Using Motion Planning to Study Protein Folding Pathways[†]

Nancy M. Amato[*]
amato@cs.tamu.edu
Dept. of Computer Science
Texas A&M University
College Station, TX 77843-3112
U.S.A.
tel: +1-979-862-2275
fax: +1-979-847-8578

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

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

[†]A preliminary version of this paper appeared in the Proc. 2001 Int. Conf. Comput. Molecular Biology (RECOMB) [47]. This research supported in part by NSF CAREER Award CCR-9624315, NSF Grants IIS-9619850, ACI-9872126, EIA-9975018, EIA-0103742, EIA-9805823, ACI-0113971, CCR-0113974, EIA-9810937, EIA-0079874, and by the Texas Higher Education Coordinating Board grant ARP-036327-017.

[*]Corresponding author.

Abstract

We present a framework for studying protein folding pathways and potential landscapes which is based on techniques recently developed in the robotics motion planning community. Our focus in this work is to study the protein folding mechanism *assuming* we know the native fold. That is, instead of performing fold prediction, we aim to study issues related to the folding process, such as the formation of secondary and tertiary structure, and the dependence of the folding pathway on the initial denatured conformation. Our work uses Probabilistic Roadmap (PRM) motion planning techniques which have proven successful for problems involving high-dimensional configuration spaces. A strength of these methods is their efficiency in rapidly covering the planning space without becoming trapped in local minima. We have applied our PRM technique to several small proteins ($\sim$ 60 residues) and validated the pathways computed by comparing the secondary structure formation order on our paths to known hydrogen exchange experimental results.

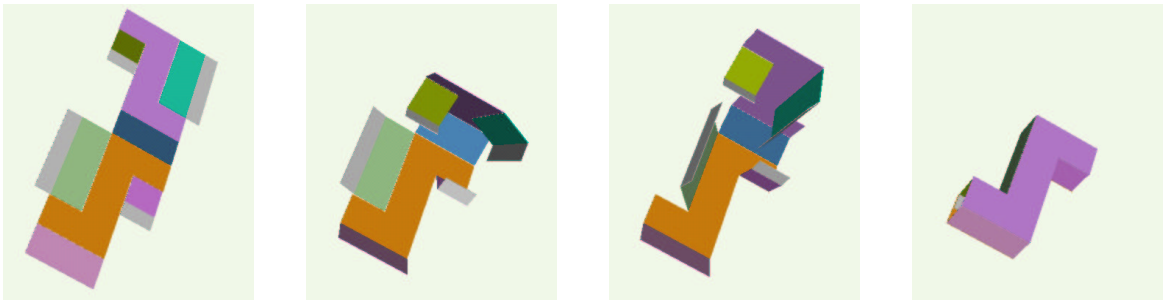
An advantage of the PRM framework over other simulation methods is that it enables one to easily and efficiently compute folding pathways from any denatured starting state to the (known) native fold. This aspect makes our approach ideal for studying global properties of the protein's potential landscape, most of which are difficult to simulate and study with other methods. For example, in the proteins we study, the folding pathways starting from different denatured states sometimes share common portions when they are close to the native fold, and moreover, the formation order of the secondary structure appears largely independent of the starting denatured conformation. Another feature of our technique is that the distribution of the sampled conformations is correlated with the formation of secondary structure, and in particular appears to differentiate situations in which secondary structure clearly forms first and those in which the tertiary structure is obtained more directly. Overall our results applying PRM techniques are very encouraging, and indicate the promise of our approach for studying proteins for which experimental results are not available.

1 Introduction

The goal of motion planning is to compute a sequence of valid intermediate states that transform a given initial state (the start) into some desired final state (the goal). While we recognize that protein folding is vastly more complicated than traditional motion planning applications in robotics, motion planning algorithms are often described using an abstraction called *configuration space (C-space)* that is sufficiently general to apply to many seemingly unrelated problems. Briefly, the configuration space, or C-space, of a movable object is the space consisting of all positions and orientations of the object. For proteins, the configuration space is also commonly referred to as conformation space. In this paper, we use configuration (space) and conformation (space) interchangably.

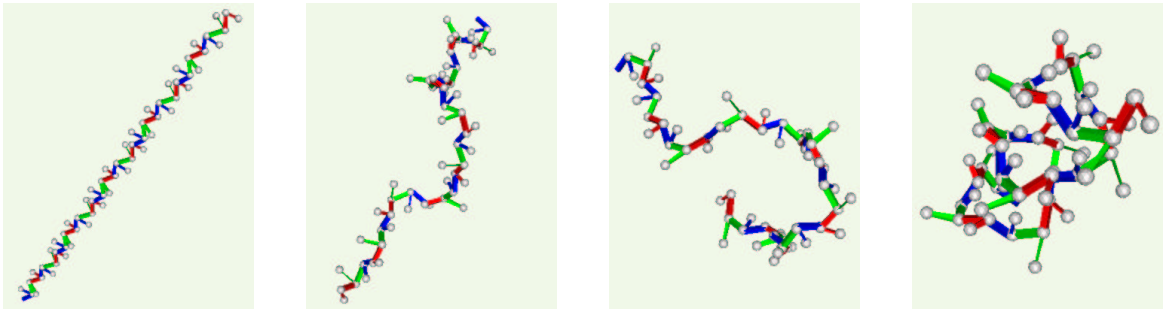
In this work, we concentrate on the application of the successful *probabilistic roadmap (PRM)* [26] motion planning method to protein folding. We have selected the PRM paradigm due to its proven success in exploring high-dimensional configuration spaces. Indeed, the PRM methodology has been used to study the related problem of ligand binding [8, 43], which is of interest in drug design. The results were quite promising and show the potential of the method for problems in computational biology and chemistry. Our success [45, 46] in applying this methodology to folding problems such as carton folding (with applications in packaging and assembly [35]), and paper crafts (studied in computational geometry [38]), provided some evidence of the feasibility of this approach for determining protein folding sequences. 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 [45]. We have obtained promising results with this PRM-based technique on several small proteins ($\sim$ 60 amino acids, which we model using $\sim$ 120 degrees of freedom).

