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

Nancy M. Amato†
amato@cs.tamu.edu

Ken A. Dill‡
dill@maxwell.ucsf.edu

Guang Song+
gsong@cs.tamu.edu

ABSTRACT

We present a novel approach for studying the kinetics of protein folding. The 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 used a PRM-based technique to study protein folding pathways of several small proteins and obtained encouraging results. 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.

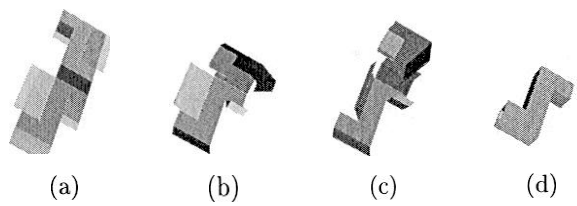


Figure 1: Snapshots of a carton folding.

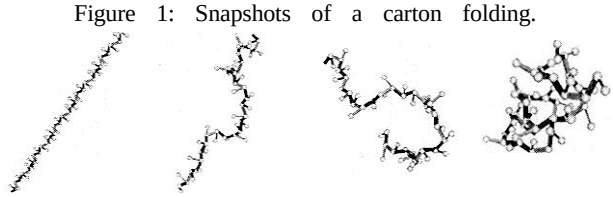


Figure 2: Snapshots of a 10 ALA chain folding

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., [26, 33]). 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 [36], 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 [35] in applying it to folding problems such as carton folding (with applications in packaging and assembly [28]), and paper crafts (studied in computational geometry [31]). For example, note the parallels between the periscope paper model folding and the small polypeptide folding in the path snapshots shown in Figures 1 and 2, respectively. In [36], 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 [36] how an analysis of the pathways contained in our roadmaps showed evidence of the two classes of folding kinetics described by Baldwin and Roses [8]. 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

*This research supported in part by NSF CAREER Award CCR-9624315, NSF Grants IIS-9619850, ACI-9872126, EIA-9975018, EIA-0103742, ACI-0113971, CCR-0113974, EIA-9810937, EIA-0079874, and by the Texas Higher Education Coordinating Board grant ARP-036327.017.

†Department of Computer Science, Texas A&M University, College Station, TX 77843-3112.

‡Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-1204.

RECOMB '02, April 18-21, 2002 Washington, D.C., USA

inferred from the ***distribution*** of the conformations contained in our roadmaps.

Related work. There are many interesting experimental results that have yet to be adequately explained or captured by theoretical models. For example, Baldwin and Rose [5] 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 topolgy. Based on this experimental observation, Baker et al. [2, 7] 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. [30] in their study of P-hairpin kinetics and was later used in the kinetics study of 20 plus proteins [29] with quite impressive results. However, despite the success of these models, there still exist many questions and uncertainties about the choices for 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 [12, 32], which would be very difficult (if not impossible) to capture with statistical mechanical models.

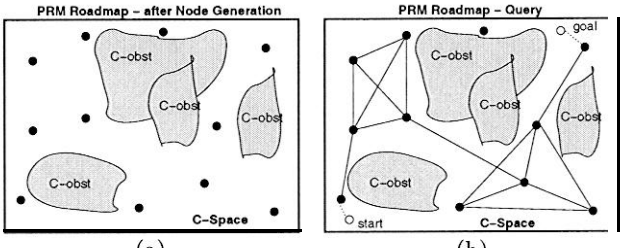
Other theoretical methods include analytical approaches [11], simulation of lattice models [11], and all-atom molecular dynamics simulations [15]. While each method has unique strengths and advantages, they all have weaknesses as well. For example, molecular dynamics simulations are computationally intensive and time-dependent, arc very sensitive to the initial conformation, and can easily result in local minima. In general, most of the proposed techniques have tremendous computational requirements because they attempt to simulate complex kinetics and thermodynamics at every point visited in conformation space.

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 more natural way to study kinetics, and should in principle 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.

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]. Typically, PRMs are used to construct a map of the feasible regions of the environment which can be used subsequently to answer many, varied motion planning queries. Briefly, PRMS work by sampling points 'randomly' from the movable object's configuration space', and retaining those that satisfy certain feasibility requirements (e.g., collision-free configurations, see Figure 3(a)). Then, these points are connected to form a graph, or roadmap, using some simple 'local' planning method to connect 'nearby' points (see Figure 3(b)). During query processing, paths connecting the start and goal configurations are extracted from the roadmap usin[3 standard graph search techniques (see Figure 3(b).)



(a) (b)

Figure 3: A PRM roadmap in C-space. A PRM roadmap: (a) after node generation, and (b) after the connection phase and being used to solve a query.

