Protein-Folding Dynamics: Overview of Molecular Simulation Techniques

Harold A. Scheraga, Mey Khalili, and Adam Liwo

2007, 58 (57-83 • DOI: 10.1146/annurev.physchem.58.032806.104614

More About This Article

Protein-Folding Dynamics: Overview of Molecular Simulation Techniques

Harold A. Scheraga, Mey Khalili, and Adam Liwo

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301; email: has5@cornell.edu

Annu. Rev. Phys. Chem. 2007. 58:57–83

First published online as a Review in Advance on October 11, 2006

The *Annual Review of Physical Chemistry* is online at http://physchem.annualreviews.org

This article's doi: 10.1146/annurev.physchem.58.032806.104614

0066-426X/07/0505-0057$20.00

Key Words

molecular dynamics, folding pathways, atomic and mesoscopic models, force fields

Abstract

Molecular dynamics (MD) is an invaluable tool with which to study protein folding *in silico*. Although just a few years ago the dynamic behavior of a protein molecule could be simulated only in the neighborhood of the experimental conformation (or protein unfolding could be simulated at high temperature), the advent of distributed computing, new techniques such as replica-exchange MD, new approaches (based on, e.g., the stochastic difference equation), and physics-based reduced models of proteins now make it possible to study protein-folding pathways from completely unfolded structures. In this review, we present algorithms for MD and their extensions and applications to protein-folding studies, using all-atom models with explicit and implicit solvent as well as reduced models of polypeptide chains.

1. INTRODUCTION

NMR: nuclear magnetic resonance

MD: molecular dynamics

Interest in the dynamics of proteins derives from its application to many properties of proteins, such as folding and unfolding, the role of dynamics in biological function, the refinement of X-ray and nuclear magnetic resonance (NMR) structures, and protein-protein and protein-ligand interactions. This review is concerned with protein-folding and -unfolding dynamics, and deals with molecular simulation techniques with emphasis on molecular dynamics (MD) and its extensions. The advantage of this approach is that, within the accuracy of the underlying potential energy functions, it provides information about the folding and unfolding pathways, the final folded (native) structure, the time dependence of these events, and the inter-residue interactions that underlie these processes. We refer the reader to the literature for experimental techniques such as fluorescence (1, 2), infrared (3, 4), and NMR (5–8) spectroscopy, as well as electron-transfer experiments (9, 10) for the study of protein-folding dynamics.

In the theoretical approach, based on empirical potential energy functions, Newton's or Lagrange's equations are solved to obtain coordinates and momenta of the particles along the folding and unfolding trajectories. Alternative approaches are based on solving Langevin's equations when the solvent is not treated explicitly. Both approaches are time-consuming and require extensive computer power to solve these equations. In fact, it is only the development of such computing power that has made it possible to solve physical problems by MD calculations.

The modern era of MD calculations with electronic computers began with the work of Alder & Wainwright (11, 12), who calculated the nonequilibrium and equilibrium properties of a collection of several hundred hard-sphere particles. By providing an exact solution (to the number of significant figures carried) of the simultaneous classical equations of motion, they were able to obtain the equation of state (pressure and volume) and the Maxwell-Boltzmann velocity distribution. Rahman (13) carried out the first MD simulations of a real system when he studied the dynamics of liquid argon at 94.4 K.

Later, Rahman & Stillinger (14) applied the MD technique to explore the physical properties of liquid water. Treating the water molecule as a rigid asymmetric rotor with an effective Ben-Naim and Stillinger pair potential version of the Hamiltonian, they computed the structural properties and kinetic behavior, demonstrating that the liquid water structure consists of a highly strained random hydrogen-bond network, with the diffusion process proceeding continuously by the cooperative interaction of neighbors.

Karplus and coworkers (15) carried out the first application of MD to proteins. However, this study did not deal with the protein-folding problem. Instead, it investigated the dynamics of the folded globular protein bovine pancreatic trypsin inhibitor. As in the work of Rahman and Stillinger, Karplus and coworkers (15) solved the classical equations of motion for all the atoms of the protein simultaneously with an empirical potential energy function, starting with the X-ray structure and with initial velocities set equal to zero. Their results provided the magnitude, correlations, and decay of fluctuations about the average structure, and suggested that the protein

interior is fluid-like in that the local atomic motions have a diffusional character. Researchers have applied this technique extensively in the refinement of X-ray and NMR structures, but because of the need to take small (femtosecond) time steps along the evolving trajectory to keep the numerical algorithm stable, it has not been successful in treating the real long-time folding of a globular protein, except for very small ones. However, many of the applications of MD of globular proteins have been made to the initial unfolding steps, followed by refolding. In applying the MD technique, one must consider numerous trajectories, rather than a single one, to cover the large multidimensional, conformational potential energy space and obtain proper statistical mechanical averages of the folding/unfolding properties.

MC: Monte Carlo
UNRES: united-residue

Since the first papers from the Karplus lab, numerous MD calculations have been carried out in the laboratories of Brooks (16), van Gunsteren (17), Levitt (18), Jorgensen (19), Daggett (20, 21), Kollman (22), Pande (23), Berendsen (24), Baker (25), McCammon (26), and others. This review is concerned with recent theoretical developments in protein-folding dynamics. Several reviews (20, 27, 28) have discussed earlier work in this field. The discussion here focuses on techniques to solve the classical equations of motion, using both an all-atom approach and an approach based on simplified models of the polypeptide chain, and to assess how far the field has progressed to be able to compute complete folding trajectories, as well as the final, native structure, based on physical principles (i.e., the interatomic potential energy and Newtonian mechanics).

2. SIMULATION TECHNIQUES

In this section, we briefly describe the equations of motion used in classical MD and the algorithms for integrating these equations. More extensive information can be found in recent review articles (26) and textbooks (29). For techniques based on Monte Carlo (MC) methods that are also used to study protein folding, we refer the reader to other review articles (30, and references therein; 31).

2.1. Equations of Motion in Molecular Dynamics

For a system of molecules treated at the classical level, Newton's equations of motion are applied, each atom treated as a point with a mass m_i (Equation 1):

$$m_i\ddot{\mathbf{r}}_i = \mathbf{F}_i \qquad i = 1, 2, \ldots, N \qquad \ddot{\mathbf{r}}_i = \frac{d^2\mathbf{r}_i}{dt^2}, \tag{1}$$

where $\mathbf{r}_i = (x_i, y_i, z_i)$ is the vector of Cartesian coordinates of the i-th atom, $\ddot{\mathbf{r}}_i$ is the corresponding acceleration, $\mathbf{F}_i$ is the vector of forces acting on the i-th atom, and N is the number of atoms. If the objects to move are more complex than atoms (e.g., α-helical segments in rigid-body dynamics or the interaction sites in coarse-grained protein models) or when the dihedral angles or other curvilinear coordinates are used, generalized Lagrange equations of motions are recommended (29) instead; we used such an approach to derive the working equations of motion of our coarse-grained united-residue (UNRES) model of polypeptide chains (32).

When the system (including the solvent) is treated at the fully atomic level, the forces are only potential forces. Otherwise, the collisions and friction forces should be introduced to mimic the collisions of the solute molecule with its environment. Such a treatment is referred to as Langevin or Brownian dynamics, and Langevin (33) presented its theoretical foundation in a paper published in 1908 on the motion of Brownian particles in a fluid. Equation 1 then becomes a stochastic differential equation with the forces on its right side expressed by Equation 2:

$$\mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r}_N) - m_i \gamma_i \dot{\mathbf{r}}_i + \mathbf{R}_i(t) \qquad i = 1, 2, \ldots, N, \tag{2}$$

where $\dot{\mathbf{r}}_i$ and γ_i are the velocity and the friction coefficient of atom i, respectively; U is the potential energy of the system ($-\nabla_{\mathbf{r}_i} U$ being the potential force acting on atom i); and $\mathbf{R}_i(t)$ is the vector of random forces arising from the collision of atom i with the molecules of the environment (solvent) that are not considered explicitly. $\mathbf{R}_i(t)$ has zero mean (Equation 3), and the $\mathbf{R}_i$s taken at different times are δ-correlated (Equation 4):

$$\langle \mathbf{R}_i(t) \rangle = 0, \tag{3}$$

$$\langle \mathbf{R}_i(t) \cdot \mathbf{R}_i(t') \rangle = 2 m_i \gamma_i k_B T \delta(t - t'), \tag{4}$$

where T is the absolute temperature of the system, k_B is the Boltzmann constant, and $\delta(x)$ is the Dirac δ-function.

