Protein Folding and Optimization Algorithms

John E. Straub
Boston University, MA, USA

Abbreviations

CBMC = configurational bias Monte Carlo; DEM = diffusion equation method; LES = locally enhanced sampling; QMA = quantum mechanical annealing.

1 INTRODUCTION

The determination of a protein's three-dimensional structure is a fundamentally important problem in biochemistry and

molecular biology.[1] Enzymological analysis can provide clues as to how a particular enzyme functions. Site-directed mutagenesis can provide evidence that certain amino acids are crucial to the enzyme's function. Equilibrium binding thermodynamics and kinetic analysis can provide clues as to how a particular reaction is catalyzed by the enzyme and how it may be bound. However, only when the three-dimensional structure of the protein is known can a complete understanding of the enzymatic mechanism be acquired.[2]

Experimental methods are available for determining protein structures. X-ray crystallographic analysis provided the first protein structures and remains the most precise method for structure determination. NMR can provide information as to the distances between pairs of atoms in the protein in solution (see ***Distance Geometry: Theory, Algorithms, and Chemical Applications***). The information provides restraints on the possible structure of the protein and can be used to obtain refined three-dimensional protein structures.

Proteins in solution are typically more mobile than those in a crystal. At equilibrium the protein in solution will fluctuate and its conformation is best represented as an 'ensemble' of structures. This is not so bad! The flexibility can leave certain portions of the protein unstructured. The mobile portions of the protein are often functionally important. Therefore, one can appreciate the disorder in the refined solution phase structure as a clue to the function of the protein and its equilibrium structure. When we seek to determine a protein's 'structure' at room temperature we must appreciate that the native state consists of an ensemble of structures.

While methods for the experimental determination of a protein's structure are increasingly automated, the difficulties faced in the experimental determination of a protein structure can be significant. Fortunately, with the advent of high speed computers, it is now possible to address the problem of protein structure prediction in computo (see ***Protein Structure Prediction in 1D, 2D, and 3D***).

Over the past 30 years, realistic atomic-level mathematical models of proteins have been developed and tested. It is now possible to analyze the energetics of conformational changes in peptides and proteins with the accuracy required to answer quantitative questions regarding the nature of short-time protein dynamics and the extent of fluctuations in a protein near its 'native' folded state.[3,4]

Rapid progress is being made in the further refinement of atomic models.[5] The next 10 years will yield a qualitative improvement in our ability to model the energetics of peptide and protein conformational equilibria (see ***Protein Force Fields***). More distant is the possibility of making obsolete the empirical energy function by performing *ab initio* molecular dynamics where the quantum mechanical ground state potential energy surface is generated 'on the fly' during a simulation. This article describes promising computational approaches which may lead to a solution of the as yet unsolved protein structure prediction problem.

1.1 Computational Models of Proteins

Realistic simulations of proteins in computo must include an accurate model of solvation (see ***Solvation: Modeling***). Atomic-level models which include explicit solvent molecules are now routinely used. Currently, it is common to carry out a computer simulation of a 1500-atom protein in a 'box' of 3000 water molecules for times exceeding 1 ns (see ***Molecular Dynamics and Hybrid Monte Carlo in Systems with Multiple Time Scales and Long-range Forces: Reference System Propagator Algorithms***). More realistic models of the solvent will include ions which are particularly important in the accurate treatment of charged hydrophilic residues on the protein's surface. While this is increasingly done using all-atom models, an important alternative is to represent the solvent as a continuum with an associated dielectric constant and ionic strength.[6,7] Reaction field techniques employ atomic level models of the solvent atoms closest to the protein while using a continuum representation of the solvent at greater distances.[8,9] Continuum solvent and reaction field methods allow one to simulate an isolated protein or infinitely dilute protein solution while Ewald sum methods treat the long-range interactions through the assumption of an infinite periodic system (see ***Molecular Dynamics: Techniques and Applications to Proteins***).[10] Another aspect of solvation, the exchange of titratable hydrogen ions between the protein and water, can be treated in equilibrium thermodynamics calculations or *ab initio* molecular dynamics.

It has been argued that the accuracy necessary to discriminate between the native and unfolded state of a protein is typically 1 kcal mol^{-1} per residue.[11] Empirical force fields have not yet reached that level of accuracy. However, it can be expected that computational models of solvated proteins will be accurate enough to predict the thermodynamic equilibrium of protein folding.

