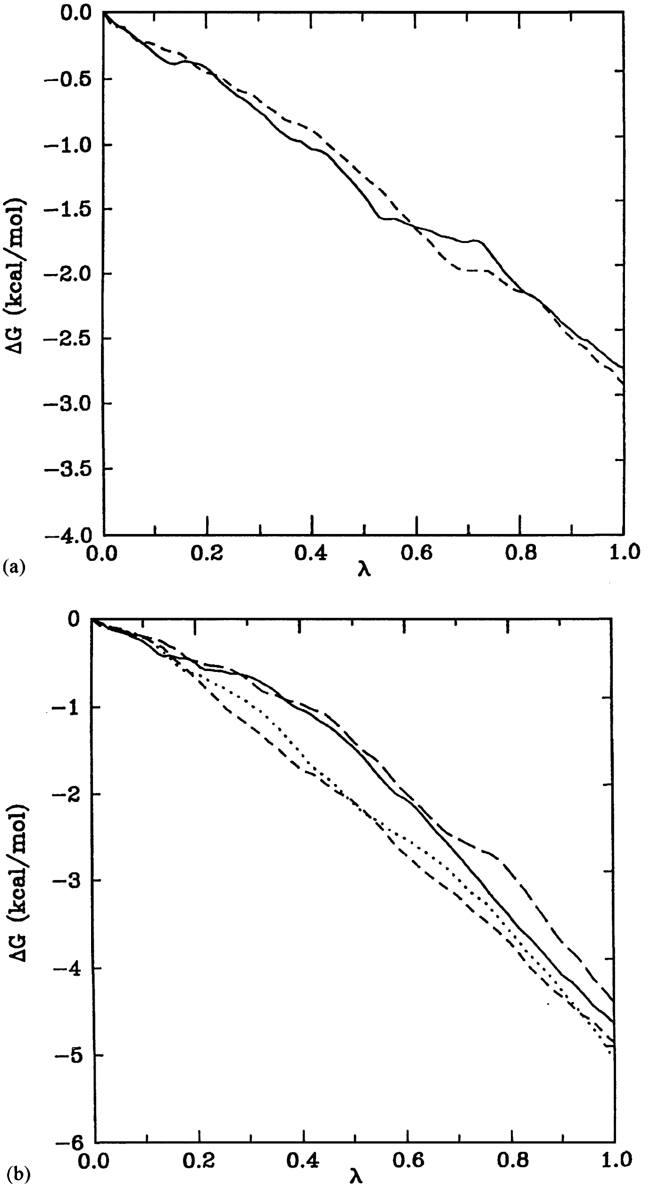


Fig. 6. Transformation between adenine ($\lambda = 0$) and 2-OH-adenine ($\lambda = 1$). (a)–(c): variation of ΔG with λ for the transformation. (a) Dodecamer duplex, (—) H → OH and (- - -) OH → H; (b) single-stranded trimer, (— and ···) H → OH and (——— and - - -) OH → H; (c) isolated base, (—) H → OH and (- - -) OH → H. (d), (e): contribution of (—) electrostatic and (- - -) van der Waals interaction to ΔG of the OH → H transformation. (d) Dodecamer duplex; (e) isolated base.

