Kinetics Analysis Methods For Approximate Folding Landscapes *

Lydia Tapia, Xinyu Tang, Shawna Thomas, and Nancy M. Amato
Parasol Lab, Dept. of Computer Science, Texas A&M University, College Station, TX 77843
{ltapia, xinyut, sthomas, amato}@cs.tamu.edu

Protein motions play an essential role in many biochemical processes. Lab studies often quantify these motions in terms of their kinetics such as the speed at which a protein folds or the population of certain interesting states like the native state. Kinetic metrics give quantifiable measurements of the folding process that can be compared across a group of proteins such as a wild-type protein and its mutants.

We present two new techniques, Map-based Master Equation solution and Map-based Monte Carlo simulation, to study protein kinetics through folding rates and population kinetics from approximate folding landscapes, models called maps. From these two new techniques, interesting metrics that describe the folding process, such as reaction coordinates, can also be studied. In this paper we focus on two metrics, formation of helices and structure formation around tryptophan residues. These two metrics are often studied in the lab through CD spectra analysis and tryptophan fluorescence experiments, respectively. The approximated landscape models we use here are the maps of protein conformations and their associated transitions that we have presented and validated previously.

In contrast to other methods such as the traditional master equation and Monte Carlo simulation, our techniques are both fast and can easily be computed for full length detailed protein models. We validate our map-based kinetics techniques by comparing folding rates to known experimental results. We also look in depth at the population kinetics, helix formation, and structure near tryptophan residues for a variety of proteins.

[1]This research supported in part by NSF Grants EIA-0103742, ACR-0081510, ACR-0113971, CCR-0113974, ACI-0326350, by the DOE, and by HP. Tapia supported in part by a NIH Molecular Biophysics Training Grant (T32GM065088) and previously supported by a Department of Education GAANN Fellowship. Thomas supported in part by a Department of Education GAANN Fellowship and previously supported by a NSF Graduate Research Fellowship and a P.E.O. Scholarship.

1 Introduction

As proteins fold to their native, functional state, they undergo critical conformational changes that effect their functionality. Some conformational changes are detrimental. For example, diseases such as Mad Cow disease or Alzheimer's disease are caused by misfolded proteins [6]. Insight into the kinetics and detailed mechanics of the folding process will help explain critical information about the protein such as its function and why it misfolds.

In lab experiments, kinetic measurements are used frequently to quantify the folding process. Numerous lab experimental methods such as Circular Dichroism (CD spectra), fluorescence studies, hydrogen-deuterium exchange, and pulse-labeling [26], give time-scaled based views of the folding process. These measurements can be used to compare the kinetics of a group of proteins. For example, often the effects of mutations can be studied in detailed through the comparison of kinetic metrics.

Simulating protein folding kinetics has been a difficult task performed on small structures through computationally expensive methods such as molecular dynamics or Monte Carlo simulations. Studies on larger proteins have recently been accomplished, but these simulations have only been done on limited proteins represented with coarse models.

In our previous work [2, 32], we studied protein folding through the application of a method that builds an approximate map of a protein's potential energy landscape. This map contains thousands of feasible folding pathways to the known native state enabling the study of global landscape properties. We obtained promising results for several proteins [32]. The pathways were validated by comparing secondary structure formation order with known experimental results. However, we were unable to study kinetic properties such as relative folding rates and population kinetics.

This work introduces new methodologies for studying the kinetics of protein folding: Map-based Master Equation (MME) and Map-based Monte Carlo (MMC) solution. These techniques provide quantitative kinetic measurements such as relative folding rates and population kinetics that we could not obtain before from our maps. In contrast to other methods such as the traditional master equation and Monte Carlo simulation, our techniques are both fast and can easily be computed for full length and detailed protein models. We also show that these two new techniques facilitate the study of interesting metrics that describe the folding process, reaction coordinates. In this paper we focus on two metrics: formation of helices and structure formation around tryptophan residues. These two metrics are often studied in the lab through CD spectra analysis and tryptophan fluorescence studies, respectively [26]. We validate our techniques by comparing folding rates to known experimental results. We also look in depth at the population kinetics and the reaction coordinates for a variety of proteins in order to correlate our simulation results with the trends seen in experiment.

We invite the community to help us enrich our publicly available database of motions and kinetics analysis by submitting to our server: http://parasol.tamu.edu/foldingserver/

2 Related Work

There are many different methods for studying protein folding kinetics. In this section we briefly introduce some of the methods, give insight into their strengths and weaknesses, and discuss the kinetics that each method provides.

Molecular Dynamics. Molecular dynamics simulates the dynamics of the folding process using Newton's classical equations of motion. The forces applied are usually approximations computed using the first derivative of an empirical potential function. Molecular dynamics studies are highly realistic and help give insight into how proteins fold in nature. They also facilitate study of the underlying folding mechanism, provide folding pathways, and identify intermediate folding states. While they give physically realistic simulations, these simulations come at a large computational cost. For example, it has taken months of supercomputer time to simulate a microsecond of a very small (36 residues) protein folding [10] using molecular dynamics! Researchers are identifying ways to counteract the cost of MD sim-

ulations. For example, the The Folding@Home distributed computing project [28] computes MD simulations with a cluster of over 30,000 computers worldwide.

Monte Carlo Simulation. Monte Carlo simulation finds a single folding trajectory [8, 14]. However, each run is computationally expensive because at each point in the conformation space search, complex kinetics and thermodynamics are simulated. Multiple runs are often done because the search is stochastic. Like molecular dynamics, Monte Carlo simulations provide highly realistic insight into the folding process.

Master Equation Kinetics. Folding kinetics have also been studied through a computation across the folding landscape. One way this has been done is through the use of lattice models that have enumerated the folding landscape, and then the master equation is computed for this landscape [7, 22, 23, 25]. One advantage of these approaches is that the transition state emerges from the dominate modes of the master equation solution. However, these models are very simplistic and do not represent real structures or sequences. Recent applications of the master equation have been able study proteins with full structures [33]. However, the enumeration of the folding landscape is limited to the formation of contact clusters, which are groupings of nearby contacts as derived from the native-state contact map.

