Using Motion Planning to Map Protein Folding Landscapes and Analyze Folding Kinetics of Known Native Structures†

Nancy M. Amato*  
Dept. of Computer Science  
Texas A&M University  
College Station, TX 77843-3112  
amato@cs.tamu.edu  
tel: +1-979-862-2275  
fax: +1-979-458-0425

Ken A. Dill  
Dept. of Pharmaceutical Chemistry  
University of California  
San Francisco, CA 94143-1204  
dill@maxwell.ucsf.edu

Guang Song  
Dept. of Computer Science  
Texas A&M University  
College Station, TX 77843-3112  
gsong@cs.tamu.edu

Keywords: motion planning, probabilistic roadmap methods, protein folding, folding pathways, kinetics.

*Corresponding Author.
†This research supported in part by NSF CAREER Award CCR-9624315, NSF Grants ACI-9872126, EIA-9975018, EIA-0103742, EIA-9805823, ACR-0081510, ACR-0113971, CCR-0113974, EIA-9810937, EIA-0079874, by the DOE ASCI ASAP program, and by the Texas Higher Education Coordinating Board grant ATP-000512-0261-2001.
Abstract

We investigate a novel approach for studying the kinetics of protein folding. Our framework has evolved from robotics motion planning techniques called probabilistic roadmap methods (PRMs) that have been applied in many diverse fields with great success. In our previous work, we presented our PRM-based technique and obtained encouraging results studying protein folding pathways for several small proteins. In this paper, we describe how our motion planning framework can be used to study protein folding kinetics. In particular, we present a refined version of our PRM-based framework and describe how it can be used to produce potential energy landscapes, free energy landscapes, and many folding pathways all from a single roadmap which is computed in a few hours on a desktop PC. Results are presented for 14 proteins. Our ability to produce large sets of unrelated folding pathways may potentially provide crucial insight into some aspects of folding kinetics, such as proteins that exhibit both two-state and three-state kinetics, that are not captured by other theoretical techniques.
1 Introduction

There are large and ongoing research efforts whose goal is to determine the native structure of a protein from its amino acid sequence (see, e.g., [24, 31]). In this paper, we assume the native structure is known, and our focus is on the study of protein folding kinetics and mechanisms. This is also a very important research area which has taken on increased practical significance with the realization that misfolded or only partially folded proteins are associated with many devastating diseases [22]. Yet, despite intensive efforts by both experimentalists and theorists to understand the behavior and mechanism of the folding process, there are still many questions to be answered.

In previous work [34], we proposed a technique for computing protein folding pathways that was based on the successful probabilistic roadmap (PRM) [21] method for robotics motion planning. We were inspired to apply this methodology to protein folding based on our success [33] in applying it to folding problems such as carton folding (with applications in packaging and assembly [26]), and paper crafts (studied in computational geometry [29]). For example, note the parallels between the periscope paper model folding and the small polypeptide folding in the path snapshots shown in Figure 1. In [34], promising results were obtained for several small proteins (∼60 amino acids) and we validated our pathways by comparing the secondary structure formation order with known experimental results.

As evidence of the insight that might be provided with our approach, we demonstrated in [34] how an analysis of the pathways contained in our roadmaps showed evidence of the two classes of folding kinetics described by Baldwin and Roses [9]. For example, we noted that in our simulations, the three alpha helices in Protein A always formed first before packing into the final tertiary structure. In contrast, Protein G (domain B1), a small protein with one alpha helix and a four strand beta sheet, seemed to form the secondary structure gradually on the way to the tertiary structure. Moreover, as we will also see, this behavior could in fact be inferred from the distribution of the conformations contained in our roadmaps.

1.1 Our contribution

In this paper, we present a refined version of our motion planning framework and describe how it can be used to map a protein’s potential and free energy landscapes. Our work provides an alternative approach that finds approximations to the folding pathways while avoiding local minima and detailed simulations. In particular, our technique can produce potential energy landscapes, free...
energy landscapes, and many folding pathways all from a single roadmap which is computed in a few hours on a desktop PC. This computational efficiency enables us to compute roadmaps containing a representative set of feasible folding pathways from many (hundreds or thousands) denatured conformations to the native state. To illustrate our technique, we analyze folding pathways in terms of secondary structure formation order for many proteins, and compare and validate them with experimental results when available.

The unique ability of our method to produce large sets of unrelated folding pathways may potentially provide crucial insight into some aspects of folding kinetics that are not captured by other theoretical techniques. In particular, the large set of unrelated folding pathways present in our roadmaps provides an opportunity to study folding kinetics by directly analyzing folding pathways. This appears to be a natural way to study kinetics, and should enable us to capture multi-state folding kinetic behaviors if they exist. For example, both two-state and three-state folding kinetics of hen egg-white Lysozyme should be present in a good roadmap. Folding pathways have not been used to study such complex behaviors since it was difficult, if not impossible, to find witnesses of mechanisms with previous simulation methods.