A major strength of PRMS is that they are quite simple to apply, even for problems with high-dimensional configuration spaces, requiring only the ability to randomly generate points in C-space, and then test them for feasibility (local connection can often be performed using multiple applications of the feasibility test).

In previous work, we proposed the PRM framework as a methodology for studying protein folding when the native structure is known [36]. 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 [27].

'The movableobject's *configuration* ***space,*** or C-space, is the set of all positions and orientations of the movable object, feasible or not [23].

2.1 Modeling proteins (C-space)

The amino acid sequence is modeled as a multi-link tree-like articulated 'robot', where flexible positions (e.g., atomic bonds) correspond to joints and rigid portions (e.g., atoms) correspond to links. Using a standard modeling assumption for proteins [37], 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); all atomic bond lengths and bond angles are assumed to be fixed. 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 two rotations at the ends don't 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 27r 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:

$$C = \{q \quad q \in S^{2n}\}. \tag{1}$$

Note that C simply denotes the set of all possible conformations, but says nothing about their feasibility. For protein folding, the validity of a point in 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 [36], sampling was biased by sampling from a selected set of normal distributions centered around the native structure. This worked well for small proteins ($\approx$ 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 (> 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.

Similar biased sampling strategies have been applied successfully in robotics applications [3, 10, 17, 18, 20, 24, 38], where oversampling in and near narrow passages in C-space is crucial for some problems. Also, as described in Section 1, Alm and Baker [7, 2] and Muñoz and Eaton [29] have used knowledge of the topology of the native state to predict the folding rates and mechanisms of some proteins.

2.2.1 Sampling strategy

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_α atoms of hydrophobic 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:

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

We set $E_{\min}$ = 50000 KJouls/mol and $E_{\max}$ = 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 [9, 34] and in our previous work on protein folding [36].

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_{frontier} nodes. We randomly pick N_{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_{frontier} in the lowest bin.) Each selected seed node q_o will be used to generate as many as N_{children} nodes ▪ these new nodes will be sampled from normal distributions with origin q_o 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 (Le., at least N_{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 than our original technique [36] in covering the conformation space for larger proteins.

2.2.2 Distribution of nodes

An interesting effect starts to emerge after node generation. The potential vs. RMSD distributions for several proteins are shown in Figure 4. Note the contrast between the distributions for all alpha (a and d) and all beta (c and f) 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