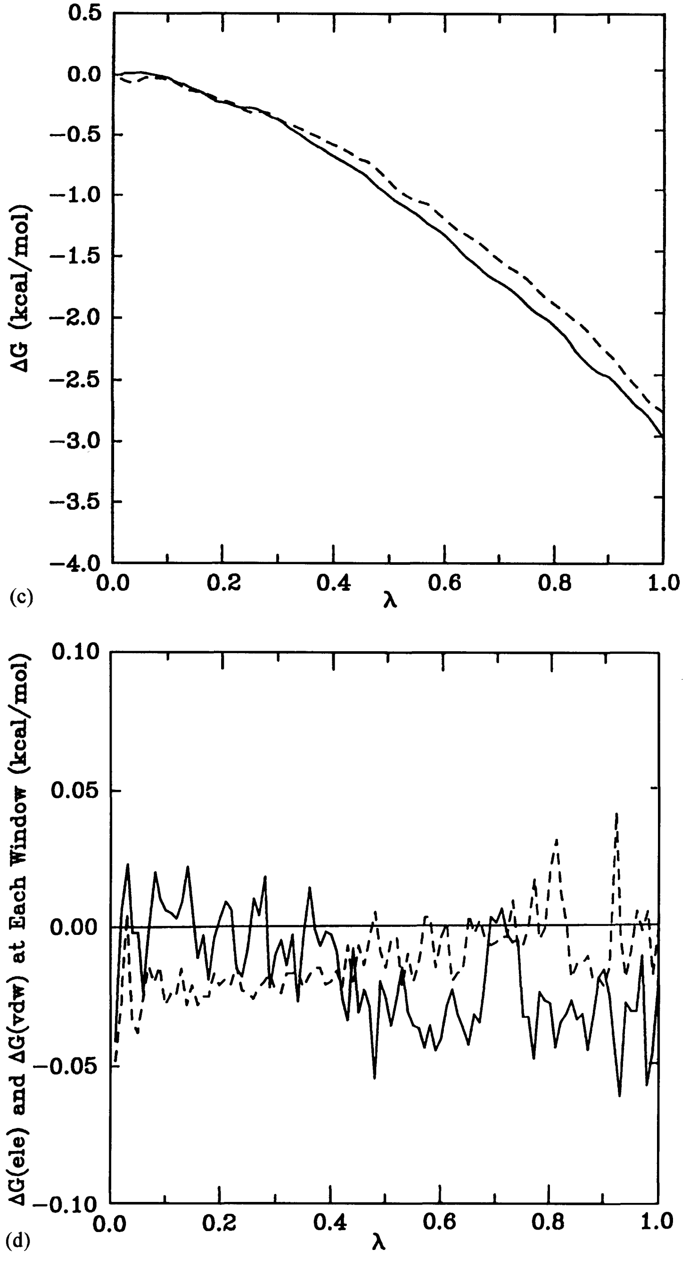


Fig. 6. (continued).

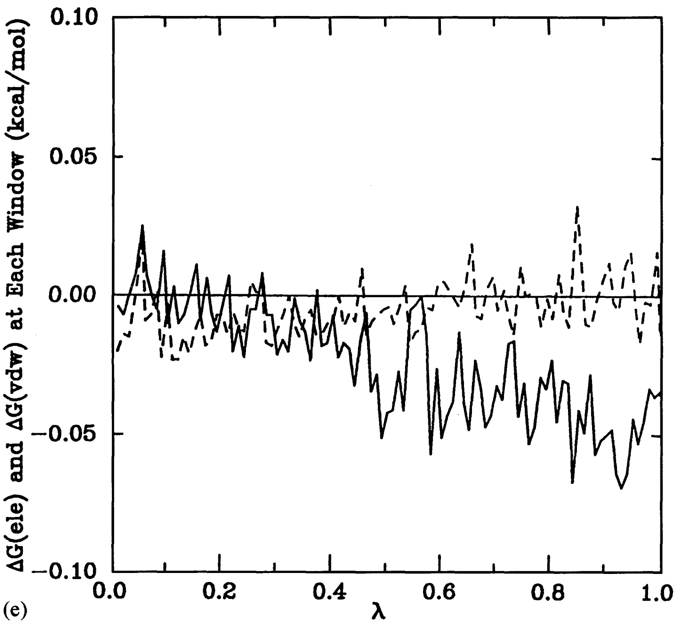


Fig. 6. (continued).

variation around the B-form structure, but a significant deviation from the B-form appears in some properties and the structure alterations are not necessarily in accord with those observed in the crystal structure.

In this study, the calculated dynamical structure maintains a stable double helix of DNA over 100 ps of MD simulation. We performed some procedures in addition to the usual one in order to obtain a stable dynamical trajectory. One was the positional restraint for the dodecamer duplex at the first half of the equilibration stage. Since the interaction between the duplex and the water added around the duplex is not necessarily adjusted well by the minimization procedure, the dynamical equilibration of the water structure is necessary before the start of molecular dynamics of the whole system. Another was the addition of a harmonic constraint for hydrogen bonds involved in Watson–Crick base pairing. This constraint is applied only to the last two base pairs on one end, not all the base pairs, so that the dynamical behavior of the perturbed residue is not influenced by the constraint. It is observed that the MD simulation with the constraint results in a more stable structure trajectory as compared to that with no constraint in which the DNA structure is sometimes distorted from a B-form structure. In addition, it can be observed that the local dynamical behavior of the two constrained base pairs is not influenced by the constraint as compared to that of two other base pairs on the opposite side. The constraint seems to prevent the process initiating the distorted state from the normal double-helix state of the duplex, although the reason is not clear.