1.2 Impossibility of Exhaustive Search

When one uses computer simulations to examine thermodynamic properties of a protein using an atomic-level model, one makes statistical assumptions regarding the relative importance of conformations of the protein. To calculate the average energy of the system, we assume that the protein will visit all possible conformations of the protein that have finite energy. Many high energy conformations make very small contributions and it is a reasonable approximation to ignore them. However, for a peptide of intermediate size or a protein, at room temperature, there are far more statistically important conformations available to the system than can be sampled by a computer simulation. Unfortunately, to rigorously make contact between the computed microscopic model and thermodynamic averages, we must demand that all possible conformations in the system are sampled. (When molecular dynamics is used to sample conformations of the protein, that demand is equivalent to satisfying the 'ergodic hypothesis'.) This sampling problem exists and is independent of the realism of the atomic-level models used to represent the protein.[1]

(Fortunately, it is empirically true that for many systems, while the sampling in a computer simulation is not exhaustive, computed thermodynamic averages can be accurate. This can be true for several reasons. The first is that there may be 'broken ergodicity' for the experimental system as well. The computation of a thermodynamic property of the native state of the protein may not require an average over unfolded structures of the protein. Secondly, experimental measurements taken over a short time will not sample all fluctuations available to the system over an infinite time. Finally, while the experimental system may fluctuate a great deal more than the simulated

model of the system, the particular observable may be insensitive to many of those fluctuations. Therefore, the fact that the sampling of conformations is incomplete should not prevent one from carefully exploring thermodynamic properties of proteins such as folding free energies using computer simulation. See ***Free Energy Simulations***.)

1.3 Minimal Models of Proteins

To address the difficulty imposed by the limitations on sampling and the time scales accessible to computer simulations of atomic-level models, 'mimimal models' of proteins have been developed.[12-18] These models are meant to preserve the most fundamental interactions that are believed to contribute to the thermodynamic and kinetic properties of proteins (see ***Protein Modeling***). These include solvophobic interactions (hydrophobic/hydrophilic partitioning), backbone hydrogen bonding, and structural constraints (imposed by backbone dihedral angles summarized in the Ramachandran potential energy surface).[19,20]

The 'minimization' of the model can proceed ad hoc guided by chemical intuition. Alternatively, minimal models may be derived by a well-defined coarse-graining transformation beginning with an all-atom representation of the system. They may also be empirically derived by proposing a general functional form and performing a non-linear fit of the parameters of the function to a set of experimental data.[21] These models are attractive in that the number of degrees of freedom and possible conformations is far fewer than the number of conformations accessible to an all-atom model.

1.4 Defining the Native Folded State of a Protein

Given a model of the protein, one is left to define a criterion for the folded state of the protein. There is much debate over whether the native folded state of the protein represents a thermodynamically dominant state or a kinetically trapped state.[22] It may be that there are examples of both types. To determine the structure of a kinetically determined folded state, it must be possible to simulate the folding of the protein from its unfolded state. In real time, this process typically takes on the order of milliseconds to seconds and is beyond the realm of current molecular dynamics simulation.[23] If the native state is thermodynamically dominant then, if the equilibrium ensemble of protein structures can be simulated, the native state could presumably be recognized from the set of structures at equilibrium.[18] Because of computational limitations, it is not possible to generate an ensemble of structures that adequately represents the equilibrium ensemble for even the smallest globular proteins. A simplifying assumption is necessary.

An assumption which greatly simplifies the problem is that the native state of the protein is well represented by the conformation of lowest free energy. An even greater simplification follows by assuming that the conformation of lowest free energy is also the protein conformation of lowest potential energy. The latter view, popularized by the pioneering work of Harold Scheraga and co-workers, is adopted in most of the work to date on protein folding.[24] The problem of protein structure prediction becomes one of global optimization where the object is to minimize the free energy or potential energy as a function of the protein's conformation. This view will be adopted throughout this article.

Independent of the exact features of the model or criterion defining the protein's folded state, the computational demands of evaluating thermodynamic and kinetic properties of these models can be formidable. At the present time, the best methods combined with the most powerful computational engines are inadequate to fold an all-atom model of a protein in computo. As such, a careful choice of the computational method is essential. The development of new computational methods is infinite in its possibilities. The field of development of conformational optimization algorithms for proteins has shown rapid progress in recent years. This rapid development of new algorithms promises to continue. This article provides a snapshot of the field of protein structure prediction as a problem of conformational optimization. There is an emphasis on the most general and fundamental methods where further development appears to be most likely. The discussion is not intended to be a comprehensive review or even a survey of the most effective methods. The reader is referred to the references for a more comprehensive discussion.

2 METHODOLOGY

This section briefly describes a variety of methods which can be effective in searching for low energy minima of a model protein's potential energy function.

2.1 Searching the Energy Landscape

The search for the protein conformation of lowest potential energy can proceed in many ways. Initially a seed conformation is generated. It may be that some experimental information is available to restrict the possible conformations. The information may be as general as a radius of gyration derived from neutron scattering or as precise as a three-dimensional atomic structure derived from diffraction studies. Once a starting structure is generated the search of the energy landscape demands a method for generating a new conformation of the protein based on a previously defined structure or set of structures. The two paradigms for such searches are the molecular dynamics (MD) and Monte Carlo (MC) algorithms.

The MD method deterministically generates a new configuration in a volume centered about a previous set of configurations by solving Newton's equation of motion using a finite difference algorithm.[25] It is important that the algorithm is time-reversible and area-preserving in phase space.[8] For this search it is necessary to know the force on all atoms or particles in the model of the protein (see ***Molecular Dynamics: Techniques and Applications to Proteins***).