1.2 Related work

Table 1 provides a summary comparison of various models for protein folding. While each method has unique strengths and advantages, they all have weaknesses as well. Both Monte Carlo simulation and molecular dynamics provide only one folding trajectory and each run is computationally intensive because they attempt to simulate complex kinetics and thermodynamics at every point visited in conformation space. Statistical mechanical models, on the other hand, assume extremely simplified molecular interactions and are limited to studying global averages of folding kinetics. Lattice models (see, e.g., [12]) have been well studied and possess great theoretical value but cannot be applied to real proteins. Our PRM approach, by constructing a roadmap that approximates the folding landscape, computes multiple folding pathways in a single run and provides a natural way to study protein folding kinetics at the pathway level. What we sacrifice is path quality, which can be improved through bigger roadmaps, oversampling, or other techniques. Apaydin et al. [6, 7] have also used PRM-based techniques to study protein folding. However, their work differs from ours in several aspects. First, the basic elements of their model are secondary structure elements (e.g., helices and strands) which are treated as rigid components. Their protein models therefore
have only a few degrees of freedom while ours, which have 2 degrees of freedom for each residue, self-avoiding walk. Secondly, while our focus is on studying detailed folding pathways and kinetics, their focus has been to compare the PRM approach with methods such as Monte Carlo simulation. They have shown that the PRM-based technique converges to the same stochastic distribution as Monte Carlo simulation, but is much faster.

There are many interesting experimental results that have yet to be adequately explained or captured by theoretical models. For example, Baldwin and Rose [9] noted that the folding kinetics of small proteins display two classes of folding behavior. In some cases, a protein folds by forming native-like secondary structure (e.g., Cytochrome C), and in other cases the protein seems to fold rapidly through a possible tertiary nucleation mechanism (e.g., CI2). Theoretical approaches capable of identifying both behaviors are needed.

Another interesting experimental result [1] suggests that the folding process for small proteins is mainly determined by native state topology. Based on this experimental observation, Baker et al. [2, 8] proposed a statistical mechanical model that uses the topology of the native state to predict folding rates and mechanisms. This insight had been made earlier by Muñoz et al. [28] in their study of β-hairpin kinetics and was later used in the kinetics study of 20 plus proteins [27] with quite impressive results. However, despite the success of these models, there exist many uncertainties related to the selection of the free energy functions and the restrictions on the structure of the conformations analyzed which strongly affect the results of the models. Finally, there is experimental data suggesting that some proteins, such as hen egg-white Lysozyme, display different kinetic behavior (e.g., two-state or three-state) along different pathways [13, 30], which cannot be captured with statistical mechanical models.

1.3 Outline

We begin in Section 2 with an overview of probabilistic roadmap motion planning methods and describe how they can be applied to compute folding pathways for proteins. Next, in Section 3 we present the potential and free energy functions used in our method. Our results for the 14 proteins studied are presented in Section 4. We conclude with some final remarks in Section 5.
2 A Probabilistic Roadmap Method for Protein Folding

Our approach to protein folding is based on the probabilistic roadmap (PRM) approach for motion planning [21]. PRMs work by constructing a connectivity graph of the feasible regions of the environment which can be used subsequently to answer many, varied motion planning queries. A detailed description of how the PRM framework can be applied to protein folding is presented in our previous work [5]. The basic idea is illustrated in Figure 2. We first sample some points in the protein’s conformation space (Figure 2(a)); generally, our sampling is biased to increase density near the known native state. Then, these points are connected to form a graph, or roadmap (Figure 2(b)). The weight assigned to a directed edge reflects the energetic feasibility of transition between the conformations corresponding to the two end points. Finally, folding paths to the native state are extracted from the roadmap using standard graph search techniques (Figure 2(c)). In the following sections we describe how to sample the nodes and weight the edges.

In previous work, we proposed the PRM framework as a methodology for studying protein folding when the native structure is known [34]. The main difference from the usual PRM application is that the collision detection feasibility test is replaced by a preference for low energy conformations. We obtained very promising results for several small proteins (e.g., proteins A and GB1, both with approximately 60 residues), and in particular, we showed that the pathways extracted from our roadmaps seemed to be in agreement with known experimental results [25].

2.1 Modeling proteins (C-space)

The amino acid sequence is modeled as an articulated ‘robot’. Using a standard modeling assumption for proteins, the only degrees of freedom (dof) in our model of the protein are the backbone’s phi and psi torsional angles, which we model as two revolute joints (2 dof), see Figure 3. Moreover, side chains are modeled as spheres and have zero dof. Thus, the model for a $k$ residue sequence will consist of $2k$ links and $2(k - 1)$ revolute joints. (The first and last rotational angles do not contribute.)

Since we are not concerned with the absolute position and orientation of the protein, a conformation of an $n + 1$ amino acid protein can be specified by a vector of $2n$ phi and psi angles, each in the range $[0, 2\pi)$, with the angle $2\pi$ equated to 0, which is naturally associated with a unit circle in the plane, denoted by $S^1$. That is, the conformation space (C-space) of interest for a protein with
\( n + 1 \) amino acids can be expressed as:

\[
\mathcal{C} = \{ q \mid q \in S^{2n} \}. \tag{1}
\]

Note that \( \mathcal{C} \) simply denotes the set of all possible conformations, but says nothing about their feasibility. For protein folding, the validity of a point in \( \mathcal{C} \) will be determined by potential energy computations.

### 2.2 Node generation

Recall that in our work we begin with the known native structure and our goal is to map the protein-folding landscape leading to the native fold. The objective of the node generation phase is to generate a representative sample of conformations of the protein. Due to the high dimensionality of the conformation space, simple uniform sampling would take too long to provide sufficiently dense coverage of the region surrounding the native structure. In our previous work [34], sampling was biased by sampling from a selected set of normal distributions centered around the native structure. This worked well for small proteins (approximately 60 residues), but was not as effective in generating unstructured conformations for larger proteins.