Conformational parameters such as backbone and glycosidic torsional angles are essentially stable and the distribution ratio among the conformations is almost constant over 100 ps of molecular dynamics, although the transition of conformation is observed in some cases. These parameter values clearly indicate that the main conformations are those of a standard B-form. However, nonstandard torsion angles are observed with the transition of conformation, the so-called B_{II} conformation [42,43] in which e and z are g^- and t, respectively, in the native dodecamer crystal structure [44], although the positions of the B_{II} conformation in the simulation do not necessarily correspond to those in the crystal structure. Since the B_{II} conformation is often observed in experiment [42,43], the dynamical transition of B_I to B_{II} is reasonable. Another correlated dynamical transition is observed in the (α,γ) pair, that is, (g^-,g^+) to (g^+,t). The conformation of (g^+,t) is not found in the native dodecamer crystal structure and is different not only from a standard B-form DNA but also from a standard A-form DNA. However, this conformational transition is consistent with the general idea about the correlated motion of the backbone suggested by Olson [56], which indicates that the correlated motion keeps the base pairing and double-helix structure of DNA stable. Further, NMR experiments [57] suggest that the DNA backbone possesses substantially greater motional freedom than the base-pair moiety, and even in single-crystal X-ray structures [58] the backbone linker is susceptible to perturbation by crystal packing forces. Thus, the transition of (α,γ) is not unreasonable.

The sugar puckers, the helical twist angles and the phosphorus atom distances, which are parameters critical in differentiating the canonical B-form from the A-form, indicate that the dodecamer definitely remains in the B-form range over the entire MD trajectory. Since the mean and rms values of these parameters do not change much as the molecular dynamics proceeds, the dodecamer is conceived as being in the equilibrated state of the B-form range. We also observed that the hydrogen bonds of the Watson–Crick paired bases remained intact, with small values of rms throughout the simulation. However, it is observed that the helix structure is bent during the course of the run. In some of the simulations tested, we observed that the helices bend and then tend to straighten out. Therefore, this bending is not necessarily an irreversible process. This result, however, indicates that there are structure changes which proceed in a timescale of over 100 ps. Although the 100 ps MD simulation samples only a part of all configuration space, it can sample well the local configuration space such as base pairings. Further, it is important to note that the various conformational parameters of the first half of the duplex structure show the same trend as the last half, indicating that the constraints for the hydrogen bonds of the two terminal base pairs do not significantly influence the molecular dynamical behavior of the duplex.

The simulation including solvent molecules explicitly shows both similar aspects and a different one as compared to those without solvent molecules. We found that the pattern of the helical twist angles showed an opposite tendency to Calladine's rules as found by Rao and Kollman in the *in vacuo* simulation. A similar trend was also observed in the MD simulations for the same duplex using the GROMOS force field [25]. Therefore, it does not seem that this phenomenon is due to the type of force field used for the simulation. The large bending as in the *in vacuo* simulation is not observed in the simulation that includes solvent water molecules until past 80 ps, although there appear several kinks in the simulation. An examination of the structures at 20 ps intervals in Fig. 1 reveals the presence of kinks at C3pG4 and C13pG14 throughout the simulation. These kinks are associated with the extension of the distance between the adjacent phosphorus atoms as shown by P4-P5 and P14-P15, respectively, in Table 6. Further, the extension is related to the unusual conformation of the backbone associated with nucleotides on either side of P4 and P14, that is, (e,z,a,g) is (g^-,t,g^+,t). These relations suggest that the kinks are formed by the transformation of backbone torsion angles, since the sugar conformation remains as C2′-endo. On the other hand, it is observed that several kinks are formed temporally during the simulation. The temporal kinks at C9pG10 and A17pA18 appear to be associated with the correlated transition in e and z, namely, the transition to the B_{II} conformation from the B_I conformation which is a standard B-DNA conformation. In contrast, the kink formed at C15pG16 may be related to the sugar conformation of C3′-endo, which is consistent with the result of model building that the DNA helix bends at the junction joining A- and B-DNA [44]. These results indicate that small kinks are easily formed by the transition of backbone torsion angles and sugar conformations. When the correlated transition of (e,z) as well as that of (a,g) is brought out, the kink becomes long-lived. Nerdal et al. [57] have proposed a solution structure for d(CGCGAATTCGCG) based on the refinement of NMR-derived distance geometry structures by NOESY spectrum bank-calculation. The proposed solution structure is not the same as the crystal structure and displays a number of kinks and an overall bending of the duplex. The presence of a kink at C3pG4 is indicated in the experiment, which is consistent with our result. It should be noticed that this kink is similar to that reported for the EcoRI restriction site DNA bound to its endonuclease [8]. It may be conceivable that the duplex in solution forms the conformation suitable for binding to the endonuclease with some probability. The MD structure at 100 ps is further bent. This bending appears to be associated with the opening between the C3-G22 base pair and the G4-C21 base pair toward the minor groove in addition to the C3pG4 kink.