Statistical Mechanical Methods. Statistical mechanical methods have also been successful in studying protein folding kinetics. These methods have provided estimates of the transition state ensemble, folding rates, and Φ-values [20, 1]. Only recently has this method been applied to larger protein structures of up to 349 residues [9]. However, these models use a very simplified energy function that depends only on the topology of the protein's native state and hence are not as accurate as the distance from the native state increases (as the protein unfolds).

SRS and PFold. Stochastic Roadmap Simulations (SRS) samples motions and studies kinetics by modeling the folding energy landscape as a network of conformations where the connection between two conformations in the network reflects the transition probability between them. In early SRS work [3], the protein structure was modeled as a sequence of rigid secondary structure pieces and the packing order of these elements was studied.

In recent work [5], SRS was shown to identify the transition state ensemble and it was used to compute folding rates and Φ-values. In order to identify the transition state ensemble, the conformation is modeled as a binary vector where each bit represents a sequence of five residues. The bit is set to 0 if the subsequence is non-native or 1 if it is native-like. All possible conformations and transitions (i.e., a single bit change) were enumerated in the model. To compute Pfold, the probability of folding, they perform random walks from every conformation until it reaches either the folded state or the unfolded state. Pfold for a given conformation is then the percentage of times a random walk from that conformation reaches the folded state before the unfolded state. Transitions are not allowed out of either the folded or the unfolded state.

In this model, Pfold helps identify the transition state ensemble. They use this ensemble to calculate relative folding rates and Φ-values. However, their model only contains a single unfolded state. Thus each conformation in their model does not represent the same volume of the energy landscape. In a more realistic model, it is unlikely that there will be a single, unique unstructured ('unfolded') state, thus making the Pfold calculation more difficult for use with more structurally accurate models.

Our Contribution. The techniques introduced in this paper, MME and MMC, provide quantitative kinetic measurements such as relative folding rates and population kinetics. Also, interesting reaction coordinates such as helix formation and structure around tryptophan residues can be monitored during the simulated folding process. In our previous work [2, 32], we provided methods for building an approximate map of a protein's potential energy landscape [2, 32] and an RNA's folding landscape [31]. We have published results from our approximate maps for proteins up to 148 residues easily built on a desktop PC [32]. These maps provide a framework for the MME and MMC techniques.

In contrast to other methods such as the tradi-

tional master equation and Monte Carlo simulation, MME calculation and MMC simulation are both fast and can easily be computed for full length and detailed protein models. The MME calculation gives insight into the folding rate, equilibrium distribution, and transition states. The MMC simulation gives a stochastic view of the folding process and allows the computation of population kinetics.

3 Roadmaps for Protein Folding

In previous work [2], we introduced an approach to protein folding that is based on the probabilistic roadmap approach for motion planning [13]. We applied our method to a large number of structures and were able to identify subtle differences in the known experimental secondary structure formation order for proteins with very similar structures [29, 32].

Our method is simple and consists of two main steps: (1) sampling conformations in the landscape and (2) making transitions between sampled conformations. In the first step, conformations (nodes) are sampled on the folding landscape, with a bias to increase density near the known native state. In the second step, connections (edges) are made between sampled conformations with similar structure. Weights are assigned to directed edges to reflect the energetic feasibility of transitioning between the two endpoint conformations. This combination of nodes and weighted edges forms a roadmap that approximates the energy landscape. This roadmap encodes thousands of folding pathways. The most energetically feasible pathways in the roadmap can be extracted using these weights.

Connections between two nodes, q_1 and q_2, are labeled with edge weights that reflect the energetic feasibility of transitioning between them. This is done by first identifying all the intermediate nodes, $q_1 = c_0, c_1, ..., c_{n-1}, c_n = q_2$, that connect q_1 to q_2. For each pair of consecutive conformations c_i and c_{i+1}, the probability P_i of transitioning from c_i to c_{i+1} depends on the difference between their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$:

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

This keeps the detailed balance between two adjacent states and enables the edge weight to be computed by summing the logarithms of the probabilities for all pairs of consecutive conformations in the sequence. With this edge weight definition, we can use simple graph search algorithms to extract the most energetically feasible pathways in the roadmap between two given states (e.g. from the unfolded state to the folded state).

Protein Model. We model the protein as an articulated linkage. Using a standard modeling assumption for proteins that bond angles and bond lengths are fixed [30], the only degrees of freedom in our model are the backbone's phi and psi torsional angles which are modeled as revolute joints with values in the range $[0, 2\pi)$.

Potential Energy Calculation. Our method is flexible and allows any potential function to be used. In this paper, we use a coarse potential function similar to [18]. We use a step function approximation of the van der Waals potential component and model side chains as spheres with zero dof. If any two spheres are too close (i.e., less than 2.4Å during sampling and 1.0Å during connection), a very high potential is returned. Otherwise, the potential is:

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

where K_d is 100 kcal/mol and $d_0 = d_c = 2$ Å as in [18]. The first term represents constraints favoring known secondary structure through main-chain hydrogen bonds and disulphide bonds, and the second term is the hydrophobic effect. The hydrophobic effect is computed as follows: if two hydrophobic residues are within 6 Å of each other, then the potential is decreased by 100 kJ/mol.

4 Map-based Kinetics Analysis

Our roadmaps give an approximate view of the protein folding landscape. In the past, we have success-

fully extracted low-energy pathways, validated secondary structure formation order, and seen general and consistent trends in reaction coordinates such as native contacts present and RMSD. However, we had not been able to extract important kinetic measures such as folding rates and population kinetics. In this section, we introduce two new techniques that enable us to extract this information from our roadmaps: Map-based Master Equation solution (MME) and Map-based Monte Carlo simulation (MMC). Unlike traditional master equation calculation and Monte Carlo simulation, these techniques run very fast and can be applied to full length detailed protein models.

The application of the MMC technique to the approximated landscape reduces the roadmap to a set of stochastic configurations and pathways. The benefit of this reduction is the ability to measure reaction coordinates, metrics that describe events during the time evolution of the folding process. In this section, we also present methods for using the MMC pathways to calculate two such reaction coordinates: helix formation and formation of structure around tryptophan residues.