In this paper, we use another biased sampling strategy which has been more successful for larger proteins (more than 100 residues). It still focuses sampling around the native state, but instead of sampling from a set of normal distributions always centered around the native state, we generate new conformations by iteratively applying small perturbations to existing conformations. This version appears to produce smoother distributions and is much faster. The whole process is illustrated in Figure 4. Similar biased sampling strategies have been applied successfully in robotics applications [3, 11, 17, 18, 20, 23, 35], where oversampling in and near narrow passages in C-space is crucial for some problems. Also, as described in Section 1, Alm and Baker [8, 2] and Muñoz and Eaton [27] have used knowledge of the topology of the native state to predict the folding rates and mechanisms of some proteins.

To ensure we obtain an adequate coverage of the conformation space, we partition conformations into sets, or bins, according to the number of native contacts present, and continue generating nodes until all bins have enough conformations. A native contact is a pair of \( C_\alpha \) atoms from residues that are within 7 Å of each other in the native state. The contact number of a conformation is defined as the number of native contacts it has. Our bins are based on the contact number and are
equal-sized (we choose a bin size of 10). The number of bins is proportional to the total number of native contacts in the native state.

We initiate the process by generating a number of conformations very close to the native state by slightly perturbing the native phi/psi angles, e.g., by sampling from a normal distribution with a small standard deviation (e.g., 1°). After the joint angles are known, the coordinates of each atom in the system are calculated, and these are then used to determine the potential energy of the conformation (see Section 3.1). A node $q$ is accepted and added to the roadmap based on its potential energy $E(q)$ with the following probability:

$$P(\text{accept } q) = \begin{cases} 
1 & \text{if } E(q) < E_{\text{min}} \\
\frac{E_{\text{max}} - E(q)}{E_{\text{max}} - E_{\text{min}}} & \text{if } E_{\text{min}} \leq E(q) \leq E_{\text{max}} \\
0 & \text{if } E(q) > E_{\text{max}} 
\end{cases}$$

We set $E_{\text{min}} = 50000$ KJouls/mol and $E_{\text{max}} = 89000$ KJouls/mol which favors configurations with well separated side chain spheres. This acceptance test, which helps us retain more nodes in low energy regions, was also used when building PRM roadmaps for ligand binding [10, 32] and in our previous work on protein folding [34].

We also compute the contact number of the accepted nodes and place them in the appropriate bins. Then, we begin an iterative process of generating more nodes – our goal is to fill all bins with at least $N_{\text{frontier}}$ nodes. We randomly pick $N_{\text{frontier}}$ nodes from the lowest filled bin (conformations with high contact number) as seed nodes for the current round. (The initial sampling phase produces at least $N_{\text{frontier}}$ in the lowest bin.) Each selected seed node $q_0$ will be used to generate as many as $N_{\text{children}}$ nodes — these new nodes will be sampled from normal distributions with origin $q_0$ and standard deviations selected by cycling through the list $\{3°, 5°, 10°, 20°, 40°\}$. Each new node that passes the acceptance test is placed in the appropriate bin for its contact number. If the next bin has enough nodes (i.e., at least $N_{\text{frontier}}$), then the next iteration moves to that bin. Otherwise, more nodes are generated using seeds from the current bin. The process continues until all bins are filled. To reduce the dependence between rounds, we use seeds from the same bin only a limited number of times. This approach is more efficient in covering the conformation space of larger proteins than our original technique [34].
2.2.1 Distribution of nodes

An interesting effect starts to emerge after node generation. The potential vs. RMSD distributions for three proteins are shown in Figure 5. Note the contrast between the distributions for the all alpha (a) and the all beta (c) proteins, even though our sampling technique does not utilize information regarding secondary structure. These distributions seem to reflect the fact that all alpha proteins tend to fold differently from all beta proteins. In particular, all alpha proteins tend to form the helices first, and then the helices pack together to form the final tertiary structure. In the figure, this packing of helices is seen as the narrow ‘tail’ in the distribution where the potential changes very little as the RMSD approaches zero. In contrast, the distributions for the all beta proteins are much smoother, indicating that the secondary and the tertiary structure may be formed simultaneously. For the mixed alpha and beta protein, the plot shares some features of the plots for the all alpha and the all beta proteins. Similar distributions were observed for all proteins we have studied.

It is important to note that this distinctive behavior does not result from our choice of potential $E(q)$. Even though our potential requires knowledge of the hydrogen bonds present in the native state (see Section 3.1), it does not distinguish between helices and beta sheets because we set the same energy for all hydrogen bonds. That is, the potential $E(q)$ does not favor one kind of secondary structure over another. One explanation for the observed behavior is that proteins tend to maximize the formation of favorable interactions while minimizing conformational entropy loss, as observed by other researchers (e.g., [16]). Here, we capture this behavior in the very early stages of our approach, i.e., after the initial sampling phase. One reason could be that since the formation of helices causes little entropy loss, the corresponding conformation space remains large, while for beta sheets, the conformation space is quickly constrained by the larger entropy losses. Therefore, beta sheets appear later, close to the native state (when the surrounding conformation space is already small and entropy loss is not as significant), while alpha helices form much earlier (since this doesn’t affect the conformation space as much). Interestingly, this is captured and reflected by our sampling.

