CHAPTER 8

Computer Simulations of Protein Folding

VIJAY S. PANDE, ERIC J. SORIN, CHRISTOPHER D. SNOW AND YOUNG MIN RHEE

Department of Chemistry and Biophysics Program, Stanford University,
Stanford, CA 94305, USA

8.1 Introduction: Goals and Challenges of Simulating Protein Folding

Computer simulation holds great promise to significantly complement experiment as a tool for biological and biophysical characterization. Simulations offer the promise of atomic spatial detail with femtosecond temporal resolution. However, the application of computational methodology has been greatly limited due to fundamental computational challenges: put simply, for much of what one would want to examine, atomistic simulations would require decades to millennia to complete. Below, we detail current methods to tackle these challenges as well as recent applications of this methodology.

8.1.1 Simulating Protein Folding

Proteins play a fundamental role in biology. With their ability to perform numerous biological functions, including acting as catalysts, antibodies, and molecular signals, proteins today realize many of the goals to which modern nanotechnology aspires. However, before proteins can carry out these remarkable molecular functions, they must perform another amazing feat – they must assemble themselves. This process of protein self-assembly into a particular

RSC Biomolecular Sciences
Protein Folding, Misfolding and Aggregation: Classical Themes and Novel Approaches
Edited by Victor Muñoz

shape, or ‘‘fold,’’ is called protein folding. Due to the importance of the folded state in the biological activity of proteins, recent interest from misfolding related diseases[1] (see Chapter 10 by Esteras-Chopo *et al.*), and a fascination with how this process occurs,[2–4] there has been much work performed in order to unravel the mechanism of protein folding[5] (see Chapter 3 by Wolynes).

While there are several questions relating to the ‘‘protein folding problem,’’ including structure prediction[6,7] and protein design (see Chapter 9 by Lehmann and co-workers), here we will concentrate on another aspect of folding: *how* do proteins fold into their final folded structure? Experimentally characterizing the detailed nature of the protein folding mechanism is considerably more difficult than characterizing the static structure. We therefore turn to the combination of experiment and atomistic models (that can readily yield the desired spatial and temporal detail), but we must in turn ask ‘‘how quantitatively predictive are these simulations?’’ The true test is statistical significance. The very act of *statistically* comparing with experiment is critical, and leads to either model validation or an indication that further model refinement is necessary.

There are two approaches one can take in molecular simulation. One direction is to perform coarse-grained simulations using simplified, or ‘‘minimalist,’’ models. These models typically either make simplifying assumptions (such as Go models, which use simplified Hamiltonians[8]), or employ coarse-grained representations (such as using alpha-carbon only models to represent the protein[9]) or potentially both. While these methods are often first considered due to their computational efficiency, perhaps an even greater benefit of simplified models is their ability to potentially yield insight into general properties involved in protein folding. However, with any model there are limitations and the cost for such potential insight into general properties of folding is the limitation of restricted applicability to any *particular* protein system.

Alternatively, one can examine more detailed models. These models typically have full atomic detail, often for both the protein and solvent alike. Detailed models have the obvious benefit of potentially greater fidelity to experiment. However, this comes at two great costs. First, the computational demands for performing the simulation become enormous. Second, the added degrees of freedom lead to an explosion of extra detail and simulation-generated data; the act of gleaning insight from this sea of data is no simple task and is often underestimated, especially in light of the more straightforward (although still often difficult) task of simply performing the simulations. We emphasize that the relevant question is not whether a given method is ‘‘correct’’ in some absolute sense (as all models have limitations), but whether the model is predictive to some degree of accuracy.

Why are detailed models worth this enormous effort in both simulation and analysis? First, quantitative comparison between theory and experiment is critical for validating simulation as well as lending interpretation to experimental results. While it is generally held that experiments will not be able to yield the detail and precision available in simulations (and that simulations may likely be the only way one can fully understand the folding mechanism[10]), without quantitative validation of simulations there is no way to know whether the simulation model or methodology are sufficiently accurate to yield a faithful reproduction of reality. Indeed, without a quantitative comparison to experiment, there is no way to decisively arbitrate the relative predictive merits of one model over another.

Second, detailed models potentially have a greater predictive power. In principle, a detailed model should allow one to start purely from the protein sequence and, by simulating the physical dynamics of protein folding, yield everything that one can measure experimentally, including folding and unfolding rates, free energies, and the detailed geometry of the folded state. In practice, the ability for detailed models to achieve these lofty goals rests both on the ability to carry out the computationally demanding kinetics simulations as well as the ability for current models (force fields) to yield sufficiently accurate representations of inter-atomic interactions.

8.1.2 What Are the Challenges for Atomistic Simulation?

First, one must consider the source of the great computational demands of molecular simulation at atomic detail. To simulate dynamics, typically one numerically integrates Newton's equations for all of the atoms in the system. By choosing models with atomic degrees of freedom, one must simulate the dynamics at the timescales of atomic motion (femtoseconds). Indeed, if the timestep involved in numerical integration is pushed too high (without constraining degrees of freedom), the numerical integration becomes unstable. This leads to the trivial problem that if one wants to reach the millisecond timescale by taking femtosecond steps, many (10^{12}) steps must be taken. While modern molecular dynamics codes are extremely well optimized and perform typically millions of steps per CPU day, this clearly falls short of what is needed (see Figure 8.1).

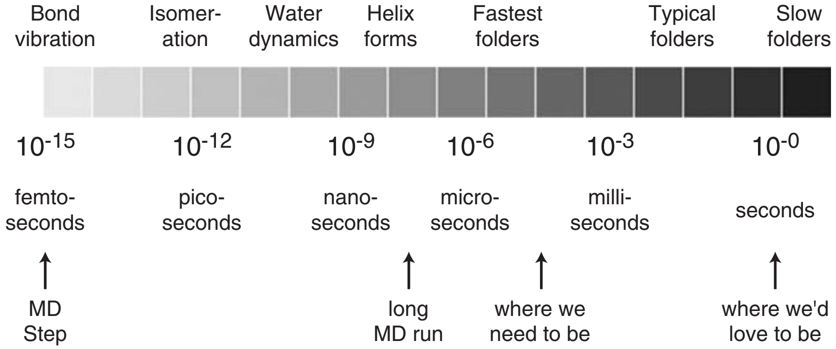

Figure 8.1 Relevant timescales for protein folding. While detailed simulations must start with femtosecond timesteps, the timescales one would like to reach are much longer, requiring billions (microseconds) to trillions (milliseconds) of iterations. Typical fast, modern CPUs can do approximately a million iterations in day, posing a major challenge for detailed simulation.

However, even if one could reach the relevant timescales, the next question is whether our models would be sufficiently accurate. In particular, would we reach the folded state, would the folded state be stable (with free energy of stability comparable to experiment), and would we reach the folded state with a rate comparable to experiment? Indeed, if one could quantitatively predict protein folding rates, free energies of stability, and structure of the relevant states at equilibrium, one would be able to predict essentially *everything* that can be measured experimentally. While rates and free energies themselves can only indirectly detail the nature of how proteins fold, clearly the ability to quantitatively predict all experimental observables is a necessary prerequisite for any successful theory or simulation of protein folding.

However, a quantitative prediction of all experimental observables is necessary but not sufficient. If a simulation could only reproduce experiments, the simulation would not yield any new insight, which is the goal of simulations in the first place. This leads to a third important challenge for simulation: gaining new insight. Indeed, as one adds detail to simulations, the burden of analysis becomes greater and greater. Atomistic simulations can easily generate gigabytes of data to be processed, but the volume of data does not reduce the inherent complexity of the physical process. A vast number of degrees of freedom from time-resolved protein and water coordinates can obscure any simple, direct analysis of the folding mechanism.

Additionally, analysis of such simulations may reflect the seemingly arbitrary state definitions used by the one performing the analysis, and great care must therefore be taken in defining the relevant states prior to data analysis. This, of course, often presents the most notable issue in interpreting simulation data, due to the sheer difficulty in collecting adequate data to define the states, and microstates, that the model would predict. As detailed below, this issue is most often overcome by employing simplified models. These models are generally built around the known or desired states prior to simulation, but suffer the obvious lack of predicting metastable, misfolded, or intermediate states that may be observable when using atomistic simulation models.