In the overdamped limit $\gamma \gg 2\omega$ (where ω is the characteristic frequency of the system), the dissipative terms ($-m_i \gamma_i \dot{\mathbf{r}}_i$) prevail over the inertial terms ($m_i \ddot{\mathbf{r}}_i$), and the latter can be neglected in Equation 1 (with forces expressed by Equation 2), resulting in the system of first-order differential equations:

$$m_i \gamma_i \dot{\mathbf{r}}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r}_N) + \mathbf{R}_i(t), \qquad i = 1, 2, \ldots, N. \tag{5}$$

Brownian dynamics is usually the method of choice for coarse-grained models of proteins. Although it is frequently used, it does not provide control over the kinetic energy because of the neglect of the inertial term, and, therefore, it cannot lead to correct values of thermodynamic properties. Complete Langevin dynamics avoids this problem.

2.2. Integrating the Equations of Motion

A system of equations given by Equation 1, together with initial coordinates and velocities, constitutes an initial-value problem. Consequently, one can use a variety of algorithms for numerical solution of the initial-value problem to integrate the equations of motion; the predictor-corrector Gear method (34) is often applied as a general-purpose algorithm in this field. An undesirable feature of the general-purpose algorithms is that they usually require high-order time derivatives to work with good accuracy. Because of the demand for low computational cost and high accuracy, a variety of specific integrators have been designed for MD algorithms of which the Verlet-type algorithms (the Verlet, velocity-Verlet, and the leap-frog algorithm) are the most common (29); all three of these algorithms are mathematically equivalent. Their most important property is the conservation of a slightly perturbed original

Hamiltonian (the shadow Hamiltonian); in other words, when the nonconservative forces are not present, the total energy oscillates about a value close to the initial energy and does not drift from the initial value, the magnitude of the oscillations increasing with increasing time step Δt.

It has been shown recently (35, 36) that this property is a consequence of the fact that all three algorithms can be derived from the Liouville formulation of the equations of motion, by splitting the Liouville propagator into parts corresponding to the momenta and coordinates. In addition, the Liouville formulation facilitated the extension of Verlet-type algorithms to Langevin dynamics (36–39). The conservation of the shadow Hamiltonian and, consequently, the avoidance of energy drift are important in the correct reproduction of thermodynamic averages (29). Therefore, although the Verlet-type algorithms are only fourth-order algorithms (i.e., the error scales as the fourth power of the integration time step, Δt), for long trajectories they are better than higher-order nonsymplectic algorithms that involve smaller errors for short trajectories but induce energy drift in the long run (29, 40). We present the velocity-Verlet algorithm below for illustration.

SHAKE: an algorithm for constrained molecular dynamics

RATTLE: a velocity version of the SHAKE algorithm

LINCS: linear constraint solver

Step 1 (updating coordinates):

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\Delta t + \ddot{\mathbf{r}}(t)\Delta t^2/2. \quad (6)$$

Step 2 (updating velocities):

$$\dot{\mathbf{r}}(t + \Delta t) = \dot{\mathbf{r}}(t) + [\ddot{\mathbf{r}}(t) + \ddot{\mathbf{r}}(t + \Delta t)]\,\Delta t/2. \quad (7)$$

For the integration algorithm to be stable, the value of the time step Δt must be an order of magnitude smaller than the fastest motions of the system. Typically, this motion is the vibration of a bond that involves a hydrogen atom with a period of the order of 10 fs, and consequently the time step is of the order of 1 fs when explicit solvent is used. When implicit solvent is used, the time step can be larger, from 2 to 5 fs (41). This is much less than the timescale of the fastest biochemically important motions such as helix formation, which takes a fraction of a microsecond, or folding of the fastest α-helical proteins, which takes several microseconds (27, 42). In one option, known as the variable step method (32, 43), the time step is reduced when hot events result in occasional significant variation of forces, but this violates time reversibility and energy conservation.

The correct procedure is to use the time-split algorithms (35, 36), which are an extension of the basic Verlet-type algorithms. In these algorithms, the forces are divided into fast-varying ones that are local (as, e.g., the bond-stretching forces) and, consequently, inexpensive to evaluate and slow-varying forces that are nonlocal forces and expensive to evaluate. Integration is carried out with a large time step for the slow forces and an integer fraction of the large time step for the fast-varying forces. Such a procedure enables the use of up to a large 20-fs time step at only a moderate increase of the computational cost (37, 40). One can achieve further effective increase of the timescale by constraining the valence geometry of the solvent molecules [the SHAKE (44), RATTLE (45), and LINCS (46) algorithms] and, yet further, by using torsional-angle dynamics (43, 47) and rigid-body dynamics (29) in which elements of structure (e.g., α-helical segments) are considered fixed. The use of simplified protein models

enables one to increase the timescale further because of averaging out fast motions that are not present at the coarse-grained level (37, 40).

NVE: constant number of particles, volume, and energy ensemble

NVT: constant number of particles, volume, and temperature ensemble

NPT: constant number of particles, pressure, and temperature ensemble

CHARMM: Chemistry at Harvard Molecular Mechanics

AMBER: assisted model building with energy refinement

GROMOS: Groningen molecular simulation

CVFF: consistent valence force field

2.3. Relationship Between the Solutions of the Equations of Motion and Ensembles

Equation 1, with conservative forces on the right-hand side, describes the motion of an isolated system, and consequently the ensembles of conformations resulting from its solutions are the microcanonical (NVE) ensembles. In reality, protein folding occurs in systems coupled to a temperature bath, and consequently we want the solution of the equations of motion to give the canonical (NVT) or isothermal-isobaric (NPT) ensembles. This remark pertains to all simulations regardless of whether the solvent is considered explicitly or implicitly. This can be accomplished by rescaling velocities to adjust the kinetic energy of the system to the required temperature, a procedure known as Berendsen's thermostat (48). However, such a treatment does not generate true canonical ensembles (49). In more sophisticated methods, known as the Nosé-Hoover (50, 51) and Nosé-Poincaré (51) thermostats, the Hamiltonian of the system is modified to correspond to temperature, and not total-energy, conservation. These methods generate canonical ensembles as demonstrated by Nosé (50, 52), and symplectic algorithms can be designed to solve the equations of motion.

If included (Equation 2), the friction and stochastic forces provide a thermostat by themselves. The temperature of the system simulated is maintained because of the appearance of the friction coefficient γ in the expression for both frictional and random forces. The frictional forces result in energy dissipation because of collisions opposing the motion, whereas the random forces result in energy gain; consequently, the solution of the system given by Equation 1 with friction and random forces included generates a canonical ensemble of conformations. Inclusion of these forces instead of the Berendsen or other thermostats may be preferable, even when explicit solvent is considered, because it results in more uniform distribution of temperature between the solute and the solvent (53).

3. ALL-ATOM APPROACH

3.1. Force Fields

Traditionally, the potential forces in Equation 2 are calculated using empirical all-atom potential functions, such as CHARMM (Chemistry at Harvard Molecular Mechanics) (54), AMBER (assisted model building with energy refinement) (41), GROMOS (Groningen molecular simulation) (55), and CVFF (consistent valence force field) (56), which include the solvent either explicitly or as a continuum (implicit-solvent treatment). We discuss the treatment of solvent in more detail in Sections 3.2 and 3.3. For a description of the all-atom force fields, we refer the reader to recent reviews (57, 58).

The functional forms of the force fields are a trade-off between accuracy in representing forces acting on atoms and low computational cost or ease of parameterization.

Thus, one usually considers only interactions between point charges in the computation of the electrostatic-interaction energy, thereby neglecting higher moments of electron-charge density and polarization effects. Harmonic functions for bond stretching and bond-angle bending are used instead of more refined ones with anharmonicity included. However, this approximation can result in unreasonably large distortions of the bond angles (59). The inherent inaccuracies also result in an incompatibility of properties obtained by using different force fields (60–62). Although the per-residue errors inherent in force-field inaccuracies amount to a fraction of a kilocalorie per mole, they translate into tens of kilocalories per mole for the entire protein. Therefore, optimization of the whole force field, as for simplified models (Section 5), needs to be performed for all-atom force fields; Fain & Levitt (63) and Schug & Wenzel (64) have already initiated this work.

TIP3P, TIP4P, and TIP5P: three-, four-, and five-point models of the water molecule

3.2. Simulations with Explicit Solvent

Explicit inclusion of water molecules provides, as realistically as possible, the kinetic and thermodynamic properties of the protein-folding process. Simulations with explicit water are carried out in a periodic box scheme; the box is usually rectangular, but other shapes are also possible (26). A less common treatment is to perform simulations in a thin layer of water around a protein molecule restrained with a weak harmonic potential (26).

