Using Stochastic Roadmap Simulation to Predict Experimental Quantities in Protein Folding Kinetics: Folding Rates and Phi-Values

Tsung-Han Chiang[*] Mehmet Serkan Apaydin[†] Douglas L. Brutlag[‡]
David Hsu[*] Jean-Claude Latombe[‡]

[*]National University of Singapore, Singapore, 117543, Singapore
[†]Duke University, Durham, NC, 27705, USA
[‡]Stanford University, Stanford, CA 94305, USA

Abstract

This paper presents a new method for studying protein folding kinetics. It uses the recently introduced Stochastic Roadmap Simulation (SRS) method to estimate the transition state ensemble (TSE) and predict the rates and the Φ-values for protein folding. The new method was tested on 16 proteins, whose rates and Φ-values have been determined experimentally. Comparison with experimental data shows that our method estimates the TSE much more accurately than an existing method based on dynamic programming. This improvement leads to better folding-rate predictions. We also compute the mean first passage time of the unfolded states and show that the computed values correlate with experimentally determined folding rates. The results on Φ-value predictions are mixed, possibly due to the simple energy model used in the tests. This is the first time that results obtained from SRS have been compared against a substantial amount of experimental data. The results further validate the SRS method and indicate its potential as a general tool for studying protein folding kinetics.

1 Introduction

Protein folding is a fundamental biological process. Starting out as a long, linear chain of amino acids, a protein molecule remarkably configures itself, or *folds*, into a unique three-dimensional structure, called the *native conformation*, in order to perform vital biological functions. There are two separate, but related problems in protein folding: structure prediction and folding kinetics. In the former problem, we are only interested in predicting the final three-dimensional structure, i.e., the native conformation, attained in the folding process. In the latter problem, we are interested in the folding process itself, e.g., the kinetics and the mechanism of folding. We have at least two important reasons for studying the folding process. First, better understanding of the folding process will help explain why and how proteins misfold and find therapies for debilitating diseases such as Alzheimer's disease or Creutzfeldt-Jakob ("mad cow") disease. Second, this will aid in the development of better algorithms for structure prediction.

In this work, we apply computational methods to study the kinetics of protein folding, specifically, to predict the folding rates and the Φ-values. The folding rate measures how fast a protein evolves from an unfolded state to the native state. The Φ-value measures the extent to which a residue of a protein attains its native state when the protein is in the transition state of the folding process. Performing such computational studies was once very difficult, due to a lack of good models of protein folding, a lack of efficient computational methods to predict experimental quantities based on theoretical models, and a lack of detailed experimental results to validate the predictions. However, important advances have been made in

recent years. On the theoretical side, the energy landscape theory [BOSW95, DC97] offers a global view of protein folding in microscopic details based on statistical physics. It hypothesizes that proteins fold in a multi-dimensional energy funnel by following a myriad of pathways, all leading to the same native conformation. On the experimental side, residue-specific measurements of the folding process (e.g., [IOF95]) provide detailed experimental data to validate theoretical predictions.

Our work takes advantage of these developments. To compute the folding rate and Φ-values of a protein, we first estimate the transition state ensemble (TSE), which is a set of high-energy protein conformations that limits the folding rate. We use the recently introduced *Stochastic Roadmap Simulation* (SRS) method [ABG+02] with a Gō-type energy function proposed in [GFG04]. SRS samples the protein conformational space and builds a directed graph, called a *stochastic conformational roadmap*. The nodes of the roadmap represent sampled protein conformations, and the edges represent transitions between the conformations. The power of such a roadmap derives from its ability to capture the stochastic nature of the folding process by compactly encoding a huge number of folding pathways, each represented as a path in the roadmap. Using the roadmap, we can efficiently compute the folding probability (P_{fold}) [DPG+98] for each sampled conformation in the roadmap and decide which conformations belong to the TSE. We then estimate folding rates and Φ-values using the set of conformations in the TSE.

To test our method, we used 16 proteins with sizes ranging from 56 to 128 residues. They all have folding rates and Φ-values determined experimentally, and have been used as a test suite in earlier work on folding kinetics prediction [GFG04]. We validated the results against the experimental data. The results show that our method predicts folding rates with accuracy better than an existing method based on dynamic programming [GFG04]. In the following, this existing method will be called the DP method, for lack of a better name. More importantly, our method provides a much more discriminating estimate of the TSE: our estimate of the TSE contains less than 10% of all sampled conformations, while the estimate by the DP method contains 85–90%. The more selective estimate better reveals the composition of the TSE and makes our method more suitable for studying the mechanisms of protein folding. We also experimented with an alternative way of estimating folding rates by computing the mean first passage time. For Φ-value prediction, the accuracy of our method varies among the proteins tested. The results are comparable to those obtained with the DP method, but both methods need to be improved in accuracy to be useful in practice.

From a methodology point of view, this is the first time that results based on P_{fold} values computed by SRS were compared against substantial amount of experimental data. Earlier work on SRS compared it with Monte Carlo simulation and showed that SRS is faster by *several orders of magnitude* [ABG+02]. The comparison with experimental data serves as a test of the methodology, and the results further validate the SRS method and indicate its potential as a general tool for studying protein folding kinetics.

The rest of this paper is organized as follows. We start with a brief review of the related work on protein folding kinetics and on the SRS method (Section 2). We then give an overview of our approach (Section 3). In the next three sections, we describe how to use the SRS method to estimate the TSE (Section 4) and predict folding rates (Section 5) and Φ-values (Section 6). Finally, we conclude with our plans for future work (Section 7).

2 Related Work

2.1 Protein Folding Kinetics

There is a large literature on estimating protein folding kinetics computationally. Many approaches have been proposed, but we can only selectively touch on a few important ones here. All-atom molecular dynamics simulation (see [DK01] for a survey) provides detailed information on folding pathways, but it is computationally expensive, even with the help of supercomputers [Tea01] or distributed computer clusters [P+03]. The alternatives include, for example, solving the master equation [CHKB98, WPD04] or es-

timating the TSE [AB99, GFG04]. For proteins with simple folding kinetics, a significant correlation was observed between the folding rate of a protein and its native-conformation topology, in particular, the *contact order* [PSB98], and this led to the belief that the fundamental physics underlying protein folding may be relatively simple [Bak00].

Recently, several related methods succeeded in predicting folding rates and Φ-values [AB99, GFG04, ME99], using simplified energy functions that depend only on the native-conformation topology of a protein. Our work also uses such an energy function, but instead of searching for rate-limiting "barriers" on the energy landscape as in [AB99, GFG04], we estimate the TSE by using SRS to compute P_{fold} values and then estimate the folding rates and Φ-values based on the energy of conformations in the TSE.