8.2 Protein Folding Models: From Atomistic to Simplified Representations

8.2.1 Atomic Force Fields

Atomistic models for protein folding typically utilize a classical force field, which attempts to reproduce the physical interaction between the atoms in the protein and solvent. The energy of the system is defined as the sum of interatomic potentials, which consist of several terms:

$$\mathrm{E} = \mathrm{E}_{\mathrm{LJ}} + \mathrm{E}_{\mathrm{Coulomb}} + \mathrm{E}_{\mathrm{bonded}} \tag{8.1}$$

The van der Waals interaction between atoms is most commonly modeled by a Lennard–Jones energy (E_{LJ})

$$E_{LJ} = \Sigma_{ij}\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}] \tag{8.2}$$

where σ_{ij} is related to the size of the atoms i and j and ε_{ij} is related to the strength of their interaction. While van der Waals attraction is relatively weak, the LJ potential also serves an important role in providing hard core repulsion between atoms. The bonded interactions modeled in E_{bonded} handle the specific stereochemistry of the molecule – in particular, the nature of the covalent bonds and steric constraints in the angles and dihedral angles of the molecule. These interactions are clearly local, but they play a very important role in determining the conformational space of the molecule; changes to the backbone dihedral potentials in such a model can lead to greatly diverging simulation results.[11] $E_{Coulomb}$ corresponds to the familiar Coulomb's law:

$$E_{Coulomb} = \Sigma_{ij}q_iq_j/r_{ij} \tag{8.3}$$

where q_i is the charge on atom i and r_{ij} is the distance between atoms i and j. To best parameterize atomic force fields, such as accounting for quantum mechanical effects between nearby atoms, some force fields also include scaling coefficients for the pairwise E_{LJ} and $E_{Coulomb}$ terms between atoms separated by three covalent bonds (so-called "1-4 scaling"), and it has recently been demonstrated that modifying these scaling terms can significantly alter simulation results.[11]

It is perhaps most natural to handle the pairwise interactions explicitly as in Equation (8.1). However, this leads to simulation codes whose performance scales as N^2, where N is the number of atoms being simulated. Clearly, this is very computationally demanding. To reduce this demand, the calculation can ideally be made to scale linearly with N. For inherently short range interactions, it is natural to do this with cutoffs and long range corrections, *i.e.* to set the potential to zero smoothly once the distance is beyond some cutoff, such as 12 Å. Such cutoff procedures have been shown to lead to qualitatively incorrect results for Coulomb interactions[12] and reaction field or Ewald-based methods have been suggested as alternatives that can obtain significantly better results.[13]

Clearly there are many parameters in the above formulas. Indeed, these numbers grow further when one considers the fact that the chemical environment of atoms causes even the same type of chemical element (*e.g.* carbon) to act very differently. For example, carbon in a hydrocarbon chain will behave fundamentally differently from carbon in an aromatic ring. In order to handle such purely quantum mechanical effects in a classical model, one creates multiple atom types (corresponding to the different relevant environments) for each physical atomic element. In this example, one would define different carbon atom types. Thus, while there are only a handful of relevant physical atoms involved (primarily carbon, hydrogen, oxygen, and nitrogen), there can be tens to hundreds of different atom types.

Although this is clearly the natural way to handle the role of chemical environment in a classical model, this leads to an explosion of parameters needed in the model, leading to a modeling challenge in the determination of these parameters. Several groups have risen to this challenge and have developed parameterizations for the force field functionals similar to the form above. Typically, these parameterizations are divided into terms for proteins (such as AMBER,[14] CHARMM,[15] and OPLS[16]) and for the solvent (such as TIP or SPC models). Additionally, these force fields are typically parameterized using a specific water model, and may also be associated with specific molecular dynamics packages. One should thus be careful in combining protein and solvent models and also not confuse atomic force fields with the molecular dynamics software for which they were derived.

8.2.2 Implicit Solvation Models

With the parameterization described above for the physical forces between atoms, one can simulate all relevant interactions: protein–protein, protein–solvent, and solvent–solvent. However, in typical simulations with solvent represented explicitly (*i.e.* directly simulating the solvent atom by atom), the number of solvent atoms is much larger than the number of protein atoms and thus the majority of the computational time (*e.g.* 90%) goes into simulating the solvent. Clearly the solvent plays an important role since the hydrophobic and dielectric properties of water are essential to protein stability.[17,18] However, an alternative to explicit simulation of water is to include these properties *implicitly* by using a continuum model of solvent properties.

Typically, these models account for hydrophobicity in terms of some free-energy price for solvent exposed area on the protein. These surface area (SA) based methods vary somewhat in terms of how the surface area is calculated as well as the energetic dependence on this exposed surface area. We stress that one should not *a priori* expect that a simpler (and perhaps less accurate) calculation of the surface area yields worse results than a more geometrically accurate SA calculation. Indeed, since SA is itself an approximation, what is important for the fidelity of the model is not the geometric accuracy of the surface area but rather whether the SA term faithfully reproduces the physical effect as judged by comparison to experiment.

The dielectric contribution of water to the free energy is in some ways a more difficult contribution for which to account. The canonical method follows the Poisson–Boltzmann (PB) equation. To demonstrate the philosophy of implementing PB calculations, consider a protein immersed in solvent where the protein and solvent are modeled as dielectric media with dielectric constants of ε_{in} and ε_{out} respectively (thus making the dielectric a function of spatial position, $\varepsilon(x, y, z)$). Also, consider that the protein will likely have charges with a spatial density $\rho_{\text{protein}}(x, y, z)$ and that there will be counter-ions in the solvent with a charge density $\rho_{\text{countert}}(x, y, z)$. In this case, we can describe the resulting electrostatic potential and charge density as

$$\begin{aligned}\nabla[\varepsilon(x, y, z)\nabla\phi] &= -4\pi\rho(x, y, z)\\ &= -4\pi[\rho_{\text{protein}}(x, y, z) + \rho_{\text{countert}}(x, y, z)]\end{aligned} \tag{8.4}$$

where the total charge density ρ(x, y, z) is comprised of both the protein and counter-ion charges. If one assumes that the counter-ion density is driven thermodynamically to its free energy minimum, we can make the "mean field"-like approximation that

$$\rho_{\text{countert}}(x, y, z) = \Sigma_I n_i q_i \exp[-q_i\phi(x, y, z)/kT] \tag{8.5}$$

where n_i is the bulk number density of counter-ion species i and q_i is its charge. Thus, this method handles counter-ions implicitly as well as aqueous solvent. Including this term leads to the so-called non-linear Poisson–Boltzmann equation. If the Boltzmann term is Taylor expanded for small $\phi(x, y, z)/kT$ (*i.e.* high temperature, low counter-ion concentration, or low potential strength), one gets the so-called linearized Poisson–Boltzmann equation.

In general, the Poisson–Boltzmann equation is considered by many to be the "gold standard" for implicit solvation calculations. It can be used for both energy and force calculation[19] and is thus suitable for molecular dynamics. However, PB calculation is also typically very computationally demanding and there has been much effort to develop more computationally tractable, empirical approximations to the PB equation. For example, Still and co-workers developed an empirical approximation to PB.[20] Based on a generalization of the Born equation for the potential of atoms, Still's Generalized Born (GB) model (and its subsequent variants from Still's group and other groups) have been shown to be both computationally tractable and quantitatively accurate for some problems, including the solvation free energy of small molecules[20] and protein folding kinetics.[21]

8.2.3 Minimalist Models

To further simplify the model, the protein force field can be generated from the experimental structure. Using the information of the native conformation, attractive parts of the LJ potentials for all non-native contact pairs can be reduced or turned off altogether. Such a potential may lead to minimized frustration for folding (*i.e.* smoothing the energy landscape by removing small energetic barriers and metastable microstates, as shown in Figure 8.2), enabling much faster folding simulations. In many cases, the model is built by considering each amino acid residue as one particle (coarse-graining) to maximize the simplification. Using explicit or implicit solvent models mentioned in the above paragraphs is technically possible, though such an approach will lose the benefit of using the minimalist model itself. Therefore, solvent effects are usually considered using Langevin dynamics (random forces imparted on each

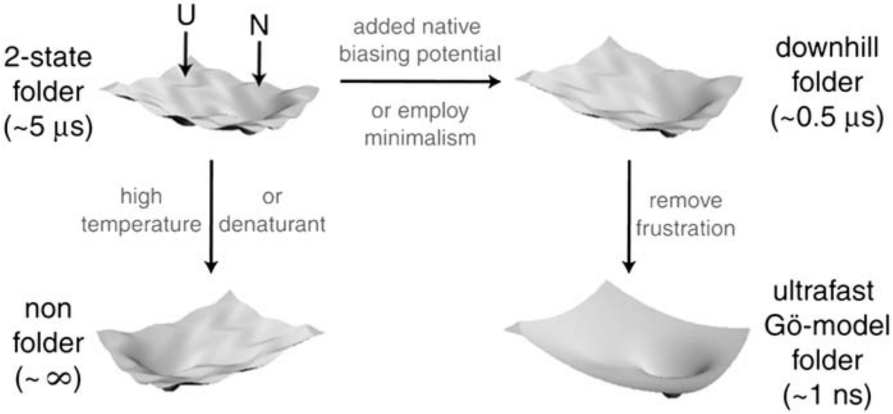

Figure 8.2 Example free-energy surface for a simple two-state folder and related surfaces derived by adding external forces or simplifications to the simulation model, demonstrating the variation in necessary simulation timescales for sampling of various models. Some sampling methods, such as REMD and umbrella sampling, make use of several landscapes by adding biasing potentials or including a large range of temperatures, while minimalist models remove landscape frustration and/or the presence of a non-native free-energy basin.

simulated body to represent solvent viscosity) or can be incorporated explicitly in the pairwise protein non-bonded interaction potential.[22]

8.2.4 How Accurate Are the Models?

Any question of accuracy must consider the desired experimental observable. One natural quantity to examine is the solvation free energy of small molecules, such as amino acid side chains[23] With recent advances in high-precision free-energy methods,[23,24] one can directly compare the models to experiment within experimental error.