There are currently a number of water models used in MD simulations. These include the ST2 model of Stillinger (65), the SPC model of Berendsen et al. (66), and Jorgensen's TIP3P, TIP4P, and TIP5P models (67). These models were parameterized assuming that a cut-off is applied to nonbonded interactions, but they are often used with Ewald summation to treat long-range electrostatics. Horn et al. (68) recently developed an extension of the TIP4P model to be used with Ewald summation termed TIP4P-Ew. All these models treat water as a rigid molecule. Although bond stretching and bond-angle bending (69), or polarization effects and many-body interactions (70), have been introduced into water models, they involve a large increase of computational expense, which has limited their use as widely as the SPC or TIP models. The water models are usually parameterized at a single temperature (∼298 K) and therefore do not correctly capture the temperature dependence of properties such as the solvent density or diffusion coefficients (68).

The presence of water molecules in the system dramatically increases the number of degrees of freedom (typically by more than 1000 degrees of freedom). Because of this limitation, along with the small values of the time step in integrating the equations of motion (of the order of femtoseconds), explicit-solvent all-atom MD algorithms can simulate events in the range of 10^{-9}s to 10^{-8}s for typical proteins and 10^{-6}s for very small proteins (20, 42). These timescales are at least one order of magnitude smaller than the folding times of proteins (10). The most impressive and the longest explicit-solvent ab initio canonical MD simulation starting from unfolded conformations is one by Duan & Kollman (22) on the villin headpiece. They observed conformations with significant resemblance to the native state in a 1-μs run. However, their simulation fell significantly short of the folding time for this protein, which is

GBSA: generalized Born surface area

~5 μs. Therefore, at present, explicit solvent MD by itself is not capable of simulating the folding pathways of proteins in real time, except for very small proteins. However, it has been combined successfully with other search methods in some interesting and ingenious algorithms to study energy landscapes and folding pathways. We discuss some of these algorithms in Section 6.

3.3. Implicit-Solvent Methods

The use of continuum representations of the solvent greatly decreases the number of degrees of freedom in the system and, consequently, the sampling time. The rigorous implicit treatment of solvent in MD involves (*a*) designing an effective potential function that describes the change of the free energy of the system on the change of the conformation of the solute molecule and (*b*) direct effect of the solvent on the dynamics of the solute molecule through collisions, which results in the appearance of net friction and random forces. We discuss point *a* in this section, whereas point *b* is discussed in Section 2.1.

The most common treatment of electrostatic interactions between the solute and solvent makes use of the generalized Born surface area (GBSA) model (71), which includes an approximation to the solution of the Poisson-Boltzmann equation for a system comprising the solute molecule immersed in a dielectric with counter-ions and also takes into account the loss of free energy owing to the formation of a cavity in the solvent; for details, we refer the reader to References 72 and 73. Simpler models have also been developed in which the free energy of solvation is expressed in terms of solvent-accessible surface areas of solute atoms (74) or solvent-excluded volumes owing to the contributions from pairs of atoms (75, 76), but they are used in other applications than MD. The GBSA model can lead to discontinuous forces because of its explicit use of molecular surface area (77); an algorithm that overcomes this shortcoming by introducing a smoothing function has been designed recently (78). Use of the GBSA model eliminates the need for the lengthy equilibration of water necessary in explicit water simulations. However, this model does not reproduce the all-atom free-energy landscape of folding and can overestimate the stability of the native state (79).

With the addition of the Berendsen thermostat or other thermostats discussed in Section 2.3, canonical simulations can be carried out with implicit-solvent models; such a treatment corresponds to a low-viscosity limit and has been applied with success to all-atom ab initio folding simulations of proteins by canonical MD. Jang and colleagues (80) were able to fold protein A, a 46-residue protein with a three-helix-bundle fold, and the villin head piece from the extended state with all-atom MD and the generalized Born model of solvation. However, ignoring solute-solvent friction makes the folding times, calculated with implicit-solvent MD simulations, the lower bounds of the true experimental folding times of proteins (81). The folding rates depend strongly on solvent viscosity (82), and the absence of viscosity in simulations leads to a fast collapse to a nonnative globule, with folding proceeding from this globule (83).

Even when friction and stochastic forces are included, the continuum models do not account for differences in protein-folding dynamics such as cooperative expulsion

of water on folding (84). Structured water plays a role in the folded state of many proteins (85). The dewetting effect around the hydrophobic group is a liquid-vapor-phase equilibrium process, which is not described accurately with the implicit-solvent models (83).

ACM: amplified collective motion

3.4. Constant pH Simulations

The most accurate treatment of solvent effects should include the proton exchange between protein ionizable groups and solvent. A number of models have been proposed for performing MD at constant pH with dynamic protonation states with explicit treatment of the solvent (86, 87). Each of these uses MC sampling to select protonation states based on calculated energy differences between the possible protonation states. Recently researchers have extended this method to include a generalized Born implicit solvation model (88, 89); this eliminates the necessity of solvent equilibration for a new protonation state. However, the typical error in the predicted pK_a values of ionizable groups is from 0.5 to 2 logarithmic units (86–89). Recently, we used this approach to explore the conformational space of an 11-residue alanine-based peptide containing basic amino acids (90). The theoretical titration curves determined in dimethylsulfoxide and in methanol were in good agreement with the experimental ones, whereas we observed more significant discrepancies for water, which we attributed to a more significant contribution of specific hydration.

4. COLLECTIVE COORDINATE ALGORITHMS

Functionally relevant motions of proteins occur along the direction of a few collective coordinates, which dominantly contribute to the atomic fluctuations (91). The use of collective coordinates results in the extraction of functionally relevant motions from the simulation results. Collective coordinates are extracted by methods such as essential dynamics (91) from a large number of MD or MC trajectories. Among the methods that use collective coordinates to increase the sampling efficiency of MD simulations are conformational flooding (92), essential dynamics sampling (93), and the amplified collective motion (ACM) method (94).

In the conformational flooding method (92), one uses collective coordinates to extract fast-moving degrees of freedom from MD trajectories produced prior to production simulations. Deep minima in the subspace spanned by the fast degrees of freedom are filled with a Gaussian biasing potential; this operation decreases the energy barriers between different minima and, consequently, produces conformational transitions beyond the time domain reached by conventional MD. Schulze et al. (92) have applied conformational flooding in a series of room-temperature simulations to accelerate molecular motions of the native fold of carbonmonoxy myoglobin and define the lower-tier hierarchy of substate structure. The computed conformational space and associated transitions coincide with previously suggested putative ligand-escape pathways and support a hierarchical description of protein dynamics and structure.

In the essential dynamics sampling method (93), the protein motions are constrained to move along the essential collective modes. The disadvantage of this

method is that the collective modes must be calculated from a predetermined set of protein conformations that have been simulated for a short period of time and represent only the local motions of the protein. Because collective modes are treated as a linear combination of Cartesian coordinates, they are conformation dependent, and the essential subspace is different when the protein conformation belongs to different local states. An example of the application of this method is the study of thermal unfolding of horse heart cytochrome *c* (95)

ACM uses the collective modes obtained by the coarse-grained/anisotropic network model (96), instead of carrying out initial MD simulations, to guide the atomic-level MD simulations. This method overcomes the limitation of the essential dynamics sampling method, which is the invariance of essential subspace in the conformational space. The motions along the collective modes are amplified by coupling them to a thermal bath at a higher temperature than the rest of the motions. Although different motions are coupled to different temperatures, the temperature averaged over all degrees of freedom of the system is usually approximately 300 K (94). ACM drives the system to escape from unfolded local minima and, as opposed to high-temperature unfolding simulations, expands the sampling region selectively without extending the accessible conformational space to high-energy regions. Therefore, it expands the accessible conformational space while still restricting the sampling within the lower-energy region of the conformational space. Zhang and colleagues (94) applied ACM to two test systems: the ribonuclease A S-peptide analog, in which they observed refolding of the denatured peptide in eight simulations out of ten, and bacteriophage T4 lysozyme, in which they observed extensive domain motions between the N and C termini.