4.1 Map-based Master Equation (MME)

The master equation calculation gives insight into the folding rate, the equilibrium distribution, and transition states. However, it requires a detailed model of the possible conformations and their associated transitions. In the past, this has been done by enumerating landscapes – feasible only for small protein models or segments.

In this work we develop a strategy for applying the master equation to the approximation of the folding landscape provided by our roadmaps. As we will show, our roadmaps provide a suitable framework to apply the master equation without requiring an enumeration of the conformation space. A major benefit of this is that the Map-based Master Equation (MME) technique enables us to apply the master equation to much larger proteins than was possible before.

Master equation formalism has been developed for folding kinetics in a number of earlier studies [12, 33]. The stochastic process of folding is represented as a set of transitions among all n conformations (states). The time evolution of the population of each state, $P_i(t)$, can be described by the following master equation:

$$dP_i(t)/dt = \sum_{i \neq j}^{n} (k_{ji} P_j(t) - k_{ij} P_i(t)) \tag{3}$$

where k_{ij} denotes the transition rate from state i to state j. Thus, the change in population $P_i(t)$ is the difference between transitions *to* state i and transitions *from* state i.

If we use an n-dimensional column vector $\mathbf{p}(\mathrm{t}) = (P_1(t), P_2(t), \ldots, P_n(t))'$ to denote the population of n conformational states, then we can construct an $n \times n$ matrix M to represent the transitions, where

$$\begin{cases} M_{ij} = k_{ji} & i \neq j \\ M_{ii} = -\sum_{i \neq j} k_{ij} \end{cases} \tag{4}$$

The master equation can be represented in matrix form:

$$d\mathbf{p}(t)/dt = M\mathbf{p}(t). \tag{5}$$

The solution to the master equation is:

$$P_i(t) = \sum_k \sum_j N_{ik} e^{\lambda_k t} N^{-1}_{kj} P_j(0) \tag{6}$$

where N is the matrix of eigenvectors N_i for the matrix M in equation 4 and Λ is the diagonal matrix of its eigenvalues λ_i. $P_j(0)$ is the initial population of conformation j.

From equation 6, we see that the eigenvalue spectrum is composed of n modes. If sorted by magnitude in ascending order, the eigenvalues include $\lambda_0 = 0$ and several small magnitude eigenvalues. Since all the eigenvalues are negative, the population kinetics will stabilize over time. The population distribution $\mathbf{p}(t)$ will converge to the equilibrium Boltzmann distribution, and no mode other than the mode with the zero eigenvalue will contribute to the equilibrium. Thus the eigenmode with eigenvalue $\lambda_0 = 0$ corresponds to the stable distribution, and its eigenvector corresponds to the Boltzmann distribution of all conformations in equilibrium.

Similarly, we see that the large magnitude eigenvalues correspond to the fast folding modes, that is, those modes which fold in a burst. Their contribution to the population will die away quickly. Similarly, the smaller the magnitude of the eigenvalue is, the more influence its corresponding eigenvector has on the global folding process. Thus, the global folding rates are determined by the slow modes.

For some folders (2-state folders), their folding rate is dominated by only one non-zero slowest mode. If we sort the eigen spectrum by ascending magnitude, there will be one other eigenvalue λ_1 in addition to eigenvalue λ_0 that is significantly smaller in magnitude than all other eigenvalues. This λ_1 corresponds to the folding mode that determines the global folding rate. We will refer to it as the *master folding mode*. Its corresponding eigenvector denotes its contribution to the population of each state. Hence, the large magnitude components of the eigenvector correspond to the states whose populations are most impacted by the master folding mode. These states are the transition states [24, 25].

We apply the master equation formalism to our roadmaps by assigning each node in our roadmap to a row (and column) in the matrix M. The transition rates are computed directly from the edge weight: $K_{ij} = K_0 e^{-W_{ij}}$. K_0 is the constant coefficient adjusted according to experimental results. We will use MME to compute the relative folding rates for several proteins with known kinetics.

4.2 Map-based Monte Carlo (MMC)

Population kinetics provides information about the time evolution of different conformational populations. In our earlier work, we simply extracted the most energetically feasible paths in the roadmap to study the folding process. However, this does not mirror the stochastic folding process and cannot be used to determine the type of kinetic information that we are interested in here. In this paper, we show how we can adapt Monte Carlo simulation and apply it directly to our roadmaps. Because the roadmap approximates the energy landscape, we can use the pathways computed by the Map-based Monte Carlo (MMC) simulation to compute population kinetics.

Applying Monte Carlo simulation to our approximated landscape allows for the study of large protein structures with only a small computational cost. Previously, the size of the protein's conformational space limited the application of Monte Carlo techniques to small proteins (e.g. all-atom 56 residue protein [27]). However, our roadmap provides a precomputed framework for this walk and greatly simplifies the computation required by Monte Carlo analysis.

In order to apply the Monte Carlo technique to our roadmap, we must ensure that the likelihood of transitioning from one neighbor to another is probabilistically biased by their Boltzmann transition probabilities. During roadmap construction, we compute edge weights that reflect the energetic feasibility to transition from one neighbor to another. We turn these edge weights into transition probabilities to perform the Monte Carlo simulation. One way to do this is to cluster the edge weights into disjoint buckets that reflect a grouping of edge weight qualities. After all edge weights are assigned a bucket, edge weights within a bucket are assigned a probability Q_{ij} reflecting their quality within the bucket. In doing so, the probability of each edge weight is assigned in a biased Gaussian fashion that favors clear discrimination of low edge weights, yet still can differentiate between edges of all weights. Then the probability to transition between two states, P_{ij} can be calculated as:

$$P_{ij} = \begin{cases} \frac{Q_{ij}}{1+\sum_{j=0}^{n-1} Q_{ij}} & \text{if } j \neq i \\ \frac{1}{1+\sum_{j=0}^{n-1} Q_{ij}} & \text{if } j = i \end{cases} \quad (7)$$

where n is the number of outgoing edges from node i. This ensures the sum of all probabilities (including the self-transition probability) out of node i is one. Note that the transition probability is dependent on the number of outgoing edges from a node. Since during roadmap construction we only attempt connections between the k closest neighbors according to some distance metric, the out-degree for all nodes is roughly similar. Thus, this transition probability calculation is fair to all nodes in the roadmap and maintains detailed balance.