Free energy studies: The transformation of 2-hydrogen to an NH_2 group results in free energy changes of -4.15, -3.76 and -1.65 kcal/mol for the duplex, the trimer and the isolated base system, respectively. As described earlier, the average free energy differences in the trimer system have been obtained by weighing the runs based on the stacking properties. Although two hydrogen-bonding hydrogens are added by this transformation, the decrease in the free energy of the isolated base in water is not as large as expected. This may be attributed to a decrease in the dipole moment from 2.29 D (9-Me-adenine) to 1.23 D (2-NH_2-9-Me-adenine). It is interesting that the electrostatic contribution for ΔG is positive rather than negative at lower values of λ. This suggests that a change in water structure is necessary to attain the preferential electrostatic interaction between the NH_2 group and the water molecules. In contrast, the free energy change of the trimer is larger and almost the same as that of the dodecamer. In this case, there is a large variation between the six simulations because of the existence of multiple configuration. It is observed that ΔG is larger (−3.36, −3.20 and −3.25 kcal/mol) when the stacking structure is preserved in the trimer and smaller (−5.18, −5.70 and −4.89 kcal/mol) when the stacking structure is distorted and other favorable interactions are formed between the perturbed adenine residue and the other residues or water molecules. The larger ΔG difference between the latter and the isolated base shows that 2-NH_2-adenine also interacts strongly with the other residues in a single-stranded state. In the duplex structure, the NH_2 group forms a hydrogen bond with the O2 of Thy19 without a change in intact structure. This leads to favorable electrostatic interaction between NH_2-Ade6 and Thy19 throughout the transformation. We compared the interaction energy of the minimized structure of the modified duplex containing NH_2-adenine with that of the intact duplex. In the energy-minimized structure, the electrostatic interaction energy between the base pair is 12.0 kcal/mol, larger than that in intact Ade-Thy base pairs (about 8 kcal/mol), but much smaller than that in intact Gua-Cyt base pairs (about 18 kcal/mol). The van der Waals interaction contributes significantly to the decrease in free energy, which arises mainly from the interaction of NH_2-Ade6 with Thy7, Thy19 and Thy20. The total interaction energy in the energy-minimized structure is −15.86 kcal/mol, about −11 kcal/mol and about −21 kcal/mol for the NH_2-Ade-Thy base pair, Ade-Thy base pairs and Gua-Cyt base pairs, respectively.

The transformation of both Ade6 and Ade18 to 2-NH_2-adenine brings about almost twice as much a decrease in ΔG as the single mutation, that is, 8.10 kcal/mol. The change in the interaction energy by the transformation shows the same tendency as that of one residue transformation of Ade6, except for the slight decrease in interaction energy between residues 6 and 18. This reflects that there is no stacking between residues 6 and 18 and the distance between two 2-NH_2 groups is longer than that between two 6-NH_2 groups. The difference in the free energy change of two 2-NH_2 group transformations between the duplex and the trimer is −0.19 kcal/mol, indicating that the duplex becomes slightly unstable with the substitution as compared to the single-stranded state in spite of the increase in the number of hydrogen bonds between the base pairs. Brennan and Gumport [59] reported the thermal stabilities of d(pGGAATTCC) and d(pGGA2,6APTTCC) in which the 2,6AP shows a 2-NH_2-adenine residue. It can be calculated from the data that ΔΔG is approximately −0.3 kcal/mol. There is a difference of 0.11 kcal/mol between the calculated and experimental ΔΔG values. However, taking into account the uncertainty of the single-stranded state due to the existence of multiple configuration and the difference of the target oligomer and the condition of the system such as pH and salt concentration, this difference appears to be relatively small and acceptable for semiquantitative discussion.

The introduction of F and Cl groups into the Ade6 residue makes the duplex slightly unstable as compared to the single-stranded state, as indicated by the ΔΔG values of 0.18 and −0.17 kcal/mol, respectively. It should be noted that in the trimer system, ΔG is considerably less as the stacking structure is disturbed, but it is larger in the isolated base system. This may reflect that the F and Cl groups do not interact favorably with the other residues in the trimer as opposed to the NH_2 group. The variations in the solvation free energies among 9-Me-adenine, 2-F-9-Me-adenine and 2-Cl-9-Me-adenine, which are 0.0, -1.09 and -2.29 kcal/mol, respectively, show good correspondence with the dipole moments of these molecules, which are 2.29, 3.09 and 4.37 D, respectively. The increase in the dipole moment seems to induce a significant increase in the electrostatic interaction between water and the modified base as shown in Figs. 4e and 5e, respectively. On the contrary, the free energy change in the duplex is due to the attractive van der Waals interaction and the repulsive electrostatic interaction between F-Ade6/Cl-Ade6 and the rest of the system. The electrostatic interaction energy between F-Ade6/Cl-Ade6 and Thy19 in the transformed duplex is smaller ((-6.61 kcal/mol)/(-5.47 kcal/mol)) than that in the unmodified duplex (-7.96 kcal/mol). This decrease is considered to be mainly due to the negative charge interactions between the F of F-Ade6/Cl of Cl-Ade6 and the O2 of Thy19. In addition, an increase in electrostatic interaction energy is observed between F-Ade6/Cl-Ade6 and the sugar part of Thy20 by 0.64/1.40 kcal/mol. On the other hand, favorable van der Waals interactions are induced between F-Ade6/Cl-Ade6 and a few bases such as Thy7, Thy19 and Thy20. Consequently, part of the repulsive electrostatic interaction is compensated by the attractive van der Waals interaction. However, the interaction energy of the modified duplex is still larger than that of the unmodified duplex by 1.48 kcal/mol for the F duplex and 3.00 kcal/mol for the Cl duplex.