2.3 Connecting the roadmap

Connection is the second phase of roadmap construction. The objective is to obtain a roadmap encoding representative, low energy paths. For each roadmap node, we first find its $k$ nearest
neighbors, for some small constant $k$, and then try to connect it to them using some simple local planner (see Figure 6(a)). After this is repeated for all roadmap nodes, we have a connectivity roadmap that is like a net laid down on the energy landscape (see Figure 6(b)). In our results, $k = 20$ and the distance metric used was Euclidean distance in $C$. We also experimented with RMSD distances, and found that the Euclidean distance was not only faster (by a factor of 5-10), but also resulted in better, denser connection.

When two nodes $q_1$ and $q_2$ are connected, the corresponding edge $(q_1, q_2)$ is added to the roadmap. Each edge $(q_1, q_2)$ is assigned a weight that depends on the sequence of conformations \{ $q_1 = c_0, c_1, c_2, \ldots, c_{n-1}, c_n = q_2$ \} on the straight line in $C$ connecting $q_1$ and $q_2$. For each pair of consecutive conformations $c_i$ and $c_{i+1}$, the probability $P_i$ of moving from $c_i$ to $c_{i+1}$ depends on the difference between their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$.

$$P_i = \begin{cases} 
e^{-\frac{\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \leq 0 \end{cases}$$

This keeps the detailed balance between two adjacent states, and enables the weight of an edge to be computed by summing the logarithms of the probabilities for consecutive pairs of conformations in the sequence. (Negative logs are used since each $0 \leq P_i \leq 1$.)

$$w(q_1, q_2) = \sum_{i=0}^{n-1} -\log(P_i),$$

In this way, we encode the energetic feasibility of transiting from one conformation to another in the edge connecting them.

### 2.4 Extracting folding pathways

The roadmap is a map of the protein-folding landscape of the protein. One way to study this landscape is to inspect and analyze the pathways it contains.

One interesting problem is to find a feasible path between a given initial conformation (e.g., any denatured conformation) and the native structure. If the start conformation is not already in the roadmap, then we can simply connect it to the roadmap just as was done for the other roadmap nodes during the connection phase (Section 2.3), and then use Dijkstra’s algorithm [14] to find the smallest weight path between the start and goal conformations. This process is described in detail in [5].
An important feature of our approach is that the roadmap contains many folding pathways, which together represent the folding landscape. For each conformation in the roadmap, there are many pathways between it and the native state. In the work presented here, we concentrate on only the shortest such path, in terms of our edge weights. (We intend to consider the k-shortest paths in future work.) Therefore, the pathways we study are the shortest paths from all conformations to the native state. Fortunately, this can be done by computing the single-source shortest-path (SSSP) tree from the native structure (see Figure 6(c)). Using Dijkstra’s algorithm, this takes $O(V^2)$ time, where $V$ is the number of roadmap nodes.

To further facilitate the analysis of the roadmap’s pathways, it is useful to reduce the number that must be analyzed by clustering ‘similar’ pathways. We do this by truncating our SSSP tree at denatured conformations, i.e., those with very little structure. While there are many possible definitions of little structure, we classify a conformation as such if it has no formed secondary structures (such as alpha helices) and no contacts between secondary structures (such as between two beta strands of a beta sheet). We determine that a structure is not present if less then 10% of the necessary native contacts for that secondary structure are present.
3 Energy Computations

Our technique uses the potential energy of a conformation during roadmap construction. We also use the free energy to assist in the analysis of our paths.

3.1 Potential energy

As previously mentioned, the way in which a protein folds depends critically on the potential energies of the conformations on the folding pathway. Our PRM framework incorporates this bias by accepting conformations based on their potential energy (Section 2.2) and by weighting roadmap edges according to their energetic feasibility (Section 2.3). Our framework is flexible enough to use any method for computing potential energies. Our current work uses a very simplistic potential. As the strengths and weaknesses of our approach are better understood, we can easily incorporate more sophisticated computations as deemed necessary. For example, we might retain a simple function for roadmap construction, but use a more sophisticated and expensive function for the query process, or for evaluating or improving already computed folding pathways.

Briefly, we use a step function approximation of the van der Waals potential component. Our approximation considers only the contribution from the side chains. Additionally, in our model of each amino acid, we treat the side chain as a single large ‘atom’ $R$ located at the $C_\beta$ atom. For a given conformation, we calculate the coordinates of the $R$ ‘atoms’ (our spherical approximation of the side chains) for all residues. If any two $R$ ‘atoms’ are too close (less than 2.4 Å during node generation and 1.0 Å during roadmap connection), a very high potential is returned.

If all the distances between all $R$ ‘atoms’ are larger than 2.4 Å, then we proceed to calculate the potential as follows:

$$U_{tot} = \sum_{restraints} K_d \{ \left[ (d_i - d_0)^2 + d_c^2 \right]^{1/2} - d_c \} + E_{hp},$$

(4)

The first term represents constraints which favor the known secondary structure through main-chain hydrogen bonds and disulphide bonds and the second term ($E_{hp}$) is the hydrophobic effect. The parameter $K_d$ is set to 100 KJ/mol, and the distances are $d_0 = d_c = 2.4 \text{Å}$, and $d_i$ is the separation between the hydrogen and oxygen atoms that form the hydrogen bond in the native structure.