2.2 Probabilistic Motion Planning and Molecular Motion

SRS is inspired by the probabilistic roadmap (PRM) methods [CLH+05], which have been highly successful for motion planning of robots with many degrees of freedom, a provably hard computational problem [Rei79]. In motion planning, the goal is to find a path for a robot to move from a start configuration to an end configuration without colliding with any obstacles. The main idea of PRM methods is to sample at random the space of all robot configurations—a space conceptually similar to a protein conformation space—and construct a graph, called a *probabilistic roadmap*, that captures the connectivity of this space. Every path in this graph represents a collision-free sequence of motions for the robot to move between the configurations corresponding to the endpoints of this path.

For molecular motion, similar roadmap graphs can be constructed to capture transitions between molecular conformations. Singh et al. introduced the PRM methods to the study of molecular motion in their work on ligand-protein binding [SLB99]. This approach has since been applied and adapted to study various aspects of protein folding, including energy profiling along dominant folding pathways [ADS02, ASBL01, SA01], the formation order of secondary structure elements [ADS02], and P_{fold} calculation [ABG+02]. It has also been used to build approximation of the space of collision-free conformations for protein loops [CSRST04] and to study RNA folding [TKT+04].

Most of the earlier work [ADS02, ASBL01, SLB99, SA01] treats the roadmap as a deterministic graph, with heuristic edge weights based on the energy difference between molecular conformations. The heuristic edge weights measure the energetic difficulty of transiting along the edges of the roadmap. Graph search techniques are then used to extract "low-energy" paths from the roadmap. These methods focus on only one or a few hypothesized important pathways and ignore all the rest. SRS is fundamentally different: a stochastic conformational roadmap is in essence a Markov chain model that captures the stochastic nature of molecular motion. It enables a global analysis of all the pathways contained in a roadmap, using tools from the Markov chain theory. It also provides a formal relationship between SRS and the well-established Monte Carlo method. Such a Markov chain model can also be combined with information from molecular dynamics simulation to provide details of protein folding at the atomistic level [RSGC05, SSP04].

In our earlier work, we used SRS to study protein folding, but the results were compared only with those obtained from Monte Carlo simulation. Here, we extend the work to compute folding rates and Φ-values and validate the results directly against experimental data.

3 Overview

The *conformation* of a protein is a set of parameters that uniquely specify the structure of the protein, e.g., the backbone torsional angles ϕ and ψ. The *conformational space* $\mathcal{C}$ contains all the conformations of a protein. If $\mathcal{C}$ is parametrized by d conformational parameters, then a conformation can be regarded as a point in a d-dimensional space.

Each conformation q of a protein has an associated energy value $E(q)$, determined by the interactions between the atoms of the protein and between the protein and the surrounding medium, e.g., the van der Waals and electrostatic forces. The energy E is a function defined over $\mathcal{C}$ and is often called the *energy landscape*. According to the energy landscape theory, proteins fold along many pathways in $\mathcal{C}$. These pathways start from unfolded conformations and all lead to the same native conformation.

To understand protein folding kinetics, we need to analyze the folding pathways and identify those conformations, called the *transition state ensemble* (TSE), that act as barriers on the energy landscape and limit the folding rate. In the simple case where there is a dominant folding pathway with a single major energy peak along the pathway, the TSE can be defined as the conformations with energy at or near the peak value. In general, there may be many pathways, and along every pathway, there may be multiple energy peaks. This makes the TSE more difficult to identify. To address this issue, Du et al. introduced the notion of $\mathrm{P_{fold}}$ [DPG+98]. In a folding process, the $\mathrm{P_{fold}}$ value of a conformation q is defined as the probability of a protein reaching the native (folded) conformation before reaching an unfolded conformation, starting from conformation q. $\mathrm{P_{fold}}$ measures the kinetic distance between q and the native conformation. From any conformation q with $\mathrm{P_{fold}}$ value greater than 0.5, the protein is more likely to fold first than to unfold first, thus q is kinetically closer to the native conformation. The TSE is defined as the set of conformations with $\mathrm{P_{fold}}$ equal to 0.5. Defining the TSE using $\mathrm{P_{fold}}$ has many advantages. In particular, $\mathrm{P_{fold}}$ is not determined by any specific pathway, but depends on all the pathways from unfolded conformations to the native conformation. It thus captures the ensemble behavior of folding.

We can compute the $\mathrm{P_{fold}}$ value for a conformation q by performing many folding simulation runs from q and count the number of times that they reach the native conformation before an unfolded one. However, a large number of simulation runs are needed to estimate the $\mathrm{P_{fold}}$ value accurately, and doing so for many conformations incurs prohibitive computational cost. The SRS method approximates the $\mathrm{P_{fold}}$ values for many conformations simultaneously in a much more efficient way. In the following, we first describe the computation of the TSE using SRS (Section 4) and then the computation of folding rates (Section 5) and Φ-values (Section 6) based on the energy of conformations in the TSE.

4 Estimating the TSE through Stochastic Roadmap Simulation

SRS is an efficient method for exploring protein folding kinetics by examining many folding pathways simultaneously. We use SRS to compute $\mathrm{P_{fold}}$ values and then determine the TSE based on the computed $\mathrm{P_{fold}}$ values.

4.1 A Simplified Folding Model

To study protein folding kinetics, we need an energy function that accurately models the interactions within a protein and the interactions between a protein and the surrounding medium at the atomic level. For this, we use the simple, but effective energy model developed by Garbuzynskiy et al. [GFG04]. This model is based on the topology of a protein's native conformation. An important concept here is that of *native contact*. Two atoms are considered to be in contact if the distance between them is within a suitably chosen threshold. A native contact between two atoms of a protein is a contact that exists in the native conformation. Given a conformation q, we can obtain all the native contacts in q by computing the pairwise distances between the atoms of the protein.

The energy model that we use divides a protein into contiguous segments of five residues each. Each segment must be either folded or unfolded completely. In other words, atoms within a folded segment must gain all their native contacts with other atoms in folded segments, while atoms within an unfolded segment are assumed to form a disordered loop and lose all their native contacts. We thus represent the conformation

of a protein by a binary vector, with 1 representing a folded segment and 0 representing an unfolded segment. In particular, the native conformation is $(1, 1, \ldots, 1)$, and the unfolded conformation is $(0, 0, \ldots, 0)$.