The decrease in free energy by the transformation of 2-H to the OH group in the isolated base system (-2.91 kcal/mol) is larger than that of the NH_2 group (-1.65 kcal/mol). This is consistent with the fact that 2-OH-9-Me-adenine has a larger dipole moment (1.80 D) and an OH group forms a stronger hydrogen bond with water molecules. In fact, the main contribution to the free energy change comes from the electrostatic interaction as shown in Fig. 6e. It should be noted that the electrostatic interaction at lower values of λ is positive as in the case of the $H \rightarrow NH_2$ transformation. In other words, a change in the water structure is necessary for the preferential electrostatic interaction, that is, hydrogen-bond formation between the OH group and the water molecules. Further, one has to take into account the keto–enol tautomerism of the base. In general, the naturally occurring bases display predominantly keto tautomeric forms. Since the keto form of 2-OH-9-Me-adenine, namely, 9-Me-isoguanine, resembles a guanine, 2-OH-9-Me-adenine is to occur predominantly as the keto form rather than the enol form in solution. Therefore, the free energy change from 9-Me-adenine to 2-OH-9-Me-adenine/9-Me-isoguanine is assumed to be smaller than -2.91 kcal/mol. In the trimer system, a large decrease in free energy is also observed due to the electrostatic contribution. This is thought to come from the electrostatic interactions between the OH group and the other groups, which stabilize the single-stranded state more than the double-stranded state. It should be noted that this free energy change is not influenced by the variation of the trimer conformation compared to other transformations studied as shown in Fig. 6b. In the duplex system, the main contribution comes from the electrostatic interaction between OH-Ade6 and Thy19, and the interaction energy is more positive than that of NH_2-Ade by 2.0 kcal/mol, but more negative than that of unmodified Ade by 2.1 kcal/mol. However, the contribution to ΔG at lower values of I (0–4) is dominated by van der Waals interactions as shown in Fig. 6a. The OH-Ade6 forms attractive van der Waals interaction with Thy7, Ade18, Thy19 and Thy20. The OH group does not necessarily form a hydrogen bond with the O2 of Thy19 at lower values of I, since the electrostatic interaction between them is not strong yet. The H of the OH group tends to rotate to a different direction from that of O2 in order to attain the appropriate van der Waals interaction. Further, the distance between the two O atoms of the OH group and the O2 of Thy19 is slightly shorter in order to form the hydrogen bond for keeping the base pair intact. The keto form cannot form the base-pairing with thymine, since only one hydrogen bond is formed, if any. $\Delta\Delta G$ is 2.08 kcal/mol, which shows that the OH duplex is much more unstable than the unmodified duplex. Further, taking the existence of the keto form into consideration, the stability of the duplex in water may become worse. Recently, Benner and co-workers [60,61] have developed a method for incorporating a new Watson–Crick base pair into duplex DNA and RNA by DNA and RNA polymerases. They reported the formation of base pairing between 2-OH-Ade and Thy in 13-mer duplex DNA using the above method. Thus, our study of base-pairing between 2-OH-Ade and Thy is relevant in addressing the stabilities of these modified base pairs.