There are large and ongoing research efforts whose goal is to determine the native folds of proteins (see, e.g., [41, 32]). In this paper, we assume we already know the native fold, and our focus is on the folding process, i.e., how the protein folds to that state from some initial state. Many researchers have remarked that knowledge of the folding pathways might provide insights into and a deeper understanding of the nature of protein folding [21, 42]. Although there have been some recent experimental advances [17], computational techniques for simulating this process are important because it is difficult to capture the folding process experimentally. In our work, we exploit the fact that the native structure of the protein is known to focus our exploration of the conformation space to regions around the native fold. Knowledge of the native state has also been used by other researchers, such as Baker and co-workers [1, 6] and Muñoz and



(a) (b) (c) (d)

Figure 1: Snapshots of a carton folding.



(a) (b) (c) (d)

Figure 2: Snapshots of a 10 ALA chain folding.

Eaton [37] who use the topology of the native state to predict the folding rates and mechanisms of some proteins. Advantages of our PRM approach are that it efficiently covers a large portion of the planning space, in this case, the conformation space, and that it enables one to easily and efficiently compute folding pathways from *any* denatured starting state (including the traditionally studied extended conformation) to the native fold. This aspect makes our approach ideal for studying global properties, such as secondary structure formation and the funnel structure of the protein's potential landscape. For example, we have found that folding pathways extracted from the same roadmap for different starting denatured states usually share common portions when they are close to the native state. Another feature of our technique is that the distribution of the sampled conformations is correlated with the formation of secondary structure, and in particular appears to differentiate situations in which secondary structure clearly forms first and those in which the tertiary structure is obtained more directly. Such global issues are difficult to simulate and study with other traditional methods, such as molecular dynamics.

The fact that a protein's three-dimensional structure is determined by its amino acid sequence was first demonstrated in Anfinsen's pioneering work [5]. Since then, many different approaches for predicting protein structure have been explored (see [49] for a review). We note that our work is different in focus from such studies because it assumes *a priori* knowledge of the native fold and concentrates on determining the folding pathways. In folding simulations for protein structure prediction, several computational approaches have been applied to this exponential–time problem, including energy minimization [33, 51], molecular dynamics simulation [31], Monte Carlo methods [14, 27], and genetic algorithms [11, 50]. Among these, molecular dynamics is most closely related to our approach. Much work has been carried out in this area [15, 16, 19, 31], which tries to simulate the true dynamics of the folding process using the classical Newton's equations of motion. The forces applied are usually approximations computed using the first derivative of an empirical potential function. The advantage of using molecular dynamics is that it helps us understand how proteins fold in nature. It also provides a way to study the underlying folding mechanism, to investigate folding pathways, and can provide intermediate folding states. However, the simulations required for this approach are computationally intensive and time-dependent. They are also

heavily dependent on the initial conformation and can easily result in local minima. Most proposed techniques have tremendous computational requirements because they attempt to simulate complex kinetics and thermodynamics. Our work provides an alternative approach that finds approximations to the folding pathways while avoiding local minima and detailed simulations.

1.1 Outline

We begin in Section 2 with an overview of probabilistic roadmap motion planning methods and describe how they can be applied to compute folding pathways for proteins. Next, in Section 3, we briefly describe our potential energy computations, and in Section 4 we discuss our approach for validating the pathways computed by our method. Our results for a 10-ALA polypeptide chain and two small proteins (the B domain of protein A and the B1 domain of protein G) are presented in Section 5. We conclude with some final remarks in Section 6.

2 A Probabilistic Roadmap Method for Protein Folding

Given a description of the environment and a movable object (the 'robot'), the motion planning problem is to find a feasible path that takes the movable object from a given start to a given goal configuration [29]. Since there is strong evidence that any complete planner (one that is guaranteed to find a solution, or determine that none exists) requires time exponential in the number of degrees of freedom (dof) of the movable object [29], attention has been focussed on randomized or probabilistic methods.

Our approach to the folding problem is based on the probabilistic roadmap (PRM) approach to motion planning [26]. Briefly, PRMs work by sampling points 'randomly' from C-space, and retaining those that satisfy certain feasibility requirements (e.g., they correspond to collision-free configurations of the movable object, see Figure 3(a)). Then, these points are connected to form a graph, or roadmap, using some simple 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 using standard graph search techniques (see Figure 3(c).)

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 (the local connection can often be performed using multiple applications of the feasibility test).

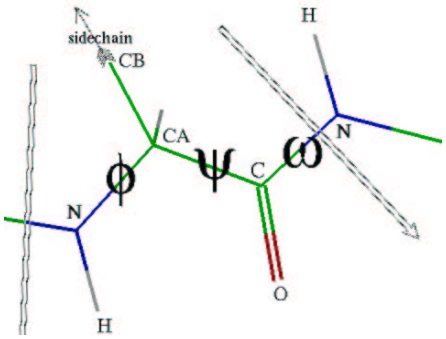
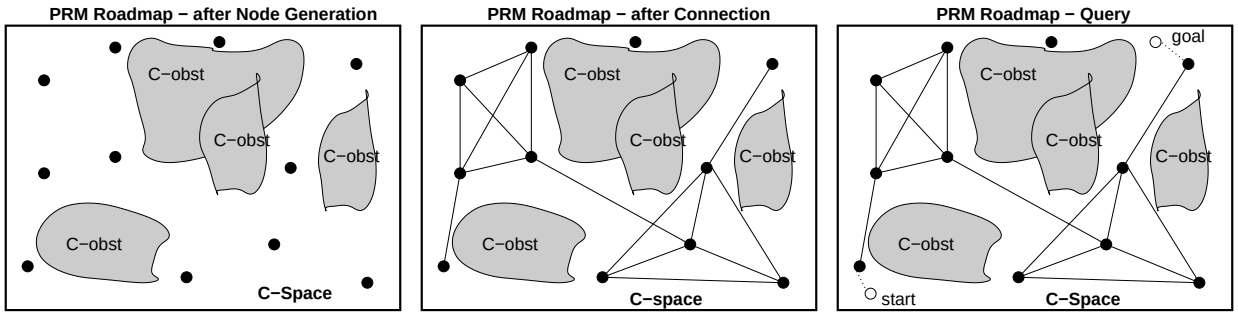


Figure 4: The ϕ and ψ angles in an amino acid [23].