Using this representation, a protein with N residues has $2^{\lceil N/5 \rceil}$ distinct conformations. To further reduce computation time, Garbuzynskiy et al. suggested a restriction which accepts only conformations with at most two unfolded regions in the middle of a protein plus two unfolded regions at the ends of the protein, where a region is defined as a sequence of contiguous five-residue segments. With a maximum of four unfolded regions, we can capture the folding and unfolding of proteins with up to roughly 100 residues [GFG04].

The free energy of a conformation q is calculated based on the number of native contacts and the length of unfolded segments in q:

$$E(q) = \varepsilon \cdot n(q) - T \cdot (2.3R \cdot \mu(q) + S(q)) \,. \tag{1}$$

In the formula above, $n(q)$ is the number of native contacts in the folded segments of q, $\mu(q)$ is the number of residues in the unfolded segments of q, and $S(q)$ is the entropy for closing the disordered loops. For the rest, which are all constants, ε is the energy of a single native contact, T is the absolute temperature, and R is the gas constant. A similar energy function has been used in the work of Alm and Baker [AB99].

Our model uses all the atoms of a protein, including the hydrogen atoms, to calculate the energy. For protein structures determined by X-ray crystallography, hydrogen atoms are missing and we added them using the Insight II program at pH level 7.0.

4.2 Constructing the Stochastic Conformational Roadmap

A stochastic conformational roadmap G is a directed graph. Each node of G represents a conformation of a protein. Each directed edge from a node q_i to a node q_j carries a weight P_{ij}, which represents the probability for a protein to transit from q_i to q_j. If there is no edge from q_i to q_j, the probability P_{ij} is 0; otherwise, P_{ij} depends on the energy difference between q_i and q_j, $\Delta E_{ij} = E(q_j) - E(q_i)$.

The transition probability P_{ij} is defined according to the Metropolis criterion, which is also used in Monte Carlo simulation:

$$P_{ij} = \begin{cases} (1/n_i) \exp(-\frac{\Delta E_{ij}}{RT}) & \text{if } \Delta E_{ij} > 0 \\ 1/n_i & \text{otherwise} \end{cases} , \tag{2}$$

where n_i is the number of outgoing edges of q_i, R is the gas constant, and T is the absolute temperature. The factor $1/n_i$ normalizes the effect that different nodes may have different numbers of outgoing edges. We also assign the self-transition probability:

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}, \tag{3}$$

which ensures that the transition probabilities from any node sums to 1.

SRS views protein folding as a random walk on the roadmap graph. If q_{U} and q_{F} are the two roadmap nodes representing the unfolded and the native conformation, respectively, every path in the roadmap from q_{U} to q_{F} represents a potential folding pathway. Thus, a roadmap compactly encodes an exponential number of folding pathways.

To construct the roadmap G using the folding model described in Section 4.1, we enumerate the set of all allowable conformations in the model (with the restriction of a maximum of four unfolded regions) and use them as the nodes of G. There is an edge between two nodes if the corresponding conformations differ by exactly one folded or unfolded segment.

4.3 Computing P_{fold}

P_{fold} measures the kinetic distance between a conformation q and the native conformation q_F. The main advantage of using P_{fold} to measure the progress of protein folding is that it takes into account all folding pathways sampled from the protein conformation space and does not assume any particular pathway *a priori*.

Recall that the P_{fold} value τ of a conformation q is defined as the probability of a protein reaching a native conformation q_F before reaching an unfolded conformation q_U, starting from q. Instead of computing τ by brute force through many Monte Carlo simulation runs, we construct a stochastic conformational roadmap and apply the first step analysis [TK94]. Let us consider what happens after a single step of transition:

- We may reach a node in a native conformation, which, by definition, has P_{fold} value 1.
- We may reach a node in an unfolded conformation, which has P_{fold} value 0.
- Finally, we may reach an intermediate node q_j with P_{fold} value τ_j.

The first step analysis conditions on the first transition and gives the following relationship among the P_{fold} values:

$$\tau_i = \sum_{q_j \in \{q_F\}} P_{ij} \cdot 1 + \sum_{q_j \in \{q_U\}} P_{ij} \cdot 0 + \sum_{q_j \notin \{q_F, q_U\}} P_{ij} \cdot \tau_j, \tag{4}$$

where τ_i is the P_{fold} value for node q_i. In our simple folding model, both the native and the unfolded conformation contains only a single conformation, but in general, they may contain multiple conformations.

The relationship in (4) gives a linear equation for each unknown τ_i. The resulting linear system is sparse and can be solved efficiently using iterative methods [ABG+02].

The largest protein that we tested has 128 residues, resulting in a total of 314,000 allowable conformations. It took SRS only about a minute to compute P_{fold} values for all the conformations on a PC workstation with a 1.5GHz Itanium2 processor and 8GB of memory.

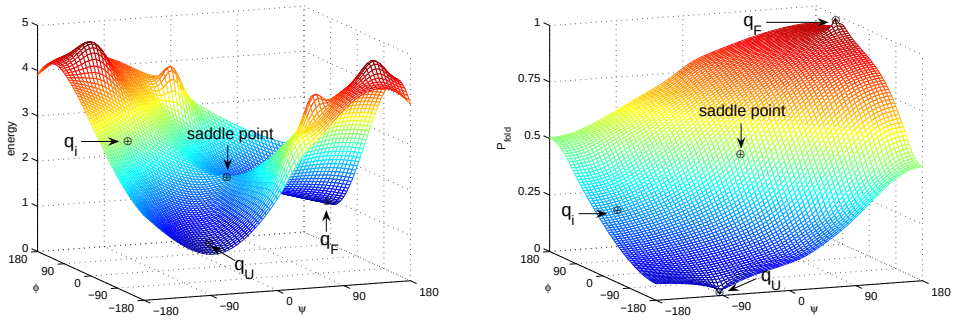
4.4 Estimating the TSE

After computing the P_{fold} value for each conformation, we identify the TSE by extracting all conformations with P_{fold} value 0.5. However, due to the simplification and discretization used in our folding model, we need to broaden our selection criteria slightly and identify the TSE as the set of conformations with P_{fold} values within a small range centered around 0.5. We found that the range between 0.45 to 0.55 is usually adequate to account for the model inaccuracy in our tests, and we used it in all the results reported below.