A detailed description of our potential can be found in [5].
3.2 Entropy and free energy

While the potential energy is used to construct the roadmap, the free energy is used to analyze
roadmap paths and allows us to estimate and compare folding rates.

There are three main components of our free energy function: the hydrogen bond interaction,
the entropy, and the hydrophobic term. The van der Waals term is not considered. Similar
approximations were used in Munoz and Eaton [27, 28] and Baker et al. [2] in their statistical
mechanical models. The strengths used for the three terms are very similar to those used by
Munoz and Eaton [28].

For the hydrogen bond interaction, we check the distance between all pairs of donors and
acceptors found in the native fold for a given conformation. If any pair of atoms is within 3.0 Å
of each other, we consider that the hydrogen bond exists. We count the total number of hydrogen
bonds formed in that conformation \(N_{hb}\), and then the hydrogen bond contribution to the free
energy is \(F_{hb} = -0.86 \text{kcal/mol} \times N_{hb}\).

For the entropy, we consider it as follows. Each time a hydrogen bond is formed, the protein
becomes more constrained and loses some entropy, or its free energy increases. For a given
conformation with \(N_{hb}\) hydrogen bonds, we calculate the entropy by first calculating the effective contact order (ECO) for each hydrogen bond. Then the total entropy loss can be written as
\[
\Delta s = \sum_i^{N_{hb}} \log ECO_i \quad \text{[16]},
\]
and the total free energy change is \(F_{entropy} = 6.0 \text{cal/mol/K} \times (300K) \times \Delta s\).

For the hydrophobic effect, for a given conformation we check the distances between the \(C_\alpha\)
atoms for all hydrophobic residues. We count the number \(N_{hydro}\) that are within 7 Å, and
determine the effect on free energy: \(F_{hydrophobic} = -2.19 \text{kcal/mol} \times N_{hydro}\).

There are at least two things reflected in this free energy function. One is that the free energy
increase by entropy loss is normally bigger than the free energy decrease due to the formation of
hydrogen bonds. However, proteins are still driven to fold because of the third term, the hydropho-
bic effect. Another point is the way entropy is calculated reflects that proteins normally prefer to
form local contacts first to save entropy loss [16].
4 Results and Discussion

In this work, our goal is to understand how proteins fold to a known native structure, or more generally, to understand the protein-folding landscape. Our focus is therefore not on fold prediction, but rather we aim to understand folding kinetics to the known native state. We hope to gain insight into the underlying folding mechanism since we desire to reproduce, or at least approach, results close to experimental observations.

In this section we investigate how well the roadmaps constructed using our PRM-based technique map the potential and free energy landscapes of the proteins. We test our method on 14 small proteins that have been the subject of other protein folding kinetics studies [2, 27]. In all cases, we construct the PRM-based roadmaps, compute the contact number of each roadmap node, and analyze all folding pathways contained in the SSSP tree as described in Section 2.4. For each such pathway, we compute the formation order of secondary structure on it, and compare this with experimental results, if available. Finally, we try to compare and contrast our results with those of previous protein folding kinetics studies [2, 27].

4.1 The proteins

In this paper, we study 14 proteins listed in Table 2. In addition to protein G (B1 domain) and protein A, which we have been working with since the beginning, we have selected 12 proteins that were studied in Munoz and Eaton’s [27] work studying the folding kinetics of small proteins. Our work is greatly motivated by theirs, as well as the work of Baker’s group [2]. For all proteins, we determine the secondary structure formation order from the paths in our roadmaps, and we also study whether our approach can be used to produce results similar to theirs [2, 27].

4.2 Running time and statistics

Traditional simulation methods usually produce folding pathways by choosing a proper force or potential to drive the protein molecule in the conformation space. Therefore, each execution produces only one folding pathway, each of which has large computational requirements. Roadmap methods sacrifice accuracy (as much as is desired, which is a user specified parameter) in favor of rapid coverage, and hence are very effective in avoiding entrapment in local minima. Roadmap construction takes 2-15 hours for the 14 proteins studied (see Table 3). However, this, plus another few minutes or so analyzing the roadmap’s connectivity graph, is all that is needed to produce (approximately)
the potential energy landscape (see Figure 5), the free energy landscape (see Figure 8), and multiple folding pathways, all in a single run.

4.3 Secondary structure formation order

In [34], we presented folding pathways for a few small proteins in terms of their secondary structure formation order. These results were obtained in a rather ad hoc fashion. Here, we describe a more rigorous method of determining formation order which is based on the formation order of the native contacts between all residues. This method determines the secondary structure formation order for a given input path. The analysis will be repeated for each pathway we wish to study. First, to determine the contacts in the native state, we compare all pairs of $C_\alpha$ atoms, and those that are within $7 \text{ Å}$ of each other in the native state are said to form a native contact. Then, when analyzing a conformation $q$, we determine the contact is present in $q$ if the corresponding $C_\alpha$ atoms are $\leq 7 \text{ Å}$ apart. For each native contact, we record the time step on our path when it appears. To determine when a secondary structure appears, we compute the average appearance time for the contacts which determine that structure. In addition to providing a more formal method of validation, computing contact formation orders provides us with a tool for performing more detailed analysis of the folding pathways.

