IN *Proc. ACM Int. Conf. on Research in Computational Molecular Biology (RECOMB)*, 2006

Predicting Experimental Quantities in Protein Folding Kinetics using Stochastic Roadmap Simulation

Tsung-Han Chiang[1], Mehmet Serkan Apaydin[2], Douglas L. Brutlag[3], David Hsu[1], and Jean-Claude Latombe[3]

[1] National University of Singapore, Singapore 117543, Singapore
[2] Dartmouth College, Hanover, NH 03755, USA
[3] 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 Φ-values for protein folding. The new method was tested on 16 proteins. Comparison with experimental data shows that it estimates the TSE much more accurately than an existing method based on dynamic programming. This leads to better folding-rate predictions. 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 success further validates the SRS method and indicates its potential as a general tool for studying protein folding kinetics.

1 Introduction

Protein folding is a crucial biological process in nature. 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 state*, 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 state, 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 conformation 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 [4, 7] 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 state. On the experimental side, residue-specific measurements of the folding process (see, e.g., [14]) 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 [3] on a folding energy landscape proposed in [12]. SRS samples the protein conformational space and builds a directed graph, called the *stochastic conformational roadmap*. The nodes of the roadmap represent sampled protein conformations, and the edges represent transitions between the conformations. The roadmap compactly encodes a huge number of folding pathways and captures the stochastic nature of the folding process. Using the roadmap, we can efficiently compute the folding probability (P_{fold}) [8] for each sampled conformation in the roadmap and decide which conformations belong to the TSE. Finally, we estimate folding rates and Φ-values using the set of conformations in the TSE.

We tested our method on 16 proteins with sizes ranging from 56 to 128 residues and validated the results against experimental data. The results show that our method predicts folding rates with accuracy better than an existing method based on dynamic programming (DP) [12]. 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 accurate estimate better reveals the composition of the TSE and makes our method more suitable for studying the mechanisms of protein folding. For Φ-value prediction, the accuracy of our method varies among the proteins tested. The results are comparable to those obtained from 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* [3]. The comparison with experimental data serves as a test of the methodology, and the success further validates the SRS method and indicates its potential as a general tool for studying protein folding kinetics.

2 Related Work

There are many approaches for studying protein folding kinetics, including all-atom or lattice molecular dynamics simulation (see [9] for a survey), solving master equations [6, 21], and estimating the TSE [1, 12]. Recently, several related methods suc-

ceeded in predicting folding rates and Φ-values [1, 12, 15], using simplified energy functions that depend only on the topology of the native state 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 [1, 12], 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.

SRS is inspired by the probabilistic roadmap (PRM) methods for robot motion planning [5]. In motion planning, our goal is to find a path for a robot to move from an initial configuration to a goal 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 that captures the connectivity of this space. Methods derived from PRM have been applied to ligand-protein docking [17], protein folding [3, 2], and RNA folding [19]. 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. SRS has also been combined with molecular dynamics simulation to study protein folding rates and mechanisms [18].

3 Overview

The *conformation* of a protein is a set of parameters that specify uniquely 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 over the energy landscape. These pathways start from unfolded conformations and all lead to the same native state.

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. For convenience, we also say that such conformations are in the transition state. 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 P_{fold} [8]. In a folding process, the P_{fold} value of a conformation q is defined as the probability of a protein reaching the folded (native) state before reaching an unfolded state, starting from conformation q. P_{fold} measures the kinetic distance between q and the folded state. From any conformation q with 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 folded state. The TSE is defined as the set of conformations

with P_{fold} equal to 0.5. Defining the TSE using P_{fold} has many advantages. In particular, P_{fold} is not determined by any specific pathway, but depends on all the pathways from unfolded states to the folded state. It thus captures the ensemble behavior of folding.

We can compute 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 folded state before an unfolded one. However, a large number of simulation runs are needed to estimate the P_{fold} value accurately, and doing so for many conformations incurs prohibitive computational cost. The SRS method approximates the 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 (Sect. 4) and then the computation of folding rates (Sect. 5) and Φ-values (Sect. 6) based on the energy of conformations in the TSE.