4.5 An Example on a Synthetic Energy Landscape

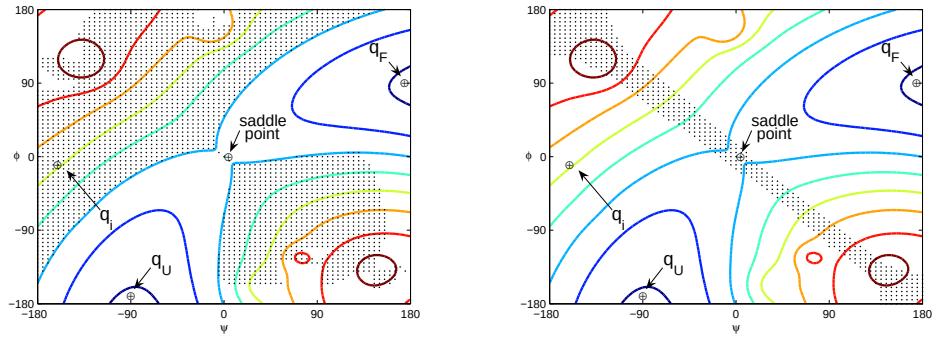
Consider a tiny fictitious protein with only two residues. Ignoring the side-chains, we can specify its conformation by two backbone torsional angles ϕ and ψ. For the purpose of illustration, instead of using the simplified energy function described in Section 4.1, this example uses a saddle-shaped energy function over a two-dimensional conformation space (Figure 1*a*) in which the two torsional angles vary continuously over their respective ranges. On this energy landscape, almost all intermediate conformations have energy levels at least as high as the unfolded conformation q_U and the native conformation q_F. This synthetic energy landscape is conceptually similar to more realistic energy models commonly used. Namely, to go from q_U to q_F, a protein must pass through energy barriers.

The computed P_{fold} values for this energy landscape is shown in Figure 1*b*. A comparison of the two plots in Figure 1 shows that the conformations with P_{fold} value 0.5 correspond well with the energy barrier that separates q_U and q_F.



(*a*) (*b*)

Figure 1. P_{fold} values for a synthetic energy landscape. (*a*) A synthetic energy landscape. (*b*) The computed P_{fold} values.



(*a*) (*b*)

Figure 2. Estimation of the TSE for the energy landscape shown in Figure 1. The conformation-space region corresponding to the TSE is shaded and overlaid on the contour plot of the energy landscape. (*a*) The DP method. (*b*) The SRS method.

5 Predicting Folding Rates

The folding rate is an experimentally measurable quantity that determines how fast the protein proceeds from the unfolded conformation to the native conformation. By observing how it varies under different experimental conditions, we can gain an understanding of the important factors that influence the folding process.

The speed at which a protein folds depends exponentially on the height of the energy barrier that must be overcome during the folding process. The higher the barrier, the harder it is for the unfolded protein to reach the native conformation and the slower the process. Because of the exponential dependence, even a small difference in the height of the energy barrier has significant effect on the folding rate. Therefore, accurately identifying the TSE is crucial when predicting the folding rate.

5.1 Methods

After identifying the TSE using the SRS method described in the previous section, we compute the folding rate in the same way as that in [GFG04]. First, we calculate E_{TSE}, the total energy of the TSE, according to

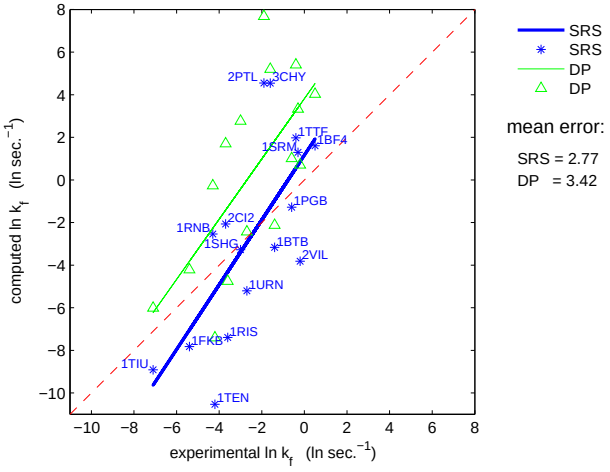


Figure 3. Predicted folding rates versus the experimentally measured folding rates.

the following equation [GFG04]:

$$\exp(-\frac{E_{\rm TSE}}{RT}) = \sum_{q \in {\rm TSE}} \exp(-\frac{E(q)}{RT}), \tag{5}$$

where the summation is taken over the set of all conformations in the TSE, R is the gas constant and T is the absolute temperature. We then compute the rate constant $k_{\rm f}$ according to the following theoretical dependence [GFG04]:

$$\ln(k_{\rm f}) = \ln(10^8) - (\frac{E_{\rm TSE}}{RT} - \frac{E(q_{\rm U})}{RT}), \tag{6}$$

where $E(q_{\rm U})$ is the energy of the $q_{\rm U}$.

5.2 Results

Using data from the Protein Data Bank (PDB), we computed folding rates for 16 proteins (see Appendix A for the list). The results are shown in Figure 3. The horizontal axis of the chart corresponds to the experimentally measured folding rates (see [GFG04] for the sources of data), and the vertical axis corresponds to the predicted values. The best-fit lines of the data are also shown. For comparison, we also computed the folding rates using the DP method [GFG04] and show the results in the same chart. Note that since the chart plots $\ln k_{\rm f}$, it basically compares the height of the energy barrier.

Figure 3 shows that both methods can predict the trend reasonably well. The best-fit line of SRS is closer to the diagonal, indicating better predictions. This is confirmed by comparing the average error in $\ln k_{\rm f}$ for the two methods.

We also examined the effect of the simplying assumptions made in the energy model proposed in [GFG04]. If a protein is divided into contiguous segments of 4 residues instead of 5 (see Section 4.1), the average error for SRS improves slightly to 2.35. However, the smaller error comes at the cost of increased computational time by 5 to 10 times, depending on the size of the protein. If a protein is divided into segments of 6 residues, the average error remains roughly the same at 2.74 for the set of proteins tested, and the computational time is reduced by 3 to 5 times. We also tried to remove the restriction of 4 unfolded regions. This has almost no effect on the computed folding rates and confirms that the simplification is reasonable for the protein size considered here.

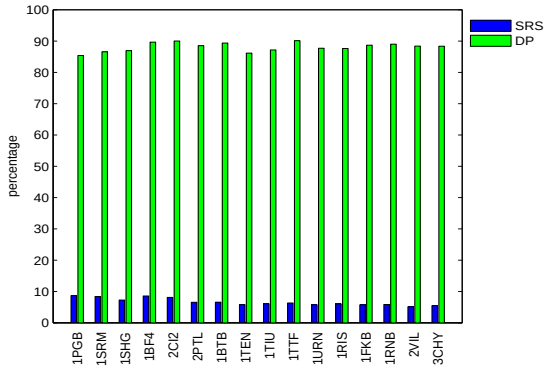


Figure 4. The percentage of conformations in the TSE.