As an example, the results for a single path for protein CI2 are shown in Figure 7. In the figure, the full contact matrix (among all residues) is shown on the right, and blow-ups of the indicated regions are shown on the left. The cells of the blow-ups contain the time step in which the indicated contact formed on the path we are analyzing. For example, blow-up I (see Figure 7) shows the contact between residues 5 and 60 appeared at time step 216 on our path. To get an approximation of the time step in which a particular structure appeared, we average the appearance time steps for all of its contacts. For protein CI2 (Figure 7), the alpha helix (group II) formed first around time step 122 (the average of the time steps in II), then beta strands 3 and 4 (group V) came together around time step 187, followed by the contacts between beta strands 2 and 3 (group IV) around time step 210, which was closely followed by the contacts between beta strands 1 and 4 (group I, formed at time step 214) and the contacts between the alpha helix and beta strand 4 (group III, formed at time step 217). One may note that in some blowups, for example (I), there are some outliers, i.e., contacts of the same secondary structure that formed significantly later than others. This could occur as follows. Suppose a hairpin of eight residues forms contacts between residues
1-8, 2-7, 3-6, and 4-5. The formation of these contacts alone defines the hairpin structure. However, it is likely that, e.g., residues 1 and 7 also form a contact and that this contact could form later and appear as an outlier.

For each protein, we performed contact formation analysis on pathways from all denatured conformations to the native state. As mentioned above, for each denatured conformation, we extract the shortest path between it and the native structure in the roadmap. Since there are thousands of conformations in the roadmap (see Table 3), literally thousands of paths have to be analyzed. Each analysis yields a secondary structure formation order for that path, the formation order appears most often (i.e. with highest percentage) among all the paths is then considered to be the secondary structure formation order of the protein. Results for the 14 proteins studied are shown in Table 4. For the proteins that are also listed in Li and Woodward’s paper describing hydrogen-exchange experimental results [25], our results seem to be in good agreement with known results. This provides some evidence of the quality of our pathways before using them to facilitate further study. Moreover, the results could be used to predict folding behaviors for proteins for which we have no experimental data. One fact clearly seen in the formation order for all proteins is that they all seem to form local contacts first, and then those with increasing sequence contact order, like a zipper process as shown in [15, 16].

4.4 Folding kinetics: the 'global' approach vs. pathway level approach

Some recent statistical mechanical models have shown impressive success in predicting folding kinetics of many small proteins [2, 27]. In this approach, they use a simple model to calculate a protein’s free energy. To reduce the number of conformations that must be tested, structure is only allowed to form in a restricted number of localized regions in the sequence (e.g., one, two or three distinct regions at any given time). Then, the free energy of the conformations is plotted with respect to the number of native contacts present. The result is a free energy profile. It was observed that for several small two-state proteins, the folding rate could be computed from these profiles. Note that these profiles are not related to any folding pathway.

In terms of our method, these profiles are roughly equivalent to a plot of our roadmap nodes showing free energy vs. contact number. That is, this plot represents a global analysis and can possibly yield average folding rates, which may be accurate for small proteins with single feature folding pathways. In Figure 8, we show the free energy distribution vs. the number of native
contacts for proteins A, GB1, CI2, and Ubiquitin. (We chose these four proteins because our results are in agreement with the experimental data on their secondary structure formation order. More plots for other proteins can be found in [4].) One can see that any folding rates that might be inferred are from the free energy averages of the conformations with the same number of native contacts. As a result, it could easily miss detailed information of the energy landscape. This is illustrated Figure 9. The averaging approach would mix the 2-state (A) and 3-state (B) folding pathways together and would produce only their statistical average.

Therefore, this averaging approach is limited and potentially will miss subtle behavior. Moreover, for a protein like Hen egg white Lysozyme, which displays two unique folding pathways, one with two-state behavior, and one with three-state behavior [13, 30], averaging techniques like the statistical mechanical model cannot discover both behaviors (see Figure 9). This is one example where it seems to be crucial to have more detailed pathway information, such as is available in our roadmaps. One experiment we tried was to cluster paths into groups based on their secondary structure formation order, and then to analyze each group separately. Figure 10 shows folding path profiles for a representative path from each group (paths with similar secondary structure formation order) for each protein. The plots in Figure 10 show the free energy profile, and the native contact profile, for proteins G, A, CI2, and Ubiquitin. Similar plots for the other proteins are available in [4]. From the figure one can see the free energy profile is quite different from pathway to pathway, suggesting that protein molecules might undergo different folding kinetics at different regions of the conformation space. We realize that such differences in the free energy profiles might also be due to the fluctuation and instability of the representative path. What is still needed is some good way of analyzing and summarizing all the pathways in our roadmaps. This will enable us to retain the important details while reducing the total volume of data. The development of such techniques is the subject of on-going work.
5 Conclusion

In this paper, we present a refined version of our motion planning framework for studying protein folding kinetics. We describe how it can be used to produce potential energy landscapes, free energy landscapes, and many folding pathways all from a single *roadmap* which is computed in a few hours on a desktop PC.

Results are presented for 14 proteins, and are compared with results obtained by other methods such as statistical mechanical models. Our ability to produce large sets of unrelated folding pathways may potentially provide crucial insight into some aspects of folding kinetics that are not captured by other theoretical techniques, such as proteins that exhibit both two-state and three-state kinetics. Thus, our technique provides a way to study folding kinetics at the pathway level.