We used the standard error for the estimation of errors in the simulation. This evaluation would be valid when the errors come from statistical variation. It is possible, however, that systematic errors are caused by the biased sampling of configuration space and the time lag between Hamiltonian change and configuration change. The convergence of free energy calculations may be estimated by the comparison between the forward and reverse simulations except for the mutation of flexible molecules such as the trimer system. The discrepancy between the two simulations is below 0.40 kcal/mol except for that of the $H \rightarrow NH_2$ transformation in the duplex system, which is 0.85 kcal/mol for a mean free energy change of -4.15 kcal/mol. These discrepancies are considered to be acceptable for the semiquantitative comparison among the free energy changes examined. In the trimer system, the ΔG value is not necessarily in agreement between the simulations, since both the initial and final states are supposed to have multiple minima which show very different configurations from each other. In such a case, it is difficult to attain convergence, unless a very long simulation is carried out, which is impractical in the present computer environment. In order to reduce this difficulty, we performed the simulations using single-stranded trimer, not dodecamer, and the six or four different structures as the initial configurations. The simulation results are classified into two groups. In the first group, the base stacking structure is maintained during the simulation, and in the second group the base stacking structure is disturbed. Thus, the mean value of four to six simulations may approximate to the true value, since the system is assumed to transfer between the two groups described above. Although a more strict estimate of the convergence as well as the improvement of force field parameters may be necessary to obtain quantitatively accurate free energy differences, this study strongly suggests that molecular dynamics and free energy perturbation can give us profound insight into the dynamical property and stability of DNA, which is more flexible than globular proteins.

Table 12 *Free energy differences for 2-Ade analogues*

System	ΔΔG (kcal/mol)	
	Standard average	Weighted average
NH_2-Adenine (6)	0.11 ± 0.46	− 0.39 ± 0.74
NH_2-Adenine (6,18)	0.42 ± 0.46	− 0.19 ± 0.74
F-Adenine	0.10 ± 0.09	0.18 ± 0.11
Cl-Adenine	− 0.26 ± 0.18	− 0.17 ± 0.13
OH-Adenine	1.98 ± 0.15	2.08 ± 0.17

The relative stabilities of the modified DNA duplexes in solvent water are summarized in Table 12. The calculated stability increases in the following order: OH duplex < F duplex < unmodified duplex < Cl duplex < NH_2 duplex. Taking into consideration the uncertainty due to the existence of multiple configuration in the trimer system, except for the OH duplex, the stability should be estimated to be almost equal. The dodecamer duplex used in this study is known to be recognized specifically through the recognition site (GAATTC) and cleaved the phosphodiester bond between the guanine and adenine residues by the EcoRI endonuclease. It has been reported [10,11] that the modification of adenine residues in the recognition site influences the cleavage reaction by the endonuclease. The introduction of an NH_2 group into the 2-position of the inner adenine residue of d(pGGAATTCC) causes a 4.6-fold decrease in k_{cat} and a 1.8-fold increase in K_m, whereas the change from the 6-position to the 2-position of the NH_2 group in d(pGGAATTCC) and d(CTGAATTCAG) results in a 3-fold decrease and a 1.3-fold increase, respectively, in k_{cat} and a 4-fold and a 12-fold increase, respectively, in K_m. In addition, the modified duplex, in which an NH_2 group is introduced into the 2-position of the outer adenine residue of d(CTGAATTCAG), is not a substrate but a competitive inhibitor for the endonuclease. On the other hand, the duplex obtained by changing the NH_2 from the 6-position to the 2-position is neither a substrate nor an inhibitor. These results suggest that the introduction of an NH_2 group into adenine residues of the recognition site influences the catalytic step and not the binding step for the endonuclease cleavage reaction. This is reasonable, since the 2-NH_2 group is not supposed to interact with the endonuclease directly as shown by the X-ray structure of the DNA–endonuclease complex [8,9]. Further, it does not seem that the inhibition of the catalytic step is induced by the function of the NH_2 group itself in the 2-position, because the substitution of 2-NH_2-purine, in which there is a 2-NH_2 group instead of a 6-NH_2 group, which interacts with the endonuclease, results in the inhibition of the binding step rather than the catalytic step. It has been suggested that the accurate discrimination between the canonical recognition site and its closely related sites by EcoRI endonuclease is dependent not only on the complementary interaction between DNA and the endonuclease mainly through the protein-base contacts [8,9] and the protein-phosphate contacts, [8,62–64], but also on the conformational change from the enzyme–substrate complex to the transition state complex [8,65,66]. In the case of the 2-NH_2-adenine duplex, the DNA conformation change seems to be associated with a decrease in cleavage rate, since the contact mode between the duplex and the endonuclease is not considered to change by the modification. This study suggests that the stability of the duplex, namely, the equilibration between the double-stranded state and the single-stranded state, is not influenced significantly by the introduction of a 2-NH_2 group. The local denaturation in the recognition site of the duplex rather than the complete one appears to be important for the conformational change to the transition state. This is also supported by the experimental results [66] that the oligodeoxynucleotides containing a mismatch within the EcoRI recognition site are rather good substrates for the endonuclease and, in some cases, they are cleaved more efficiently than those containing the canonical sequence. We are in the process of studying the dynamic behavior of modified duplexes and the result will be published elsewhere.