5.3 Accuracy in Estimating the TSE

It is interesting to note in Figure 3 that DP consistently predicts higher k_{f} compared to SRS. Since a higher k_{f} corresponds to lower energy barrier along the folding pathway, the TSE identified by DP must have lower energy. This is significant in terms of the accuracy of folding rate prediction and suggests that an important difference exists between the TSE estimated by SRS and that estimated by DP.

The difference between SRS and DP in estimating the TSE becomes more apparent when we compare the percentage of sampled conformations that belong to the TSE. Figure 4 shows that the TSE estimated by SRS includes less than 10% of all allowable conformations. In contrast, the TSE estimated by DP includes, surprisingly, 85-90%. Closer inspection reveals that the TSE computed by SRS is mostly a subset of the TSE computed by DP. Combining this observation with the better prediction accuracy of SRS, we conclude that the additional 80% or so conformations identified by DP are not only unnecessary, but also negatively affect folding rate prediction.

Although it is difficult to know the true percentage of conformations that should belong to the TSE, careful examination of the DP method shows that it indeed may include in the TSE many conformations that are suspicious. This is best illustrated using the example in Figure 1*a*. According to the DP method, a conformation q belongs to the TSE, if q has the highest energy along the folding pathway that has the lowest energy barrier among all pathways that go through q. This definition tries to capture the intuition that q is the location of minimum barrier on the energy landscape. For the energy landscape shown in Figure 1, the globally lowest energy barrier is clearly the conformation q_{s} at the saddle point. So q_{s} belongs to the TSE. For any other conformation q, there are two possibilities. When $E(q) < E(q_{\mathrm{s}})$, any path through q must have a barrier higher than or equal to $E(q_{\mathrm{s}})$, and q cannot possibly achieve the highest energy along the path. Thus, q does not belong to the TSE. The problem arises when $E(q) \geq E(q_{\mathrm{s}})$. In this case, to place q in the TSE, all it takes is to find a path that goes through q and does not pass through any other conformation with energy higher than $E(q)$. This can be easily accomplished on the saddle-shaped energy landscape for most conformations with $E(q) \geq E(q_{\mathrm{s}})$, e.g., the conformation q_i indicated in Figure 1. Including such conformations in the TSE seems counter-intuitive, as they do not constitute a barrier on the energy landscape.

As we have seen in Section 4.5, the SRS method includes in the TSE only those conformations near the barrier of the energy landscape, but the DP method includes many additional conformations, some of which are far below the energy of the barrier (see Figure 2 for an illustration). Therefore, the TSE estimated by DP tends to have lower energy than the TSE estimated by SRS, resulting in over-estimated folding rates.

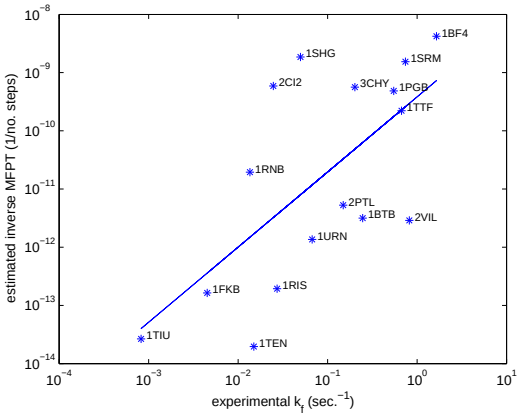


Figure 5. Estimated inverse mean first passage time versus experimentally measured folding rates (k_{f}). For the entire test set of 16 proteins, the correlation is 0.73. The line indicates the best-fit line against data.

5.4 Mean First Passage Time

An alternative way of estimating the folding rate is to compute the mean first passage time (MFPT), a quantity inversely proportional to the folding rate. To do this, we can perform many simulation runs. Each simulation starts from an unfolded conformation and stops as soon as it reaches the native conformation. We count the total number of steps that the simulation takes and use its average over all the simulation runs as an estimate of MFPT.

We can easily perform such simulations on a stochastic conformational roadmap by running many independent random walks starting from the node q_{U}, using the edge transition probabilities to decide which node to go next, and stopping when a random walk first reaches the node q_{F}. The average length of the random walks then serves as an estimate of MFPT. Instead of explicitly running many simulations, which are computationally expensive, we can use the SRS framework to compute the average number of steps s_i that it takes to reach the native conformation from any arbitrary node q_i in the roadmap. We again use the first step analysis for this and establish a system of linear equations for s_i, similar to (4):

$$s_i = 1 + \sum_{q_j \in \{q_{\mathrm{F}}\}} P_{ij} \cdot 0 + \sum_{q_j \notin \{q_{\mathrm{F}}\}} P_{ij} \cdot s_j \qquad \text{for every } q_i. \tag{7}$$

Solving this linear system gives s_i for every q_i, including that for q_{U}, i.e., the estimate for MFPT.

Using (7), we estimated the MFPT for the same 16 proteins tested in Section 5.2. The results are plotted in Figure 5 against experimentally measured folding rates. For the entire test set of 16 proteins, the correlation is 0.73. By removing a single outlier, the correlation improves substantially to 0.83. In a related study [GGF05], Monte Carlo simulation was used to estimate MFPT. Monte Carlo simulation has some potential advantages over SRS, as it does not require restricting the number of conformations by, for example, dividing a protein into contiguous segments of five residues that fold or unfold together. However, due to the high computational cost of Monte Carlo simulation, the number of simulation runs had to be limited to 50, and each run had a cutoff of 10^8 steps [GGF05]. In terms of the correlation with experimental data, our results are comparable to those based on Monte Carlo simulations. However, Monte Carlo simulations were able to finish on only 10 proteins in the test set [GGF05], whereas SRS computed the MFPT for all 16 proteins in about 40 minutes. The 6 additional proteins tend to have longer MFPT than the rest. This confirms that they are indeed difficult for Monte Carlo simulation and further demonstrates the computational advantage of SRS.