The protein folding problem has a few notable differences from usual PRM applications. First, the traditional collision-free constraint is replaced by a preference for low energy conformations. In particular, the way in which a protein folds depends on the potential energy of the conformations at each step in the process. In general, the lower the potential, the more stable the conformation, and the native state of the protein is thought to be the global minimum. In the folding process, transitions from configurations with higher potential to configurations with lower potential are favored. The protein will, however, pass through local minima and maxima states when folding. Second, in PRM applications, it is often considered sufficient to find *any* feasible path connecting the start and goal configurations. For protein folding, however, we are interested in the *quality* of the path, and in particular, we are searching for energetically favorable paths.

2.1 C-spaces of folding objects

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 [49], we consider all atomic bond lengths and bond angles to be constants, and consider only phi/psi torsional angles, which we model as two revolute joints (2 dof), see Figure 4. Side chains are modeled as spheres and are given no degrees of freedom. Thus, for an amino acid sequence with k amino acids (often referred to as *residues*), our model will consist of $2k$ links and $2(k-1)$ revolute joints.



(a) (b) (c)

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

The joint angle of a revolute joint takes on values in $[0, 2\pi)$, with the angle 2π equated to 0, which is naturally associated with a unit circle in the plane, denoted by S^1. Assuming some position and orientation for one of the links (the base), the positions of each of the remaining links can be specified by the *joint angle* between the link and some adjacent link. Thus, since we are not concerned with the absolute position and orientation of the object in the environment (i.e., we can use any nominal position for the base link), a *configuration* of an n joint tree-like articulated object can be specified by a vector of n joint angles. That is, the configuration space of interest for a protein with $n+1$ amino acids can be expressed as:

$$\mathcal{C} = \{q \mid q \in S^{2n}\}. \quad (1)$$

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

2.2 Node generation

A configuration $q \in \mathcal{C}$ can be generated by assigning each joint angle a value in its allowable range. 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). The node q is accepted and added to the roadmap based on its potential energy $E(q)$ with the following probability:

$$P(\text{accept } q) = \begin{cases} 1 & \text{if } E(q) < E_{\min} \\ \frac{E_{\max} - E(q)}{E_{\max} - E_{\min}} & \text{if } E_{\min} \le E(q) \le E_{\max} \\ 0 & \text{if } E(q) > E_{\max} \end{cases} \quad (2)$$

This acceptance criterion was also used when building PRM roadmaps for ligand binding in [43]. This filtering helps us to generate more nodes in low energy regions. In our case, we set $E_{\min} = 50000$ KJouls/mol and $E_{\max} = 89000$ KJouls/mol. A configuration with overlapping side chains, for example, has higher potential and is thus more likely to be rejected during node generation.

2.2.1 Gaussian sampling – Concentrated sampling near the native state

Due to the high dimensionality of the conformation space, simple uniform sampling would have to be very dense to cover the conformation space sufficiently to reliably characterize the important features of the potential energy landscape. One idea is to use the Ramachandran plot [40] to bias our sampling. This is appealing, since it shows the distribution of the ϕ and ψ angles for the residues. However, an argument similar to the Levinthal paradox shows that the resulting space is still too large to be sampled efficiently.

Fortunately, in our case, a more focussed sampling strategy can be devised based on the fact that we assume that the native fold is known *a priori*.[1] Thus, we can take advantage of this knowledge and design a sampling strategy biased around the native fold with the goal of characterizing the potential landscape leading to the native fold. In particular, we select a set of normal distributions around the native fold and sample from these distributions. The set of standard deviations (STDs) we use in the results presented here are {5°, 10°, 20°, 40°, 80°, 160°}. The small STDs capture the detail around the goal, while the larger STDs ensure adequate roadmap coverage of the conformation space. Our simulation results pre-

[1] We would like to remind the reader that the focus of the work presented here is **not** to predict native folds, but rather to study folding pathways and potential funnels leading to a known native fold.

sented in Section 5 show that this Gaussian sampling strategy is very useful.

Similar biased sampling strategies have been applied successfully in robotics applications [2, 10, 20, 22, 25, 30, 52], where oversampling in and near narrow passages in C-space is crucial for some problems. In recent work, Baker and co-workers [6, 1] and Muñoz and Eaton [37] have used knowledge of the topology of the native state to predict the folding rates and mechanisms of some proteins.

2.3 Connecting the roadmap

The second phase of the algorithm is roadmap connection. For each node, we first find its k nearest neighbors in the roadmap 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 $\mathcal{C}$. 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 $\mathcal{C}$ connecting the two roadmap nodes [3]. If there are still multiple connected components in the roadmap after this stage (which is generally the case, and is in fact sometimes unavoidable, see, e.g., [9, 12]), other techniques will be applied to try to connect different connected components (see [2] for details).