5. SIMPLIFIED MODELS

Reduced (mesoscopic, coarse-grained) models of proteins, in which each amino-acid residue is represented by only a few interaction sites, in principle offer an extension of the timescale of simulations compared with that of all-atom models (30, 97). The development of these models started with the pioneering work of Levitt (98), who derived a coarse-grained potential for a polypeptide chain by averaging the all-atom energy surface. Reduced models can be divided as follows: general models of protein-like polymers, knowledge-based models, physics-based models, and models biased toward the native structure or elements of the native structure (99). Recently, Shih et al. (100) have extended the coarse-grained treatment to protein-lipid systems. There is no explicit solvent in the simplified models and, consequently, researchers use Langevin or Brownian dynamics in simulations to mimic the nonconservative forces from the solvent.

The general models usually assume a single interaction site per residue (31, 101, 102), and the potentials used are Lennard-Jones-type or simpler contact potentials reflecting hydrophobicities of interacting residues. These potentials are not meant to reproduce the detailed features of the protein energy surface, but rather to study general properties of folding. Nevertheless, simulations with general reduced models have contributed significantly to our understanding of the events that occur in the

folding process and of foldability criteria (31, 102). These models are most commonly applied in lattice simulations (102).

In the Gō-like models (103), the potentials for those pairs of residues in contact in the native structure are attractive and those between other pairs repulsive (38, 104). Such potentials provide minimum frustration but are specific for a given sequence and a given native structure (105). Use of these models is based on the assumption that the native-state topology largely determines folding, whereas nonnative interactions are of secondary importance. Examples of application are studies of the kinetics and sequence of folding events (38, 104), and the thermal (106) and mechanical (107) unfolding of proteins. The model developed by Sorenson & Head-Gordon (108) contains a bias toward native secondary structure by the choice of the potentials for rotation about the $C^{\alpha}\cdots C^{\alpha}$ virtual bonds. Interactions between side chains depend on their hydrophobicities, as opposed to a Gō-like model. Researchers have used this model to study the folding kinetics of ubiquitin-like sequences (108), as well as protein L and G (99) by Langevin dynamics. He & Scheraga (109) used a simpler model with secondary-structure bias to study the folding of model β-sheet sequences.

PDB: Protein Data Bank

The knowledge-based potentials are derived from protein structural databases (30, 110, 111), both in their form and parameterization. A large part of these is based on the Boltzmann principle (110) to derive the components (30, 110, 112) from the relevant distribution and correlation functions calculated from the Protein Data Bank (PDB) (113). The other approach is the derivation of the potentials for threading; the principle is to locate the native-like structures as the lowest in energy (111) among a large number of decoys derived from the PDB. The potentials based on the Boltzmann principle also require calibration on known protein structures to obtain an appropriate balance of different energy terms. The knowledge-based potentials are used mostly for fold recognition and other knowledge-based methods of protein-structure prediction. Nevertheless, Kolinski and colleagues (114, 115) used the potentials derived by Kolinski & Skolnick (30) in protein-folding simulations using lattice MC dynamics and unfolding simulations of proteins.

We define the final category, the physics-based reduced models, as those that have connection to physics in the derivation and the functional form, although they are usually parameterized using information from the PDB or structures of selected proteins. The earliest model was the one developed by Levitt (98); however, it was not used in actual protein-folding simulations. The model developed by Wolynes' group (116, and references therein) is a partially physics-based one. It assumes a detailed description of the backbone and united side chains and can, therefore, be termed semimesoscopic. It contains a knowledge-based part known as an associative-memory Hamiltonian, which is based on the set of correlations between the sequence of the protein under study and the set of sequence-structure patterns in a set of memory proteins. The parameters of the force field have been determined by potential-function optimization, which is based on energy-landscape theory (116), according to which the ratio of the folding (T_f) to the glass-transition (T_g) temperature is maximized. This ratio is approximated by the ratio known as the Z-score, $Z = \delta E_s/\Delta E$, where δE_s is the difference between the energy of the native and nonnative states (the stability gap), and ΔE is the standard deviation of the energy of the nonnative states. A

similar model is that developed by Takada and colleagues (117), which does not have any knowledge-based part of the Hamiltonian. This model was applied in studying folding mechanisms and in protein-structure prediction (117).

The physics-based UNRES model developed in our laboratory (118–120) assumes two centers of interaction per residue: the united peptide group located in the middle between two consecutive C^{α} atoms and the united side chain attached to the C^{α} atom by a virtual bond. As opposed to the simplified models described above, which use arbitrary expressions for energy or analogs from all-atom force fields, the UNRES effective energy function is derived as a restricted free energy of the virtual chain (118); the degrees of freedom not present in the model (the solvent degrees of freedom, the internal degrees of freedom of the side-chain angles, and angles of rotation of the peptide groups) have been integrated out because they are assumed to vary much faster than the coarse-grain degrees of freedom. The force field was optimized using a method based on a hierarchy of folding developed in our laboratory (119). When implemented in MD, UNRES was able to fold a 75-residue protein in 4 h on average with a single processor (120). Comparison with all-atom MD revealed that UNRES offers a 4000-fold speed-up relative to all-atom simulations with explicit solvent and more than a 200-fold speed-up compared with simulations with implicit solvent (37). Consequently, real-time folding simulations are readily possible.

6. SOME ASPECTS AND EXTENSIONS OF MOLECULAR DYNAMICS

6.1. The Choice of Reaction Coordinate

A reaction coordinate is an abstract one-dimensional coordinate that represents progress along a reaction pathway. For more complex reactions (e.g., protein folding), this choice can be difficult. Free energy is often plotted against a reaction coordinate to illustrate the energy landscape or potential energy surface associated schematically with the reaction.

A good reaction coordinate is one that can distinguish between the unfolded and folded and the unfolded/near-folded ensemble successfully. Projecting the trajectories onto one or several reaction coordinates, such as the fraction of native contacts (Q), can produce a landscape that shows a clear difference between the native and the unfolded states. But in general, the folding transitions cannot be projected onto two dimensions without overlap of kinetically distinct conformations. We recently observed either a two-state or three-state folding behavior for protein A (121), depending on the choice of the reaction coordinate (Q or C^{α} RMSD). However, researchers have achieved accurate projections of simulations onto appropriate reaction coordinates, which agreed with the experiment. For example, Onuchic and colleagues (122) used the Gō model for reversible folding of CI2, Src SH3, barnase, RNase H, and CHe Y, and their results matched experiment.

Radhakrishnan & Schlick (123) developed the transition-path sampling method for all-atom MD in which a number of MD trajectories are focused near the

conformational-transition path, and they applied it to map out the entire closing conformational profile of RNA polymerase. They found that there is a sequence of conformational checkpoints involving subtle protein-residue motion that may regulate fidelity of the polymerase repair or replication process.

6.2. Multiple Trajectories

In MD, one usually generates a statistical ensemble [e.g., a canonical ensemble (NVT)], and the quantity of interest is an ensemble average (e.g., the average structure of the native basin). To obtain a good ensemble average, many trajectories have to be simulated so that the statistical errors owing to insufficient sampling are minimized. However, owing to the computational expense of MD simulations, this is not possible using direct all-atom simulations with explicit solvent and difficult using implicit solvent. An example of a multitrajectory all-atom study with implicit solvent is the work by Duan and coworkers (124) on the folding of the Trp-cage (a 20-residue peptide). They generated 14 trajectories and derived the kinetic equations for folding. They found two folding routes: a fast one that proceeded directly to the native state and a slow one that proceeded through a misfolded intermediate. The calculated folding time was in good agreement with the experimental value.

More trajectories can be run in a given amount of time, and consequently more reliable folding statistics can be collected when using simplified models of polypeptide chains (Section 5). For example, using their coarse-grained potential biased toward native secondary structure, Brown & Head-Gordon (99) have calculated the folding pathways, the folding temperature, thermodynamic characteristics of folding, kinetic rate, denatured-state ensemble, and transition-state ensemble of protein L by a reduced representation of proteins and Langevin dynamics simulations of 1000 trajectories, and we recently determined the folding kinetics for protein A by running 400 Langevin dynamics trajectories of this protein from the extended state, with our united-residue potential (UNRES) (121).