4.2.1 Helix Formation

The protein folding process can be monitored in the lab through the formation of local portions of the three-dimensional structure of the protein. These local segments, commonly helices and strands, are the secondary structure of the protein. In the lab, the average formation of secondary structure can be measured through the technique of far-UV CD spectroscopy. At far-UV wavelengths (190-250 nm) the chromophore is the peptide bond and the resulting signal from CD spectroscopy appears when the peptide bond is located in a regular folded environment. It is common to monitor the formation of a specific type of secondary structure during the folding process by performing CD spectroscopy at a certain wavelength. One of the most common measurements is done at the wavelength of 220nm where the formation of helices can be monitored.

There are many ways to measure helix formation in silico. In statistical mechanical simulations, the protein backbone is modeled by a sequence of dihedral angles, one angle between each pair of residues [9]. Helix formation has been measured from these simulations by summing the individual angle change between conformations. Unlike the single angle per residue model, our model consists of two angles that can be independently similar or dissimilar. Given this independence and a more complex protein model, we explored alternative ways of defining the formation of helices. Also unlike the statistical mechanical model, our pathways and configurations are extracted stochastically through the MMC technique.

In the results presented in this paper, we used a measurement of helix formation that calculates the native contact formation in helices, $H(t)$, as a function of time step, t, from the MMC simulation:

$$H(t) = \frac{\sum_{ij} H_{ij}(t)}{H(native)} \qquad \text{where } i,j \in helix \quad (8)$$

The contribution of a single contact, $H_{ij}(t)$, is equal to 1 if the residue pair (i, j) forms a native contact in the configuration at time step t. In order to compare results across proteins, the values of $H(t)$ are normalized by the number of contacts at helices measured at the protein's native state, $H(native)$. Thus, 1 represents the full formation of the helix structures in a configuration and 0 represents no helix structure formed.

4.2.2 Tryptophan Structure Formation

The protein folding process can also be studied in the lab by monitoring the fluorescence of certain amino acids. The fluorescence yield of these amino acids is determined by their local environment given the configuration of the protein. While all aromatic amino acids are known to fluoresce under certain conditions, the tryptophan residue is often favored for experiments because of its high fluorescence yield.

Even though tryptophan rarely occurs in proteins, it is common to mutate a protein to make fluorescence studies possible. Tryptophan can be introduced into the structure where fluorescence yield is optimized through site-directed mutagenesis. For example, they are often placed in the core of the protein and away from polar amino acids that detract from their yield.

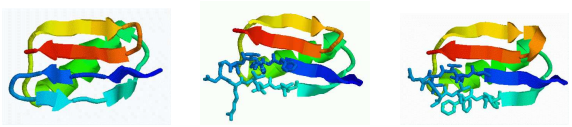
In order to monitor the local environment of the tryptophan residues, we explore the effect of native contacts. As tryptophans are involved in native contacts, their local environment becomes more similar to the environment in the native state. At that native structure, we expect their fluorescence to be maximized. A similar approach was used in [9]. However, unlike [9], our pathways and configurations are extracted stochastically through MMC.

In the results presented in this paper, we use a measurement of tryptophan structure formation that calculates the native contact formation tryptophan residues, $Trp(t)$, as a function of time step, t, from the MMC simulation.

$$Trp(t) = \frac{\sum_{ij} Trp_{ij}(t)}{Trp(native)} \qquad \text{where } i,j \in tryptophan \quad (9)$$

The contribution of a single contact, $Trp_{ij}(t)$, is equal to 1 if the residue pair (i, j) forms a native contact in the configuration at time step t and either i or j is a tryptophan. This is a simple measure and could be modified for more complex local environ-

ments impacting fluorescence yield. In order to compare results across proteins, the values of $Trp(t)$ are normalized by the number of contacts in the native state involving tryptophans, $Trp(native)$. Thus, a value of 1 represents the full formation of the structure involving the tryptophan residues, and a value of 0 represents no tryptophan structure formed.



(a) (b) (c)

Figure 1: Ribbons diagrams for (a) protein G and its variants (b) NuG1 and (c) NuG2. The mutated hairpin is in wireframe for NuG1 and NuG2.

5 Experimental Results

In this section we present results demonstrating how we can extract kinetics information from our roadmaps. We show that our Map-based Master Equation (MME) can accurately compute the relative folding rates of protein G and two of its variants. Then we use our Map-based Monte Carlo (MMC) simulation to investigate the folding population kinetics of the native state for several small proteins studied in our previous work [32]. When available, the helix formation and tryptophan contact formation calculated during the folding process of these proteins is also shown. It would be computationally prohibitive to apply the traditional Monte Carlo simulation or Master equation calculation to these proteins and detailed protein model, hence we cannot compare to them.

5.1 Relative Folding Rates by MME

One interesting protein to study is protein G (Figure 1 (a)). Protein G is a small two-state folder composed of a central α-helix and two β-hairpins. Nauli et al. [21] created two mutants of protein G to alter its folding behavior to switch the hairpin formation order while maintaining the same secondary and tertiary structure, NuG1 (Figure 1 (b)) and NuG2 (Figure 1 (c)). They also show that these two mutants fold 100 times faster than protein G.

We used our new MME to compute the relative folding rates of these two proteins on roadmaps that reached stable secondary structure formation order. In the results shown here, the potential values were normalized to fall between 0 and 1 for the fastest computation of the master equation solution. Table 1 gives the magnitudes of the 10 smallest eigenvalues for each protein. Figure 2 shows the magnitudes of the 5 smallest eigenvalues. Recall that the smallest non-zero eigenvalues represent the rate-limiting barrier in the folding process. Therefore, they have the largest impact on the global folding rate. As seen in the magnitude of the second eigenvalue in Figure 2, protein G folds much slower than the two mutants, NuG1 and NuG1. Also, NuG1 and NuG2 fold at very similar rates. This matches what has been been seen in lab experiments. While in previous work [32] we were able to accurately identify the hairpin formation order of protein G and mutants NuG1 and NuG2, we were unable to study the change in folding rate.