In the primitive MC method[8] a trial configuration is randomly generated from an existing reference structure. If the energy of the trial configuration is less than or equal to the energy of the reference configuration, it is accepted as the new configuration. If the trial configuration is higher in energy by an amount ΔU, it is accepted with a probability $\exp(-\Delta U/kT)$ where T is the temperature and k is Boltzmann's constant (see ***Monte Carlo Simulations for Liquids***).

Some algorithms exist between the MD and MC algorithmic paradigms. These methods include noisy or stochastic MD, where a random force is added to the gradient of the potential, and hybrid[26] and smart MC[8,25] where the force is used to inform the choice of a trial move in the MC method (see

Molecular Dynamics and Hybrid Monte Carlo in Systems with Multiple Time Scales and Long-range Forces: Reference System Propagator Algorithms). Other variations include modifications of the potential in MD or the probability of acceptance employed in the MC method. A sampling of the more successful algorithms is presented below.

2.1.1 Simulated Annealing using MD and MC

Suppose we hope to find the lowest point on the surface of the Earth. We might start with a coarse photograph of the Earth to recognize low lying basins. More detailed images of these basins could then provide better estimates of the depth of the basins. Finally, ground exploration could give a precise measure of the elevation of a particular site. In simulated annealing the coarse-grained search is carried out with an initial dynamics at high temperature (a bit like flying high over the potential surface). The system is then slowly cooled. As the temperature is lowered, the system will spend increasingly long periods of time in low lying energy basins. At low enough temperatures, the system will be found with highest probability in the lowest lying energy minimum. If the temperature is lowered slowly enough, in proportion to the inverse logarithm of the simulation length, the system will sample the equilibrium distribution at every temperature. It can then be guaranteed that the system will find the global energy minimum. The 'simulated annealing' protocol is readily implemented using MD or MC methods to search the conformational space (see ***Macromolecular Structure Calculation and Refinement by Simulated Annealing: Methods and Applications***).

The simulated annealing method is named for the analogy with the physical process of annealing. High temperature equilibration followed by slow cooling removes strains in the system. Because of the practical limits on computation, for all but the simplest systems, a faster than optimal annealing schedule must be used. In such cases, some strains will be removed, but there is no guarantee that the global energy minimum will be located.

Ultimately, it is the effectiveness of an optimization algorithm that determines its usefulness.[27] Simulated annealing is an optimization method that provides a guarantee, even in what is for most systems at present a computationally unrealizable limit, that the global minimum will be found. Moreover, the clear analogy with the physical process of annealing makes the algorithm more than a 'black box' method. Finally, it is easily implemented in an MD or MC simulation program.

The general nature of the simulated annealing method makes it possible to generalize the method of search in ways that lead to significant improvements over standard MD- and primitive MC-based searches. Some of these techniques are summarized in this section.

2.1.2 Standard MC

In the primitive MC method,[8] the probability of making a transition from one conformation, x, to another, x', consists of two parts. The first is the probability of making a trial move from conformation x to conformation x' which is $T(x \rightarrow x')$. The second is the probability of accepting the trial move $A(x \rightarrow x')$. The underlying demand that the system should sample the equilibrium probability distribution of being in a given conformation x with probability $\rho(x)$ is satisfied by the condition of 'detailed balance' which we write

$$\rho(x)T(x \rightarrow x')A(x \rightarrow x') = \rho(x')T(x' \rightarrow x)A(x' \rightarrow x) \qquad (1)$$

The generated 'walk' through conformational space should be non-repeating (irreducible) and each move should depend only on the current position and not on the history of the walk (Markovian).

In the standard Metropolis MC algorithm, the trial move is generated uniformly within a region of space such that $T(x \rightarrow x') = T(x' \rightarrow x) = \text{constant}$. To sample the Boltzmann probability of the system visiting a conformation x' relative to another conformation x, given by $\rho(x')/\rho(x) = \exp(-\Delta U/kT)$, the acceptance probability must be given by

$$p = \min[1, \exp(-\Delta U/kT)] \qquad (2)$$

Here $\Delta U = U(x') - U(x)$ is the change in potential energy associated with the move. The 'walk' through conformation space generated in such a way is Markovian.

It is possible to generate a class of 'smart' MC algorithms by replacing the trial move uniformly generated within a give region with a trial move based on details of the potential surface that improve the sampling of conformational space. For example, by biasing the move in the direction of the force, it is more likely that the move will be accepted. We can vary the trial move distribution as long as (1) we continue to satisfy the condition of detailed balance and (2) the walk through conformation space is non-repeating and Markovian. The following sections describe a number of effective MC search methods based on this idea.