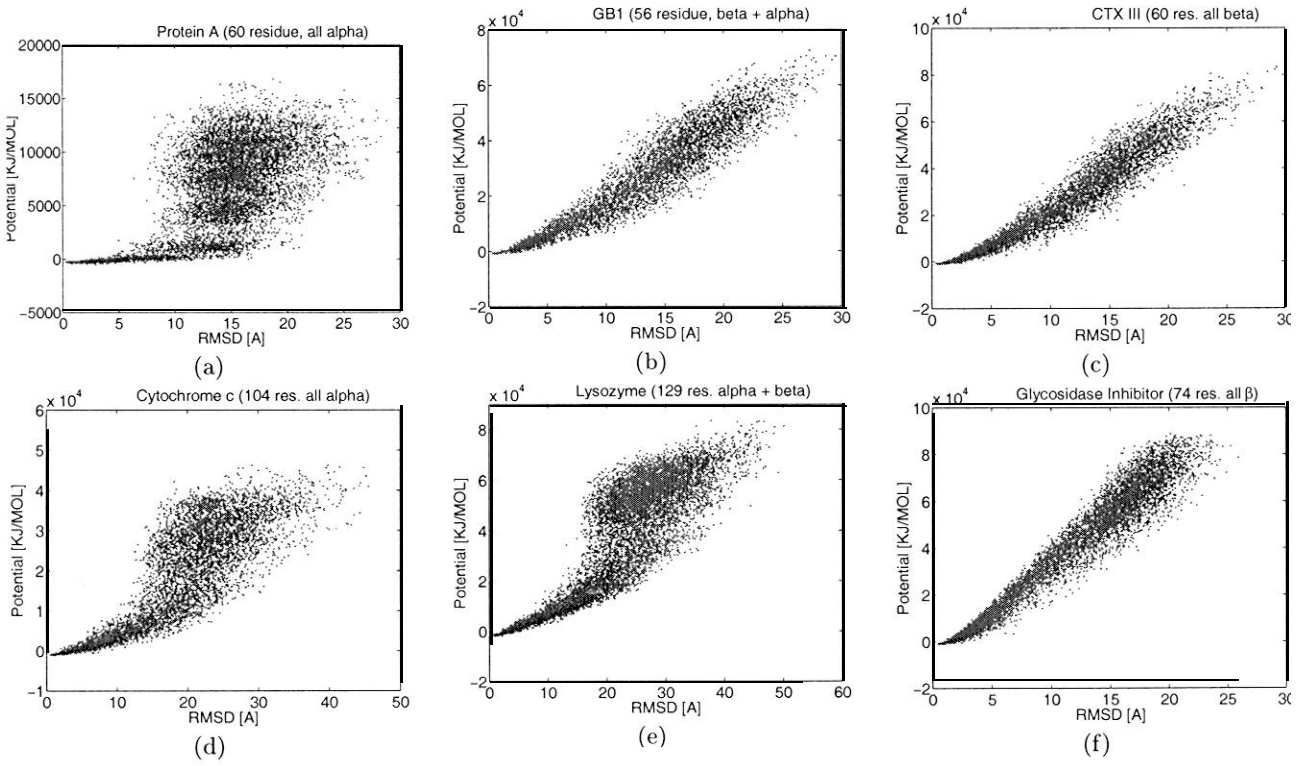


Figure 4: The potential vs RMSD distribution for proteins (a) A, (b) GB1, (c) CTX III, (d) Cytochrome c, (e) hen egg white Lysozyme, and (f) α-Amylase Inhibitor. The two proteins in the first (left) column are all alpha proteins, the middle column contains mixed alpha and beta proteins, and the third (right) column contains all beta proteins.

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 proteins, the plots share some features of the plots for both the all alpha and the all beta proteins. And moreover, the degree of similarity seems to be related to the proportion of the protein composed of a particular secondary structure. For example, hen egg-white Lysozyme (e), whose secondary structure is mainly alpha, has a similar distribution to the all alpha Cytochrome C (d), and the distribution for protein GB1 (b), which is more beta than alpha, is similar to the all beta protein CTX III (c).

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. 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.

Each connection attempt performs feasibility checks for n intermediate conformations between the two corresponding nodes as determined by the chosen local planner (the number of such conformations is determined by the desired resolution which may be set by the user). In our simulations, we use the common straight-line local planner, which interpolates without bias along the straight line in C connecting the two roadmap nodes [4].

When two nodes q_1 and q_2 are connected by the local planner, 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 probabil-

ity P_i of moving from c_i to c_{i+1} depends on the difference between their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$.

$$P_i = \begin{cases} e^{\frac{-\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \leq 0 \end{cases} \quad (2)$$

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 $0 \leq P_i \leq 1$.)

$$w(q_1, q_2) = \sum_{i=0}^{n-1} -log(P_i), \quad (3)$$

By assigning the weights in this manner, we can find the most energetically feasible path in our roadmap when performing queries. A similar weight function, with different probabilities, was used in [34].

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.

The resulting roadmap can be used 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 [13] to find the smallest weight path between the start and goal conformations.

One important feature of our approach is that the roadmap contains many folding pathways, which together represent the folding landscape. We can extract many such paths by computing the single-source shortest-path (SSSP) tree from the native structure. Using Dijkstra's algorithm, this takes $O(V^2)$ time, where V is the number of roadmap nodes.

To 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 and/or improving already computed folding pathways.

We now describe the simple potential energy function we used. We start with:

$$U_{\text{tot}} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + \sum_{atom\ pairs} (A/r_{ij}^{12} - B/r_{ij}^{6}), \quad (4)$$

which is similar to the potential used in [25]. The first term represents constraints which favor the known secondary structure through main-chain hydrogen bonds and disulphide bonds. The parameter K_d is set to 100 KJ/mol, and the distances are $d_0 = d_c = 2\text{Å}$, and d_i is the separation between the hydrogen and oxygen atoms. The second term corresponds to the van der Waals interaction among the atoms. The parameters for the van der Waals interaction can be found in [25], which encodes strong preference for interactions between oxygon and hydrogen atoms.

However, even for relatively small proteins with about 60 residues, there are nearly one thousand atoms. Non-hydrogen atoms also number in the hundreds. Therefore, performing all pairwise van der Waals potential calculations (the second summation) can be computationally intensive. To reduce this cost, 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_β atom; R is modeled with a fixed-size rigid sphere and is treated as an "extended carbon atom" for the van der Waals interaction in [25]. 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 then 2.4 Å during node generation and 1.0 Å during roadmap connection), a very high potential is returned. (See [6] for more details.) The side chain is chosen for this purpose because it mainly reflects the geometric configuration of a residue. By doing this, the computational cost is reduced by two orders of magnitude. Our results indicate that enough accuracy seems to be retained to capture the main features of the interaction for the proteins we study.