For explicit solvent models,[24] the solvation free energies of small molecule analogs to amino acid side chains show a systematic shift (towards being less soluble). This would lead to an artificial stabilization of proteins (since the unfolded state would be less stable) and could have a significant impact on predicted protein–protein and protein–ligand free energies. These results suggest natural force-field improvements; recent work in this direction removes this systematic shift, leading to models with zero mean error with solvation free energy experiments and a surprisingly low RMSD ($\sim 0.4\,\text{kcal mol}^{-1}$).[24]

How accurate are implicit solvent models? While the GB models are somewhat empirical, they have been shown to agree reasonably well with PB calculations. More importantly, GB models have been able to accurately predict experimental results, such as the solvation free energy of small molecules.[20,25] In the end, experiment must of course be the final arbiter of any theoretical method. Moreover, while PB is on much firmer mathematical footing (*i.e.* one can derive it directly from the Poisson equation), one must consider that PB itself is empirical in nature in some respects. The concept of a dielectric is macroscopic; it is an approximation to apply this macroscopic concept to the microscopic world of small molecules and proteins (hundreds to thousands of atoms). However, the success of PB as a predictive tool demonstrates the validity (or, at the very least, predictive power) of such methods and approximations.

8.3 Sampling: Methods to Tackle the Long Timescales Involved in Folding

Simulating the mechanism of protein folding is a great computational challenge due to the long timescales involved. Below, we briefly summarize some methods that have been used to address this challenge. As in any computational method, each has its own limitations and it is natural to consider the regime of applicability of each method (Figure 8.2).

8.3.1 Tightly Coupled Molecular Dynamics (TCMD)

To simulate molecular dynamics (MD) one typically integrates Newton's equations numerically for the atoms in the system with femtosecond timesteps to include the fast timescales of atomic motion. Thus, to reach the millisecond timescale, many (10^{12}) steps must be taken. While modern molecular dynamics codes are extremely well optimized and perform typically 10^6 steps per CPU day, this clearly falls short of what is needed. Using multiple CPUs in a tightly coupled fashion to speed a single trajectory is appealing, but is an inefficient use of CPU power (*i.e.* one does not get a 100× speed increase with 100 CPUs) and thus has not been widely used to get beyond the nanosecond timescale, with the notable exception of Duan and Kollman's single 1 μs trajectory of the villin headpiece.[26]

8.3.2 Replica Exchange Molecular Dynamics (REMD)

Replica Exchange Molecular Dynamics[27–31] has become a powerful technique to explore the free-energy landscapes of proteins, with speed increases[32] of roughly 10× over traditional MD. Moreover, REMD efficiently parallelizes with only slightly coupled networking required. However, REMD achieves its speed increase by using a non-physical form of kinetics (in temperature replica space). This method yields a Boltzmann-weighted ensemble after sufficient convergence,[32] but the trajectories cannot themselves be used to predict any direct kinetic properties, although aspects related to the kinetics (such as possibly kinetically relevant intermediates) can be inferred from the resulting free-energy landscapes.[29]

8.3.3 High-temperature Unfolding

While folding times are very long from a simulation point of view, unfolding (especially under high denaturation conditions) can be very fast – on the nanosecond timescale.[33–35] Under extreme denaturing conditions (*e.g.* ~400 K temperature), one would expect the folded state to become only metastable, with a low barrier to unfolding. Daggett and Levitt[33] first took advantage of this scenario, and Daggett's group has subsequently pioneered this method to examine a variety of proteins and compare their results to experiment, especially with a comparison of ϕ values calculated at high-temperature folding *vs.* experimental measurements.[10,36] One note of caution is that the transition state character is dependent on temperature. For example, the Gruebele lab has found temperature-sensitive phi-values.[37] Of particular significance of the impact of this approach has been the ability to closely connect simulation predictions to experiment. However, applying extreme temperatures to models developed under ambient/biological temperatures (*i.e.* 300 ± 10 K) must be done with caution: it has recently been shown that even force fields that appear to be extremely accurate for the system studied fail to reproduce experimentally observed temperature-dependent trends at high and low temperatures.[11] While it is possible to study protein unfolding under conditions that approximate experiment, simulations to date trade authentic recapitulation of the experimental kinetics in favor of computational tractability.

8.3.4 Low-viscosity Simulation Coupled with Implicit Solvation Models

This is another common means to try to tackle long timescales.[38–41] In regular simulations with implicit solvent model, one typically uses the Langevin equation for dynamics and employs a damping term consistent with water-like viscosity. However, water is relatively viscous and such simulations can be very costly. Instead, many groups have proposed the use of viscosities only 1/100 to 1/1000 that of water (or even no viscosity at all). While lowering the viscosity greatly speeds the kinetics,[38] the effect of such non-physical modeling inherently assumes a potential risk of altering not only the rate but also the nature of the overall kinetics of the system.[42] Assuming simulation convergence, the correct thermodynamics should be obtained, but it must also be understood that the thermodynamics will be based on the model, and therefore may also miss microstates that are coupled with properties of the solvent (such as, in this case, viscosity).

8.3.5 Coarse-grained and Minimalist Models

These kinds of models have played a large role in recent simulations of protein folding.[6,22,43] The idea is to largely trade chemical complexity for computational tractability. Coarse-grained models allow one to directly address a range of hypotheses relating to general properties of folding. However, in their generality, by construction they may lack the ability to access more detailed questions of folding (depending on the nature of the question of interest). Whereas detailed models cannot, in general, be used to collect ensemble statistics for large biomolecular systems, this is not true for minimalist models, and a recent study used such a model to make a direct connection between individual folding pathways and the bulk observed folding mechanism for a system consisting of ~5000 atoms.[44]

8.3.6 Path Sampling

Given an initial trajectory between the unfolded and folded regions, which can be generated *via* high-temperature unfolding or similar means, this method generates an ensemble of different pathways that join the unfolded and folded regions. For example, Bolhuis and co-workers determined the formation order of hydrogen bonds and the hydrophobic core in a β-hairpin.[45] Using the fluctuation-dissipation theorem,[46] it is possible to calculate folding rates from these ensembles. More recently, a new method called transition interface sampling[47] introduced an alternate method to calculate transition rates. Since path-sampling methods are very computationally demanding, it is interesting to consider whether one can construct an algorithm that can more efficiently utilize simulation data (*e.g.* folding trajectories) in order to predict folding rates and mechanisms.

8.3.7 Graph-based Methods

Graph-based methods sample configuration space and connect nearby points with weights according to their transition probabilities. From these graphs, it is possible to calculate such properties as most probable path, p_{fold} values[48] as well as to analyse the order in which secondary structures form.[49] However, the graph representation of protein-folding pathways does not solve the sampling problem, but recasts it, and sampling any continuous, high-dimensional space is still a difficult challenge. Previous graph-based methods have sampled configuration space uniformly (*i.e.* choosing conformations at random) or used sampling methods biased towards the native state. Clearly, as the protein size increases, it becomes very difficult to sample the biologically important conformations with random sampling.

8.3.8 Markovian State Model Methods[50–53]

These methods have recently shown promise to allow for an atomically detailed model with quantitative prediction of kinetics. They can take advantage of the benefits of many of the methods above, such as in the generation of initial nodes, as well as build upon the methods of path sampling and graph-based methods to use short paths to predict complex kinetics.

8.4 Validation of Simulation Methodology: Protein Folding Kinetics

To study protein folding kinetics – and especially compare theory to experiment – it is natural to ask which quantities should be compared. The most experimentally accessible quantitative observables of two-state proteins are the folding and unfolding rates from which one can obtain the thermodynamic stability. Thus, it is important to validate any simulation method through quantitative comparison to experiment with proper statistics. As rates and free energies are the natural quantitative experimental measurements, relative or absolute prediction of these quantities is necessary for a direct connection to experiment and a true assessment of theoretical methodology.