References

1. Rao, S.N., Singh, U.C. and Kollman, P.A., J. Am. Chem. Soc., 108(1986)2058.
2. Remers, W.A., Rao, S.N., Singh, U.C. and Kollman, P.A., J. Med. Chem., 29(1986)1256.
3. Hausheer, F.H., Singh, U.C. and Colvin, O.M., Anti-Cancer Drug Des., 5(1990)159.
4. Langley, D.R., Doyle, T.W. and Beveridge, D.L., J. Am. Chem. Soc., 113(1991)4395.
5. Singh, S.B., Hingerty, B.E., Singh, U.C., Greenberg, J.P., Geacintov, N.E. and Broyde, S., Cancer Res., 51(1991)3482.
6. Hausheer, F.H., Singh, U.C. and Saxe, J.D., J. Am. Chem. Soc., submitted.
7. Frederick, C.A., Grable, J., Melia, M., Samudzi, C., Jen-Jacobson, L.J., Wang, B.-C., Greene, P., Boyer, H.W. and Rosenberg, J.M., Nature, 309(1984)327.
8. McClarin, J.A., Frederick, C.A., Wang, B.-C., Greene, P., Boyer, H.W., Grabale, J. and Rosenberg, J.M., Science, 234(1986)1526.
9. Kim, Y., Grable, J.C., Love, R., Greene, P.J. and Rosenberg, J.M., Science, 249(1990)1307.
10. Brennan, C.A., Van Cleve, M.D. and Gumport, R.I., J. Biol. Chem., 261(1986)7270.
11. McLaughlin, L.W., Benseler, F., Graeser, E., Piel, N. and Scholtissek, S., Biochemistry, 26(1987)7238.
12. Fliess, A., Wolfes, H., Rosenthal, A., Schwellnus, K., Blocker, H., Frank, R. and Pingoud, A., Nucleic Acids Res., 14(1986)3463.
13. Seela, F. and Kehne, A., Biochemistry, 26(1987)2232.
14. Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, S., Italkura, K. and Dickerson, R.E., Proc. Natl. Acad. Sci. USA, 78(1981)2179.
15. Levitt, M., Cold Spring Harbor Symp. Quant. Biol., 47(1983)251.
16. Tidor, B., Irikura, K.K., Brooks, B.R. and Karplus, M., J. Biomol. Struct. Dyn., 1(1983)231.
17. Singh, U.C., Weiner, S.J. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 82(1985)755.
18. Rao, S.N., Singh, U.C. and Kollman, P.A., Isr. J. Chem., 27(1986)189.
19. Siebel, G.L., Singh, U.C. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 82(1985)6537.
20. Van Gunsteren, W.F., Berendsen, H.J., Guersten, R.G. and Zwinderman, H.R., Ann. New York Acad. Sci., 482(1986)287.
21. Rao, S.N. and Kollman, P.A., Biopolymers, 29(1990)517.