Index	Protein G	NuG1	NuG2
1	9.44e-15	4.59e-14	3.78e-14
2	2.56e-2	2.13e+0	6.68e+0
3	2.97e+0	2.25e+0	7.53e+0
4	6.69e+0	2.48e+0	8.31e+0
5	6.85e+0	3.80e+0	8.90e+0
6	6.97e+0	4.33e+0	10.29e+0
7	7.10e+0	5.77e+0	11.40e+0
8	7.48e+0	7.59e+0	12.16e+0
9	7.70e+0	8.86e+0	12.21e+0
10	9.64e+0	9.91e+0	12.64e+0

Table 1: Magnitudes of the 10 smallest eigenvalues computed by MME for protein G and its mutants NuG1 and NuG2.

Figure 3 shows the performance of MME for roadmaps ranging in size from 2000 to 15000 nodes. The running time of MME scales linearly with roadmap size (i.e., the size of the landscape model). Thus, MME has an advantage over the traditional

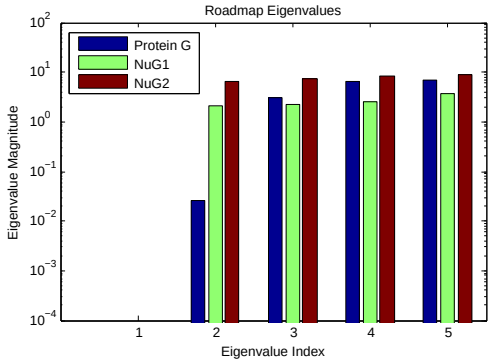


Figure 2: Eigenvalue comparison between protein G and mutants NuG1 and NuG2 computed by MME. NuG1 and NuG2 are experimentally known to fold 100 times faster than protein G [21].

master equation solution. While traditional master equation solution is usually applied to a fully enumerated landscape, MME is only computationally limited by the size of the approximated landscape model. Here we have shown that this approximated model can be a subset of the entire configuration space. This enables us to study larger proteins with more detailed models than can be handled by traditional techniques.

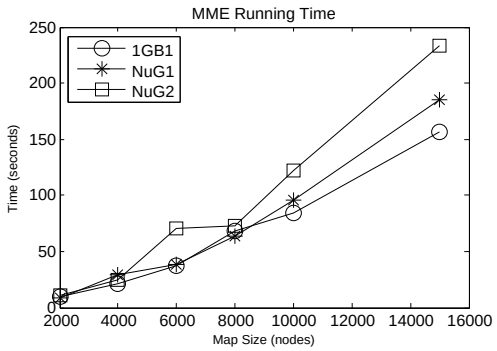


Figure 3: Running time of MME for protein G, and its variants NuG1 and NuG2 as a function of roadmap size. MME scales linearly with the size of the landscape model/map.

5.2 Folding Kinetics by MMC

We can also study the folding process by computing the population kinetics of the native state with our new MMC simulation. A single roadmap encodes thousands of folding pathways. Previously, we extracted folding pathways by finding the most energetically feasible pathways in the roadmap. While this provided useful information about high level folding events such as the temporal ordering of secondary structure which we could validate against experiment, we could not use the deterministically extracted pathways to infer kinetic information. By instead extracting pathways stochastically using MMC, we can now compute population kinetics for different states. For example, we can compare the population kinetics of the unfolded state and the folded state.

We computed the population kinetics of several two-state folders studied in our previous work [32] (see Table 2). In that work, we were able to produce roadmaps whose secondary structure formation order matched native state out-exchange experiments and pulsed-labeling experiments when available [19]. We use the same roadmaps here, but are able to supplement our previous results by using MMC to compute the population kinetics of the folded state and of the unfolded state. Table 2 also displays the MMC analysis time. In all cases, the analysis took less than 1 hour on a 2.4 GHz desktop PC with 512 MB RAM.

Figure 4 displays the results for several proteins studied. MMC was run for 500 iterations and 50,000 time steps. Our experience shows that this provided population kinetics with small variance. These proteins are similar in size (ranging from 53 to 86 residues) and varying secondary structure makeup. We study all α proteins, all β proteins, and mixed α and β proteins.

Notice that the population kinetics of the native state for the all α proteins (Figure 4(a,b)) shows a gradual growth at a constant rate. The all β proteins (Figure 4(c)) and mixed proteins (Figure 4(d–f)), however, display a steep climb in their population kinetics and then plateau. We believe this is due to nucleation effects (e.g., that each native contact does not have the same probability of forming) present in structures containing β-sheets. For exam-

Protein Name	PDB ID	Length	SS	Nodes	Edges	MMC Time (m)
Dv Rubredoxin (RdDv)	1rdv	52	2α+3β	4000	206440	20.83
Murine Epidermal GF (mEGF)	1egf	53	3β	4000	199600	19.94
Cp Rubredoxin (RdCp)	1smu	54	3α+3β	6000	200072	22.19
Protein G, domain B1 (Protein G)	1gb1	56	1α+4β	4000	198588	20.71
Protein A, domain B (Protein A)	1bdd	60	3α	6000	276342	23.12
Acyl-coenzyme A Binding Protein (ACBP)	2abd	86	5α	18000	953900	35.94

Table 2: Proteins studied and MMC analysis time.

ple, a contact near the turn of a β-hairpin (i.e., with lower effective contact order) has a greater probability to form early while more non-local native contacts such as those at the end of the hairpin have a lower probability to form early. Their formation probability increases as the protein folds/nucleates. This is commonly referred to as a "zipping" process [11]. Conversely, most contacts in an α-helix are local (i.e., have a low effective contact order) thus their formation probabilities are all similar and constant throughout the folding process.