8.4.1 Low-viscosity Simulations

We now consider rate predictions made using atomistic potentials based on various approximations of the physics of inter-atomic interactions (including especially solvent-mediated interactions). Caflisch and co-workers have pioneered long atomistic folding simulations using simple, computationally efficient implicit solvent models. By using low (or no) viscosity in their simulations, they accelerate the timescales involved in folding and are able to observe multiple folding transitions in single trajectories. Though not guaranteeing ensemble level convergence, such reversible folding transitions are strong evidence that sampling is sufficient for useful thermodynamic analysis.

For example, two secondary structural motifs were studied by Caflisch *et al.*: the α-helical Y(MEARA)$_6$ peptide,[54] and Beta3s, a three-stranded antiparallel β-sheet.[55] Surprisingly, the helical peptide, which was shown to contain more helical content (and thus helical stability) than the (AAQAA)$_3$ peptide, folded much more slowly at 300 K, with a mean folding time of ~80 ns. For Beta3s, a mean folding time of 31.8 ns was predicted at 360 K, and a following study predicted a folding time of 39 ns at 330 K,[56] both significantly faster than the ~5 μs timescale reported by De Alba *et al.* at lower temperatures.[57] Increased sampling of Beta3s in four additional simulations of length 2.7 μs or greater extended the predicted folding time using this model to ~85 ns at 330 K. Additional simulations were also conducted to study the folding of the Beta3s mutant with the two sets of turn GS residues replaced with PG pairs,[38] with the mutant folding three times faster than Beta3s. These inverse folding times thus remain rather high.

Dynamics at low viscosity helps tackle an important challenge of molecular simulations. It is therefore natural to examine the strengths and weaknesses of

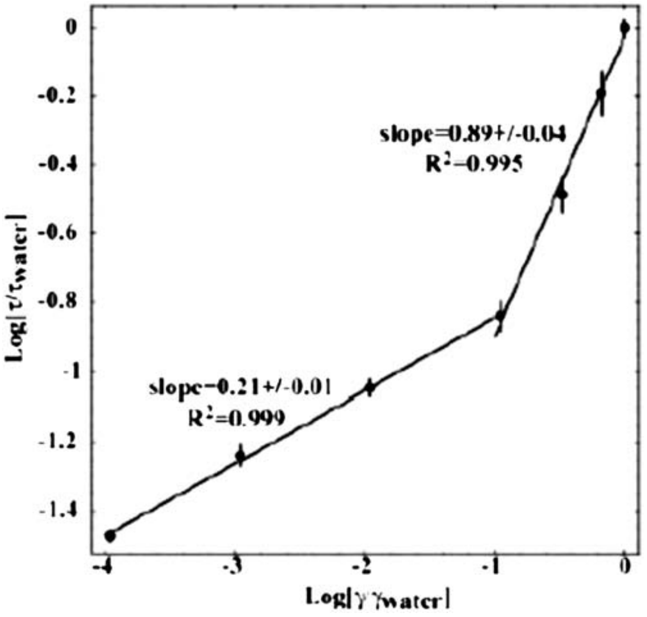

Figure 8.3 Viscosity dependence of the folding time of the Tryptophan Cage molecule in implicit solvent. The folding times and associated errors were calculated using the maximum-likelihood approach. Folding times and viscosities are given relative to the folding time in water and the viscosity of water, respectively. The error bars given are error propagated on the basis of the Cramer–Rao errors for the individual folding times.

this method. A non-linear relationship between folding time and viscosity was reported by Zagrovic *et al.* for the folding kinetics of a 20-residue tryptophan-cage mini-protein in the GB/SA implicit solvent model of Still *et al.*[20] under a range of solvent viscosities.[42] Figure 8.3 plots the observed relationship between inverse rate ($\tau = 1/k$) and viscosity ($1/\gamma$) relative to the case for water-like viscosity (*i.e.* $\gamma_{\mathrm{water}} = 91\ \mathrm{ps}^{-1}$).[58] In the figure it is apparent that linear scaling of the folding time with solvent viscosity holds for viscosities as low as $\sim 1/10$ that of water. However, below this point the folding time scales as $t \sim \gamma^{1/5}$. While applying such scaling rules to the rate predictions of Caflisch and co-workers described above (in low viscosity) would clearly bring their values closer to experimentally established rates for these systems, the precise effect of low viscosity for each of these systems remains unclear.

8.4.2 Estimating Rates with a Two-state Approximation

Including water-like viscosity significantly increases the required sampling time, yet allows absolute folding kinetics to be measured directly. To this end, Pande and co-workers have applied distributed computing to sample trajectory space stochastically and extract rates from an ensemble dynamics (ED) perspective.[21] Two-state behavior is the central concept upon which rates are extracted *via* ED; dwell times in free energy minima of the conformational space are significantly longer than transition times (*i.e.* barrier crossing is much faster than the waiting period). The probability of crossing a barrier separating states A and B by time t is thus given by

$$P(t) = 1 - \mathrm{e}^{-kt} \tag{8.6}$$

where k is the folding rate. In the limit of $t \ll 1/k$, this simplifies to $P(t) \approx kt$ and the folding rate (according to the Poisson distribution) is given by

$$k = \frac{\mathrm{N}_{\mathrm{folded}}}{t \cdot \mathrm{N}_{\mathrm{total}}} \pm \frac{\sqrt{\mathrm{N}_{\mathrm{folded}}}}{t \cdot \mathrm{N}_{\mathrm{total}}} \tag{8.7}$$

For example, if 10 000 simulations are run for 20 ns each and 15 of them cross a given barrier, we obtain a predicted rate of $k = 0.075(\pm 0.019)\,\mu\mathrm{s}^{-1}$, corresponding to a folding time of $13.3(\pm 3.4)\,\mu\mathrm{s}$. In this way, we can use many short trajectories to investigate the folding behavior of polymers that fold on the microsecond timescale: as we've shown previously, using M processors to simulate folding results in an M-times speedup of barrier crossing events.[59] When $t > 1/k$, as is the case for helix formation and other fast processes, ensemble convergence to absolute equilibrium can be established, and the complete kinetics and thermodynamics can be extracted simultaneously.[60]

In several recent studies, Pande and co-workers have utilized implicit solvent models while maintaining water-like viscosity *via* a Langevin or stochastic dynamics integrator with an inverse relaxation time γ. In the first study,[61] they introduced a method of "coupled ensemble dynamics" as a means to simulate the ensemble folding of the C-terminal β-hairpin of Protein G (1GB1) using the GB/SA continuum solvent model of Still *et al.*[20] and the OPLS united atom force field[16] with water-like viscosity. A total sampling time of $\sim 38\,\mu\mathrm{s}$ was obtained, with a calculated inverse folding rate of $4.7(\pm 1.7)\,\mu\mathrm{s}$, in good agreement with the experimentally determined value of $6\,\mu\mathrm{s}$.[62]

Other hairpin structures have been studied by the Pande group more recently, both in an effort to gain insight into hairpin folding dynamics and for a more thorough comparison to experimental measurements. They reported folding and unfolding rates for three Trp zipper β-hairpins[63] using the methodology described above, including TZ1 (PDBID 1LE0), TZ2 (PDBID 1LE1), and TZ3 (PDBID 1LE0 with G6 replaced by D-proline). The relative inverse folding rates are in good agreement with experimental fluorescence and IR measurements provided by experimental collaborators. Unfolding rates were also predicted with relatively strong agreement.

Beyond these investigations of simple hairpin subunits, several small proteins were studied using an implicit solvent methodology. The first, a 20-residue mini-protein known as the Trp cage, was shown to have an experimental folding time of $\sim 4\,\mu\mathrm{s}$. From simulations (totaling $\sim 100\,\mu\mathrm{s}$) the folding rate was estimated based on a cutoff parameter in alpha carbon RMSD space: $k_{fold}(3.0\,\text{Å}) = (1.5\,\mu\text{s})^{-1}$, $k_{fold}(2.8\,\text{Å}) = (3.1\,\mu\text{s})^{-1}$, $k_{fold}(2.7\,\text{Å}) = (5.5\,\mu\text{s})^{-1}$, $k_{fold}(2.6\,\text{Å}) = (6.9\,\mu\text{s})^{-1}$, and $k_{fold}(2.5\,\text{Å}) = (8.7\,\mu\text{s})^{-1}$. While the predicted folding time roughly agreed with the experimental value, the calculations illustrated the dependence of rates upon definition of the native state, as was described above (to minimize this dependence cutoffs must be chosen along an optimal reaction coordinate). Post analysis of ensemble folding data is not necessarily trivial unless many folding events are present and a stable native ensemble is easily distinguished from decoys with similar topology. Similar rate predictions were made for two mutants of the 23-residue BBA5 mini-protein and compared to temperature jump measurements made in the Gruebele laboratory.[64] A single mutation replaced F8 with W, which acts as the fluorescent probe, while the double mutant also included a replacement of V3 with Y. The agreement between simulation predictions and experimental measurements was excellent for the double mutant at 6 μs and 7.5(±3.5) μs respectively. The agreement was less striking in the case of the single mutant, where experiment offered an upper limit of 10 μs and simulation predicted 16 μs, with a range of 7 to 43 μs based on the alpha carbon RMSD cutoff used (still a notably accurate prediction).