Pande and coworkers (125) designed a method based on simulating multiple trajectories at the all-atom level that enables one not only to study folding pathways but also to estimate rate constants. The method is based on the observation that, with the assumption that crossing of a single barrier obeys a single-exponential kinetics, the probability for a system to cross the free-energy barrier for the first time is increased M times if M parallel trajectories are simulated. With $M \approx 10{,}000$, the chance to observe a folding event is amplified to a real timescale of simulations. Pande and coworkers use worldwide distributed computers to carry out parallel simulations; the task is ideal for such a computation scheme because no synchronization is required. The barrier crossing on a given trajectory is detected by monitoring the change of heat capacity computed from the energy variance along the trajectory; the appearance of a maximum signals barrier crossing. Because folding events occur only on a small fraction of trajectories in real simulation time (i.e., $t \leq 1/k$, where k is the first-order rate constant), one can assume that $f(t) \approx kt$, where $f(t)$ is the fraction of folded conformations over all trajectories. Consequently, the rate constant corresponding to a

given event can be estimated from Equation 8:

$$k = \frac{N_{folded}}{t \cdot N_{total}} \pm \frac{\sqrt{N_{folded}}}{t \cdot N_{total}}, \quad (8)$$

where N_{folded} and N_{total} are the number of folded structures and the total number of structures collected in all trajectories, respectively, and $\pm$ denotes the error estimate (125), which was estimated based on the fact that, consistent with the first-order kinetics, N_{folded} obeys the Poisson distribution.

REMD: replica-exchange molecular dynamics

The above procedure is sufficient for two-state folders with single-barrier crossing. For multistate folders, the final configurations from the trajectory for which the crossing of the first barrier has occurred are distributed over all processors, and the calculation is restarted (125). Then, the same procedure is used to estimate the rate constant corresponding to crossing the next barrier. However, the method is not as successful as for small two-state folders because the time for initial equilibration and for diffusion across the barrier scales with chain length (126). Also, the fraction of folded conformations does not obey the Poisson distribution for too short simulation time (127), and the simulations are also quite sensitive to the choice of the unfolded ensemble (128).

6.3. Replica-Exchange Molecular Dynamics

In this method, pioneered for proteins by Sugita & Okamoto (129), a number of MC or MD simulations are started at different temperatures. After a certain time, the temperatures are exchanged between trajectories, a decision made by using a Metropolis criterion. This assures that the sampling follows a Boltzmann distribution at each temperature. The kinetic trapping at lower temperatures is avoided by exchanging conformations with higher-temperature replicas. Each pair of replicas must have an overlapping energy distribution (129); this means that the number of replicas scales as $N^{3/2}$, where N is the chain length (129). If care is taken to reach equilibrium (130), replica-exchange molecular dynamics (REMD) is a powerful tool for sampling the folding landscape and is easily parallelizable. For these reasons, REMD has wide applicability and has been used with implicit-solvent MD simulations (131) and simplified models (132) to study protein folding. However, it does not provide direct information about kinetics, as opposed to regular multiple-trajectory simulations (133).

Among the well-known explicit solvent REMD simulations is that of Berne and colleagues (134), who applied it to obtain the free-energy landscape of a β-hairpin, and that of Garcia & Onuchic (135), who used it to determine the free-energy landscape of protein A. Both sets of researchers obtained convergence to the equilibrium distribution with quantitative determination of the free-energy barrier of the folding.

6.4. High-Temperature Unfolding Simulations of Proteins

The key assumption invoked in this method, pioneered by Daggett & Levitt (18), is the principle of microscopic reversibility, which states that the unfolding of a protein is the reverse of the folding process. The protein-unfolding simulations are usually run at

temperatures above 498 K, at which the native structure is lost in a few nanoseconds. Simulations start from the native crystal or NMR structure, which improves the probability of sampling a relevant region of the conformational space. Care must be taken to set the water density to that of liquid water at high temperatures, so the excess pressure is reduced and the water remains liquid for up to 498 K (136).

The features of the unfolding process such as the transition-state ensemble and the unfolded ensemble obtained by this method have shown remarkable agreement with those determined experimentally by Φ-value analysis in a number of studies. This method has been applied successfully to study a number of proteins (such as BPTI, myoglobin, protein A, ubiquitin, the SH3 domain, and the WW domain) and to study processes from amyloidogenesis to domain swapping. For details of the results, we refer the reader to the review by Day & Daggett (20).

Importantly, the free-energy landscape and the transition-state ensemble can be altered by thermal or chemical denaturant, and there may be significant differences between high-temperature and physiological free-energy landscapes (137, 138). At high temperatures, the transition state shifts toward the native state, and at very high temperatures, the rapid unfolding events are irreversible. However, for a wide range of temperatures, the nature of the protein-unfolding transition is temperature independent (139). Furthermore, Shea & Brooks (28) have shown that the principle of microscopic reversibility does not hold under strongly nonequilibrium conditions.

6.5. Folding Dynamics from the Free-Energy Landscape

Shea & Brooks (28) pioneered the landscape approach to folding dynamics. The free-energy landscape is obtained from the equilibrium population distribution of the protein. This distribution can be obtained only by an umbrella-sampling method, in which an additional harmonic potential is added to the Hamiltonian of the system to bias the sampling. The starting conformations are generated by unfolding simulations. Specific structures are selected from this ensemble as starting conformations for umbrella sampling under a set of desired thermodynamic constraints (such as the fraction of native contacts or radii of gyration). After each constrained trajectory has been simulated for a long enough time, the conformations from each trajectory are clustered to generate the density of states from which the free energy is calculated. Brooks and colleagues have applied their method successfully to determine the free-energy landscapes and folding dynamics for protein A, GB1, and Src-SH3, with good agreement with experiment (28).

However, this method assumes that the degrees of freedom orthogonal to the reaction coordinate equilibrate quickly, which might not always be the case. Also, the simulation time needed for large chain movement could significantly exceed the length of a typical umbrella-sampling simulation used in this method (140).

6.6. Stochastic Difference Equation Method

In this method, pioneered by Elber and colleagues (141), the classical action over the folding pathway(s) is minimized, which is an alternative to solving Newton's equations

of motion with a small time step. As opposed to solving Newton's equations, both the initial unfolded conformations (Y_u) and the final folded conformation (Y_f) must be known. The action is defined as an integral over the trajectory length, l, as given by Equation 9:

$$S = \int_{Y_u}^{Y_f} \sqrt{2(E-U)}dl, \tag{9}$$

where S is the action, E is the total energy, U is the potential energy of the system, and dl is the step length. This equation is discretized, so minimization of S of Equation 9 is equivalent to minimizing the following target function:

$$T = \sum_i \left(\frac{\partial S/\partial Y_i}{\Delta l_{i,i+1}} \right)^2 \Delta l_{i,i+1} + \lambda \sum_i (\Delta l_{i,i+1} - \langle \Delta l \rangle)^2 , \tag{10}$$

where $\Delta l_{i,i+1}$ is the step length, and λ is the strength of a penalty function that restrains the step length to the average length.

Ghosh et al. (142) used the stochastic difference equation method to study the folding pathways of protein A. They performed the minimization of T of Equation 10 for 130 initial unfolded conformations obtained by thermal unfolding of the experimental structure of this protein. They found that the C-terminal α-helix is the most stable one and forms first, and the formation of secondary- and tertiary-structure elements is strongly coupled with the somewhat earlier formation of secondary structure. This observation is consistent with some experimental and simulation results (37, 143) but contradicts others (144, 145), according to which the middle helix forms first. In a recent extensive all-atom MD study, A. Jagielska & H.A. Scheraga (submitted manuscript) have shown that the relative rates of formation of the three α-helices depend on temperature; consequently, a different order can be observed under different conditions of simulation and experiment.

6.7. Quantum-Classical Molecular Dynamics

The classical equations of motion (Section 2.1) are valid when chemical reactions are not involved because the typical amplitudes of motions are much smaller than the corresponding thermal De Broglie wavelengths. Furthermore, some biological processes (such as oxygen binding to hemoglobin, enzymatic reactions, and the light-induced charge transfer in the photosynthetic reaction centers) involve quantum effects such as a change in chemical bonding, noncovalent intermediates, tunneling of proton and electron, and dynamics on electronically excited states that cannot be modeled with the classical formulas. One can handle processes involving proton transfer by introducing a special potential function for the proton(s) exchanged between the proton-acceptor atoms (146). For a general purpose, a hybrid approach known as QM/MM has been designed (147), in which the system is partitioned into a small core (within which the actual chemical reaction occurs) and the surroundings. The core is treated at the quantum-mechanical level, whereas the surroundings are treated at the classical level. The electrostatic potential from the surroundings contributes to the Hamiltonian of the core part.

7. CONCLUSIONS AND OUTLOOK

The recent developments of MD both in theory (more accurate force fields, reduced models of proteins and other polymers, and more efficient and stable integration algorithms) and computational techniques (use of distributed computing) provide a convincing reason to believe that ab initio simulations of protein-folding dynamics are at hand. Despite successful folding simulations on relatively small proteins using an all-atom scheme (22, 80), the robust approach is likely to involve the use of physics-based reduced models of proteins that require less than one processor day of simulations per trajectory (120). However, in-depth understanding of protein-folding pathways must ultimately involve atomic details. In our view, the robustness and low cost of simplified models and accuracy of atomic ones can be combined by using a hybrid approach in which folding trajectories are first calculated using mesoscopic models and then are converted to all-atom trajectories using, for example, the physics-based approach developed in our laboratory (148, 149).