When two nodes q_1 and q_2 are connected by the local planner, the corresponding edge is added to the roadmap. We associate a weight with each edge (q_1, q_2). The weight is computed by examining 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 $\mathcal{C}$ connecting q_1 and q_2. For each pair of consecutive conformations c_i and c_{i+1}, the probability P_i of moving from c_i to c_{i+1} depends on the difference between their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$.

$$P_i = \begin{cases} e^{\frac{-\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \leq 0 \end{cases} \quad (3)$$

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 all pairs of consecutive conformations in the sequence. (Negatives of the logs are used since each $0 \leq P_i \leq 1$.)

$$w(q_1, q_2) = \sum_{i=0}^{n-1} -log(P_i), \quad (4)$$

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 [43].

2.4 Querying the roadmap

The resulting roadmap can be used to find a feasible path between given start and goal conformations. This is done a bit differently than usual for PRMs. Usually, attempts are made to connect the start and the goal to the same connected component of the roadmap. If this succeeds, a path is returned, otherwise failure is reported.

For our protein folding problems, we already have the goal (the native state) in the roadmap (since it is known *a priori* and is exploited during sampling as described in Section 2.2.1). For the starting conformation (any denatured state), we connect it into the roadmap, just as was done for the other roadmap nodes during the connection phase (Section 2.3). Dijkstra's algorithm [13] is then used to find the smallest weight path between the start and goal conformations. If the potential of some intermediate node is too large (as compared to some predetermined maximum), we remove the offending edge from the roadmap and repeat the process. This could occur because we can elect to enforce different, stricter thresholds in the query phase than were used in the roadmap construction phase. We have found that enforcing different requirements in the construction and query phases often decreases construction time significantly while impacting the query time only slightly [48]. Moreover, adding the start to the roadmap in this manner facilitates our search for the lowest weight path and augments the roadmap after each query is performed (this has been noted as a possible optimization for regular PRM applications as well).

2.4.1 Path smoothing

Since the nodes are generated randomly and connected using straight–line connections, the path returned by the query could possibly be improved by targeted local deformations. This process is often called smoothing in the robotics literature, and it is widely recognized that paths computed using PRMs

benefit from smoothing [26]. Basically, we attempt to 'push,' or deform, the given path into a better path. We used this strategy successfully in Computer Aided Design (CAD) applications to transform invalid user-collected paths into valid paths [7].

There exist many possible resampling strategies. We have applied the following simple method. We resample around all the nodes on the query path that have higher potential than some user specified threshold. For each such node c, we generate k neighboring nodes N_c (we used $k = 10$). If all nodes in N_c have higher potential than c, we stop. Otherwise, we let c' be the node in N_c with lowest potential, and repeat the process by generating neighbors of c'. We repeat this process for some fixed number of iterations. Essentially, this can be viewed as an approximate gradient descent operation. After all nodes have been processed, we connect the new nodes into the roadmap and then perform the query again.

Van der Waals Parameters				
Bond Type	A	B	ε	r^0
H..H	290	1.07	0.0010	2.8525
O..O	145,834	328	0.18479	3.1005
N..N	3952850	2556	0.41315	3.8171
C..C	3075695	953	0.07382	4.3150
A..A	1200965	425	0.03763	4.2202
H..O	2913	241	5.0	1.7

Table 1: Van der Waals Parameters. -ε is the minimum potential energy at separation r^0, which is the equilibrium radius. They are presented here for comparative purposes. The atom types are defined as follows: O is oxygen, N is nitrogen, H is hydrogen, C denotes extended carbon atoms, and A denotes carbon atoms in a carbonyl or carboxyl group. For the interactions of other atom pairs, we use the geometric means of the A and B values of the atoms involved, for example, $A_{O..N} = (A_{O..O} * A_{N..N})^{1/2}$[31].

3 Potential Energy Computations

In our model of each amino acid, we treat the side chain as a single large 'atom' R, located at the C_β atom. Therefore, as shown in Figure 4, each amino acid thus modelled consists of six atoms: one nitrogen (N), one hydrogen (H), one oxygen (O), two carbons (C and C_α), and C_β/R. For example, for the 10 alanine polypetide chain (10-ALA) example we studied, R is composed of the C_β atom and three hydrogen atoms. It is treated as an "extended carbon atom" for the van der Waals interaction in [31] (see also the caption of Table 1).

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

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + \sum_{atom\ pairs} (A/r_{ij}^{12} - B/r_{ij}^{6}), \quad (5)$$

which is similar to the potential used in [31]. 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 $d_0 = d_c = 2\mathring{A}$, and d_i is the separation between hydrogen atom and oxygen atom. The second term corresponds to the van der Waals interaction among the atoms. The parameters for the van der Waals interaction are listed in Table 1, which encodes strong perference for interactions between oxygon and hydrogen atoms.

We used this potential *only* for our 10-ALA polypeptides and no restraints were set (i.e., the first term in Equation (5) is 0). In this case, the potential is therefore the van der Waals potential plus implicit hydrogen bonds.

However, even for relatively small proteins (around 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. 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. 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.

In particular, if the minimum distance is less than 1.0 Å, we return a very large value; if the minimum distance is greater than 1.0 Å but less than 2.4 Å, we return a value of larger than $E_{\max}^{\text{gen}}$, but smaller than $E_{\max}^{\text{con}}$, where $E_{\max}^{\text{gen}}$ and $E_{\max}^{\text{con}}$ are the maximum thresholds for node generation and node connection,

respectively. Therefore, a conformation where the minimum distance between any pair of R atoms is less than 1.0 Å is always invalid, while a conformation where the minimum distance is 1.0-2.4 Å is invalid during node generation, but valid during node connection. This relaxation during connection allows proteins to go through higher potentials during the connection phase [43, 44]. Currently, we set $E_{\max}^{\text{gen}} =$ 89000 KJ/mol and $E_{\max}^{\text{con}} =$ 300,000 KJ/mol for all proteins studied except 10-ALA, for which we set them as 500 and 600 KJ/mol, repectively.