2.1.3 Tsallis Statistical MC

One general method for choosing the trial move is to generate many conformations of the protein according to a distribution $\tau(x)$. The trial move can then be sampled randomly from the distribution of conformations $\tau(x)$ such that $T(x \rightarrow x') = \tau(x')$ and similarly $T(x' \rightarrow x) = \tau(x)$. Note that in this form it is no longer true that $T(x \rightarrow x') = T(x' \rightarrow x)$. Not to worry! As long as the detailed balance criterion is satisfied we can expect to sample the correct equilibrium distribution for the protein $\rho(x)$. The new acceptance probability takes the form

$$p = \min[1, \exp(-\Delta U/kT)\tau(x)/\tau(x')] \qquad (3)$$

Andricioaei and Straub[28] have recently employed a similar acceptance probability where the trial step is sampled from a distribution function of a form proposed by Tsallis.[29] In 'Tsallis statistics', the standard Gibbs entropy $S = -k \int dx\, \rho(x) \ln \rho(x)$ is modified to the form $S_q = k \int dx (1 - \rho_q(x)^{q-1})/(q-1)$ which is equal to the Gibbs entropy formula in the limit that $q = 1$. The equilibrium probability distribution functions take the form

$$\tau(x) = \frac{1}{Z_q}\left[1 - (1-q)\frac{U(x)}{kT}\right]^{1/(1-q)} \qquad (4)$$

where Z_q is a normalization constant (the generalization of the partition function). When $q > 1$ the distribution function $\tau(x)$ is significantly more delocalized than the Gibbs–Boltzmann statistical probability $\rho(x)$. For example, when the potential is harmonic the Gibbs–Boltzmann distribution function is a

Gaussian while the Tsallis distribution function ($q = 2$) is a Cauchy-Lorentz distribution.

In simulations of proteins there are often high free energy barriers partitioning the conformation space. By sampling either continuously or occasionally according to the delocalized distribution $\tau(x)$ it is possible for the MC walk to overcome these barriers and to sample more effectively the conformational space of the protein. This sampling method may be applied to an MC simulated annealing search in a straightforward manner.

Alternatively, in a simulated annealing protocol, the system can be initially simulated at high temperature and $q > 1$ using $p = \min[1, \tau(x')/\tau(x)]$. As the temperature is lowered to zero, the value of q is lowered to 1. During the run, there is enhanced sampling of the conformational space. At the end of the annealing run, the physically relevant Gibbs-Boltzmann statistical probability is recovered. This method has been applied to the global optimization of a tetrapeptide.

2.1.4 *Jump-walking MC*

In the jump-walking MC of Franz et al.[30] a trial move is generated with probability P_J from a probability distribution generated at a high temperature T_J as $T(x \to x') = \tau(x') = \exp[-U(x')/kT_J]/Z(T_J)$. The new conformation is accepted with probability

$$p = \min[1, \exp(-\Delta U/kT)\tau(x)/\tau(x')] \tag{5}$$

Otherwise, with probability $(1 - P_J)$, the system makes a standard trial move at temperature T.

This method has the advantage of mixing the enhanced global sampling of a high temperature MC walk with the effective local sampling of a low temperature MC walk. It has been shown to be quite effective and can readily be applied to the simulated annealing of proteins and peptides.

2.1.5 *Multicanonical MC*

The difficulty in overcoming high free energy barriers on the conformational potential energy surface lies in the fact that the canonical ensemble probability of sampling any conformation x of energy E is $\rho(E) = n(E)\exp[-E/kT]/Z(T)$. An important innovation has been made by Berg. To improve the sampling of high energies it is possible to make the probability of visiting a conformation of energy E independent of the energy. This corresponds to a 'multicanonical' probability distribution function $\rho(E) \propto n(E)w(E) =$ constant which implies that $w(E) \propto 1/n(E)$.[31] The multicanonical MC walk is thus a one-dimensional random walk in energy. The system is able to readily overcome barriers of any energy. This makes it a very attractive method for application to biomolecular systems.[32,33]

In practice, one must generate an estimate of the density of states $n(E)$. This can be done by performing a series of MC runs at high temperatures. The high temperature walks sample much of conformational space including regions with high energy barriers. From the runs, the density of states $n(E)$ can be extracted. If a symmetric trial distribution is used and $T(x \to x') = T(x' \to x)$ then the acceptance probability for making a move from x to x' can be written

$$p = \min[1, n[E(x)]/n[E(x')]] \tag{6}$$

Equilibrium averages in the canonical ensemble may be calculated by reweighting the probability $\rho(E) = n(E)\exp[-E/kT]/Z(T)$. More details on the computation of $n(E)$ for proteins over a focused region of energy and the application to simulations of helix-coil transitions can be found in the literature.[34]

2.1.6 *Configurational Bias MC (CBMC)*

A promising method for the effective search of conformational space using MC methods has been proposed by Siepmann and co-workers and De Pablo (see ***Monte Carlo Simulations for Polymers***).[8,35] The method is based on an MC method where the trial move distribution is biased towards regions where the acceptance probability will be greatest. A bias of this kind is important in polymeric systems with topological constraints. An example is the constraints imposed on the bonds, angles, and torsions of peptides and proteins. The CBMC technique has been applied to small peptides in vacuum by Deem and Bader.[36] Extension of the method to simulations of peptides employing implicit or continuum models of solvation should be straightforward.