Future work includes more detailed analysis of pathways, such as grouping similar pathways, and calculation of protein folding rates.
6 Acknowledgment

We would like to thank Chris Sewell for helping with the experiments. We would also like to thank Jean-Claude Latombe for pointing out the connection between box folding and protein folding, and Marty Scholtz, Michael Levitt, and Vijay Pande for useful suggestions.
References


Figure 1: Snapshots of a carton (top) and a 10 alanine polypeptide chain (bottom) folding.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Folding Landscape</th>
<th>#Pathways Produced</th>
<th>Path Quality</th>
<th>Time Dependent</th>
<th>Compute Time</th>
<th>Folding Kinetics</th>
<th>Native Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular dynamics</td>
<td>No</td>
<td>1</td>
<td>Good</td>
<td>Yes</td>
<td>long</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Monte Carlo</td>
<td>No</td>
<td>1</td>
<td>Good</td>
<td>Yes</td>
<td>long</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Statistical Model</td>
<td>Yes</td>
<td>0</td>
<td>N/A</td>
<td>No</td>
<td>short</td>
<td>Average</td>
<td>Yes</td>
</tr>
<tr>
<td>PRM-Based</td>
<td>Yes</td>
<td>Many</td>
<td>Approx</td>
<td>No</td>
<td>short</td>
<td>Multiple</td>
<td>Yes</td>
</tr>
<tr>
<td>Lattice Model</td>
<td>Not used on real proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: A comparison of protein folding models.
Figure 2: A PRM roadmap for protein folding shown imposed on a visualization of the potential energy landscape: (a) after node generation (note sampling is denser around N, the known native structure), (b) after the connection phase, and (c) using it to extract folding paths to the known native structure.

Figure 3: A three amino acid segment of a protein. The phi (ϕ) and psi (ψ) angles in the middle amino acid are labeled. The figure is taken from [19].
Figure 4: An illustration of our iterative perturbation sampling strategy shown imposed on a visualization of the potential energy landscape, where \( N \) denotes the native structure.

Figure 5: The potential vs RMSD distribution for proteins (a) A (all alpha helix), (b) GB1 (mixed alpha/beta), and (c) CTX III (all beta).

Figure 6: (a,b) Roadmap connection and (c) extraction of folding pathways shown imposed on a visualization of the potential energy landscape, where \( N \) denotes the native structure.
Table 2: Proteins are listed in ascending order of number of residues (res). The number of alpha helices (α) and beta strands (β) is listed in the ss column.
<table>
<thead>
<tr>
<th>PDB</th>
<th>res</th>
<th>nodes</th>
<th>edge</th>
<th>time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GB1</td>
<td>56</td>
<td>5126 (5506)</td>
<td>70k</td>
<td>3.71</td>
</tr>
<tr>
<td>1BDD</td>
<td>60</td>
<td>5471 (9106)</td>
<td>104k</td>
<td>7.03</td>
</tr>
<tr>
<td>1SHG</td>
<td>62</td>
<td>5427 (5502)</td>
<td>59k</td>
<td>2.89</td>
</tr>
<tr>
<td>1COA</td>
<td>64</td>
<td>7975 (8407)</td>
<td>104k</td>
<td>6.87</td>
</tr>
<tr>
<td>1SRL</td>
<td>64</td>
<td>8755 (8822)</td>
<td>111k</td>
<td>4.95</td>
</tr>
<tr>
<td>1CSP</td>
<td>67</td>
<td>6735 (6852)</td>
<td>72k</td>
<td>4.67</td>
</tr>
<tr>
<td>1NYF</td>
<td>67</td>
<td>6219 (6332)</td>
<td>70k</td>
<td>3.42</td>
</tr>
<tr>
<td>1MJC</td>
<td>69</td>
<td>5990 (6142)</td>
<td>62k</td>
<td>4.30</td>
</tr>
<tr>
<td>2AIT</td>
<td>74</td>
<td>8246 (8477)</td>
<td>92k</td>
<td>7.11</td>
</tr>
<tr>
<td>1UBQ</td>
<td>76</td>
<td>8357 (10667)</td>
<td>119k</td>
<td>9.44</td>
</tr>
<tr>
<td>1PKS</td>
<td>79</td>
<td>7685 (10257)</td>
<td>95k</td>
<td>9.32</td>
</tr>
<tr>
<td>1PBA</td>
<td>81</td>
<td>8085 (10747)</td>
<td>114k</td>
<td>10.40</td>
</tr>
<tr>
<td>2ABD</td>
<td>86</td>
<td>7330 (12577)</td>
<td>149k</td>
<td>14.20</td>
</tr>
<tr>
<td>1BRN</td>
<td>110</td>
<td>6601 (10607)</td>
<td>108k</td>
<td>15.80</td>
</tr>
</tbody>
</table>

Table 3: Running time for constructing roadmaps for 14 proteins and statistics for each roadmap, including the number of nodes in the same connected component as the native structure, the total nodes (in parenthesis), and the number connections (edges) in the roadmap.
Figure 7: Protein CI2. The full contact matrix (right) and blow-ups (left) showing the time steps when the contacts appear on our path. Blow-ups I, II, III, IV and V correspond to the beta 1-4 contacts, the alpha helix contacts, the contacts between alpha and beta 4, the beta 2-3 contacts, and the beta 3-4 contacts, respectively.