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

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + E_{hydrophobic}, \quad (6)$$

The first term is exactly the same as in Equation (5), 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 information is obtained from a program called "DSSP" [24], 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.

4 Validating Folding Pathways

For the protein folding pathways found by our PRM framework to be trusted, we must find some way to validate the method with known results. Even though the folding pathways provided by PRMs cannot be explicitly associated with actual timesteps, they do provide us with a temporal ordering. Therefore, we could study the following features:

- The intermediate or transition states on the pathway, and the order in which they are obtained.
- Secondary structure formation order.

Folding intermediates have been an active research area over the last few years, even though there is still debate about whether a protein must go through particular intermediate states to reach the native conformation, see, e.g., [36]. (This is thought to be true for some, but not all, proteins.) Therefore, one possibility is to compare our folding pathways with experimental results known about folding intermediates. If it is useful to identify intermediate states, and the PRM technique is shown to be successful in determining them, then this approach could prove to be a valuable tool for studying protein structure formation.

The formation order of secondary structures is also an interesting feature to study. For example, a technique to extract secondary structure formation order and its relation to tertiary structure formation would be a valuable tool for investigating issues such as whether secondary structure always forms before the tertiary structure, or if instead tertiary structure is formed in a one-stage transition. In this paper, we focus on validating our folding pathways by comparing the order in which the secondary structures form in our paths to results for some small proteins that have been determined by hydrogen exchange (native state out-exchange and pulse labeling) experiments [34]. These results are presented in Section 5.4.

5 Results and Discussion

We now describe protein folding results obtained using our PRM-based approach. In this paper we can only show path snapshots; movies showing the folding process can be found on our webpage [4].

5.1 Implementation details

The PRM code we used was our group's C++ motion planning library [2], which implements a variety of PRM variants. The experiments were performed on an Intel Pentium III 550 MHz PC.

5.2 Proteins studied

We first study a polypeptide with ten Alanine amino acids. This is a small enough structure that we can study it in detail, and for which we can afford more precise potential energy computations. This model is from the IMB Jena Image Library of Biological Macromolecules [23].

We next present results for two small proteins. The structures for all proteins studied were obtained from the indicated pdb file from the Protein Data Bank [39]. Protein GB1 (streptococcal protein G, immunoglobulin-binding domain B1, pdb: 1GB1) has

56 amino acids (112 dof) and consists of one alpha helix and one four strand beta sheet. Beta strands 1 and 2 form the N-terminal hairpin, and beta strands 3 and 4 form the C-terminal hairpin. Protein A (Staphylococcus Aureus Protein A, immunoglobulin-binding B domain, pdb: 1BDD) has 60 amino acids (120 dof) and consists of three alpha helices. Illustrations of proteins GB1 and A are shown in Figures 12 and 13.

In addition, we study in less detail several other proteins: CTX III (60 amino acids, pdb: 2CRT), Cytochrome C (104 amino acids, pdb: 1HRC), hen egg white Lysozyme (129 amino acids, pdb: 1UIH), and α-Amylase Inhibitor (74 amino acids, pdb: 1HOE).

5.3 Roadmap construction: sampling and efficiency

Table 2 shows roadmap statistics for the 10-ALA chain and proteins A and GB1. As can be seen, the largest roadmaps constructed for the proteins took about 6 hours on a Pentium III 550 MHz PC. Given the roadmap (which is constructed once during preprocessing), the time to extract a folding path (the query time) from any random starting conformation is feasible (less than 7 minutes for our largest problems).

Since the roadmap nodes reflect the conformation space of a given protein (particularly near the native state), it is instructive to examine the distribution of the sampled nodes. Figure 5(a-c) shows the ϕ vs. ψ distributions for the 10-ALA chain, and for proteins GB1 and A, respectively. The effect of our graduated normal sampling distribution (Section 2.2.1) around the native fold is clearly seen from the dense sampling around the α and/or β regions.

Figure 5(d-f) shows the node distributions in terms of their RMSD distance from the goal and their potential. These plots provide a glimpse of the shape of the funnel nature of the potential landscape around the native fold, where all dimensions have been collapsed into the RMSD distance. It is interesting to note the differences in the plots for protein GB1 and protein A. As discussed in more detail in Section 5.6, we believe these distributions can provide insight into folding behavior.

5.4 Validation of folding pathways

While the paths encoded in our roadmaps cannot be associated with any real time, they do give a temporal ordering for the conformations on the pathway. Thus, we can attempt to validate our pathways by comparing the secondary structure formation order on our paths with experimental results providing this information. In particular, we use hydrogen exchange experimental results (see [34]) which have been used to indicate which secondary structure components are the last to unfold (the slow exchange core, identified by native state out-exchange experiments) or the first to form (the folding core, identified by pulsed-labeling experiments). There is some disagreement in the community as to whether the slow exchange core is also the folding core, but we have focussed on examples in which there is agreement.

For this purpose we study proteins GB1 and A. In both cases, the extended amino acid chain is the starting denatured conformation. It is connected to the roadmap, and then we compute the minimum weight path in the roadmap connecting it to the native three-dimensional structure. Snapshots of folding paths found by our planner for protein GB1 and protein A are shown in Figure 7 and Figure 8, respectively. We then examined our paths to determine the order in which secondary structure appears. This was done both visually and more formally by determining when residues in the helix or beta sheet were within the appropriate distance of each other to form the structure (i.e., when the necessary *contacts* are formed [18]).

In general, our results are very encouraging. For both proteins studied, the formation order of the secondary structures on our paths seems to agree with the experimental results. Thus, while further investigation and tuning of the PRM technique for proteins is still needed, our preliminary findings show that this motion planning approach is a potentially valuable tool. For example, it could be used to study the secondary structure formation order for proteins where this has not yet been determined experimentally.