If all the distances between all R 'atoms' are larger than 2.4 A, then we proceed to calculate the potential as follows (we don't have van der Waals term):

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d,\} + E_{hp}, \quad (5)$$

The first term is exactly the same as in Equation (4), i.e., it represents constraints which favor the known secondary structure through main-chain hydrogen bonds and disulphide bonds. The hydrogen bond and disulphide bond in-

formation is obtained from a program called "DSSP" [19], and is then passed to our code as part of the input. The second term is the hydrophobic effect and is considered in the following simplistic way. We assign a hydrophobicity value of 1 to all non-polar amino acids, and 0 to the rest. When the sidechains (the R 'atoms' to be exact) of any two non-polar amino acids come within a distance of R_h, the potential is decreased by E_h. In our case, we set R_h = 6 Å and $E_h = 20KJ/Mol$. The hydrophobicity information of a given protein is also passed to our code.

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 [29, 30] and Baker et al. [2] in their statistical mechanical models. The strength for the three terms is very similar to those used by Munoz and Eaton [30].

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 are 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 $F_{\text{hb}} = -0.86kcal/mol * N_{\text{hb}}$.

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 [16]. As = $\sum_i^{N_{\text{hb}}} \log \text{ECO}_i$, and then the total free energy change is: $F_{\text{entropy}} = 6.0cal/mol/K * (300K) * \Delta s$.

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

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 hydrophobic 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 and to gain insight into the underlying folding mechanism since we desire to reproduce, or at least approach, results close to experimental observations.

Description of Proteins Studied			
Name	pdb	res	ss
Protein G domain B1	1GB1	56	$1\alpha+4\beta$
Staphylococcus Protein A	1BDD	60	3α
SH3 domain a-spectrin	1SHG	62	58
CI2	1COA	64	$1\alpha+4\beta$
SH3 domain src	1SRL	64	5β
CspB form *Bacillus subtilis*	1CSP	67	7β
SH3 domain fyn	1NYF	67	5β
CspA	1MJC	69	7β
Tendamistat	2AIT	74	7β
Ubiquitin	1UBQ	76	$1\alpha+5\beta$
SH3 domain PI3 kinase	1PKS	79	$1\alpha+5\beta$
procarboxipeptidase A2	1PBA	81	$3\alpha+3\beta$
ACBP bovine	2ABD	86	5α
Barnase	1BRN	110	$3\alpha+7\beta$

Table 1: Proteins are listed in ascending order of number of residues (res). The number of alpha helices (α) and beta strands (β) are listed in SS column.

Running Time and Roadmap Statistics				
PDB	res	nodes	edge	time (h)
1GB1	56	5126 (5506)	70k	3.71
1BDD	60	5471 (9106)	104k	7.03
1SHG	62	5427 (5502)	59k	2.89
1COA	64	7975 (8407)	104k	6.87
1SRL	64	8755 (8822)	111k	4.95
1CSP	67	6735 (6852)	72k	4.67
1NYF	67	6219 (6332)	70k	3.42
1MJC	69	5990 (6142)	62k	4.30
BAIT	74	8246 (8477)	92k	7.11
1UBQ	76	8357 (10667)	119k	9.44
1PKS	79	7685 (10257)	95k	9.32
1PBA	81	8085 (10747)	114k	10.40
2ABD	86	7330 (12577)	149k	14.20
1BRN	110	6601 (10607)	108k	15.80

Table 2: Running time for constructing roadmaps for 14 proteins and statistics for each roadmap, including nodes in the same connected component as the native structure, total nodes (in parenthesis), and the number connections (edges) among the nodes.

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, 29], 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, 29].

4.1 The Proteins

In this paper, we study 14 proteins listed in Table 1. 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 [29] work studying the folding kinetics of small proteins.