SUMMARY POINTS

1. MD is an invaluable tool for studying protein folding and dynamics as well as thermodynamics *in silico*. Because of computational cost, ab initio folding simulations with explicit water are limited to peptides and very small proteins, whereas simulations of real-size proteins are confined to high-temperature unfolding and refolding, or dynamics of the experimental structure.
2. The use of mesoscopic models with physics-based potentials (e.g., UNRES) is a reasonable trade-off between computational cost and accuracy, and enables one to carry out ab initio folding simulations in real time. This approach appears preferential to the use of collective coordinates, rigid-body, or dihedral-angle dynamics, which provide less speed-up.
3. When water is not considered explicitly, it is imperative to account for protein-water interactions through the introduction of implicit-solvent models. However, water also influences the dynamics through collisions with the protein molecule; this effect can be handled by introducing Langevin dynamics. Care must be taken when implicit-solvent simulations are carried out in a non-Langevin mode (i.e., without introducing friction and random forces) because the simulated events then occur in too short a time.
4. Symplectic algorithms for integrating the equations of motion are most reliable because they provide stable trajectories in the long run and control of the accuracy of the solution by monitoring the property that should be conserved.
5. Protein folding must be analyzed in terms of time evolution of ensembles and not of a single molecule; therefore, parallel simulations must be run to discern all possible folding pathways. As for now, this is possible only for

mesoscopic models. Also, more than one observable must be used to monitor the progress of folding. Multiple-trajectory simulations are easily parallelizable and can, therefore, take full advantage of distributed computing.

6. Extensions of MD such as replica-exchange MD enhance the scope of simulations in terms of size and timescale, although this comes at the expense of losing detailed information about the folding pathways.

FUTURE ISSUES

1. Currently, force fields are not perfect (even the all-atom ones). It is possible to obtain different results with different force fields. Therefore, improving force fields (both the all-atom and the reduced ones, and the water potentials) is a priority.
2. Because use of reduced models is the most reasonable option for large-scale ab initio folding simulations, and atomistically detailed results are required in most applications, there is a need to design a method for converting coarse-grained trajectories to all-atom trajectories.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (GM14312), the National Science Foundation (MCB00-03722), the NIH Fogarty International Center (TW7193), and grant 3 T09A 032 26 from the Polish Ministry of Education and Science. This research was conducted by using the resources of (*a*) our 392-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (*b*) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (*c*) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, (*d*) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (*e*) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

LITERATURE CITED

1. Haran G. 2003. Single-molecule fluorescence spectroscopy of biomolecular folding. *J. Phys. Condens. Matter* 15:R1291–317
2. Baryshnikova EN, Melnik BS, Finkelstein AV, Semisotnov GV, Bychkova VE. 2005. Three-state protein folding: experimental determination of free-energy profile. *Protein Sci.* 14:2658–67
3. Vu DM, Myers JK, Oas TG, Dyer RB. 2004. Probing the folding and unfolding dynamics of secondary and tertiary structures in a three-helix bundle protein. *Biochemistry* 43:3582–89

4. Wang T, Lau WL, DeGrado WF, Gai F. 2005. T-jump infrared study of the folding mechanism of coiled-coil GCN4-p1. *Biophys. J.* 89:1–8
5. Collins ES, Wirmer J, Hirai K, Tachibana H, Segawa S, et al. 2005. Characterisation of disulfide-bond dynamics in non-native states of lysozyme and its disulfide deletion mutants by NMR. *Chembiochem* 6:1619–27
6. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. 2005. Simultaneous determination of protein structure and dynamics. *Nature* 433:128–32
7. Dyson HJ, Wright PE. 2005. Elucidation of the protein folding landscape by NMR. *Meth. Enzymol.* 394:299–321
8. Munoz V, Ghirlando R, Blanco FJ, Jas GS, Hofrichter J, Eaton WA. 2006. Folding and aggregation kinetics of a β-hairpin. *Biochemistry* 45:7023–35
9. Nishida S, Nada T, Terazima M. 2005. Hydrogen bonding dynamics during protein folding of reduced cytochrome *c*: temperature and denaturant concentration dependence. *Biophys. J.* 89:2004–10
10. Lee JC, Chang I-J, Gray HB, Winkler JR. 2002. The cytochrome *c* folding landscape revealed by electron-transfer kinetics. *J. Mol. Biol.* 320:159–64
11. Alder BJ, Wainwright TE. 1957. Phase transition for a hard-sphere system. *J. Chem. Phys.* 27:1208–9
12. Alder BJ, Wainwright T. 1958. Molecular dynamics by electronic computers. In *Proc. Int. Symp. Transp. Process. Stat. Mech.*, ed. I Prigogine, pp. 97–131. New York: Intersciences
13. Rahman A. 1964. Correlations of the motion of atoms in liquid argon. *Phys. Rev.* 2:A405–11
14. Rahman A, Stillinger FH. 1971. Molecular dynamics study of liquid water. *J. Chem. Phys.* 55:3336–59
15. McCammon JA, Gelin BR, Karplus M. 1977. Dynamics of folded proteins. *Nature* 267:585–90
16. Brooks CL III. 1992. Characterization of native apomyoglobin by molecular dynamics simulations. *J. Mol. Biol.* 227:375–80
17. Mark AE, van Gunsteren WF. 1992. Simulation of the thermal-denaturation of hen egg-white lysozyme: trapping the molten globule state. *Biochemistry* 31:7745–48
18. Daggett V, Levitt M. 1993. Protein unfolding pathways explored through molecular-dynamics simulations. *J. Mol. Biol.* 232:600–19
19. Tirado-Rives J, Jorgensen WL. 1993. Molecular dynamics simulations of the unfolding of apomyoglobin in water. *Biochemistry* 32:4175–84
20. Day R, Daggett V. 2003. All-atom simulations of protein folding and unfolding. *Adv. Protein Chem.* 66:373–803
21. Daggett V. 2006. Protein folding simulations. *Chem. Rev.* 106:1898–916
22. Duan V, Kollman PA. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–44
23. Pande VS, Rokhsar DS. 1999. Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G. *Proc. Natl. Acad. Sci. USA* 96:9062–67

20. Presents a comprehensive review of the possibilities and limitations of all-atom MD and its application to studying protein-unfolding and -refolding pathways.

22. Describes the first ab initio simulation of folding of a small α-helical protein by all-atom MD with explicit water.

24. Roccatano D, Amadei A, diNola A, Berendsen HJC. 1999. A molecular dynamics study of the 41–56 β-hairpin from B1 domain of protein G. *Protein Sci*. 8:2130–43
25. Tsai J, Levitt M, Baker D. 1999. Hierarchy of structure loss in MD simulations of src SH3 domain unfolding. *J. Mol. Biol*. 291:215–25
26. Adcock SA, McCammon JA. 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev*. 106:1589–615
27. Brooks CL III, Karplus M, Pettitt BM. 1988. *Proteins: A Theoretical Perspective of Dynamics: Structure and Thermodynamics*. New York: Wiley
28. Shea JE, Brooks CL III. 2001. From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem*. 52:499–535
29. Frenkel D, Smit B. 2000. *Understanding Molecular Simulation: From Algorithms to Applications*. New York: Academic
30. Kolinski A, Skolnick J. 2004. Reduced models of proteins and their applications. *Polymer* 2:511–24
31. Shakhnovich E. 2006. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev*. 106:1559–88
32. Khalili M, Liwo A, Rakowski F, Grochowski P, Scheraga HA. 2005. Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J. Phys. Chem. B* 109:13785–97
33. Langevin P. 1908.The theory of Brownian movement. *C. R. Acad. Sci*. 146:530–33
34. Gear CW. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*. Cliffs, NJ: Prentice Hall
35. Martyna GJ, Tuckerman ME, Tobias DJ, Klein ML. 1996. Explicit reversible integrators for extended systems dynamics. *Mol. Phys*. 87:1117–57
36. Ciccotti G, Kalibaeva G. 2004. Deterministic and stochastic algorithms for mechanical systems under constraints. *Philos. Trans. R. Soc. London Ser. A* 362:1583–94
37. Khalili M, Liwo A, Jagielska A, Scheraga HA. 2005. Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model α-helical systems. *J. Phys. Chem. B* 109:13798–810
38. Cieplak M, Hoang TX, Robbins MO. 2002. Kinetics nonoptimality and vibrational stability of proteins. *Proteins* 49:104–13
39. Barth E, Schlick T. 1998. Overcoming stability limitations in biomolecular dynamics. I. Combining force splitting via extrapolation with Langevin dynamics in LN. *J. Chem. Phys*. 109:1617–32
40. Rudnicki WR, Bakalarski G, Lesyng B. 2000. A mezoscopic model of nucleic acids. Part 1: Lagrangian and quaternion molecular dynamics. *J. Biomol. Struct. Dyn*. 17:1097–108
41. Pearlman DA, Case DA, Caldwell JW, Ross SW, Cheatham TE III, et al. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to

26. Includes a description of algorithms, force fields, and techniques going beyond MD such as QM/MM and constant pH simulations.