2.1.7 *Standard MD*

It is also possible to search conformational space using a numerical solution of Newton's equations of motion ($F = ma$).[8,25] The numerical solution can be achieved using a variety of finite difference algorithms such as the Verlet algorithm

$$x(t + dt) = 2x(t) - x(t - dt) + dt^2F(x)/m \tag{7}$$

where $F(x)$ is the force on the system at configuration x, (which is typically equal to the gradient of an empirical potential energy function $V(x)$), $x(t)$ is the configuration of the system at time t, and dt is the time step used in the integration. Given the system's configuration at times t and $t - dt$, the new configuration of the system can be determined at time $t + dt$. This equation of motion can be used to simulate a protein system at constant energy as a way of searching conformational space or simulating protein folding. Below we describe a number of methods which build on the MD paradigm to provide enhanced sampling of conformational space.

2.1.8 *MD using Extended Lagrangian Methods*

Following the extended Lagrangian paradigm proposed by Andersen,[8] the sampling in an MD simulation can be enhanced by adding fictitious degrees of freedom to the molecular system. The additional degrees of freedom can encourage the system to access conformations only rarely visited by a standard MD trajectory. The method is rigorous in that exact thermodynamic averages can be computed from the fictitious dynamics. It is also possible to parameterize the extended Lagrangian such that as the temperature is lowered in a simulated annealing run, the additional degrees of freedom decouple from the system and the final energy of the actual system is recovered.

The Hamiltonian energy function of the protein is $H_0(r, p)$. The variable w can be added with an associated fictitious mass m so that the transformed Hamiltonian is

$$H(r, p; w, p_w) = H_0(r, p) + \frac{1}{2m}p_w^2 + U(w) \tag{8}$$

The potential $U(w)$ constrains the fluctuation of w about some well-defined mean value. Parameters in the protein energy function $H_0(r, p)$ can then be made to depend implicitly on w. In the work of Berne these additional variables include the atomic diameters which are allowed to fluctuate around average values.[37] It has been shown that, by allowing the particle diameters to fluctuate, the topological frustration associated with conformational changes in compact states of biomolecular systems can be partially relieved. This leads to a significant increase in the rate of conformation space sampling over the molecular dynamics run and a higher probability of finding low lying energy minima on the potential energy surface.

2.1.9 *MD in Four Dimensions*

In a spirit similar to the extended Lagrangian method, an additional degree of freedom can be added to the position of all particles in the system. The position of an atom becomes (x, y, z, w) and the Euclidean metric distance between two particles is taken in a four-dimensional space

$$r = (x^2 + y^2 + z^2 + w^2)^{1/2} \tag{9}$$

The potential energy function of the protein $U_0(r)$ is modified to be

$$U(r) = U_0(r) + \kappa w^2 \tag{10}$$

where all distances and gradients are then computed using the modified metric. The additional external potential term is added to restrain the value of w to fluctuate about zero. When $w = 0$ the standard three-dimensional protein potential is recovered. Newton's equation of motion is simply generalized using a four-dimensional gradient. The resulting dynamics have been shown to provide an enhancement of the system's ability to sample conformational space over standard MD simulations.[38]

2.2 Searching a Deformed Energy Landscape

In the previous sections we concentrated on methods which can effectively search the rugged potential energy surface of a protein. A complementary approach is to deform the potential energy surface to make it less rugged.[27,39] The 'smoothed' potential has lower barriers acting as partitions to a thorough search.[40]

A very general type of smoothing transformation applied to the potential energy surface $U(x)$ is defined by the integral transform

$$V_\varepsilon(x_0) = \int dx\, U(x) S_\varepsilon(x, x_0) \tag{11}$$

where $S_\varepsilon(x, x_0)$ is a windowing function peaked about x_0 with characteristic width ε. As the smoothing length scale ε is increased the smoothed potential function $V_\varepsilon(x_0)$ has fewer and fewer minima. The smoothed potential can then be more easily searched by standard algorithms. In special cases, the smoothing length scale can be increased to the point that a single energy minimum survives. This sole surviving minimum can be found using a local energy minimization technique. If the smoothing function $S_\varepsilon(x, x_0)$ is carefully chosen, the energy minimum on the smoothed surface $V_\varepsilon(x_0)$ will maintain the symmetry of the global minimum of the true potential energy function $U(x)$.

In this section a number of energy minimization methods are presented which rest, at some level, on a smoothing transformation of the potential surface.