Figure 8: Free energy landscape for proteins A, G, CI2, and Ubiquitin. The line in the middle is the average free energy for all conformations with the same number of native contacts. Provided that native contacts is a good reaction coordinate, this could be used to study folding kinetics as done by other researchers.
Table 4: The secondary structure formation order on dominant pathways in our roadmaps for 14 proteins and some validations. We show the formation order of both single secondary structures (such as alpha helices) and also for contacts between two secondary structures (such as two beta strands). The parenthesis indicate there was no clear order among them. The last column shows comparisons of our results with those from hydrogen-exchange experiments [25].

<table>
<thead>
<tr>
<th>pdb</th>
<th>res. #</th>
<th>secondary structure formation order</th>
<th>exp.[25]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GB1</td>
<td>56</td>
<td>$\alpha_1, \beta_3$-$\beta_4, \beta_1$-$\beta_2, \beta_1$-$\beta_4$</td>
<td>Agreed</td>
</tr>
<tr>
<td>1BDD</td>
<td>60</td>
<td>$\alpha_2, \alpha_3, \alpha_1, \alpha_2$-$\alpha_3, \alpha_1$-$\alpha_3$</td>
<td>Agreed</td>
</tr>
<tr>
<td>1SHG</td>
<td>62</td>
<td>$(\beta_2$-$\beta_3, \beta_3$-$\beta_4), (\beta_1$-$\beta_2, \beta_1$-$\beta_5)$</td>
<td>N/A</td>
</tr>
<tr>
<td>1COA</td>
<td>64</td>
<td>$\alpha_1, \beta_3$-$\beta_4, \beta_2$-$\beta_3, \beta_1$-$\beta_4, \alpha_1$-$\beta_4$</td>
<td>Agreed</td>
</tr>
<tr>
<td>1SRL</td>
<td>64</td>
<td>$\beta_4$-$\beta_5, \beta_3$-$\beta_4, \beta_2$-$\beta_3, \beta_1$-$\beta_5, \beta_1$-$\beta_2$</td>
<td>N/A</td>
</tr>
<tr>
<td>1CSP</td>
<td>67</td>
<td>$\beta_5$-$\beta_6, \beta_2$-$\beta_3, \beta_3$-$\beta_4, (\beta_1$-$\beta_3, \beta_4$-$\beta_6, \beta_5$-$\beta_7), \beta_1$-$\beta_5$</td>
<td>N/A</td>
</tr>
<tr>
<td>1NYF</td>
<td>67</td>
<td>$\beta_3$-$\beta_4, \beta_2$-$\beta_3, \beta_1$-$\beta_2$</td>
<td>N/A</td>
</tr>
<tr>
<td>1MJC</td>
<td>69</td>
<td>$(\beta_5$-$\beta_6, \beta_2$-$\beta_3), (\beta_3$-$\beta_4, \beta_1$-$\beta_3), (\beta_1$-$\beta_3, \beta_4$-$\beta_6, \beta_5$-$\beta_7)$</td>
<td>N/A</td>
</tr>
<tr>
<td>2AIT</td>
<td>74</td>
<td>$(\beta_4$-$\beta_5, \beta_1$-$\beta_2, \beta_3$-$\beta_4), (\beta_3$-$\beta_7, \beta_2$-$\beta_6, \beta_1$-$\beta_4), \beta_1$-$\beta_5, \beta_1$-$\beta_6$</td>
<td>N/A</td>
</tr>
<tr>
<td>1UBQ</td>
<td>76</td>
<td>$\alpha_1, \beta_3$-$\beta_4, \beta_1$-$\beta_2, \beta_3$-$\beta_5, \beta_1$-$\beta_5$</td>
<td>N/A</td>
</tr>
<tr>
<td>1PKS</td>
<td>79</td>
<td>$\alpha_1, \beta_2$-$\alpha_1, \beta_3$-$\beta_4, (\beta_1$-$\beta_2, \beta_2$-$\beta_3), \beta_1$-$\beta_5$</td>
<td>N/A</td>
</tr>
<tr>
<td>1PBA</td>
<td>81</td>
<td>$(\alpha_3 \alpha_1 \alpha_1$-$\alpha_2) \alpha_1$-$\beta_1, \beta_1$-$\beta_2, \beta_1$-$\beta_3$</td>
<td>N/A</td>
</tr>
<tr>
<td>2ABD</td>
<td>86</td>
<td>$\alpha_3, \alpha_4$-$\alpha_5, \alpha_2, \alpha_4, \alpha_0, \alpha_5, \alpha_2$-$\alpha_3, \alpha_2$-$\alpha_4, \alpha_1$-$\alpha_4$</td>
<td>N/A</td>
</tr>
<tr>
<td>1BRN</td>
<td>110</td>
<td>$\alpha_1, \alpha_3, \alpha_2, \beta_1$-$\alpha_2, \alpha_2$-$\alpha_3, \beta_5$-$\beta_6, \beta_2$-$\beta_3, \beta_4$-$\beta_3, \beta_6$-$\beta_7, \beta_3$-$\beta_4, \beta_1$-$\beta_2$</td>
<td>Not sure</td>
</tr>
</tbody>
</table>
Figure 9: Statistical mechanical models which compute only global statistics cannot identify multiple kinetics behavior. In this example, the averaging of the 2-state and 3-state behaviors results in an incorrect 2-state profile.
Figure 10: Free energy profiles, plotted in solid lines, for four folding pathways for proteins G, A, CI2, and Ubiquitin. The number of native contacts present in each conformation on the path is shown with dashed lines; note that they are not monotonically increasing.