29. Includes a description of all MD algorithms used in biomolecular simulations.

36. Discusses comprehensively symplectic algorithms and stochastic algorithms for MD derived by using the Liouville formalism.

simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91:1–41
42. Kubelka J, Hofrichter J, Eaton WA. 2004. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* 14:76–88
43. Gibson KD, Scheraga HA. 1990. Variable step molecular-dynamics: an exploratory technique for peptides with fixed geometry. *J. Comput. Chem.* 11:468–97
44. Ryckaert J, Ciccotti G, Berendsen H. 1977. Numerical integration of Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comp. Phys.* 27:327–41
45. Andersen H. 1983. RATTLE: a velocity version of the SHAKE algorithm for molecular dynamics calculations. *J. Comp. Phys.* 52:24–34
46. Hess B, Bekker H, Berendsen HJC, Fraajie JGEM. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–72
47. Chen J, Im W, Brooks C III. 2005. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J. Comput. Chem.* 26:1565–78
48. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–90
49. Morishita T. 2000. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *J. Chem. Phys.* 113:2976–82
50. Nosé S. 1984. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* 52:255–68
51. Hoover WG. 1985. Canonical dynamics: equilibrum phase-space distribution. *Phys. Rev. A* 31:1695–97
52. Nosé S. 2001. An improved symplectic integrator for Nosé-Poincaré thermostat. *J. Phys. Soc. Jpn.* 70:75–77
53. Paterlini MG, Ferguson DM. 1998. Constant temperature simulations using the Langevin equation with velocity Verlet integration. *Chem. Phys.* 236:243–52
54. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217
55. Berendsen HJC, van der Spoel D, van Drunen R. 1995. GROMACS: a message-passing parallel molecular dynamics implementation. *Comp. Phys. Commun.* 91:43–56
56. Ewig CS, Thacher TS, Hagler AT. 1999. Derivation of class II force fields. 7. Nonbonded force field parameters for organic compounds. *J. Phys. Chem. B* 103:6998–7014
57. van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, et al. 2006. Biomolecular modeling: goals, problems, perspectives. *Angew. Chem. Int. Ed. Engl.* 43:4064–92
58. Mackerell AD Jr. 2004. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* 25:1584–604

59. Roterman IK, Lambert MH, Gibson KD, Scheraga HA. 1989. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. ϕ-ψ maps for N-acetyl alanine N'-methyl amide: comparisons, contrasts and simple experimental tests. *J. Biomol. Struct. Dyn*. 7:421–53
60. Yeh IC, Hummer G. 2002. Peptide loop closure kinetics from microsecond molecular dynamics simulation in explicit solvent. *J. Am. Chem. Soc*. 124:6563–68
61. Sorin EJ, Pande VS. 2005. Exploring helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J*. 88:2472–93
62. Williams S, Causgrove TP, Gilmanshin R, Fang KS, Callender RH, et al. 1996. Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* 35:691–97
63. Fain B, Levitt M. 2003. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proc. Natl. Acad. Sci. USA* 100:10700–5
64. Schug A, Wenzel W. 2006. An evolutionary strategy for all-atom folding of the 60-amino-acid bacterial ribosomal protein L20. *Biophys. J*. 90:4273–80
65. Stillinger FH, Rahman A. 1974. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys*. 60:1545–67
66. Berendsen HJC, Grigera JR, Straatsma TP. 1987. The missing term in the effective polar potentials. *J. Phys. Chem*. 91:6269–71
67. Jorgensen WL, Chandrasekhar J, Madura JD, Impy RW, Klein ML. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*. 79:926–35
68. Horn HW, Swope WC, Pitera JW, Madura JD, Dick TJ, et al. 2004. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* 120:9665–78
69. Ferguson DM. 1995. Parameterization and evolution of a flexible water model. *J. Comput. Chem*. 16:501–11
70. Sprik M, Klein ML. 1988. A polarisable model for water using distributed charge sites. *J. Chem. Phys*. 89:7556–60
71. Still WC, Tempczyk A, Hawley RC, Hendrickson T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112: 6127–29
72. Ferrara P, Apostolakis J, Caflisch A. 2002. Evolution of a fast implicit solvent model for molecular dynamics simulations. *Proteins* 46:24–33
73. Bashford D, Case D. 2000. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem*. 51:129–52
74. Vila J, Williams RL, Vásquez M, Scheraga HA. 1991. Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins* 10:199–218
75. Stouten PFW, Frömmel C, Nakamura H, Sander C. 1993. An effective solvation term based on atomic occupancies for use in protein simulations. *Mol. Simul*. 10:97–120
76. Augspurger JD, Scheraga HA. 1996. An efficient, differentiable hydration potential for peptides and proteins. *J. Comput. Chem*. 17:1549–58

77. Wawak RJ, Gibson KD, Scheraga HA. 1994. Gradient discontinuities in calculations involving molecular surface area. *J. Math. Chem*. 15:207–32
78. Im W, Lee M, Brooks C III. 2003. Generalized Born model with a simple smoothing function. *J. Comput. Chem*. 24:1691–702
79. Bursulaya B, Brooks CL III. 2002. Comparative study of folding free energy landscape of a three-stranded β-sheet protein with explicit and implicit solvent models. *J. Phys. Chem. B* 104:12378–83
80. Jang S, Kim E, Shin S, Pak Y. 2003. Ab initio folding of helix bundle proteins using molecular dynamics simulations. *J. Am. Chem. Soc.* 125:14841–46
81. Ferrara P, Apostolakis J, Caflisch A. 2000. Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B* 104:5000–10
82. Plaxco KW, Baker D. 1998. Limited internal friction in the rate-limiting step of a two-state protein folding reaction. *Proc. Natl. Acad. Sci. USA* 95:13591–96
83. Ten Wolde PR, Chandler D. 2002. Drying-induced hydrophobic polymer collapse. *Proc. Natl. Acad. Sci. USA* 99:6539–43
84. Cheung MS, Garcia AE, Onuchic JN. 2002. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. USA* 99:685–90
85. Teeter MM. 1991. Water-protein interactions: theory and experiment. *Annu. Rev. Biophys. Biophys. Chem*. 20:577–600
86. Baptista AM, Teixeira VH, Soares CM. 2002. Constant-pH molecular dynamics using stochastic titration. *J. Chem. Phys*. 117:4184–200
87. Burgi R, Kollman PA, van Gunsteren WF. 2002. Simulating proteins at constant pH: an approach combining molecular dynamics and Monte Carlo simulation. *Proteins* 47:469–80
88. Walczak AM, Antosiewicz JM. 2002. Langevin dynamics of proteins at constant pH. *Phys. Rev. E* 66:051911
89. Mongan J, Case D, McCammon J. 2004. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem*. 25:2038–48
90. Makowska J, Baginska K, Makowski M, Jagielska A, Liwo A, et al. 2006. Assessment of two theoretical methods to estimate potentiometric titration curves of peptides: comparison with experiment. *J. Phys. Chem. B* 110:4451–58
91. Amadei A, Linssen A, Berendsen H. 1993. Essential dynamics of proteins. *Proteins* 17:412–25
92. Schulze B, Grubmuller H, Evanseck J. 2000. Functional significance of hierarchical tiers in carbonmonoxy myoglobin: conformational substates and transitions studied by conformational flooding simulations. *J. Am. Chem. Soc*. 122:8700–11
93. Amadei A, Linssen A, de Groot B, Aalten D, Berendsen H. 1996. An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dyn*. 13:615–25
94. Zhang Z, Shi Y, Liu H. 2003. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys. J*. 84:3583–93