The folding rate predictions based on MFPT are comparable to those based on the free energy of the TSE in accuracy, but are slightly weaker. The MFPT method does not compute the TSE, and thus does not

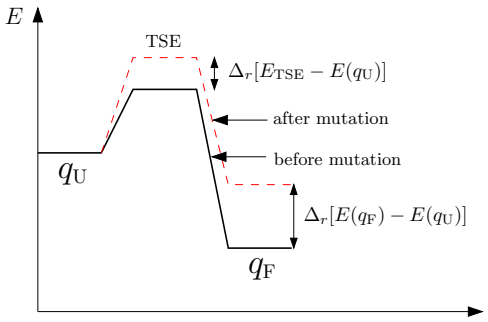


Figure 6. Φ-value.

require us to manually set a P_{fold} range to delineate the TSE, as the TSE method does The MFPT method is also potentially more effective in handling complex folding mechanisms with multiple transition states. On the other hand, while the TSE method gives a direct estimate of the folding rate, MFPT is measured in the number of simulation steps. One way of converting this quantity to a measure of folding rate is to calibrate against the experimentally measured average time that it takes for a single amino acid to fold or unfold (see [IF01]).

6 Predicting Φ-values

Φ-value analysis is the only experimental method for determining the transition-state structure of a protein at the resolution of individual residues [Fer99]. Its main idea is to mutate carefully selected residues of a protein, measure the resulting energy changes, and infer from them the structure of the protein in the transition state. Here, we would like to predict Φ-values computationally.

6.1 Methods

The Φ-value indicates the extent to which a residue has attained the native conformation when the protein is in the transition state of the folding process. More precisely, the Φ-value of a residue r is defined as

$$\Phi_r = \frac{\Delta_r[E_{\mathrm{TSE}} - E(q_{\mathrm{U}})]}{\Delta_r[E(q_{\mathrm{F}}) - E(q_{\mathrm{U}})]}, \tag{8}$$

where $\Delta_r[E_{\mathrm{TSE}} - E(q_{\mathrm{U}})]$ is the change in the free energy difference between the TSE and the unfolded conformation q_{U} as a result of mutating r. Similarly, $\Delta_r[E(q_{\mathrm{F}}) - E(q_{\mathrm{U}})]$ is the mutation-induced change in the free energy difference between the native conformation q_{F} and the unfolded conformation q_{U}. See Figure 6 for an illustration. A Φ-value of 1 indicates that the mutation of residue r affects the free energy of the transition state as much as the free energy of the native conformation, relative to the free energy of the unfolded conformation. So, in the transition state, r must have fully attained the native conformation, according to free energy considerations. Similarly, a Φ-value of 0 indicates that in the transition state, the residue remains unfolded. A fractional Φ-value between 0 and 1 indicates that the residue has only partially attained its native conformation. By analyzing the Φ-value of each residue of a protein, we can elucidate the structure of the TSE.

Using (1) and (5), we can simplify (8) and obtain the following expression for the Φ-value of residue r:

$$\Phi_r = \frac{\sum_{q \in \mathrm{TSE}} P(q) \cdot \Delta_r n(q)}{\Delta_r n(q_{\mathrm{F}})}, \tag{9}$$

where $P(q)$ is the Boltzmann probability for conformation q and $\Delta_r n(q)$ is the change in the number of native contacts for conformation q as a result of mutating r.

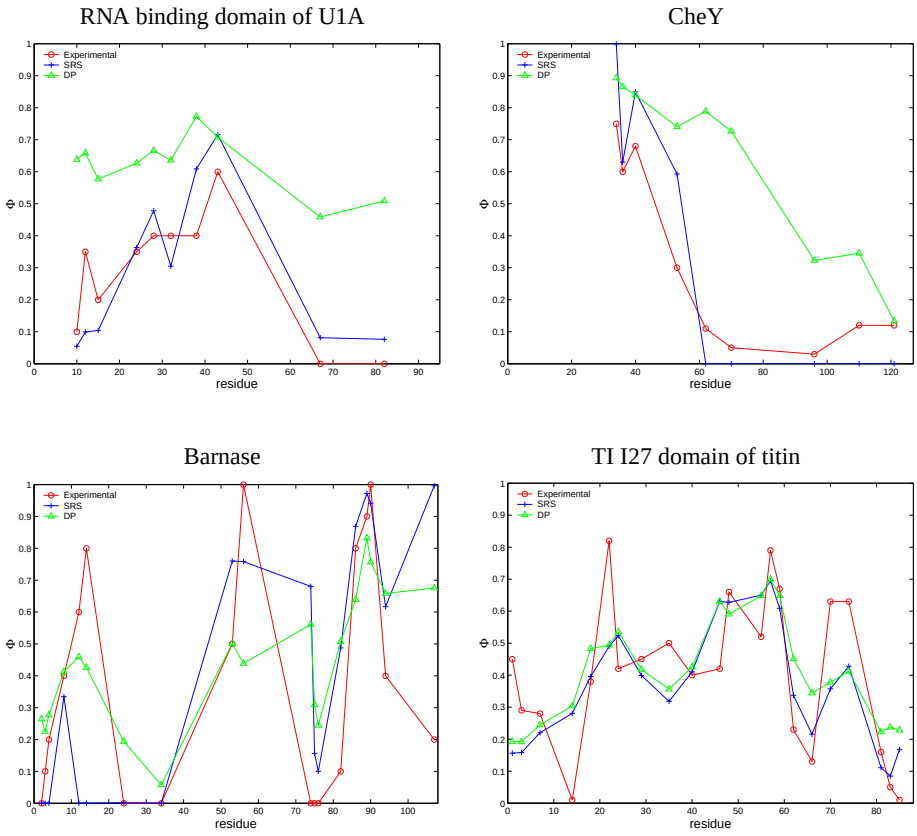


Figure 7. Φ-value predictions for four proteins.

6.2 Results on Φ-value Prediction

The Φ-value is more difficult to predict than the folding rate, because it is a detailed experimental quantity and requires an accurate energy model for prediction. We computed Φ-values for 16 proteins listed in Appendix A, but got mixed results. Figure 7 shows a comparison of the Φ-values computed by SRS and DP and the Φ-values measured experimentally. The sources of the experimental data are available in [GFG04]. In general, our Φ-value predictions based on X-ray crystallography structures are better than those based on NMR structures. When compared with DP, SRS is much better for some proteins, such as CheY and the RNA binding domain of U1A, both of which have X-ray crystallography structures. For the other proteins, the results are mixed. In some cases (e.g., barnase), our results are slightly better, and in others (e.g., TI I27 domain of titin), slightly worse. Table 1 shows the performance of SRS and DP over the 16 proteins tested. Since Φ-values range between 0 and 1, the errors are fairly large for both SRS and DP. To be useful in practice, more research is needed for both methods.