5.4.1 Visual identification of secondary structures

Protein GB1 has 56 residues (112 dofs), and consists of an alpha helix and a four strand beta sheet. The beta sheet is composed of the N-terminal and C-terminal beta hairpins. Pulse-labeling experimental results [28, 34] indicate that the alpha helix and the C-terminal hairpin form first and are protected during hydrogen-deuterium exchanges. Native state out-exchange results also indicate the helix is in the slow exchange core. The path found by our method is consistent with these results. For example, from the snapshots shown in Figure 8, one can clearly see that alpha helix in the middle of the polypeptide forms first.

Protein Folding Roadmap Construction Statistics								
Model	**dof**	**Gen**	**Con**	**#N sam**	**#N BigCC**	**#edges**	**Query**	**#N path**
polypeptide	20	22	271	500	491	6034	2.2	7
10-ALA		88	1245	2000	1992	24847	50	10
		440	10607	10000	9988	128713	263	8
Protein	112	60	703	500	499	6263	13	6
GB1		241	3020	2000	1998	25681	50	9
		1226	21941	10000	9997	130774	394	6
Protein	120	90	750	500	497	6148	16	9
A		362	3246	2000	1990	25044	57	8
		1765	22768	10000	9975	127854	357	7

Table 2: Roadmap construction statistics for Protein GB1 and Protein A. 'Gen', 'Con' and 'Query' represent the node generation, connection and query times in seconds, resp. '#N sam' is the number of sampled nodes. '#N BigCC' is the number of the nodes in the biggest connected component of the roadmap, '#edges' is the total number of edges, and '#N path' is the number of roadmap nodes in the final folding path.

Protein A has 60 residues (120 dofs), and consists of three alpha helices. The pulse-labeling results [34] show that the three alpha helices form at about the same time and are in the folding core. Our paths seem to be consistent with these results, as seen in the path snapshots in Figure 7.

5.4.2 Contact distance identification of secondary structures

As a more formal means of validating when secondary structures form in our paths, we have analyzed them in terms of the time steps in which the necessary contacts are formed between hydrophobic residues, e.g., as in [18]. 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 Å of each other are said to form a native contact. Then, when analyzing a conformation q, if the corresponding C_α atoms are ≤ 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, contact distances provide us with a tool for performing more detailed analysis of the folding pathways.

Contact distance analysis was performed on the paths for proteins GB1 and A. The results are shown in Figures 9 and 10. In the figures, the full contact matrix (among hydrophobic residues) is presented 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.

Illustrations of proteins GB1 and A labeling the alpha helices and beta strands are shown in Figures 12 and 13, respectively. For protein GB1 (Figure 9), 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. For protein A (Figure 10), alpha helix 3 (IV) formed around time step 151, alpha helix 1 (I) formed around time step 157, alpha helix 2 (III) formed around time step 161, the contacts between helices 2 and 3 (V) formed around time step 195, and the contacts between helices 1 and 3 (II) formed around time step 200. Thus, in both cases, the contact analysis further validated our paths with the known experimental results [34].

5.5 Sensitivity to Sampling Density

An important consideration for a sampling based method like PRM is to decide how many samples are sufficient to accurately map the interesting portion of the conformation space. One way to address this question is to analyze paths from the same initial conformation to the native state in roadmaps of different size. In Figure 5(g-i) we analyze the potential energy profiles of the folding paths for the 10-ALA chain, protein GB1, and protein A, respectively, for three different sized roadmaps. We expect that as the