One of the most notable simulation studies to date was the *tour-de-force* 1-μs trajectory of the villin headpiece conducted by Duan and Kollman.[26] Following the methods described above, Pande and co-workers have simulated the ensemble folding of this 36-residue three-helix bundle (PDBID 1VII) using the GB/SA continuum solvent and the OPLS united atom force field in water-like viscosity.[65] With over 300 μs of simulation time, the folding time was predicted to be 5 μs (1.5–14 μs using alpha carbon RMSD cutoffs of 2.7–3 Å, as described above), which was compared to the 11-μs folding time derived from NMR lineshape analysis. A follow-up study by Eaton and co-workers tested the prediction using temperature-jump fluorescence and found the folding time to be 4.3(±0.6) μs, thereby validating the rate prediction.

To study the formation of more complex protein structure, Pande and co-workers reported unbiased folding simulations of the 23-residue mini-protein BBA5 in explicit solvent.[66] Ten thousand independent MD simulations of the denatured conformation of BBA5 solvated in TIP3P water resulted in an aggregate simulation time of over 100 μs. This sampling yielded 13 complete folding events which, when corrected for the anomalous diffusion constant of the TIP3P model, results in an estimated folding time of 7.5(±4.2) μs. This is in excellent agreement with the experimental folding time of 7.5(±3.5) μs reported by Gruebele and co-workers.[64]

Folding of the villin headpiece was first attempted by Duan and Kollman in 1998.[26] Using TIP3P explicit solvent, their single 1-μs simulation did not show complete folding, which is not surprising given the ~5-μs folding time for that protein. Pande and co-workers have recently reported folding of this protein using the TIP3P water model and the AMBER-GS force field at 300 K,[67,68] thus increasing the maximum sequence size of proteins for which simulated folding has been observed with MD. With a total sampling time of nearly 1 ms, a folding time of 10(±1.7) μs was predicted using a particle mesh Ewald treatment of long range electrostatics. Identical simulations using a reaction field treatment yielded 9.9(±1.5) μs. These values are somewhat slower than the 4.3(±0.6) experimental folding time, which might be due to the slow equilibration previously observed for helix formation under the AMBER-GS potential.[60]

What are the limitations of this two-state method? The direct observation of folding kinetics presents difficulties, especially for larger proteins or those without single exponential behavior. For example, folding ensembles generated from a single unfolded model attempt to populate the unfolded ensemble *and* observe folding. However, the timescale involved for the initial equilibration and the timescale necessary for chain diffusion across the folding barrier scale dramatically with chain length.[69] These factors make it increasingly difficult to observe both equilibration and folding for large proteins. In addition, Paci *et al.* have shown that folding events in extremely short trajectories can proceed from high-energy initial conformations.[41] Deviations from two-state behavior can also make interpretation of ensemble kinetics difficult,[70] and, given the short timescale of current folding simulations (10–1000 ns), any obligate intermediate with an appreciable dwell time (1–100 ns) may represent a sufficient deviation. In a downhill folding scenario, the principal limitation of the ensemble dynamics approach is the potentially lengthy and temperature-dependent timescale for protein conformational diffusion.[71] Fortunately, these challenges may not be intractable: the timescale for downhill equilibration to a relaxed unfolded ensemble may require long simulations,[72] but should be much faster than folding. Also, the detection of intermediates and multiple pathways can be accomplished by the comparison of folding and unfolding ensembles. Finally, these concerns may also be addressed with new Markovian State Model methods,[51–53,73] described in more detail below.

Regardless of the relatively strong agreement between ensemble simulations in implicit solvent and experimental rate measurements, several factors must be considered in interpreting such simulation results. Lacking a discrete representation of water, these studies ignore the potential role that aqueous solvent might play in the folding process. Furthermore, the compact nature of the relaxed unfolded state ensembles observed using the GB/SA solvent model may pose problems for the folding of larger proteins, such as trapping in compact unfolded conformations.

8.4.3 Markovian State Models (MSMs)

The two-state methods described and applied above work well if there are no intermediate states accumulating on timescales comparable to the trajectory length or longer (*e.g.* greater than 20–100 ns) and if the chains are relatively short (*e.g.* less than 50 residues). However, as one examines the folding of larger and more complex proteins, the two-state approximation will surely eventually break down and reaching even just the relaxation time for a given chain will become a challenge. Also, even if the folding is two state, the simple diffusion of the polymer chain (which scales like the number of residues squared or cubed) will start to require very long trajectories. In anticipation of these problems, we have proposed a new method: Markovian state models.[51–53,73]

Markovian state models transform simulation data gathered from MD trajectories into a kinetic model that includes transition time data. As opposed to traditional transition path sampling analysis,[45,47,74] this method would incorporate all of the simulated data into the results, therefore potentially yielding an increase in efficiency. Our MSM model assumes first-order Markovian transitions between states: simply put, we assume that the next state visited during dynamics will depend solely on the current state and not on previous states visited. Moreover, from an MSM, one can easily calculate any kinetic quantity which can be related to some structural property, such as p_{fold}[75] for all configurations sampled and the mean first passage time (MFPT) from the unfolded state to the folded state. This method also provides a compact representation of the pathways in the system, useful for understanding the mechanisms involved in folding. MSM methods improve on the current graph-based techniques by sampling points using molecular dynamics (MD), thereby greatly increasing the probability that the configurations that are included are kinetically relevant. In addition, the simulation time between points inherently captures transition times, making the direct calculation of folding rates possible.

Early results from MSM methods appear to be promising. Results on a beta hairpin[51] and the villin headpiece and protein A[68] find quantitative agreement with experimental folding times, allowing for a quantitative prediction of timescales considerably longer than the individual trajectories used to construct the MSM. Moreover, these methods do not assume two-state behavior and thus can serve as a test of the two-state approximation; the agreement with two-state behavior in these methods supports employing the two-state method in simple proteins, although it is likely that the two-state approximation will break down for larger, more complex proteins or proteins that have unusual kinetics, such as putative downhill folders.

8.4.4 Other Approaches

While the studies described above offer insight into the most elementary events in protein folding, a number of studies have recently been published on the formation and/or denaturation of larger protein structures. Daggett and coworkers have reported *unfolding* rate predictions using explicit solvent models with direct experimental comparisons. The 61-residue engrailed homeodomain (En-HD) forms a three-helix bundle similar to the villin headpiece and is known to undergo thermal denaturation at 373 K with a half-life predicted by long extrapolation of experimental kinetic data at lower temperatures of 4.5 to 25 ns. Mayor *et al.* simulated the thermally induced unfolding of En-HD using the F3C water model[76] in ENCAD[77] at this temperature with an unfolding rate on the tens of nanoseconds timescale.[10,78] The time needed to reach the putative transition state at 75 and 100 °C, 60 ns and 2 ns respectively, was roughly consistent with the extrapolated experimental unfolding rates (precise rates cannot be extracted from a single unfolding event due to the stochastic nature of protein dynamics).

Bolhuis simulated the folding of the C-terminal β-hairpin of protein G using the transition interface sampling method described above to extract transition kinetics.[45] At 300 K, with an equilibrium constant of ~ 1, the predicted folding time of 5 μs using the TIP3P explicit solvent is in good agreement with the experimental rate of 6 μs[62] as well as the rate predicted by Zagrovic *et al.* using an implicit solvent.[61] The observed agreement suggests that path sampling will be useful in future simulation studies to elucidate the kinetics and mechanisms inherent to protein folding, and it will be interesting to see such methods applied to larger, more complex systems.