80. Describes the first simulation of all-atom folding of protein A using AMBER with the GB implicit-solvent model.

95. Roccatano D, Daidone I, Ceruso MA, Bossa C, Di Nola A. 2003. Selective excitation of native fluctuations during thermal unfolding simulations: horse heart cytochrome *c* as a case study. *Biophys. J.* 84:1876–83
96. Atilgan A, Durell S, Jernigan R, Demierl M, Keskin O, Bahar I. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–15
97. Nielsen SO, Lopez CF, Srinivas G, Klein ML. 2004. Coarse grain models and the computer simulation of soft materials. *J. Phys. Condens. Matter* 16:R481–512
98. Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107
99. Brown S, Head-Gordon T. 2004. Intermediates and the folding of proteins L and G. *Protein Sci.* 13:958–70
100. Shih AY, Arkhipov A, Freddolino PL, Schulten K. 2006. A coarse grained protein-lipid model with application to lipoprotein particles. *J. Phys. Chem. B* 110:3674–84
101. Klimov DK, Thirumalai D. 1997. Viscosity dependence of the folding rates of proteins. *Phys. Rev. Lett.* 79:317–20
102. Veitshans T, Klimov D, Thirumalai D. 1996. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* 2:1–22
103. Taketomi H, Ueda Y, Gō N. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific interunit interactions. *Int. J. Pept. Protein Res.* 7:445–59
104. Hoang TX, Cieplak M. 2000. Molecular dynamics of folding of secondary structures in Gō-type models of proteins. *J. Chem. Phys.* 112:6851–62
105. Karanicolas J, Brooks CL III. 2003. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* 334:309–25
106. Cieplak M, Sulkowska J. 2005. Thermal unfolding of proteins. *J. Chem. Phys.* 123:194908
107. Cieplak M. 2006. Protein unfolding in a force clamp. *J. Chem. Phys.* 123:194901
108. Sorenson JM, Head-Gordon T. 2002. Toward minimalist models of larger proteins: a ubiquitin-like protein. *Proteins* 46:368–79
109. He SQ, Scheraga HA. 1998. Brownian dynamics simulations of protein folding. *J. Chem. Phys.* 108:287–300
110. Sippl MJ. 1993. Boltzmann principle, knowledge-based mean fields and protein-folding: an approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.* 7:473–501
111. Meller J, Elber R. 2001. Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins* 45:241–61
112. Crippen GM. 1996. Easily searched protein folding potentials. *J. Mol. Biol.* 260:467–75
113. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–42

114. Kmiecik S, Kurcinski M, Rutkowska A, Gront D, Kolinski A. 2006. Denatured proteins and early folding intermediates simulated in a reduced conformational space. *Acta Biochim. Pol.* 53:131–43
115. Kolinski A, Klein P, Romiszowski P, Skolnick J. 2003. Unfolding of globular proteins: Monte Carlo dynamics of a realistic reduced model. *Biophys. J.* 85:3271–78
116. Wolynes PG. 2005. Energy landscapes and solved protein-folding problems. *Philos. Trans. R. Soc. London Ser. A* 363:453–64
117. Fujitsuka Y, Takada S, Luthey-Schulten ZA, Wolynes PG. 2004. Optimizing physical energy functions for protein folding. *Proteins* 54:88–103
118. Liwo A, Czaplewski C, Pillardy J, Scheraga HA. 2001. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* 115:2323–47
119. Oldziej S, Liwo A, Czaplewski C, Pillardy J, Scheraga HA. 2004. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape. 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B* 108:16934–49
120. Liwo A, Khalili M, Scheraga HA. 2005. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA* 102:2362–67
121. Khalili M, Liwo A, Scheraga HA. 2006. Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. *J. Mol. Biol.* 355:536–47
122. Clementi C, Nymeyer H, Onuchic JN. 2000. Topological and energetic factors: What determined the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–53
123. Radhakrishnan R, Schlick T. 2004. Orchestration of cooperative events in DNA synthesis and repair mechanism unraveled by transition path sampling of DNA polymerase β's closing. *Proc. Natl. Acad. Sci. USA* 101:5970–75
124. Chowdhury S, Lee MC, Duan Y. 2004. Characterizing the rate-limiting step of Trp-cage folding by all-atom molecular dynamics simulations. *J. Phys. Chem. B* 108:13855–65
125. Pande VS, Baker I, Chapman J, Elmer S, Kaliq S, et al. 2003. Atomistic protein folding simulations on the submillisecond timescale using world-wide distributed computing. *Biopolymers* 68:91–109
126. Krieger F, Fierz B, Bieri O, Drewello M, Kiefhaber T. 2003. Dynamics of unfolded polypeptide chains as model for the earliest steps in protein folding. *J. Mol. Biol.* 332:265–74
127. Paci E, Cavalli A, Vendruscolo M, Caflisch A. 2003. Analysis of the distributed computing approach applied to the folding of a small β peptide. *Proc. Natl. Acad. Sci. USA* 100:8217–22
128. Marianayagam NJ, Fawzi NL, Head-Gordon T. 2005. Protein folding by distributed computing and the denatured state ensemble. *Proc. Natl. Acad. Sci. USA* 102:16684–89

120. Describes the application of the physics-based UNRES mesoscopic potential to MD and successful ab initio folding of a number of proteins.

125. Describes the use of multiple-trajectory simulations to speed up simulated folding; the approach can readily be used to study two-state folders.

129. Describes the application of the replica-exchange MD method to sample the conformational space and calculate thermodynamic properties of proteins.

129. Sugita Y, Okamoto Y. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–51

130. Rhee YM, Pande VS. 2003. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* 84:775–86

131. Felts A, Harano Y, Gallicchio E, Levy R. 2004. Free energy surfaces of β-hairpin and α-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins* 56:310–21

132. Nanias M, Czaplewski C, Scheraga HA. 2006. Replica exchange and multicanonical algorithms with the coarse-grained united-residue (UNRES) force field. *J. Chem. Theor. Comput.* 2:513–28

133. Earl DJ, Deem MW. 2005. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7:3910–16

134. Zhou RH, Berne BJ, Germain R. 2001. The free energy landscape for β hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA* 98:14931–36

135. Garcia AE, Onuchic JN. 2003. Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. USA* 100:13898–903

136. Daggett V, Fersht AR. 2003. Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* 28:18–25

137. Dinner AR, Karplus M. 1999. Is protein unfolding the reverse of protein folding? A lattice simulation analysis. *J. Mol. Biol.* 292:403–19

138. Finkelstein AV. 1997. Can protein unfolding simulate protein folding? *Protein Eng.* 10:843–45

139. Cavalli A, Ferrara P, Caflisch A. 2002. Weak temperature dependence of the free energy surface and folding pathways of structured proteins. *Proteins* 47:305–14

140. Snow CD, Sorin EJ, Rhee YM, Pande VS. 2005. How well can simulations predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* 34:43–69

141. Elber R, Ghosh A, Cardenas A. 2002. Long time dynamics of complex systems. *Acc. Chem. Res.* 35:396–403

142. Describes the application of the minimum-action principle to minimize the action of the system to determine protein-folding pathways.

142. Ghosh A, Elber R, Scheraga HA. 2002. An atomically detailed study of the folding pathways of protein A with the stochastic difference equation. *Proc. Natl. Acad. Sci. USA* 99:10394–98

143. Kolinski A, Skolnick J. 1994. Monte-Carlo simulations of protein-folding. 2. Application to protein A, ROP, and crambin. *Proteins* 18:353–66

144. Sato S, Religa TL, Fersht AR. 2006. Φ-analysis of the folding of the B domain of protein A using multiple optical probes. *J. Mol. Biol.* 360:850–64

145. Boczko EM, Brooks CL III. 1995. First-principles calculation of the folding free energy of a 3-helix bundle protein. *Science* 269:393–96

146. Bala P, Grochowski P, Nowinski K, Lesyng B, McCammon JA. 2000. Quantum-dynamical picture of a multistep enzymatic process: reaction catalyzed by phospholipase A_2. *Biophys. J.* 79:1253–62

147. Olsson MHM, Parson WW, Warshel A. 2006. Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis. *Chem. Rev.* 106:1737–56

148. Kaźmierkiewicz R, Liwo A, Scheraga HA. 2002. Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte Carlo method. *J. Comput. Chem*. 23:715–23
149. Kaźmierkiewicz R, Liwo A, Scheraga HA. 2003. Addition of side chains to a known backbone with defined side-chain centroids. *Biophys. Chem*. 100:261–80. Erratum. *Biophys. Chem*. 106:91

Annual Review of
Physical Chemistry

Volume 58, 2007

Contents