4 Estimating the TSE Using Stochastic Roadmap Simulation

SRS is an efficient method for exploring protein folding kinetics by examining many folding pathways simultaneously. We use SRS to compute P_{fold} values and then determine the TSE based on the computed 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. [12]. This model is based on the topology of a protein's native state. 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 state. 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 the 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 folded (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. With a maximum of four unfolded regions, we can capture the folding and unfolding of proteins with up to roughly 100 residues [11].

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 [1].

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:

$$P_{ij} = \begin{cases} (1/n_i)\exp(-\frac{\Delta E_{ij}}{k_{\mathrm{B}}T}) & \text{if } \Delta E_{ij} > 0 \\ 1/n_i & \text{otherwise} \end{cases},$$

where n_i is the number of outgoing edges of q_i, k_{B} is the Boltzmann 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:

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

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_{F} and q_{U} are the two roadmap nodes representing the folded and the unfolded 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 Sect. 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 $\mathrm{P_{fold}}$

$\mathrm{P_{fold}}$ measures the kinetic distance between a conformation q and the native state q_{F}. The main advantage of using $\mathrm{P_{fold}}$ 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 $\mathrm{P_{fold}}$ value τ of a conformation q is defined as the probability of a protein reaching the native state q_{F} before reaching the unfolded state 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 [20]. Let us consider what happens after a single step of transition:

- We may reach a node in the folded state, which, by definition, has $\mathrm{P_{fold}}$ value 1.
- We may reach a node in the unfolded state, which has $\mathrm{P_{fold}}$ value 0.
- Finally, we may reach an intermediate node q_j with $\mathrm{P_{fold}}$ value τ_j.

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

$$\tau_i = \sum_{q_j \in \{q_{\mathrm{F}}\}} P_{ij} \cdot 1 + \sum_{q_j \in \{q_{\mathrm{U}}\}} P_{ij} \cdot 0 + \sum_{q_j \notin \{q_{\mathrm{F}}, q_{\mathrm{U}}\}} P_{ij} \cdot \tau_j, \tag{2}$$

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

The relationship in (2) gives a linear equation for each unknown τ_i. The resulting linear system is sparse and can be solved efficiently using iterative methods [3].

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 $\mathrm{P_{fold}}$ values for all the conformations on a PC workstation with a 1.5GHz Itanium2 processor and 8GB of memory.

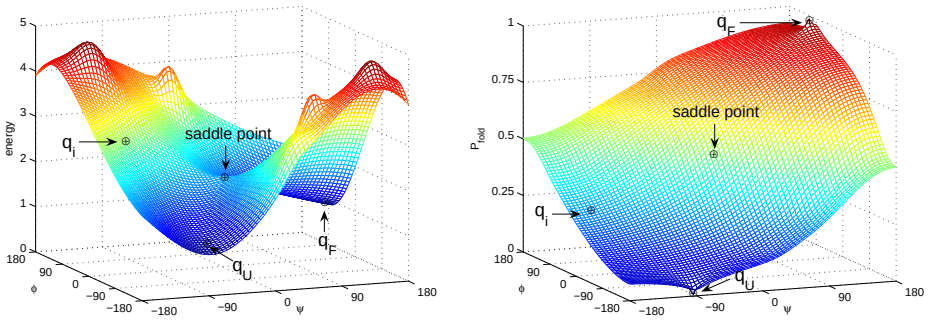
4.4 Estimating the TSE

After computing the $\mathrm{P_{fold}}$ value for each conformation, we identify the TSE by extracting all conformations with $\mathrm{P_{fold}}$ 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 $\mathrm{P_{fold}}$ 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