22. Zielinski, T.J. and Shibata, M., Biopolymers, 29(1990)1027.
23. Srinivasan, J., Withka, J.M. and Beveridge, D.L., Biophys. J., 58(1990)533.
24. Hausheer, F.H., Singh, U.C., Palmer, T.C. and Saxe, J.D., J. Am. Chem. Soc., 112(1990)9468.
25. Swaminathan, S., Ravishanker, G. and Beveridge, D.L., J. Am. Chem. Soc., 113(1991)5027.
26. For a review of recent applications using free energy perturbation methods, see:
    a. Beveridge, D.L. and DiCupa, F.M., Annu. Rev. Biophys. Biophys. Chem., 18(1989)431.
    b. Van Gunsteren, W.F., Protein Eng., 2(1988)5.
    c. Karplus, M. and Petsko, G.A., Nature, 347(1990)631.
27. Pearlman, D. and Kollman, P., Biopolymers, 29(1990)1193.
28. Dang, L.X., Pearlman, D. and Kollman, P., Proc. Natl. Acad. Sci. USA, 87(1990)4630.
29. Ross, W., Hardin, C., Tinoco, I., Rao, S., Pearlman, D. and Kollman, P., Biopolymers, 28(1989)1939.
30. Zwanzig, R.W., J. Chem. Phys., 22(1954)1420.
31. Singh, U.C., Brown, F.K., Bash, P.A. and Kollman, P.A., J. Am. Chem. Soc., 109(1987)1607.
32. Rao, B.G. and Singh, U.C., J. Am. Chem. Soc., 111(1989)3125.
33. Singh, U.C. and Kollman, P.A., QUEST (version 1.0), University of California, San Francisco, CA, 1986.
34. Hehre, W.J., Stewart, R.F. and Pople, J.A., J. Chem. Phys., 51(1969)2657.
35. Singh, U.C. and Kollman, P.A., J. Comput. Chem., 5(1984)129.
36. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P.K., J. Am. Chem. Soc., 106(1984)765.
37. Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7(1986)230.
38. Allinger, N., J. Am. Chem. Soc., 99(1977)8127.
39. Jorgensen, W.L., Chandrasekhar, J. and Madura, J.D., J. Chem. Phys., 79(1933)926.
40. a. AMBER (version 3.3.) is a fully vectorized version of AMBER(3.0) with coordinate coupling, intra/inter decomposition and the option to include the polarization energy as part of the total energy by Singh et al. (Singh, U.C., Weiner, P.K., Caldwell, J.W. and Kollman, P.A., University of California, San Francisco, CA, 1986).
    b. Galaxy 2.0, A Multipurpose Molecular Modeling and Drug Design Package. Copyright (1995), AM Technologies Inc., San Antonio, TX.
41. Van Gunsteren, W.F. and Berendsen, H.J.C., Mol. Phys., 34(1977)1311.
42. Prive, G., Heinemann, U., Chandrasegaran, S., Kan, L., Kopka, M. and Dickerson, R., Science, 238(1987)498.
43. Cruse, W.B.T., Salisbury, S., Brown, T., Cosstick, R., Eckstein, F. and Kennard, O., J. Mol. Biol., 192(1986)891.
44. Saenger, W., Principles of Nucleic Acid Structure, Springer, New York, NY, 1984.
45. Westof, E. and Sundaralingam, M., J. Am. Chem. Soc., 105(1983)970.
46. Callandine, C., J. Mol. Biol., 161(1982)343.
47. Dickerson, R.E. and Drew, H.R., J. Mol. Biol., 149(1981)761.
48. Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, K., Itakura, K. and Dickerson, R.E., Proc. Natl. Acad. Sci. USA, 78(1981)2179.
49. Rao, S.N. and Kollman, P., Biopolymers, 29(1990)517.
50. Altona, C. and Sundaralingam, M., J. Am. Chem. Soc., 95(1973)2333.
51. Altona, C., In Sundaralingam, M. and Rao, S.T. (Eds.) Structure and Conformation of Nucleic Acids and Protein–Nucleic Acid Interactions, University Park Press, Baltimore, MD, 1975, pp. 613–629.

52. Olsthoorn, C.S.M., Haasnoot, C.A.G. and Altona, C., Eur. J. Biochem., 106(1980)85.
53. Powell, J.T., Richards, E.G. and Gratzer, W.B., Biopolymers, 11(1972)235.
54. Reich, C. and Tinoco Jr., I., Biopolymers, 19(1980)833.
55. Olsthoorn, C.S.M., Bostelaar, L.J., De Rooij, J.F.M., Van Boom, J.H. and Altona, C., Eur. J. Biochem., 115(1981)309.
56. Olson, W., In Neidle, S. (Ed.) Topics in Nucleic Acid Structure, Part 2, Macmillan, London, 1982, pp. 1–76.
57. Nerdal, W., Hare, D.R. and Reid, B.R., Biochemistry, 28(1989)10008.
58. Dickerson, R.E., Goodsell, D.S., Kopka, M.L. and Pjura, P.E., J. Biomol. Struct. Dyn., 5(1987)557.
59. Brennan, C.A. and Gumport, R.I., Nucleic Acids Res., 13(1985)8665.
60. Switzer, C., Moroney, S.E. and Benner, S.A., J. Am. Chem. Soc., 111(1989)8322.
61. Piccirilli, J.A., Krauch, T., Moroney, S.E. and Benner, S.A., Nature, 343(1990)33.
62. Lu, A.-L., Jack, W.E. and Modrich, P., J. Biol. Chem., 256(1981)13200.
63. Becker, M.M., Lesser, D., Kurpiewski, M., Baranger, A. and Jen-Jacobson, L., Proc. Natl. Acad. Sci. USA, 85(1988)6247.
64. Yanofsky, S.D., Love, R., McClarin, J.A., Rosenberg, J.M., Boyer, H.W. and Greene, P.J., Proteins, 2(1987)273.
65. Lesser, D.R., Kurpiewski, M.R. and Jen-Jacobson, L., Science, 250(1990)776.
66. Thielking, V., Alves, J., Fliess, A., Maass, G. and Pingoud, A., Biochemistry, 29(1990)4682.

Indexes

Author index

Subject index