In order to contrast the population kinetics of the folded state, we also studied the population kinetics of the unfolded ensemble (Figure 4). For this study, we defined the unfolded ensemble as those states with few native contacts (relative to the number of contacts in the native state). There is a clear relationship between the kinetics of the unfolded state to that of the folded state. For example, in protein A (Figure 4(a)) the population of the native state increases slowly as the population of the unfolded state ensemble decreases slowly. On the other hand, folding processes that reach folded equilibrium quickly also see a quick decrease in the population of the unfolded state ensemble.

A nice feature of the MMC technique is that it allows us to study stochastic events during the protein folding process. For the proteins studied above through population kinetics, we also examined the structural metrics of helix formation and formation of structure around tryptophan residues (see Figure 5). From the combined information in these three plots, we can deduce characteristics of the folding process. In rest of this section, we compare the individual kinetic results produced by MMC to previous lab and simulation studies for each protein.

Protein A. The B domain of protein A, containing 3 α-helices, has been the focus of many experimental studies. It does not contain a tryptophan naturally, but has been mutated so that tryptophan fluorescence can be studied [14]. It has also been studied by lattice-based Monte Carlo technique [15]. However, this lattice model only used a coarse representation of the backbone carbon-αs to model the structure. In lab and simulation studies, protein A has demonstrated formation of helix structure followed by the packing of the helices in the final folded structure [19]. Our population kinetics (Figure 4(a)) and helix formation (Figure 5(a)) plots show similar trends. While the folding process begins early on (as indicated by continual growth in helix formation beginning at time step 1), it takes at least 100 time steps for any conformation to reach the native state. This suggests that helices are formed before any conformation reaches a shape close to the native state, as seen in experiment.

ACBP. A similar process is observed in the other all α protein, Acyl-coenzyme A Binding Protein (ACBP). This protein has five helices and two tryptophans in the core of the protein. The folding of ACBP has been studied in the lab through tryptophan fluorescence, and it has been shown that it has a fast, two-state folder [16]. From our MMC kinetics, we see that ACBP exhibits similar properties as the other all α protein, protein A: continual formation of helix contacts (Figure 5(b)) and reaching the native state after the formation of many helix contacts (Figure 4(b)). However, since ACBP has two tryptophans

in the core of the protein, we see a quick increase in the formation of these contacts (Figure 5(c)) around the same time we see the native state beginning to be populated, around time step 100. This could correspond to the packing of the structure and the formation of long-range interactions in the core of the protein.

mEGF. Since the protein murine epidermal growth factor (mEGF) has no helical structure, we do not plot its helix formation. While it does have two tryptophans, they are on the tail of the protein and do not make substantial contacts with the rest of the protein.

Protein G. The B1 domain of protein G has been the focus of many lab studies from CD spectra analysis and tryptophan fluorescence [21] to hydrogen exchange and pulse labeling experiments [19]. Much of the focus on the folding process of protein G has been on the folding order of its two sets of strands. However, it is known that the helix forms before the final stages of the folding process [19]. It is never the last secondary structure element to form. In our MMC results, we see a similar ordering. Figure 5(d) shows that the helix forms quickly and is 80% formed by time step 100. By this time step, less than 20% of the protein has reached a native like conformation (Figure 4(d)). The tryptophan contact formation (Figure 5(e)) continues through the folding process with continual packing around the protein core (where the tryptophan is located).

RdCp and RdDv. Cp Rubredoxin (RdCp) and Dv Rubredoxin (RdDv) are two Rubredoxins from mesophilic organisms. While their population kinetics are similar (Figure 4(e,f)), some small details can be elucidated from the reaction coordinates studied. For RdDv, that has been studied by high-temperature MD simulations [17], we see two jumps in the population kinetics (about 50% then 90% native-like). This could be due to the early packing of protein around the hydrophobic core, as seen in the continually increasing tryptophan structure formation (Figure 5(i)). The single tryptophan is in the core of the protein. After the core is formed, the helix finishes making a final set of contacts (Figure 5(h)). This corresponds with the second jump in the population kinetics to 90% native-like (Figure 4(f)). The behavior of opening the helix loop and then unfolding the core was also seen in MD simulation [17]. RdCp was shown through tryptophan fluorescence and far-UV CD experiments to have a simple two-state kinetic and no known intermediate [4]. We also see this in our simulations. The helix formation (Figure 5(f)) and tryptophan contact formation (Figure 5(g)) show cooperative and continual growth until the native state is fully populated.

6 Conclusion

We proposed and explored new analysis tools to study protein folding kinetics: Map-based Master Equation solution (MME) and Map-based Monte Carlo Simulation (MMC). With these new methods, we can compute folding rates and extract population kinetics of various states. We validated our folding rates against known experimental data. The MME approach was able to produce relative folding rates for three proteins, matching what has been seen in lab experiments. Our population kinetics were also able to identify clear kinetic differences in proteins of different structure. Through the combination of population kinetics and helix and tryptophan structure formation information, we are able to elucidate important characteristics in the folding process. For example, our results on Protein A show that helix structure forms early, before packing of core of the protein. This is also what has been seen in lab experiment. In another case, DvRD, the hydrophobic core is formed before the helices. This behavior was also seen in MD simulations.

A important benefit of these approaches is that it enables us to study the kinetics of much larger proteins that can be handled by traditional master equation methods or Monte Carlo simulation. We believe these new tools are valuable tools for discovering important features of protein folding kinetics.

7 Acknowledgments

We would like to thank Annette Stowasser for her initial work and exploration of measuring tryptophan

contact formation and its relation to fluorescence simulation. We would also like to thank Dr. Mauricio Lasagna of the Reinhart Lab at Texas A&M University for sharing his expertise of lab-based experimentation of tryptophan fluorescence.

References

[1] E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA*, 96(20):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. *J. Comput. Biol.*, 10(3-4):239–256, 2003. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2002.

[3] M. Apaydin, A. Singh, D. Brutlag, and J.-C. Latombe. Capturing molecular energy landscapes with probabilistic conformational roadmaps. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 932–939, 2001.