2.2.1 *Diffusion Equation Method (DEM)*

Scheraga and co-workers proposed an iterative minimization method based on the search of a smoothed potential energy surface. In their DEM[41] they employed a Gaussian smoothing function

$$S_\varepsilon(x, x_0) = (2\pi\varepsilon^2)^{-1/2} \exp[-(x - x_0)^2]/2\varepsilon^2$$

Using the standard form of the empirical energy function of a protein, $V_\varepsilon(x_0)$ cannot be analytically evaluated. However, it is possible to 'refit' the potential using exponential and Gaussian functions so that the integrals required in the smoothing transformation can be performed. The DEM protocol is as follows.

(i) Increase the smoothing length scale ε so that few (or one) minima survive on $V_\varepsilon(x_0)$.
(ii) Isolate the surviving energy minima.
(iii) Reduce the smoothing length scale $\varepsilon \leftarrow \varepsilon - \delta\varepsilon$.
(iv) Track each minimum to its new position on the less smoothed surface.
(v) If $\varepsilon > 0$, return to (iii).

This method has been applied to atomic clusters, water clusters, peptides, and proteins with some success.[27,41] While it does not represent a general solution to the multiple minima problem, the DEM is an important paradigm for protein conformational optimization.

2.2.2 *The Method of Bad Derivatives*

One of the most difficult problems related to the DEM is the calculation of the transformed potential energy surface. The smoothing transformation can only be performed approximately and the resulting smoothed potential function is quite complicated. Computation of the gradient of the potential can be time consuming. It has been demonstrated that, if the Gaussian smoothing function is replaced by a 'top hat' or impulse function, the gradient of the smoothed potential $V_\varepsilon(x_0)$ can be written

$$\nabla V_\varepsilon(x_0) = \frac{1}{2\varepsilon}[U(x_0 + \varepsilon) - U(x_0 - \varepsilon)] \tag{12}$$

This gradient of the smoothed potential is a simple function of the untransformed potential $U(x)$. The form of the gradient is that of a three-point finite difference approximation to the derivative of the smoothed potential. Therefore, the smoothing is carried out implicitly while no explicit transformation of the potential function is required. This formula is exact and there is no limitation on the size of ε. Therefore, the minimization method consists of following the protocol of the DEM using bad derivatives. Results for atomic clusters and small peptides indicate that the method is as effective as the DEM.[42] More importantly, it is directly applicable to a much larger class of functions including the Boltzmann distribution.

2.2.3 *Mean-field Gaussian Density Techniques*

A limitation of the standard simulated annealing methods is that the system is represented as a single configuration (a single point in configurational space). Moves are informed by the local gradient of the potential surface. When the global minimum is sought, it is desirable to make use of non-local information about the potential surface and higher order derivatives.

An approximate approach is to represent the system using a continuous Gaussian density distribution in conformation or phase (position and momentum) space. Equations of motion of the center and width of the density can be derived. The center of the packet responds to the gradient of the potential energy function averaged over the density distribution. When the density distribution is a Gaussian function, the potential takes the form of a smoothed potential where the smoothing function is a Gaussian. The width of the density distribution responds dynamically to the curvature of the smoothed potential surface.

A simulated annealing protocol can be followed using the Gaussian density dynamics. The initial temperature is set to a high value and the distribution spreads over the potential surface. This is equivalent to increasing the initial width of the smoothing length scale in the DEM. As the system dynamics evolve, the temperature is lowered according to a cooling schedule. This will cause the distribution, on average, to have a narrower width. Eventually the temperature approaches zero and the distribution will become localized in a low lying potential energy minimum.

This method has been applied in a number of forms where the system dynamics is defined by the Liouville equation[43] or the Bloch equation.[44] In the latter case, the dynamical simulated annealing (real time dynamics where the system temperature is adjusted according to a cooling schedule) is replaced by a thermodynamical annealing where the density distribution is evolved by direct integration in temperature. The results are significantly improved over those derived by the DEM. These methods demonstrate the intimate relationship between simulated annealing and potential smoothing paradigms.[45]

2.2.4 *Locally Enhanced Sampling (LES)*

Elber and co-workers have developed a powerful method for global energy optimization known as LES. The method is based on the representation of certain sets of atoms as a swarm of 'copies.' For example, in a peptide one possible LES system would consist of multiple copies of each side chain attached to a single copy of the backbone. Each side chain copy feels the full force of the backbone; the backbone feels the mean force of the swarm of side chain copies. While the method provides an approximate dynamics for the system, the global energy minimum in the LES approximation is the exact global energy minimum for the system (since the lowest energy configuration corresponds to having all copies sitting atop one another in the optimal configurations).

In the LES approximation, barriers on the potential energy surface are reduced in magnitude in a manner similar to the DEM and mean field Gaussian density methods discussed above. (In fact, if the swarm is constrained to a Gaussian distribution, and there are a large enough number of copies in the swarm, the methods will be identical.) One advantage of the LES method is that no explicit smoothing transformation of the potential energy function is required. Moreover, the distribution of the swarm may be strongly anisotropic in space. This may provide an advantage over a Gaussian smoothing method that employs a spherically symmetric smoothing kernel. The LES method has been applied to conformational optimization[46] and homology modeling[47] with good results.