Peptides and mini-proteins allow for complete and accurate sampling of folding and unfolding events *via* simulation at biologically relevant temperatures. Pande and co-workers recently studied the helix-coil transition in two 21-residue α-helical sequences and demonstrated complete equilibrium ensemble sampling for multiple variants of the AMBER force field,[11] as shown in Figure 8.4, thus allowing quantitative assessment of the potentials studied. Observing that the previously published AMBER variants resulted in poor equilibrium helix-coil character in comparison to experimental measurements,

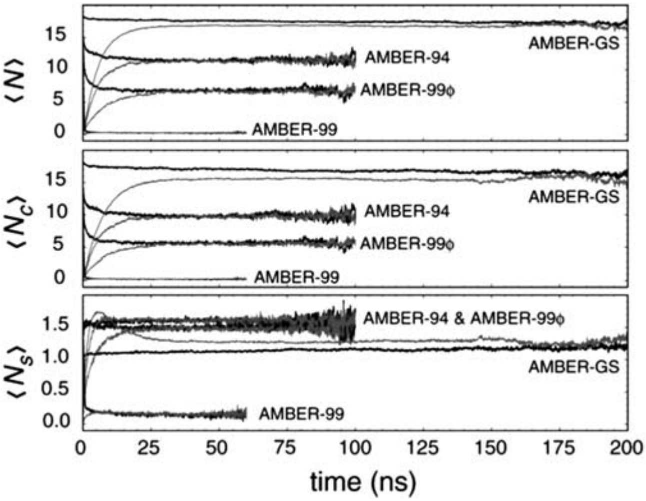

Figure 8.4 Time evolution and convergence of F_s peptide folding ensembles under the AMBER-94, AMBER-GS, AMBER-99, and AMBER-99ϕ potentials. The plots include, from top to bottom, the mean α-helix content, mean contiguous helical length, and mean number of helical segments per conformation according to classical LR counting theory. Native ensembles that converge with corresponding gray folding ensembles are shown in black. Signal noise in the longer time regime is due to fewer simulations reaching that timescale (additional data at long times have been removed for visual clarity).

they tested a new variant denoted AMBER-99ϕ and showed that it more adequately captured the helix-coil dynamics. Based on a multi-state Markovian-based analysis, a primary relaxation time of 151 ns was reported using the more accurate AMBER variant, which agreed well with the 160(±50) ns measured experimentally by Williams *et al.*[79]

Minimalist models have also continued to garner attention recently. It is usually not feasible to obtain direct kinetics information from Go-like models due to difficulty in interpreting the timestep in Go model simulations in terms of a physically measurable quantity. However, it was recently reported that Go model simulations can still be useful in predicting folding timescales of various proteins if the time and temperature are scaled properly to experimental measurements.[80] One caveat in this approach will be the necessity of a rather large training set to obtain a calibration data for such scaling. However, considering the tractability for simulation of large systems using minimalist models, it will be interesting to see whether such an approach can be generally applied for other systems.

8.5 Predicting Protein Folding Pathways

8.5.1 Kinetics Simulations

The folding pathway is arguably the most interesting prediction associated with folding simulations. As our ability to observe long-timescale transitions improves, it becomes increasingly important to clearly communicate the observed mechanism. Qualitative descriptions of the folding pathway can only be loosely interpreted in comparison to experiment. First, as mentioned above, results derived from folding simulations can be sensitive to data analysis. For example, Swope and co-workers produced several folding mechanisms for the hairpin from protein G by varying their hydrogen bond definition.[52,73] Second, there are potential semantic issues; a researcher might frame their discussion of β-hairpin folding in terms of zippering, secondary *versus* tertiary contacts, or diffusion-collision versus nucleation-condensation.

The order of "events" is a natural description of a mechanism, but an optimal description of mechanism should account for heterogeneity as well as the interplay between secondary and tertiary contacts. An excellent and recent example comes from protein A. Fersht and co-workers have qualitatively compared several published simulation predictions of the protein A folding pathway to experiment.[81,82] None of the published atomistic simulations were completely consistent with experiment, emphasizing the need for improved simulation predictions of the folding pathway, and improved quantitative means for comparing pathway predictions.

The collaborative effort between the Fersht experimental laboratory and the Daggett simulation laboratory has shed light on an entire family of unfolding mechanisms. The homeodomains, small three-helix proteins, exhibit a spectrum of folding processes, from concurrent secondary and tertiary structure formation (nucleation-condensation mechanism) to sequential secondary and tertiary formation (framework mechanism).[83] They present putative transition state conformations (two each at 373 and 498 K for En-HD; seven at 498 K for c-Myb; and two at 498 K for hTRF1) from high-temperature unfolding for En-HD, c-Myb, and hTRF1, and estimate β_T values (0.83, 0.83, 0.8 respectively) that roughly agree with the experimental β_T values (0.83, 0.79, 0.90). Excluding the mutation of two charged residues, correlation coefficients of 0.79 and 0.74 for En-HD and c-Myb were obtained between the S and Φ values. Gianni *et al.* report that folding of En-HD resembles the diffusion collision mechanism more than c-Myb and hTRF1 because the helices are nearly fully formed in the transition state. They do state that movements from diffusion-collision to nucleation-condensation are not detected simply by the helical content of the folding transition states but through analysis of whether the secondary and tertiary structures are formed simultaneously.[83] Given this strategy we feel it is particularly important to generate a statistically meaningful number of transitions to judge the relative timing of events between related molecules.

Through the two-state approximation and distributed computing, the Pande laboratory has examined the folding of several small, two-state proteins. The mechanism found varied with the protein studied. It remains to be seen if a more comprehensive mechanistic survey of many small, two-state proteins will reveal underlying mechanistic similarities or model dependencies. In several cases, distributed computing allowed direct comparison of the performances of different force fields. For example, simulations of the C-terminal beta hairpin of protein G[35,61] found that the initial states of folding were the hydrophobic collapse of the small hydrophobic core, followed by formation of hydrogen bonds.

Simulations of a small zinc finger fold (BBA5) found a different mechanism:[64] the secondary structure formed first and then independently collided to form the folded state, analogous to what one would expect from a diffusion-collision model; this is perhaps not surprising in hindsight, considering that BBA5[84] is a *de novo* designed protein and its independent elements may be more stable than in typical proteins. Finally, simulations of the villin headpiece found a different mechanism, in which formation of the rough topology was found early, following by the locking in of the side chains.[65]

It is interesting and important to consider the role of force-field variation in the determination of the folding mechanism. Moreover, beyond the force field used to describe protein-protein interactions, one may also expect variations due to the water model chosen, and differences between minimalist models and more detailed, full atomic models. A natural way to quantitatively examine these differences in mechanism is through a correlation of pfold values.[85] As the pfold value gives a quantitative measure of the location of a given conformation along the folding pathway (pfold near 0 means that the conformation is kinetically close to the unfolded state and pfold close to 1 means it is kinetically close to the folded state), a correlation of pfold values between two different models (force fields, solvent models, *etc.*) and yield a quantitative comparison between the kinetic mechanisms that would be predicted. AQ1

Upon comparing several different types of explicit water models, implicit water models, and minimalist models (all-atom and Cα Go models), Rhee and Pande[85] found that different explicit models yielded quantitatively similar folding mechanisms. Comparing explicit solvent to implicit solvent models found some greater variation, consistent with other types of comparisons between explicit and implicit solvent.[86,87] When comparing to minimalist models, little correlation was found, indicating that for the protein studied (BBA5), minimalist models could not recapitulate the dynamics described by more detailed models and, moreover, minimalist models did not agree with each other (there was a large discrepancy between all-atom and Cα Go models). While it still remains to be seen if these results will hold for larger, more complex proteins (indeed, BBA5 is a small, human-designed protein and thus may be unusual), these results suggest that there may indeed be differences, as well as laying out a quantitative method for making such comparisons in the future.

8.5.2 Thermodynamics Simulations

The success of thermodynamic methods in the prediction of the relevant folding pathways rests on sampling the entire available phase space. This is because the dominant pathways can be correctly identified only when the relative importance of various intermediates are known. Two major bottlenecks naturally emerge for a correct sampling of the vast phase space: the high dimensionality of protein configuration space and the kinetic trapping during simulations. The followings will revisit well-known methods that try to overcome these difficulties.

In the original landscape approach as pioneered by Brooks and co-workers,[6] the free-energy landscape or potential of mean force (PMF) is generated from the equilibrium population distribution. Because it is excessively time consuming to reach equilibrium for high-dimensional protein molecules with conventional molecular dynamics, simulations are performed with umbrella sampling. An additional potential (usually a quadratic or "umbrella" potential) is added to the original Hamiltonian of the system to bias the sampling. By adjusting the bias, the size of the available conformational space can be reduced to expedite the equilibration within the biased Hamiltonian. A series of biased simulations are recombined afterwards to remove the bias in a mathematically strict way using the weighted histogram analysis method.[88] The population distribution P(q) then can be converted to the free energy with $F(q) = -\ln P(q)$. With this approach, Brooks and co-workers have obtained the free-energy landscape and folding dynamics of an α-helical protein (Protein A[89]), an $\alpha\beta$ mixed protein (GB1[90,91]), and a mostly β protein (src-SH3[92]) with numerous successful comparisons to experiment. We refer the reader to an excellent review.[6]

Umbrella sampling studies produce informative free-energy landscapes but assume that degrees of freedom orthogonal to the surface equilibrate quickly. The molecular dynamics time needed for significant chain movement could significantly exceed the length of typical umbrella sampling simulations (which are each typically on the nanosecond timescale). However, in spite of this caveat, umbrella sampling approaches have been very successful. One explanation for this success lies in the choice of initial conditions: umbrella sampling simulations employ initial coordinates provided by high-temperature unfolding trajectories. This is a recurring theme: without lengthy simulations, the initial conformations are crucially important, and it appears that unfolding produces reasonable initial models.