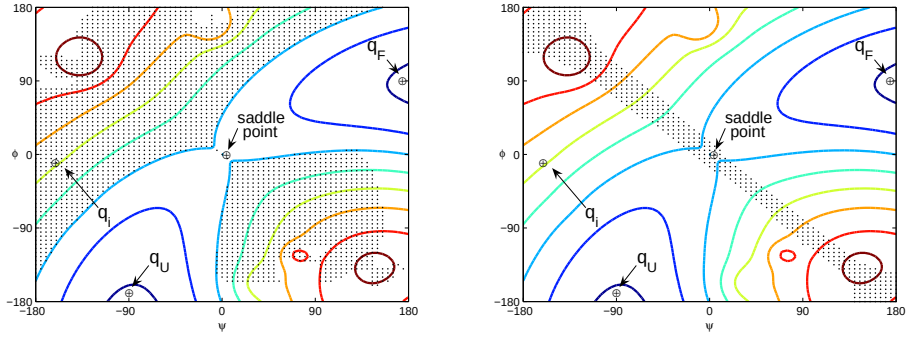
Consider a tiny fictitious protein with only two residues. Its conformation is specified by two backbone torsional angles ϕ and ψ. For the purpose of illustration, instead of using the simplified energy function described in Sect. 4.1, this example uses a saddle-shaped energy function over a two-dimensional conformation space (Fig. 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 $\mathrm{P_{fold}}$ values for this energy landscape is shown in Fig. 1*b*. A comparison of the two plots in Fig. 1 shows that the conformations with $\mathrm{P_{fold}}$ value 0.5 correspond well with the energy barrier that separates q_{U} and q_{F}.



(*a*) (*b*)

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



(*a*) (*b*)

Fig. 2. Estimation of the TSE for the energy landscape shown in Fig. 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 state to the folded state. 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 folded state 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 for predicting the folding rate.

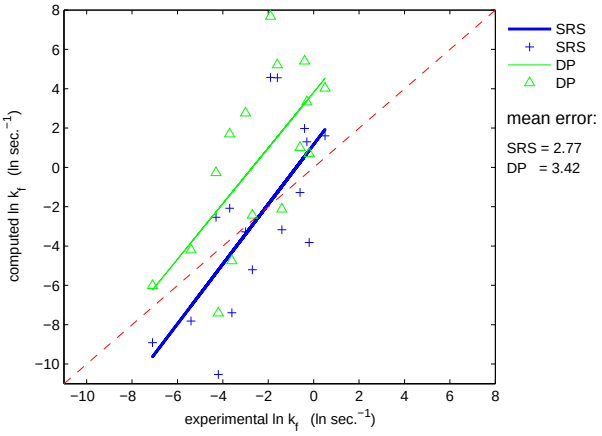


Fig. 3. Predicted folding rates versus the experimentally measured folding rates.

5.1 Methods

After identifying the TSE using the SRS method described in the previous section, we compute the folding rate the same way as that in [12], for the purpose of easy comparison. First, we calculate E_{TSE}, the total energy of the TSE, according to the following relationship [12]:

$$\exp(-\frac{E_{\mathrm{TSE}}}{RT}) = \sum_{q \in \mathrm{TSE}} \exp(-\frac{E(q)}{RT}), \tag{3}$$

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_{f} according to the following theoretical dependence [12]:

$$\ln(k_{\mathrm{f}}) = \ln(10^8) - (\frac{E_{\mathrm{TSE}}}{RT} - \frac{E(q_{\mathrm{U}})}{RT}), \tag{4}$$

where $E(q_{\mathrm{U}})$ is the energy of the q_{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 Fig. 3. The horizontal axis of the chart corresponds to the experimentally measured folding rates (see [12] 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 [12] and show the results in the same chart. Note that since the chart plots $\ln k_{\mathrm{f}}$, it basically compares the height of the energy barrier.

Fig. 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 $\ln k_{\mathrm{f}}$ for the two methods.

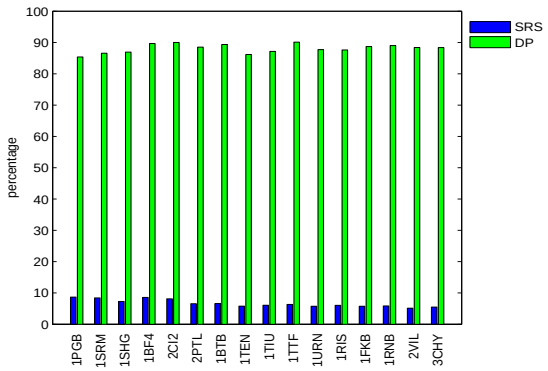


Fig. 4. The percentage of conformations in the TSE.

It is interesting to note 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.