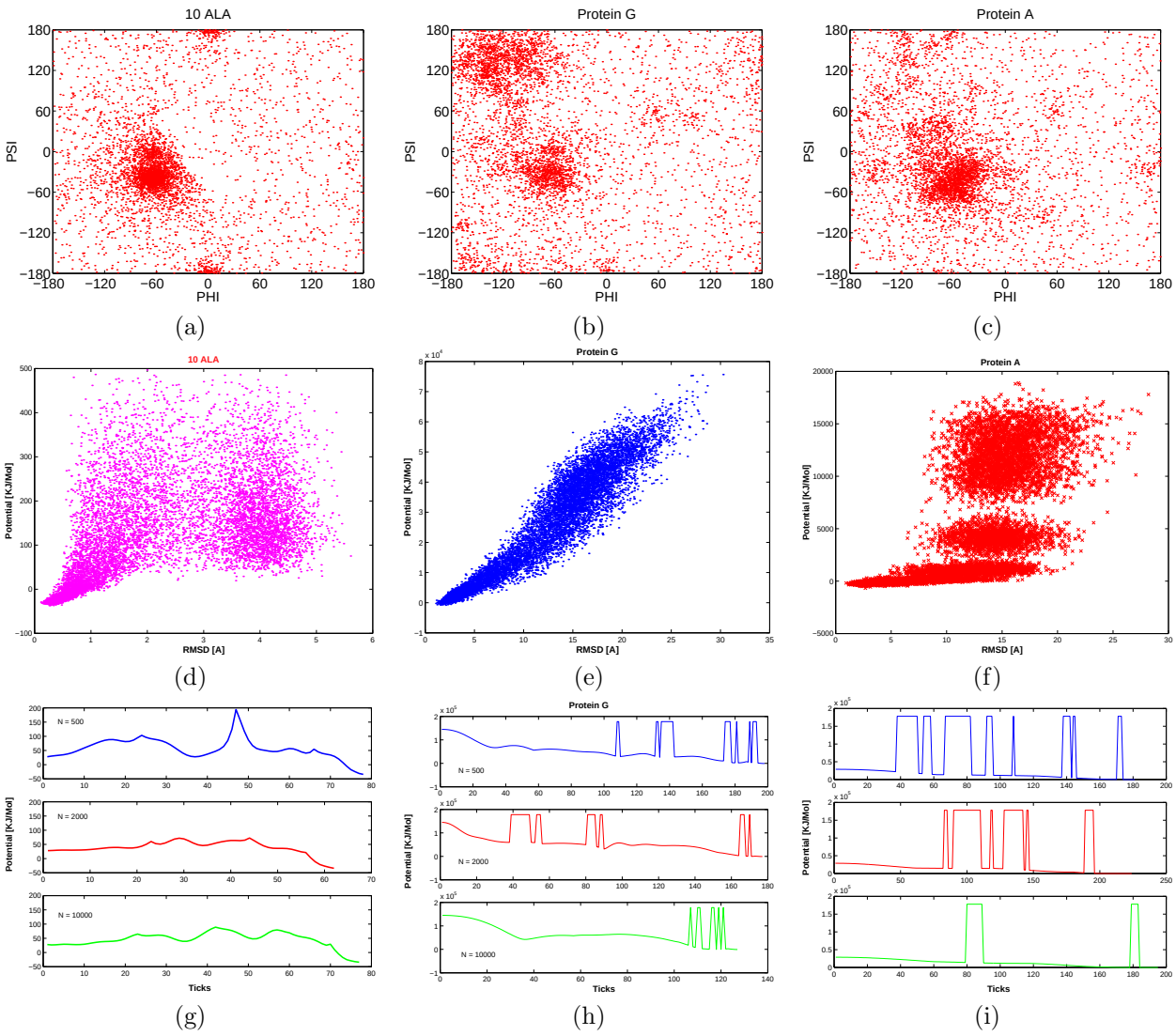
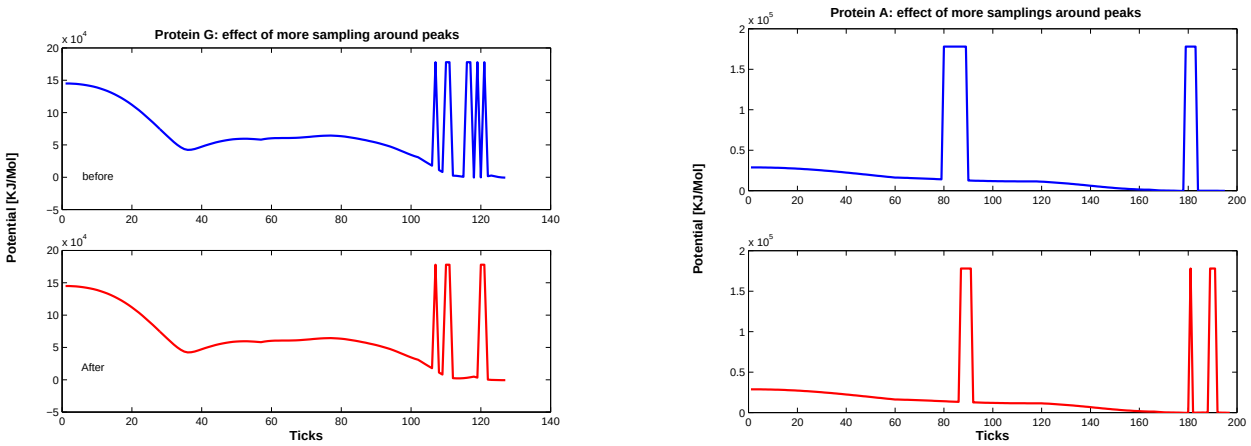


Figure 5: (a-c) Phi/Psi plots, (d-f) Potential/RMSD plots, and (g-i) path profiles for the 10-ALA chain (a,d,g), protein GB1 (b,e,h), and protein A (c,f,i).

number of nodes sampled increases (the sampling is denser), our roadmaps will contain better and better approximations of the natural folding path. Our results support this belief, and moreover, enable us to estimate how many nodes should be sampled. For example, we can see in the plots that as the number of nodes, N, is increased, the paths seem to become smoother, having fewer and smaller peaks in their profiles. When no further improvement is noted, the sampling could be determined to be sufficient.

5.5.1 Refining folding pathways by resampling

Another interesting point is the similarity among the paths for all roadmap sizes. In particular, they all have a peak (or peaks) in the potential profile near the native state (the goal). Some researchers believe such energy barriers around a folding state are crucial for a stable fold. Also, the profiles clearly show that the peak(s) right before the final fold is contributed by the van der Waals interaction (the high potential shown is the maximum value used in our step function approximation for the van der Waals component). This is consistent with the tight packing of atoms in the native fold.



(a) (b)

Figure 6: Potential energy profiles for paths before (top) and after (bottom) resampling for (a) protein GB1 and (b) protein A.

The similarity among the paths for different sized roadmaps also implies that they may share some common conformations, or subpaths, and this knowledge could be used to bias our sampling around these regions, hopefully further improving the quality of the paths. Indeed, as can be seen in Figure 6, resampling around the local maxima as described in Section 2.4.1 does indeed prove beneficial. In the figure, the top plot shows the energy profile of the original query path, and the bottom plot shows the same after resampling. We note that while the peaks were not removed entirely, they were generally reduced. We expect that more resampling would further smooth the paths, but it would not be expected to completely flatten them due to the energy barriers that are thought to surround the native fold. As previously discussed, this resampling is a useful way to compensate for the simple sampling strategy and the rather naive straight-line roadmap connections. Postponing this optimization until after the initial query is performed enables us to target our resampling efforts to only the necessary regions of the conformation space.

5.6 Potential landscapes and secondary structure formation

Each protein has a unique amino acid sequence and a unique fold. Therefore, in principle, each protein also has a unique folding behavior which will differ from other proteins in terms of folding rate, secondary and tertiary structure formation, whether it has intermediate states (conformations on the folding route which are present for extended periods of time), etc. It is interesting, however, to think about clustering proteins according to one or more of these behaviors. For example, proteins in which the secondary structure forms first, as sub-units, before they pack together to reach the tertiary structure as opposed to proteins in which secondary and tertiary structure are formed simulateneously. As another example, the length of the amino acid sequence strongly affects the folding rate – small proteins tend to fold rapidly, while larger proteins may take more time. In terms of potential landscapes, proteins that fold similarly might have potential landscapes that share certain characteristics. For example, the potential landscape of a rapidly folding protein might be relatively smooth with few local minima. We believe our framework may be a valuable tool for studying such issues.