[4] S. Cavagnero, Z. H. Zhou, M. W. W. Adams, and S. I. Chan. Unfolding mechanism of rubredoxin from pyrococcus furiosus. *Biochemistry*, 37:3377–3385, 1998.

[5] T.-H. Chiang, D. Hsu, M. S. Apaydin, D. L. Brutlag, and J.-C. Latombe. Predicting experimental quantities in protein folding kinetics using stochastic roadmap simulation. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 410–424, 2006.

[6] F. Chiti and C. Dobson. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333–366, 2006.

[7] M. Cieplak, M. Henkel, J. Karbowski, and J. R. Banavar. Master equation approach to protein folding and kinetic traps. *Phys. Rev. Lett.*, 80:3654–3657, 1998.

[8] D. Covell. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.*, 14(4):409–420, 1992.

[9] P. Das, C. Wilson, G. Fossati, P. Wittung-Stafshede, K. Matthews, and C. Clementi. Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment. *Proc. Natl. Acad. Sci. USA*, 102:14569–14574, 2005.

[10] Y. Duan and P. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[11] K. M. Fiebig and K. A. Dill. Protein core assembly processes. *J. Chem. Phys*, 98(4):3475–3487, 1993.

[12] N. G. V. Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, New York, 1992.

[13] L. E. Kavraki, P. Svestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

[14] A. Kolinski and J. Skolnick. Monte Carlo simulations of protein folding. *Proteins Struct. Funct. Genet.*, 18(3):338–352, 1994.

[15] A. Kolinski and J. Skolnick. Monte Carlo simulations of protein folding ii. application to protein a, rop, and crambin. *Proteins Struct. Funct. Genet.*, 18(3):353–366, 1994.

[16] B. B. Kragelund, P. Hojrup, M. S. Jensen, C. K. Schjerling, E. Juul, J. Knudsen, and F. M. Poulsen. Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.*, 256:187–200, 1996.

[17] T. Lazaridis, I. Lee, and M. Karplus. Dynamics and unfolding pathways of a hyperthermophilic

and mesophilic rubredoxin. *Protein Sci.*, 6:2589–2605, 1997.

[18] M. Levitt. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, 170:723–764, 1983.

[19] R. Li and C. Woodward. The hydrogen exchange core and protein folding. *Protein Sci.*, 8(8):1571–1591, 1999.

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

[21] S. Nauli, B. Kuhlman, and D. Baker. Computer-based redesign of a protein folding pathway. *Nature Struct. Biol.*, 8(7):602–605, 2001.

[22] S. Ozkan, I. Bahar, and K. Dill. Tranisition states and the meaning of ϕ-values in protein folding kinetics. *Nat. Struct. Biol.*, 8(9):765–769, 2001.

[23] S. Ozkan, K. Dill, and I. Bahar. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilizaiton model. *Protein Sci.*, 11:1958–1970, 2002.

[24] S. B. Ozkan, K. A. Dill, and I. Bahar. Fast-folding protein kinetics, hidden intermediates, and the seuential stabilization model. *Protein Sci.*, 11:1958–1970, 2002.

[25] S. B. Ozkan, K. A. Dill, and I. Bahar. Computing the transition state population in simple protein models. *Biopolymers*, 68:35–46, 2003.

[26] H. Roder, K. Maki, and H. Cheng. Early events in protein folding explored by rapid mixing methods. *Chem. Rev.*, 106:1836–1861, 2006.

[27] J. Shimada and E. I. Shakhnovich. The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. *Proc. Natl. Acad. Sci. USA*, 99(17):11175–11180, 2002.

[28] M. Shirts and V. Pande. Screen savers of the world unite. *Science*, 290:1903–1904, 2000.

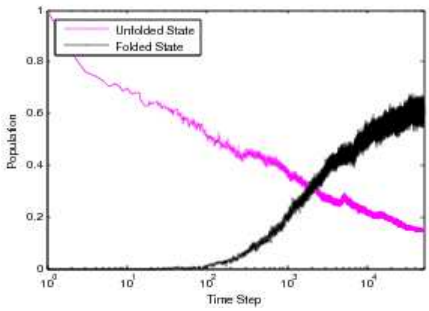
[29] G. Song, S. Thomas, K. Dill, J. Scholtz, and N. Amato. A path planning-based study of protein folding with a case study of hairpin formation in protein G and L. In *Proc. Pacific Symposium of Biocomputing (PSB)*, pages 240–251, 2003.

[30] M. J. Sternberg. *Protein Structure Prediction.* OIRL Press at Oxford University Press, 1996.

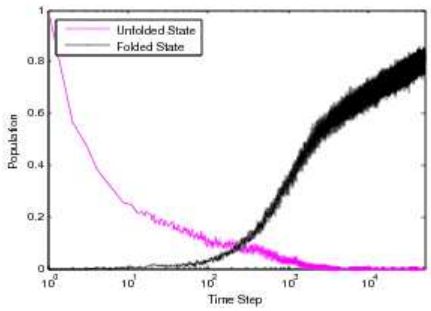
[31] X. Tang, S. Thomas, L. Tapia, and N. M. Amato. Tools for simulating and analyzing rna folding kinetics. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, 2007.

[32] S. Thomas, X. Tang, L. Tapia, and N. M. Amato. Simulating protein motions with rigidity analysis. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 394–409, 2006.

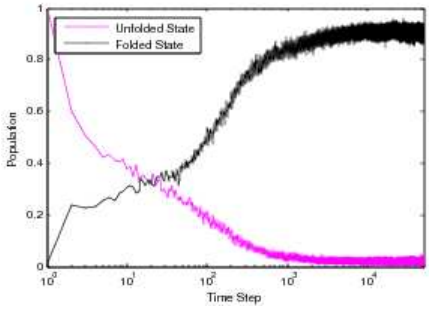
[33] T. Weikl, M. Plassini, and K. Dill. Coopertivity in two-state protein folding kinetics. *Protein Sci.*, 13:822–829, 2004.