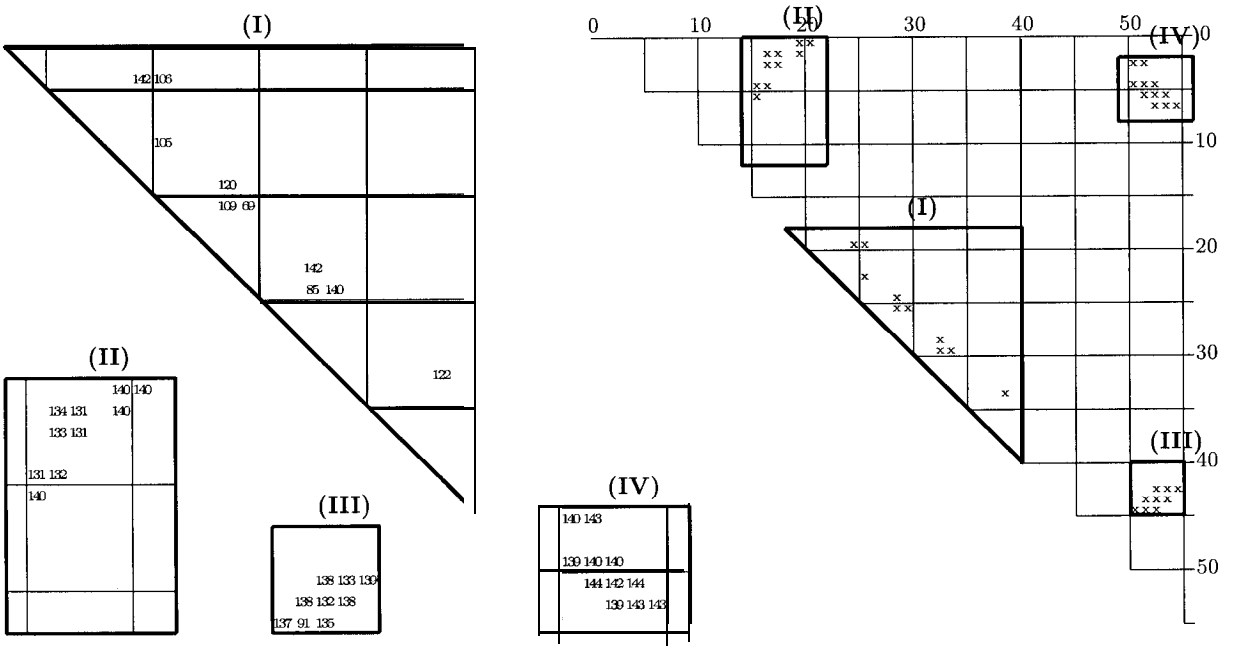


Figure 5: Protein GB1. The full contact matrix (right) and blow-ups (left) showing the time steps when the contacts appear on our path. Blow-ups I, II, III, and IV correspond to the alpha helix contacts, the beta l-2 contacts, the beta 3-4 contacts, and the beta l-4 contacts, respectively.

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, 29] when only considering the global free energy landscape structure.

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 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 2). 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 4), the free energy landscape (see Figure 6), and multiple folding pathways, all in a single run.

4.3 Secondary structure formation order

In [36], 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 hydrophobic residues as in [16]. First, to determine the hydrophobic contacts in the native state, we compare all pairs of C_α atoms of hydrophobic residues, and those that are within 7 A of each other are said to form a native contact. Then, when analyzing a conformation q, if the corresponding C_α atoms are $\leq$ 7 Å apart, we determine the contact is present in q; 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.

Contact formation analysis was performed on the paths for proteins GB1 and A. The results for protein GB1 are shown in Figure 5. (See [5] for Protein A results.) In the figures, the full contact matrix (among hydrophobic 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 in our path. For example, for protein GB1, blow-up I shows the contact between residues 34 and 38 appeared at time step 122 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 GB1 (Figure 5), the alpha helix (I) formed around time step 114 (the average of the time steps in I), the C-terminal hairpin (III, beta strands 3 and 4) formed around time step 131, the N-terminal hairpin (II, beta strands 1 and 2) formed around time step 135, and the two hairpins come together (IV, contacts form between beta strands 1 and 4) around time step 141. One may note that in some blowups, for example (I) and (III), 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 outliers.

Results for the 14 proteins studied are shown in Table 3.

Secondary Structure Formation Order and Validation				
pdb	res.	#	secondary structure formation order	exp. [27]
1GB1	II	56	al., β3-β4., β1-β2., β1-β4	Agreed
1BDD		60	α2, α3, α1, α2-α3, **al-a3**	Agreed
1SHG		62	(β2-β3 β3-β4), (β1-β2β1-β5)	N/A
1COA		64	α1, β3-β4, β2-β3, β1-β4, α1-β4	Agreed
1SRL		64	β4-β5, β3-β4, β2-β3, β1-β5, β1-β2	N/A
1CSP		67	β5-β6, β2-β3, β3-β4, (β1-β3 β4-β6 β5-β7), β1-β5	N/A
1NYF		67	β3-β4, β2-β3, β1-β2	N/A
1MJC		69	(β5-β6 β2-β3), (β3-β4 β1-β3), (β1-β5 β4-β6 β5-β7)	N/A
2AIT		74	(β4-β5 β1-β2 β3-β4), (β3-β7, β2-β6, β1-β4), β1-β5, β1-β6	N/A
1UBQ		76	α1, β3-β4, β1-β2, β3-β5, β1-β5	**Agreed**
1PKS		79	α1, β2-α1, β3-β4, (β1-β2 β2-β3), β1-β5	N/A
1PBA		81	(α3 α1 α1-α2) α1-β1, β1-β2, β1-β3	N/A
2ABD		86	α3, α4-α5, α2, α4, α0, α5, α2 α3, α2 α4, α1 α4	N/A
1BRN		110	α1, α3, α2, β1-α2, α2-α3, β5-β6, β2-β4, β4-β5, β6-β7, β3-β4, β1-β2	Not sure