2.2.5 *Quantum Mechanical Annealing (QMA)*

An important variation on the classical simulated annealing and potential smoothing methods is QMA.[27] The system is represented quantum mechanically using wave packets,[48] a variational wave function, or a diffusion MC approach[49] to solve the time-independent Schrödinger equation

$$-\frac{h^2}{2m}\nabla_x^2\varphi(x) + V(x)\varphi(x) = E\varphi(x) \tag{13}$$

The value of Planck's constant h (which controls the relative importance of the kinetic and potential energies) is initially increased to a large value. This is equivalent to raising the temperature in classical annealing. For large values of Planck's constant, the kinetic energy dominates the energy of the system which tends to delocalize over the potential surface. An estimate of the ground state wave function is then found. The value of Planck's constant is then reduced and a new estimate of the ground state wave function is found. This is equivalent to cooling the system in a simulated annealing procedure. Eventually, the value of Planck's constant is reduced to zero (or its physical value). If the representation of the wavefunction is complete, the system will isolate the non-degenerate global energy minimum of the potential surface. Otherwise, the wave function will be localized in a low lying minimum on the potential surface (which may be the global minimum).

This QMA method has been applied to atomic clusters[48,49] and model proteins[44] with some success. There remain many unexplored possibilities for the further development of the paradigm of QMA using a variety of other representations of the system.

3 CONCLUSIONS

There is at this time no general solution to the protein structure prediction problem. There are many heuristic methods that have been developed to 'predict' protein structures. Many of these methods are based on computational 'tricks' for identifying a folded state that lead one far from the physical problem of identifying a thermodynamically dominant state of the system. This article has focused on methods that stay close to the physical problem with the sense that it is from these general and rigorous methods that further progress is most likely to stem.

4 RELATED ARTICLES

AMBER: A Program for Simulation of Biological and Organic Molecules; CHARMM: The Energy Function and Its Parameterization; Conformational Sampling; Conformational Search: Proteins; Distance Geometry: Theory, Algorithms, and Chemical Applications; ECEPP: Empirical Conformational Energy Program for Peptides; Force Fields: A General Discussion; Force Fields: CFF; Free Energy Simulations; GROMOS Force Field; Macromolecular Structure

Calculation and Refinement by Simulated Annealing: Methods and Applications; Molecular Dynamics and Hybrid Monte Carlo in Systems with Multiple Time Scales and Long-range Forces: Reference System Propagator Algorithms; Molecular Dynamics: Techniques and Applications to Proteins; Molecular Surfaces and Solubility; Monte Carlo Simulations for Liquids; Monte Carlo Simulations for Polymers; Protein Force Fields; Protein Modeling; Protein Structure Prediction in 1D, 2D, and 3D; Simulated Annealing; Solvation: Modeling.