Even though umbrella sampling can expedite the sampling by simulating multiple trajectories at the same time, kinetic trapping or slow orthogonal degrees of freedom may still dominate within each umbrella potential. A number of techniques have been developed to overcome this kinetic trapping. Mitsutake *et al.* have provided an excellent review of these generalized ensemble methods.[93] We will focus on replica exchange molecular dynamics (REMD), which has been widely used in protein-folding simulations. In this approach, a number of simulations ("replicas") are performed in parallel at different temperatures. After a certain time, conformations are exchanged with a Metropolis probability. This criterion ensures that the sampling follows the canonical Boltzmann distribution at each temperature. Kinetic trapping at lower temperatures is avoided by exchanging conformations with higher-temperature replicas. This method is easier to apply than other generalized ensemble methods because it does not require *a priori* knowledge of the population distribution.

After Sugita and Okamoto demonstrated its effectiveness with a gas-phase simulation of the pentapeptide Met-enkephalin,[27] Sanbonmatsu and Garcia obtained the free-energy surface of the same system using explicit water.[28] With 16 parallel replicas they observed enhanced sampling (at least $\sim 5\times$) compared to conventional constant temperature molecular dynamics. Because the method is quite simple and because it is trivially parallelized in low-cost cluster environments, it gained wide application rapidly. Berne and co-workers applied this method to obtain a free-energy landscape for β-hairpin folding in explicit water using 64 replicas with over 4000 atoms.[94] With the equilibrium ensemble and the free-energy landscape in hand, they reported that the β-hairpin population and the hydrogen-bond probability were in agreement with experiments, and proposed that the β strand hydrogen bonds and hydrophobic core form together during the folding pathway.

If care is taken to fully reach equilibrium,[32] REMD becomes powerful for elucidating the folding landscape. For example, Garcia and Onuchic applied the method to a relatively large system, protein A.[29] With 82 replicas for more than 16 000 atoms with temperatures ranging from 277 to 548 K, and with $\sim$13 ns molecular dynamics simulations for each replica, they reported convergence to the equilibrium distribution with quantitative determination of the free-energy barrier of folding.

8.6 Conclusions

In the end, an understanding of complex biophysical phenomena will require computer simulation at some level. Most likely, experimental methods will never yield the level of detail that can be reached even today with computer simulations. However, the great challenge for simulations is to prove their validity. Thus, it is naturally the combination of powerful simulations with quantitative experimental validation that will elucidate the nature of how proteins fold.

How well do protein folding kinetics simulations currently compare with experiment? While prediction of relative rates (*e.g.* demonstrating a correlation between experimental and predicted rates) is valuable, prediction of the absolute rate without free parameters is a more stringent test. Though calculation of absolute rates is computationally demanding, we expect such absolute comparisons to become more common (for increasingly complex proteins) with the advent of new methods and increasing computer power. Finally, we stress that a quantitative prediction of rates is not sufficient to guarantee the validity of a model. The ability of fairly different models to quantitatively predict folding rates strongly suggests that more experimental data are needed to further validate simulation. Additionally, several coarse-grained calculations have been employed to study folding and unfolding rates.[80,95,96]

It is also interesting to look to what's on the near horizon. New advances in computational methods have already enabled single trajectories to reach the microsecond timescale routinely, without using a supercomputer, either by using multi-core PCs[97] or streaming processors, such as Graphics Processing Units (GPUs) or the Cell Processor in PS3s.[98] With microsecond length trajectories, fast-folding proteins can now be examined directly, with thousands of trajectories over multiple microsecond timescales directly enabling a full statistical comparison of kinetics between simulation and experiment.[97] Moreover, recent advances in force fields should allow for a significant increase in accuracy, especially with new advances in polarizable force fields.[99,100] The combination of the more advanced computational methods, with modern polarizable force fields, and the sampling power of Markovian state models should yield a potent combination to accurately predict folding properties on the microsecond to millisecond timescale for small, single-domain proteins in the very near future, and likely beyond to the second timescale in the next decade.

The ability to quantitatively predict rates, free energies, and structure from simulations based on physical force fields reflects significant progress made over the last five years. It also draws attention to a new challenge. Even the prediction of experimental observables, such as rates, within experimental uncertainty does not prove that the simulations will yield correct insight into the mechanism of folding. Indeed, recent work suggests that computational models can both agree with experiment, but disagree with each other.[66] Also, observing that a particular residue appears to participate in a non-native contact does not necessarily imply that mutating this position will accelerate folding; for example, Zagrovic and Pande[65] found non-native interactions in their simulations, but did not predict that removing this would necessarily alter the rate (indeed, the simulations performed could not predict a rate change in this case and thus this result is not necessarily in disagreement with the experiment.[101]

However, these sorts of comparisons greatly underscore the need for direct, quantitative comparison between experiment and theory over a broad range of observables as this is the only way to unambiguously test simulation predictions. We must therefore push the link between simulation and experiment further by connecting the two with new observables, multiple techniques, and increasingly strict quantitative comparison and validation of simulation methods. Without more detailed experiments, we may not be able to sufficiently test current simulation methodology and the trustworthiness of refined simulations may remain unclear. Nonetheless, the ability to predict rates, free energies, and structure of small proteins is a significant advance for simulation, likely heralding even more significant advances over the next five years.

References

1. C. M. Dobson, *Trends Biochem. Sci.*, 1999, **24**(9), 329–232.
2. C. M. Dobson and A. Sali *et al.*, *Angew. Chem. Int. Ed. Engl.*, 1998, **37**, 868–893.
3. V. S. Pande and A. Grosberg *et al.*, *Curr. Opin. Struct. Biol.*, 1998, **8**(1), 68–79.
4. V. Grantcharova and E. J. Alm *et al.*, *Curr. Opin. Struct. Biol.*, 2001, **11**(1), 70–82.
5. M. Levitt, *Nature Str. Biol.*, 2001, **8**, 392–393.
6. J. E. Shea and C. L. Brooks III, *Annu. Rev. Phys. Chem.*, 2001, **52**, 499–535.
7. D. Baker and W. A. Eaton, *Curr. Opin. Struct. Biol.*, 2004, **14**(1), 67–69.
8. H. Abe and N. Go, *Biopolymers*, 1981, **20**, 1013.
9. V. S. Pande and D. S. Rokhsar, *Proc. Natl. Acad. Sci. USA*, 1998, **95**(4), 1490–1494.
10. U. Mayor and N. R. Guydosh *et al.*, *Nature*, 2003, **421**(6925), 863–867.
11. E. J. Sorin and V. S. Pande, *J. Comput. Chem.*, 2005, **26**(7), 682–690.
12. P. J. Steinbach and B. R. Brooks, *J. Comput. Chem.*, 1994, **15**(7), 667–683.
13. I. G. Tironi and R. Sperb *et al.*, *J. Chem. Phys.*, 1995, **102**(13), 5451–5459.
14. W. D. Cornell and P. Cieplak *et al.*, *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.
15. B. R. Brooks and R. E. Bruccoleri *et al.*, *J. Comp. Chem*, 1983, **4**, 187–217.
16. W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1988, **110**, 1657–1666.
17. R. Zhou and X. Huang *et al.*, *Science*, 2004, **305**(10), 1605–1609.
18. E. J. Sorin and Y. M. Rhee *et al.*, *J. Mol. Biol.*, 2006, **356**, 248–256.
19. J. A. Grant and B. T. Pickup *et al.*, *J. Comput. Chem.*, 2000, **22**(6), 608–640.
20. D. Qiu and P. S. Shenkin *et al.*, *J. Phys. Chem. A*, 1997, **101**, 3005–3014.
21. V. S. Pande and I. Baker *et al.*, *Biopolymers*, 2003, **68**, 91–109.
22. M. S. Cheung and A. E. Garcia *et al.*, *Proc. Natl. Acad. Sci. USA*, 2002, **99**(2), 685–690.