Table 3: 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 (like an a) and also for contacts between** two **secondary structures (such as two** β **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 [27].**

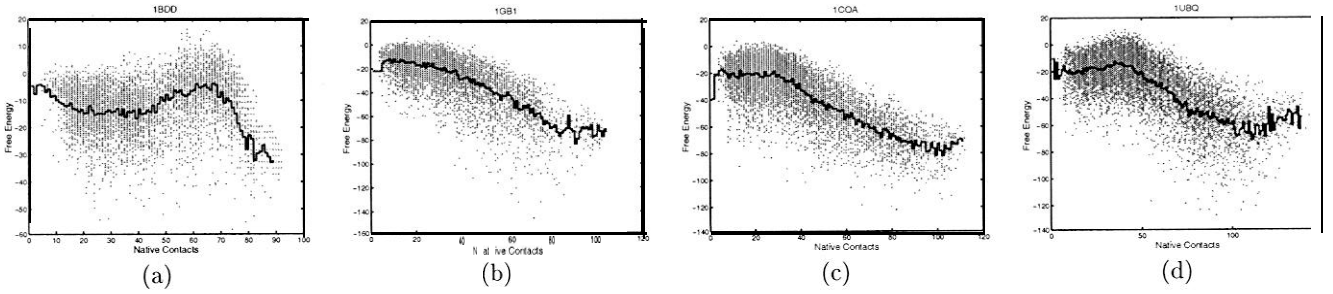


Figure 6: Free energy landscape for proteins A, G, CI2, and Ubiquitin. The line in the middle is the average free energy for all conformations with the same number of native contacts. Provided that native contacts is a good reaction coordinate, this could be used to study folding kinetics as done by other researchers.

Here, we list the potential secondary structure formation order for all proteins. The formation order for each protein shown in the table is the dominant one found. For the proteins that are also listed in Li and Woodward's paper describing hydrogen-exchange experimental results [27], our results seem to be in good agreement with known results. This gives us some confidence of our pathways before using them to facilitate further study. Moreover, the results could **be used** to check or predict for those proteins without 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 [14, 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, 29]. 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 6, we show the free energy distribution vs. native contacts for proteins A, GB1, CI2, and Ubiquitin. We chose these four since we have their secondary structure formation order from experimental data. In all cases, the results are in agreement with our simulations. (More plots for other proteins can be found in [5].) One can see that any folding rates that might be inferred clearly averages of the conformations with same number of native contacts. As a result, it could easily miss detailed information of the energy landscape.

Therefore, this averaging approach is limited and potentially will miss subtle behavior. Moreover, for proteins like Hen egg white Lysozyme, which displays two unique folding pathways, one with two-state behavior, and one with three-state behavior [12, 32], averaging techniques like the statistical mechanical model would not discover both behaviors. This is one example where it seems to be crucial to have more detailed pathway information, such as is available in our roadmaps.