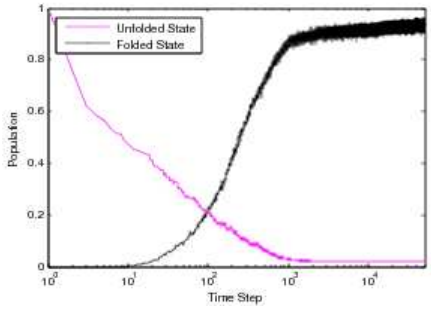
(a) Protein A: Population Kinetics



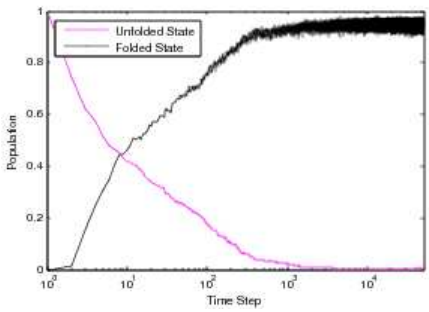
(b) ACBP: Population Kinetics



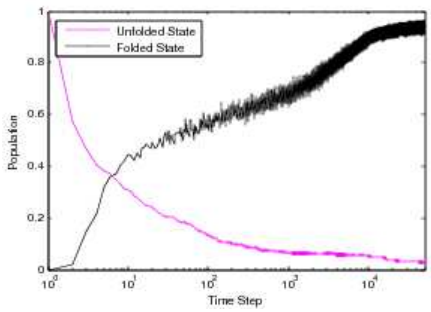
(c) mEGF: Population Kinetics



(d) Protein G: Population Kinetics

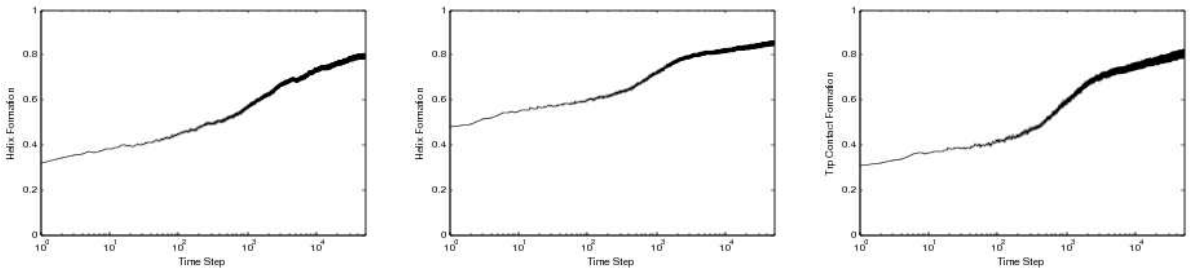


(e) RdCp: Population Kinetics

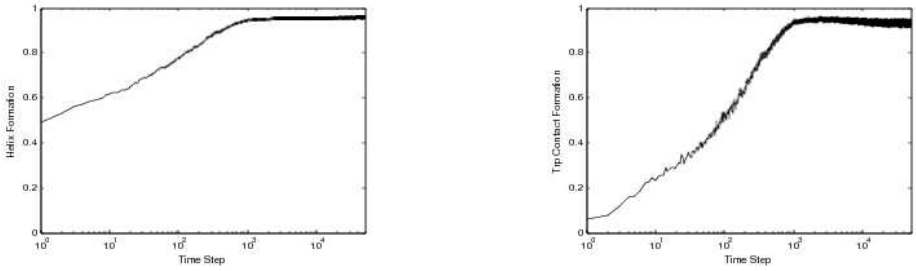


(f) RdDv: Population Kinetics

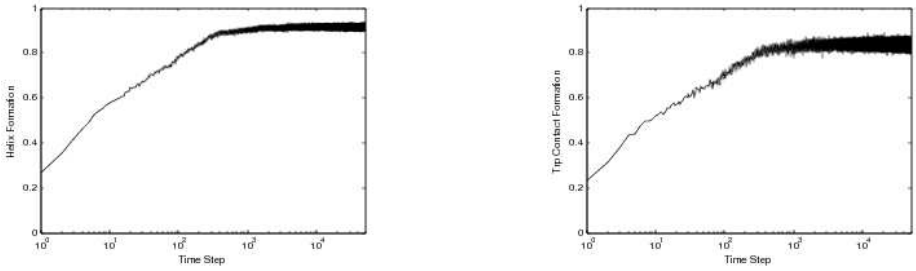
Figure 4: Population kinetics from MMC simulations for proteins in Table 2 of varying structure: (a,b) α, (c) β, (d–f), mixed.



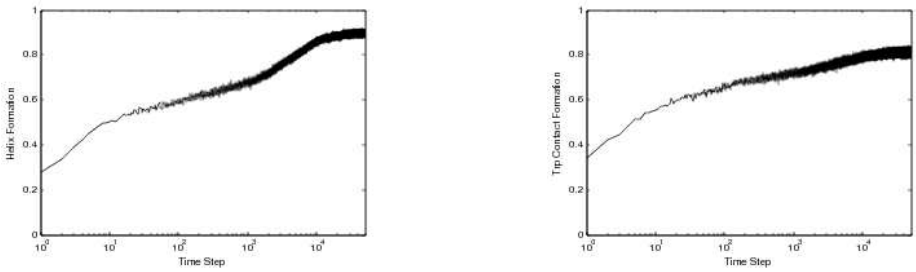
(a) Protein A: Helix Formation (b) ACBP: Helix Formation (c) ACBP: Tryptophan Contact Formation



(d) Protein G: Helix Formation (e) Protein G: Tryptophan Contact Formation



(f) RdCp: Helix Formation (g) RdCp: Tryptophan Contact Formation



(h) RdDv: Helix Formation (i) RdDv: Tryptophan Contact Formation

Figure 5: Reaction coordinates calculated from MMC simulations for proteins in Table 2 of varying structure: (a–c) α and (d–i) mixed. Tryptophan contact formation is not displayed for protein A because it does not contain any tryptophan residues. Note that mEGF (all β) is not displayed because it lacks α-helices and does not contain any tryptophan residues in the folding core.