23. M. R. Shirts and J. Pitera *et al.*, *J. Chem. Phys.*, 2003.
24. M. R. Shirts and V. S. Pande, *J. Chem. Phys.*, 2005, **122**(13), 134508.
25. V. Tsui and D. A. Case, *Biopolymers*, 2001, **56**, 275–291.
26. Y. Duan and P. A. Kollman, *Science*, 1998, **282**, 740–744.
27. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.
28. K. Y. Sanbonmatsu and A. E. Garcia, *Proteins*, 2002, **46**(2), 225–234.
29. A. E. Garcia and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA*, 2003, **100**(24), 13898–13903.
30. S. Gnanakaran and H. Nymeyer *et al.*, *Curr. Opin. Struct. Biol.*, 2003, **13**(2), 168–174.
31. A. K. Felts and Y. Harano *et al.*, *Proteins*, 2004, **56**(2), 310–321.
32. Y. M. Rhee and V. S. Pande, *Biophys. J.*, 2003, **84**, 775–786.
33. V. Daggett and M. Levitt, *J. Mol. Biol.*, 1993, **232**, 600–619.
34. D. O. Alonso and V. Daggett, *J. Mol. Biol.*, 1995, **247**, 501–520.
35. V. S. Pande and D. S. Rokhsar, *Proc. Natl. Acad. Sci. USA*, 1999, **96**(16), 9062–9067.
36. V. Daggett and A. R. Fersht, *Trends Biochem. Sci.*, 2003, **28**, 18–25.
37. J. Ervin and M. Gruebele, *J. Biol. Phys.*, 2002, **28**(2), 0092–0606.
38. P. Ferrara and A. Caflisch, *J. Mol. Biol.*, 2001, **306**(4), 837–850.
39. J. Gsponer and A. Caflisch, *J. Mol. Biol.*, 2001, **309**(1), 285–298.
40. A. Caflisch, *Trends Biotechnol.*, 2003, **21**(10), 423–425.
41. E. Paci and A. Cavalli *et al.*, *Proc. Natl. Acad. Sci. USA*, 2003, **100**(14), 8217–8222.
42. B. Zagrovic and V. Pande, *J. Comput. Chem.*, 2003, **24**(12), 1432–1436.
43. J. N. Onuchic and Z. Luthey-Schulten *et al.*, *Annu. Rev. Phys. Chem.*, 1997, **48**, 545–600.
44. E. J. Sorin and B. J. Nakatani *et al.*, *J. Mol. Biol.*, 2004, **337**(4), 789–797.
45. P. G. Bolhuis, *Proc. Natl. Acad. Sci. USA*, 2003, **100**(21), 12129–12134.
46. AQ2 D. Chandler, *Introduction to Modern Statistical Mechanics,* 1987.
47. P. G. Bolhuis, *Biophys. J.*, 2005, **88**(1), 50–61.
48. M. S. Apaydin and D. L. Brutlag *et al.*, *J. Comput. Biol.*, 2003, **10**(3–4), 257–281.
49. G. Song and S. Thomas *et al.*, *Proceedings of the Pacific Symposium on Biocomputing*, 2003.
50. G. Hummer, *J. Chem. Phys.*, 2004, **120**(2), 516–523.
51. N. Singhal and C. D. Snow *et al.*, *J. Chem. Phys.*, 2004, **121**(1), 415–425.
52. W. C. Swope and J. W. Pitera *et al.*, *J. Phys. Chem. B*, 2004, **108**(21), 6571–6581.
53. N. Singhal and V. S. Pande, *J. Chem. Phys.*, 2005, **123**(20), 204909.
54. A. Hiltpold and P. Ferrara *et al.*, *J. Phys. Chem. B*, 2000, **104**, 10080–10086.
55. P. Ferrara and A. Caflisch, *Proc. Natl. Acad. Sci USA*, 2000, **97**(20), 10780–10785.
56. A. Cavalli and P. Ferrara *et al.*, *Proteins Struct. Funct. Genet.*, 2002, **47**(3), 305–314.
57. E. De Alba and J. Santoro *et al.*, *Protein Science*, 1999, **8**, 854–865.

58. S. Yun-yu and L. Wang *et al.*, *Mol. Simul.*, 1988, **1**, 369–383.
59. M. R. Shirts and V. S. Pande, *Phys. Rev. Lett.*, 2001, **86**(22), 4983–4987.
60. E. J. Sorin and V. S. Pande, *Biophys. J.*, 2005, **88**(4), 2472–2493.
61. B. Zagrovic and E. J. Sorin *et al.*, *J. Mol. Biol.*, 2001, **313**, 151–169.
62. V. Munoz and P. A. Thompson *et al.*, *Nature*, 1997, **390**(6656), 196–198.
63. C. D. Snow and L. Qiu *et al.*, *Proc. Natl. Acad. Sci. USA*, 2004, **101**(12), 4077–4082.
64. C. Snow and H. Nguyen *et al.*, *Nature*, 2002, **420**, 102–106.
65. B. Zagrovic and C. Snow *et al.*, *J. Mol. Biol.*, 2002, **323**, 927–937.
66. Y. M. Rhee and E. J. Sorin *et al.*, *Proc. Natl. Acad. Sci. USA*, 2004, **101**(17), 6456–6461.
67. G. Jayachandran and V. Vishal *et al.*, *J. Chem. Phys.*, 2006, **124**(16), 164902.
68. G. Jayachandran and V. Vishal *et al.*, *J. Struct. Biol.*, 2007, **157**(3), 491–499.
69. F. Krieger and B. Fierz *et al.*, *J. Mol. Biol.*, 2003, **332**(1), 265–274.
70. A. R. Fersht, *Proc. Natl. Acad. Sci. USA*, 2002, **99**(22), 14122–14125.

AQ3

71. Naganathan *et al.*, 2007.
72. N. J. Marianayagam and N. L. Fawzi *et al.*, *Proc. Natl. Acad. Sci. USA*, 2005, **102**(46), 16684–16689.
73. W. C. Swope and J. W. Pitera *et al.*, *J. Phys. Chem. B*, 2004, **108**, 6582–6594.
74. C. Dellago and P. G. Bolhuis *et al.*, *J. Chem. Phys.*, 1998, **108**, 1964–1977.
75. R. Du and V. S. Pande *et al.*, *J. Chem. Phys.*, 1998, **108**, 334–350.
76. M. Levitt and M. Hirshberg *et al.*, *J. Phys. Chem. B*, 1997, **101**, 5051–5061.
77. M. Levitt, *ENCAD, Energy Calculations and Dynamics*, Palo Alto, CA, Molecular Applications Group, 1990.
78. U. Mayor and C. M. Johnson *et al.*, *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 13518–13522.
79. S. Williams and T. P. Causgrove *et al.*, *Biochemistry*, 1996, **35**, 691–697.
80. L. L. Chavez and J. N. Onuchic *et al.*, *J. Am. Chem. Soc.*, 2004, **126**, 8426–8432.
81. S. Sato and T. L. Religa *et al.*, *Proc. Natl. Acad. Sci. USA*, 2004, **101**(18), 6952–6956.
82. P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, 2004, **101**(18), 6837–6838.
83. S. Gianni and N. R. Guydosh *et al.*, *Proc. Natl. Acad. Sci. USA*, 2003, **100**(23), 13286–13291.
84. J. J. Ottesen and B. Imperiali, *Nat. Struct. Biol.*, 2001, **8**(6), 535–539.
85. Y. M. Rhee and V. S. Pande, *Chem. Phys. Lett.*, 2006, **323**, 66–77.
86. J. Wagoner and N. A. Baker, *J. Comput. Chem.*, 2004, **25**(13), 1623–1629.
87. J. A. Wagoner and N. A. Baker, *Proc. Natl. Acad. Sci. USA*, 2006, **103**(22), 8331–8336.
88. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.*, 1989, **63**(12), 1195–1198.
89. E. M. Boczko and C. L. Brooks III, *Science*, 1995, **269**(5222), 393–396.

90. F. B. Sheinerman and C. L. Brooks, *J. Mol. Biol.*, 1998, **278**, 439–456.
91. F. B. Sheinerman and C. L. Brooks, *Proc. Natl. Acad. Sci. USA*, 1998, **95**(4), 1562–1567.
92. J. E. Shea and J. N. Onuchic *et al.*, *Proc. Natl. Acad. Sci. USA*, 2002, **99**(25), 16064–16068.
93. A. Mitsutake and Y. Sugita *et al.*, *Biopolymers*, 2001, **60**(2), 96–123.
94. R. H. Zhou and B. J. Berne *et al.*, *Proc. Natl. Acad. Sci. USA*, 2001, **98**(26), 14931–14936.
95. V. Munoz and W. A. Eaton, *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 11311–11316.
96. D. N. Ivankov and A. V. Finkelstein, *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 8942–8944.
97. D. L. Ensign and P. M. Kasson *et al.*, *J. Mol. Biol.*, 2007, **374**(3), 806–816.
98. E. Elsen and M. Houston *et al.*, *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, 2006, **188**.
99. P. E. Lopes and G. Lamoureux *et al.*, *J. Phys. Chem. B*, 2007, **111**(11), 2873–2885.
100. M. J. Schnieders and N. A. Baker *et al.*, *J. Chem. Phys.*, 2007, **126**(12), 124114.
101. J. Kubelka and W. A. Eaton *et al.*, *J. Mol. Biol.*, 2003, **329**(4), 625–630.