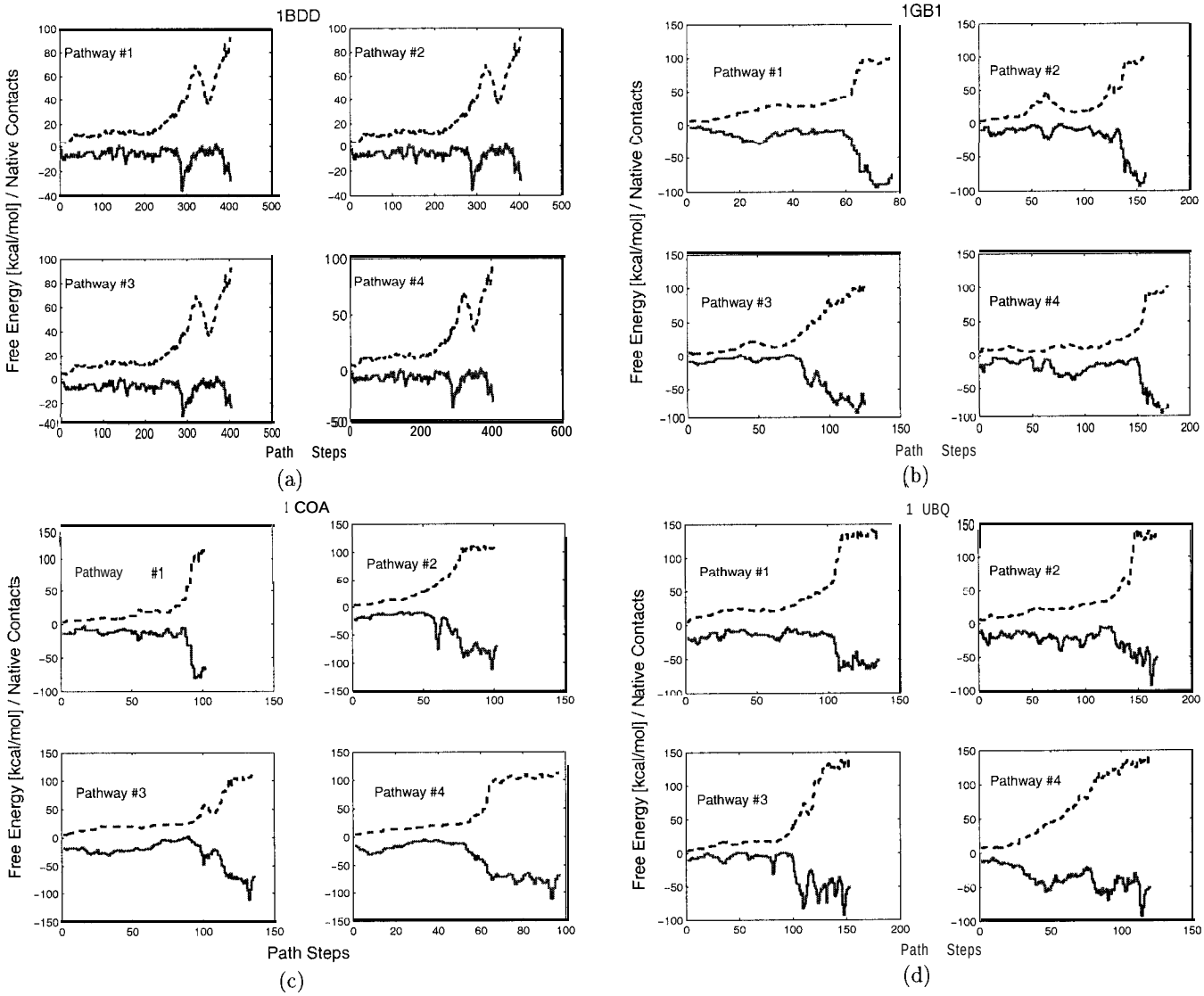


Figure 7: Free energy profiles, plotted in solid lines, for four folding pathways for proteins G, A, C12, and Ubiquitin. The number of native contacts on the paths are shown with dashed lines; note they are not monotonically increasing.

In Figure 7, the free energy profile, as well as the native contacts, are plotted for proteins G, A, C12, and Ubiquitin. Similar plots for the other proteins are available in [5]. 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. What is still needed is some good way of analyzing and summarizing all the pathways in our roadmaps which 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 also 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 (e.g., 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 to us the connection between box folding and protein folding, Marty Scholtz, Michael Levitt, and Vijay Pande for useful suggestions

7. REFERENCES

[1] E. Alm and D. Baker. Matching theory and experiment in protein folding. *Curr. Op. Str. Biol.*, 9:189–196, 1999.

[2] E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA*, 96:11305–11310, 1999

[3] N. M. Amato, 0. B. Bayazit, L. K. Dale, C. V. Jones, and D. Valleio. OBPRM: An obstacle-based PRM for 3D workspaces. In *Proc. Int. Workshop* on Algorithmic Foundations *of Robotics (WAFR)*, pages 155-168, 1998.

[4] N. M. Amato. 0. B. Bavazit. L. K. Dale. C. V. Jones. and D. Vallejo. Cdoosing good distance metrics and local planners for probabilistic roadmap methods. *IEEE* Trans. *Robot. Automat.*, 16(4):442–447, August 2000. Preliminary version appeared in ICRA 1998, pp. 630–637.

[5] N. M. Amato and G. Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. Technical "Report 01-001, PARASOL Lab, Dept. of Computer Science, Texas A&M University, Oct 2001.

[6] N. M. Amato and G. Song. Using motion planning to study protein folding pathways. Journal *of* Computational *Biology*, 2002. To appear. Preliminary version appeared in RECOMB'01.

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

[8] R.L. Baldwin and G.D. Rose. Is protein folding hierarchic? i. local structure and peptide folding. Trends *Biochem Sci.*, 24:26–33, 1999.