5.3 Accuracy in Estimating the TSE

The difference between SRS and DP in estimating the TSE becomes more apparent when we compare the percentage of sampled conformations that are present in the TSE. Fig. 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 Fig. 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 Fig. 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

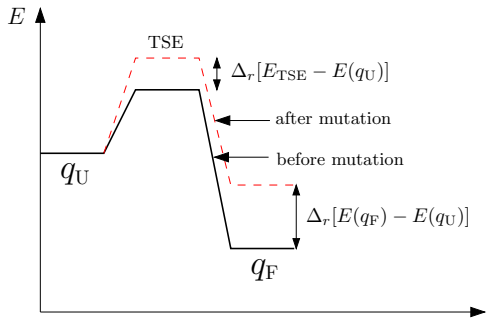


Fig. 5. Φ-value.

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 Fig. 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 Sect. 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 Fig. 2 for an illustration). Therefore, the TSE estimated by DP tend to have lower energy than the TSE estimated by SRS, resulting in over-estimated folding rates.

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 [10]. 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:

$$\Phi_r = \frac{\Delta_r[E_{\mathrm{TSE}} - E(q_{\mathrm{U}})]}{\Delta_r[E(q_{\mathrm{F}}) - E(q_{\mathrm{U}})]}, \tag{5}$$

where $\Delta_r[E_{\mathrm{TSE}} - E(q_{\mathrm{U}})]$ is the change in the energy difference between the TSE and the unfolded state 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 energy difference between the native state q_{F} and the unfolded state q_{U}. See Fig. 5 for an illustration. A Φ-value of 1 indicates that the mutation of residue r affects the energy of the transition state as much as the energy of the native state, relative to the energy of the unfolded state. So, in the transition state,

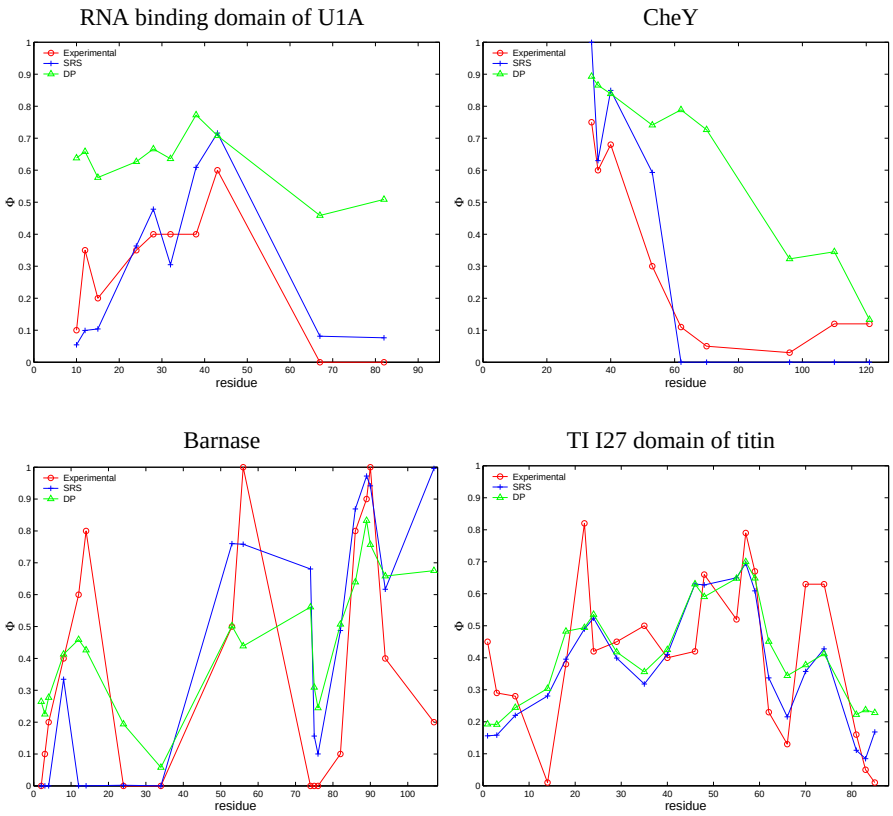


Fig. 6. Φ-value predictions for four proteins.

r must have fully attained the native conformation, according to energy considerations. Similarly, a Φ-value of 0 indicates that in the transition state, the residue remains unfolded. A fractional Φ-value 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 (3), we can simplify (5) and obtain the following expression for the Φ-value of residue r:

$$\Phi_r = \frac{\sum_{q \in \mathrm{TSE}} P(q) \cdot \Delta_r n(q)}{\Delta_r n(q_{\mathrm{F}})}, \tag{6}$$

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.