5 REFERENCES

1. G. Fasman (ed.), 'Prediction of Protein Structure and the Principles of Protein Conformation', Plenum, New York, 1989.
2. A. Fersht, 'Enzyme Structure and Mechanism', Freeman, New York, 1985.
3. J. A. McCammon and S. C. Harvey, 'Dynamics of Proteins and Nucleic Acids', Cambridge University Press, Cambridge, 1987.
4. C. L. Brooks, III, M. Karplus, and B. M. Pettitt, 'Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics', Wiley, New York, 1988.
5. (a) S. W. Rick, S. Stuart, and B. J. Berne, *J. Chem. Phys.*, 1994, **101**, 6141-6161; (b) S. W. Rick, S. Stuart, J. Bader, and B. J. Berne, *J. Mol. Liq.*, 1995, **65/66**, 31-40; (c) S. Rick and B. J. Berne, *J. Am. Chem. Soc.*, 1996, **118**, 672-679.
6. B. Honig and A. Nicholls, *Science*, 1995, **268**, 1144-1149.
7. R. E. Bruccoleri, J. Novotny, M. E. Davis, and K. A. Sharp, *J. Comput. Chem.*, 1997, **18**, 268-276.
8. D. Frenkel and D. Smit, 'Understanding Molecular Simulations', Oxford University Press, Oxford, 1996.
9. D. Beglov and B. Roux, *J. Chem. Phys.*, 1994, **100**, 9050-9063.
10. B. J. Berne and J. E. Straub, *Curr. Opin. Struct. Biol.*, 1997, **7**, 181-189.
11. R. A. Abagyan, in 'Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications', eds. W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, 1997, Vol. 3, ESCOM, Leiden.
12. S. Miyazawa and R. L. Jernigan, *Macromolecules*, 1985, **18**, 535-552.
13. J. Skolnick and A. Kolinski, *Science*, 1990, **250**, 1121-1125.
14. K. A. Dill, *Curr. Opin. Struct. Biol.*, 1993, **3**, 99-103.
15. P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Science*, 1995, **267**, 1619-1620.
16. K. Yue and K. A. Dill, *Protein Sci.*, 1996, **5**, 254-261.
17. D. Thirumalai and S. A. Woodson, *Acc. Chem. Res.*, 1996, **29**, 433-439.
18. R. A. Friesner and J. R. Gunn, *Annu. Rev. Biophys. Biomol. Struct.*, 1996, **25**, 315-342.
19. G. E. Schultz and R. H. Schirmer, 'Principles of Protein Structure', Springer, New York, 1979.
20. T. E. Creighton, 'Proteins: Structures and Molecular Properties', Freeman, New York, 1984.
21. G. M. Crippen and M. E. Snow, *Biopolymers*, 1990, **29**, 1479-1489.
22. J. D. Honeycutt and D. Thirumalai, *Biopolymers*, 1992, **32**, 695-709.
23. M. Karplus and E. I. Shakhnovich, in 'Protein Folding', ed. T. E. Creighton, Freeman, New York, 1992.
24. H. A. Scheraga, *Rev. Comput. Chem.*, 1992, **3**, 73-142.
25. M. P. Allen and D. J. Tildesley, 'Computer Simulation of Liquids', Oxford University Press, Oxford, 1991.
26. S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, *Phys. Lett. B*, 1987, **195**, 216-221.
27. J. E. Straub, in 'Recent Developments in Theoretical Studies of Proteins', ed. R. Elber, World Scientific, Singapore, 1996, pp. 137-196.
28. I. Andricioaei and J. E. Straub, *Phys. Rev. E*, 1996, **53**, R3055-R3058.
29. C. Tsallis, *J. Stat. Phys.*, 1988, **52**, 479-487.
30. D. D. Franz, D. L. Freeman, and J. D. Doll, *J. Chem. Phys.*, 1990, **93**, 2769-2784.
31. B. A. Berg and T. Neuhaus, *Phys. Lett. B*, 1991, **267**, 249-253.
32. U. H. E. Hannsmann, Y. Okamoto, and F. Eisenmenger, *Chem. Phys. Lett.*, 1996, **259**, 321-330.
33. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, 1994, **98**, 4940-4948.
34. (a) U. H. E. Hannsmann and Y. Okamoto, *Physica A*, 1994, **212**, 415-437; (b) Y. Okamoto and U. H. E. Hannsmann, *J. Phys. Chem.*, 1995, **99**, 11276-11287.
35. J. I. Siepmann and D. Frenkel, *Mol. Phys.*, 1992, **75**, 59-70.
36. M. Deem and J. Bader, *Mol. Phys.*, 1996, **87**, 1245-1260.
37. Z. Liu and B. J. Berne, *J. Chem. Phys.*, 1993, **99**, 6071-6077.
38. R. C. van Schaik, W. F. van Gunsteren, and H. J. C. Berendsen, *J. Comput.-Aided Mol. Design*, 1992, **6**, 97-112.
39. T. Head-Gordon and F. H. Stillinger, *Biopolymers*, 1993, **33**, 293-303.
40. J. E. Straub, A. Rashkin, and D. Thirumalai, *J. Am. Chem. Soc.*, 1994, **116**, 2049-2063.
41. (a) L. Piela, J. Kostrowicki, and H. A. Scheraga, *J. Phys. Chem.*, 1989, **93**, 3339-3346; (b) J. Kostrowicki and H. A. Scheraga, *J. Phys. Chem.*, 1992, **96**, 7442-7449; (c) J. Kostrowicki and H. A. Scheraga, in 'Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding', eds. P. M. Pardalos, D. Shalloway, and G. Xue, American Mathematical Society, Providence, RI, 1996, pp. 123-132.
42. I. Andricioaei and J. E. Straub, *Comput. Phys.*, 1996, **10**, 449-454.
43. J. Ma, D. Hsu, and J. E. Straub, *J. Chem. Phys.*, 1993, **99**, 4024-4035.
44. P. Amara, D. Hsu, and J. E. Straub, *J. Phys. Chem.*, 1993, **97**, 6715-6721.
45. (a) D. Shalloway, in 'Recent Advances in Global Optimization', eds. C. A. Floudas and P. M. Pardalos, Princeton University Press, Princeton, NJ, 1992, pp. 433-477; (b) B. W. Church, M. Oresic, and D. Shalloway, in 'Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding', eds. P. M. Pardalos, D. Shalloway, and G. Xue, American Mathematical Society, Providence, RI, 1996, pp. 41-64.
46. A. Roitberg and R. Elber, *J. Chem. Phys.*, 1991, **95**, 9277-9287.
47. C. Keasar and R. Elber, *J. Phys. Chem.*, 1995, **99**, 11550-11556.
48. P. Amara and J. E. Straub, *J. Phys. Chem.*, 1995, **99**, 14840-14852.
49. A. B. Finnila, M. A. Gomez, C. Sebenik, C. Stenson, and J. D. Doll, *Chem. Phys. Lett.*, 1994, **219**, 343-348.