[9] 0. B. Bayazit, G. Song, and N. M. Amato. Ligand binding with OBPRM and haptic user input: Enhancing automatic motion planning with virtual touch. In *Proc. IEEE Int.* Conf. *Robot. Autom. (ICRA)*, pages 954-959. 2001. This work was also presented as a poster at RECOMB'01.

[10] V. Boor, M. H. Overmars, and A. F. van der Stappen. The Gaussian sampling strategy for probabilistic roadmap planners. In *Proc. IEEE Int.* Conf. *Robot.* Autom. *(ICRA)*, pages 1018-1023, 1999.

[11] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Is protein folding hierarchic? i. local structure and peptide folding. Protein *Struct. Funct.* Genet, 21:167–195, 1995.

[12] C.M.Dobson, A. Sali, and M. Karplus. Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed.*, 37:868–893, *1998*.

[13] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to algorithms*. MIT Press and McGraw-Hill Book Company, 6th edition, 1992.

[14] K. A. Dill, K. M. Fiebig, and H. S. Chan. Cooperativity in protein-folding kinetics. *Proc.* Natl. *Acad. Sci. USA*, 90:1942–6, 1993.

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

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

[17] L. Han and N. M. Amato. A kinematics-based probabilistic roadmap method for closed chain systems. In Algorithmic and Computational *Robotics* - New Directions *(WAFR 2000)*, pages 233-246, 2000.

[18] D. Hsu, L. Kavraki, J-C. Latombe, R. Motwani, and S. Sorkin. On finding narrow passages with probabilistic roadmap planners. In *Proc. Int.* Workshop on *Algorithmic* Foundations *of Robotics (WAFR)*, 1998.

[19] W. Kabsch and C. Sander. *Biopolymers*, 22:2577–2637, 1983.

[20] L. Kavraki. Random *Networks* in Configuration Space *for* Fast *Path* Planning. PhD thesis, Stanford Univ, Computer Science Dept., 1995.

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

[22] P.T. Lansbury. Evolution of amyloid: What normal protein folding may tell us about fibrillogenesis and disease. *Proc. Natl. Acad.* Sci. *USA*, 96:3342–3344, 1999.

[23] J. C. Latombe. *Robot* Motion Planning. Kluwer Academic Publishers, Boston, MA, 1991.

[24] S.M. LaValle, J.H. Yakey, and L.E. Kavraki. A probabilistic roadmap approach for systems with closed kinematic chains. In *Proc. IEEE* Int. Conf. *Robot. Autom. (ICRA)*, 1999.

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

[26] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai. Protein folding: the endgame. Annu. *Rev.* Biochem., 66:549–579, 1997.

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

[28] L. Lu and S. Akella. Folding cartons with fixtures: A motion planning approach. In *Proc. IEEE Int.* Conf. *Robot.* Autom. *(ICRA)*, pages 1570-1576, 1999.

[29] 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*, 96:11311–11316, 1999.

[30] V. Muiioz, E. R. Henry, J. Hoferichter, and W. A. Eaton. A statistical mechanical model for b-hairpin kinetics. *Proc. Natl. Acad.* Sci. *USA*, 95:5872–5879, 1998.

[31] J. O'Rourke. Folding and unfolding in computational geometry. In *Proc. Japan* Conf. *Discrete Comput.* Geom. '98, pages 142-147, December 1998. Revised version submitted to LLNCS.

[32] S. E. Radford and C. M. Dobson. Insights into protein folding using physical techiniques: studies of lysozyme and α-lactalbumin. *Phil. Trans. R. Soc. Lond.*, B348:17, 1995.

[33] G. N. Reeke, Jr. Protein folding: Computational approaches to an exponential-time problem. Ann. *Rev. Comput. Sci.*, 3:59–84, 1988.

[34] A.P. Singh, J.C. Latombe. and D.L. Brutlaa. A motion planning-approach to flexible ligand binding. In *7th Int. Conf.* on *Intelligent Systems* for Molecular *Biology (ISMB)*, pages 252-261, 1999.

[35] G. Song and N. M. Amato. A motion planning approach to folding: From paper craft to protein folding. In *Proc. IEEE* Int. Conf. *Robot.* Autom. *(ICRA)*, pages 948-953, 2001.

[36] G. Song and N. M. Amato. Using motion planning to study protein folding pathways. In *Proc.* Int. Conf. *Comput.* Molecular *Biology (RECOMB)*, pages 287-296, 2001.

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

[38] S. A. Wilmarth, N. M. Amato. and P. F. Stiller. MAPRM: A probabilistic roadmap planner with sampling on the medial axis of the free space. In *Proc. IEEE* Int. Conf. *Robot.* Autom. *(ICRA)*, pages 1024–1031, 1999.