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. Fig. 6 shows

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

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 [12]. 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.

6.3 Results on the Order 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 progress of folding in the transition state only. We can extend this method to observe the details of the folding process, in particular, the order of native structure formation, by plotting the progression of each residue with respect to P_{fold}.

Each plot in Fig. 7 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 order of native structure formation that we observed is consistent with that obtained by Alm et al. [1].

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 pro-

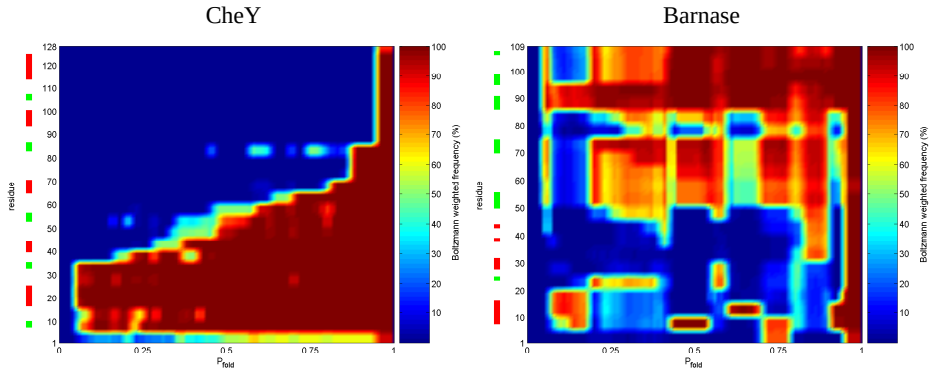


Fig. 7. Sequence of secondary structure formation. The colored bar on the left of each plot indicates secondary structures, red for helices and green for strands.

tein folding process and the limited experimental data available. The reliance on other simulation results for verification is almost inevitable.

7 Conclusion

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 [16]. 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 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 success of SRS on these difficult prediction problems further validates the SRS method and indicates its potential as a general tool for studying protein folding kinetics.

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. We also plan to improve Φ-value prediction by using a better energy model and to predict other experimental quantities, such as hydrogen-exchange protection factors [13].

Acknowledgements M. S. Apaydin's work at Dartmouth 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 grants CCR-0086013 and DMS-0443939, and by a Stanford BioX Initiative grant.

References

1. E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Nat. Acad. Sci. USA*, 96:11305–11310, 1999.
2. 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 Computational Biology (RECOMB)*, pages 2–11, 2002.
3. 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 Computational Biology (RECOMB)*, pages 12–21, 2002.
4. 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.
5. 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.
6. 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.
7. K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10–19, 1997.
8. R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.
9. Y. Duan and P. A. Kollman. Computational protein folding: From lattice to all-atom. *IBM Systems J.*, 40(2):297–309, 2001.
10. A. Fersht. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W.H. Freeman & Company, New York, 1999.
11. A. V. Finkelstein and A. Y. Badretdinov. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Folding and Design*, 2(2):115–121, 1997.
12. S. O. Garbuzynskiy, A. V. Finkelstein, and O. V. Galzitskaya. Outlining folding nuclei in globular proteins. *J. Mol. Biol.*, 336:509–525, 2004.
13. V. J. Hilser and E. Freire. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J. Mol. Biol.*, 262(5):756–772, 1996.
14. 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.
15. V. Muñoz and William A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Nat. Acad. Sci. USA*, 96:11311–11316, 1999.
16. B. Nölting. *Protein Folding Kinetics: Biophysical Methods*. Springer, 1999.
17. 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.
18. N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better Markovian state models: Predicting the foloding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chemical Physics*, 121(1):415–425, 2004.
19. 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 Computational Biology (RECOMB)*, pages 252–261, 2004.

20. H. Taylor and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, New York, 1994.
21. T. 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