Again we looked at the effect of the simplifying assumptions made in the model, as we did for the folding rate computation. They have little effect (less than 5%) on the results.

6.3 Results on the Progress of Native Structure Formation

An important advantage of using P_{fold} as a measure of the progress of folding is that P_{fold} takes into account all sampled folding pathways and is not biased towards any specific one. We have seen how to use P_{fold} to estimate Φ-values, which give an indication of the extent of folding in the transition state only. We can extend this method to observe the details of the folding process, in particular, the progress of native structure

Table 1. Performance of SRS and DP in Φ-value prediction. For each protein, the average error of computed Φ-values is calculated. The table reports the mean, the minimum, and the maximum of average errors over the 16 proteins tested.

Method	Mean	Min	Max
SRS	0.21	0.11	0.32
DP	0.24	0.13	0.35

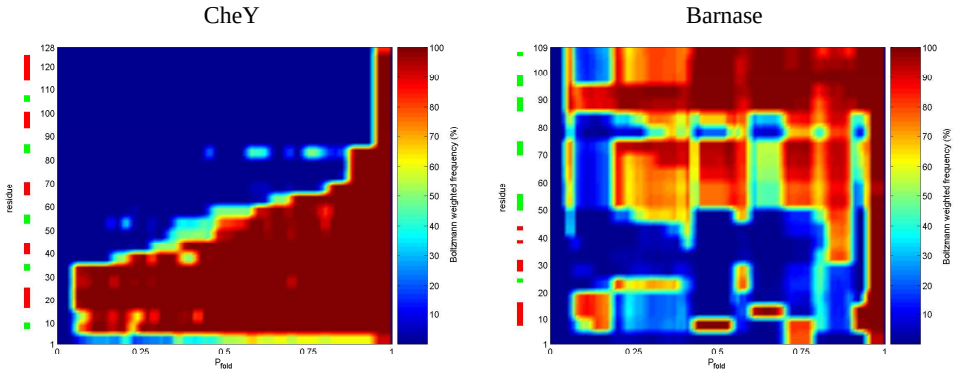


Figure 8. Progress of native structure formation. The colored bar on the left of each plot indicates secondary structures, red for helices and green for strands.

formation, by plotting the progression of each residue with respect to P_{fold}.

Each plot in Figure 8 shows the frequency with which a residue achieves its native conformation in a Boltzmann weighted ensemble of conformations with approximately same P_{fold} values. For CheY, residues 1 to 40 gain their native conformation very early in the folding process. The coherent interactions between neighboring residues is consistent with the mainly helical secondary structure of these residues. Residues 50 to 80 are subsequently involved in the folding nucleus as folding progresses. The folding of barnase is more cooperative and involves many regions of the protein simultaneously. Residues 50 to 109 dominate the folding process early on, and the simultaneous progress of different regions corresponds to the formation of the β sheet. The helical residues 1 to 50 gain native conformation very late in the folding. The progress of native structure formation that we observed is consistent with that obtained by Alm et al. [AB99].

The accuracy of Φ-value prediction gives an indication of the reliability of such plots. We made similar plots for the other proteins. Although we were able to see interesting trends for some of the other proteins, the plots are not shown here, because of the low correlation of their Φ-value predictions to experimental values. Verifying the accuracy of such plots directly is difficult, due to the limited observability of the protein folding process and the limited experimental data available. In the future, we plan to derive quantitative information on the order of secondary structure formation (see, e.g., [ADS02]) in order to allow better comparison with experimental data.

7 Discussion

This paper presents a new method for studying protein folding kinetics. It uses the Stochastic Roadmap Simulation method to compute the P_{fold} values for a set of sampled conformations of a protein and then estimate the TSE. The TSE is of great importance for understanding protein folding, because it gives insight into the main factors that influence folding rates and mechanisms. Knowledge of the structure of the TSE

may be used to re-engineer folding in a principled way [NÖ9]. One main advantage of SRS is that it efficiently examines a huge number of folding pathways and captures the ensemble behavior of protein folding. Our method was tested on 16 proteins. The results show that our estimate of the TSE is much more discriminating than that of the DP method. This allows us to obtain better folding-rate predictions. We also used SRS to efficiently compute MFPT and observed good correlation with the folding rates. We have mixed results in predicting Φ-values. One likely reason is that Φ-value prediction requires a more detailed model than the one that we used. The results that SRS achieved on these difficult prediction problems further validate the SRS method and indicate its potential as a general tool for studying protein folding kinetics.

We are working on several fronts to further improve SRS for folding kinetics prediction. Currently, our method requires each contiguous protein segment of five residues to fold or unfold together. Following the suggestion of Garbuzynskiy et al. [GFG04], we also impose a constraint on the number of unfolded regions allowed in a conformation. All these are intended to reduce the number of conformations that must be examined and keep the computational cost low. We plan to remove these assumptions and compute larger roadmaps, in order to see whether this improves the predictions. Recent work on simplifying Markov chains by removing nodes from the roadmap while retaining key properties of transition probabilities and stationary distributions seems well-suited for enabling the computation over large roadmaps [CB06, GP01, SPS04].

Another important issue is to design a better free energy function in order to improve the accuracy of Φ-value prediction. As in earlier work [AB99, GFG04, ME99], we have obtained limited success in the prediction of Φ-values. The energy function proposed in [GNS02] could be a good candidate, as it has been tested on a large database of mutants.

Our method makes the simplifying assumption that the unfolded conformation is a coil, meaning that no native contacts exist in the unfolded conformation. In fact, the early stages of protein folding are not well characterized, partly due to the difficulty of obtaining experimental data in these stages. Recently, there have been successful attempts to determine the structure of the denatured ensemble from NMR data [RMM+05, WD06]. This information can be used to better characterize the unfolded conformations and improve our prediction results.

Finally, most of the 16 proteins that we studied fold via a relatively simple two-state transition mechanism. It would be interesting to further test our method on more complex proteins, such as those that fold via an intermediate.

Acknowledgements M. S. Apaydin's work at Duke was supported by the following grants to Bruce R. Donald: NIH grant R01-GM-65982 and NSF grant EIA-0305444. D. Hsu's research is partially supported by grant R252-000-145-112 from the National University of Singapore. J.C. Latombe's research is partially supported by NSF grant DMS-0443939, and part of this work was completed, while was a visiting professor at the National University of Singapore, supported by the Kwan Im Thong Hood Cho Temple Professorship.

References

[AB99] E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. National Academy of Sciences USA*, 96(20):11305–11310, 1999.

[ABG+02] M.S. Apaydin, D.L. Brutlag, C. Guestrin, D. Hsu, and J.C. Latombe. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. In *Proc. ACM Int. Conf. on Research in Computational Biology (RECOMB)*, pages 12–21, 2002.

[ADS02] N.M. Amato, K.A. Dill, and G. Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. In *Proc. ACM Int. Conf. on Research in Computational Biology (RECOMB)*, pages 2–11, 2002.

[ASBL01] M.S. Apaydin, A.P. Singh, D.L. Brutlag, and J.C. Latombe. Capturing molecular energy landscapes with probabilistic conformational roadmaps. In *Proc. IEEE Int. Conf. on Robotics & Automation*, pages 932–939, 2001.

[Bak00] D. Baker. A surprising simplicity to protein folding. *Nature*, 405(6782):39–42, 2000.

[BOSW95] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Genetics*, 21(3):167–195, 1995.

[CB06] C. Chennubhotla and I. Bahar. Markov methods for hierarchical coarse-graining. In *Proc. ACM Int. Conf. on Research in Computational Biology (RECOMB)*, pages 379–393, 2006.

[CHKB98] M. Cieplak, M. Henkel, J. Karbowski, and J.R. Banavar. Master equation approach to protein folding and kinetic traps. *Phys. Rev. Let.*, 80:3654, 1998.

[CLH+05] H. Choset, K.M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L.E. Kavraki, and S. Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementations*, chapter 7. The MIT Press, 2005.

[CSRST04] J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran. Geometric algorithms for the conformational analysis of long protein loops. *J. Comput. Chem.*, 25(7):956–967, 2004.

[DC97] K.A. Dill and H.S. Chan. From Levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10–19, 1997.

[DK01] Y. Duan and P.A. Kollman. Computational protein folding: From lattice to all-atom. *IBM Systems J.*, 40(2):297–309, 2001.

[DPG+98] R. Du, V. Pande, A.Y. Grosberg, T. Tanaka, and E. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.

[Fer99] A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W.H. Freeman & Company, New York, 1999.

[GFG04] S.O. Garbuzynskiy, A.V. Finkelstein, and O.V. Galzitskaya. Outlining folding nuclei in globular proteins. *J. Mol. Biol.*, 336:509–525, 2004.

[GGF05] O.V. Galzitskaya, S.O. Garbuzynskiy, and A.V. Finkelstein. Theoretical study of protein folding: outlining folding nuclei and estimation of protein folding rates. *Journal of Physics: Condensed Matter*, 17:S1539–S1551, 2005.

[GNS02] R. Guerois, J.E. Nielsen, and L.L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, 320(2):369–387, July 2002.

[GP01] A. Gambin and P. Pokarowski. A new combinatorial algorithm for large markov chains. In V.G. Ghanza et al., editors, *Computer Algebra in Scientific Computing (CASC 2001)*, pages 195–212. Springer-Verlag, 2001.

[IF01] D.N. Ivankov and A.V. Finkelstein. Theoretical study of a landscape of protein folding-unfolding pathways. Folding rates at midtransition. *Biochemistry*, 40(33):9957–9961, Aug 2001.

[IOF95] L.S. Itzhaki, D.E. Otzen, and A.R. Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, 254(2):260–288, 1995.

[ME99] V. Muñoz and W.A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. National Academy of Sciences USA*, 96(20):11311–11316, 1999.

[N99] B. Nölting. *Protein Folding Kinetics: Biophysical Methods.* Springer, 1999.

[P+03] V.S. Pande et al. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68:91–109, 2003.

[PSB98] K.W. Plaxco, K.T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277(4):985–994, 1998.

[Rei79] J.H. Reif. Complexity of the mover's problem and generalizations. In *Proc. IEEE Symp. on Foundations of Computer Science*, pages 421–427, 1979.

[RMM+05] T.L. Religa, J.S. Markson, U. Mayor, S.M.V. Freund, and A.R. Fersht. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437:1053–1056, October 2005.

[RSGC05] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch. Estimation of protein folding probability from equilibrium simulations. *J. Chemical Physics*, 122:184901, 2005.

[SA01] G. Song and N.M. Amato. Using motion planning to study protein folding pathways. In *Proc. ACM Int. Conf. on Research in Computational Biology (RECOMB)*, pages 287–296, 2001.

[SLB99] A.P. Singh, J.C. Latombe, and D.L. Brutlag. A motion planning approach to flexible ligand binding. In *Proc. Int. Conf. on Intelligent Systems for Molecular Biology*, pages 252–261, 1999.

[SPS04] W.C. Swope, J.W. Pitera, and F. Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *J. Phys. Chem. B*, 108:6571–6581, 2004.

[SSP04] N. Singhal, C.D. Snow, and V.S. Pande. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chemical Physics*, 121(1):415–425, 2004.

[Tea01] IBM Blue Gene Team. Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, 40(2):310–327, 2001.

[TK94] H. Taylor and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, New York, 1994.

[TKT+04] X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N.M. Amato. Using motion planning to study RNA folding kinetics. In *Proc. ACM Int. Conf. on Research in Computational Biology (RECOMB)*, pages 252–261, 2004.

[WD06] L. Wang and B.R. Donald. A data-driven, systematic search algorithm for structure determination of denatured or disordered proteins. In *IEEE Computer Society Computational Systems Bioinformatics Conference(CSB)*, 2006. In Press.

[WPD04] R.R. Weikl, M. Palassini, and K.A. Dill. Cooperativity in two-state protein folding kinetics. *Protein Sci.*, 13(3):822–829, 2004.

A The List of Proteins Used for Testing

For each protein used in our test, the table below lists its name, PDB code, the number of residues, and the experimental method for structure determination.

Protein	PDB code	No. Res.	Exp. Meth.
B1 IgG-binding domain of protein G	1PGB	56	X-ray
Src SH3 domain	1SRM	56	NMR
Src-homology 3 (SH3) domain	1SHG	57	X-ray
Sso7d	1BF4	63	X-ray
CI-2	2CI2	65	X-ray
B1 IgG-binding domain of protein L	2PTL	78	NMR
Barstar	1BTB	89	NMR
Fibronectin type III domain from tenascin	1TEN	89	X-ray
TI I27 domain of titin	1TIU	89	NMR
Tenth type III module of fibronectin	1TTF	94	NMR
RNA binding domain of U1A	1URN	96	X-ray
S6	1RIS	97	X-ray
FKBP-12	1FKB	107	X-ray
Barnase	1RNB	109	X-ray
Villin 14T	2VIL	126	NMR
CheY	3CHY	128	X-ray