Consider, for example, the potential vs. RMSD distribution plot for protein A shown in Figure 5(f). The energies for the nodes with RMSD approximately in the range 0-10 Å form a very narrow corridor, which rapidly spreads out as the RMSD values increase. When we examine snapshots of the path shown in Figure 7, we see that even at conformations RMSD 10 Å from the native fold, the three helices have already formed (see Figure 7 (c)). Thus the narrow corridor from 10 Å to the native fold (0 Å) is only the packing of the secondary structures to form the tertiary structure. The potential changes very little in this range because the restraints and the hydophobic terms in our potential function (Equation 6) are mostly already satisfied at this point.

In contrast, the potential vs. RMSD distribution plot for protein G shown in Figure 5(e) has a rather

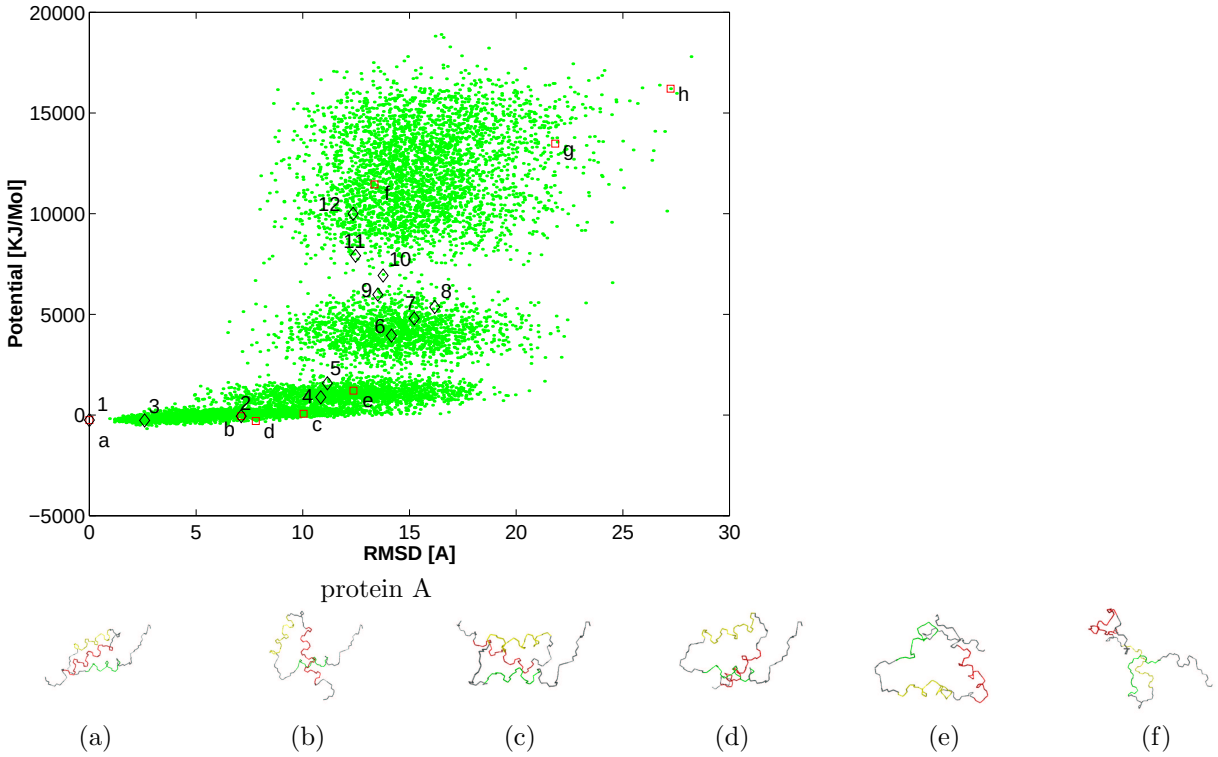


Figure 7: Folding from different starting conformations for protein A. Folding snapshots from the extended conformation are also shown in reverse order, labels running from a to f. The folding process from a random conformation to the goal is shown with numerical labels in the Potential vs RMSD plot (snapshots not shown).

different structure from that of protein A. This is consistent with our paths (and the pulse-labeling results [34, 28]), where the secondary structures form at different times (the alpha helix first, and then the beta sheet). Moreover, the beta sheet appears to obtain its tertiary position directly, and thus we do not have the clearly separable phase of the packing of secondary structures as was seen for protein A. This is reflected in the smooth funnel nature of the potential vs. RMSD plot.

The intriguing differences noted above in the distributions of the roadmap nodes in the potential vs. RMSD plots led us to conjecture that these distributions are related to the shape of the folding funnels which influence overall folding behavior. In particular, we believe that the energy distributions suggest the formation order of the secondary and the tertiary structure, and moreover, that changes in the distributions might indicate different stages of the folding process.

To investigate this issue, we considered six proteins: two proteins containing only alpha helices (A and Cytochrome C), two proteins containing both helices and beta strands (GB1 and hen egg white Lysozyme), and two proteins containing only beta stands (CTX III and α-Amylase Inhibitor). The potential vs. RMSD distributions for all six proteins are shown in Figure 11; the all alpha proteins are on the left, the all beta proteins are on the right, and the mixed proteins are in the middle.[2]

It is interesting to note the contrast between the landscapes of the all alpha and the all beta proteins, even though the same technique, which does not utilize information regarding the secondary structure, was applied to all of them. These distributions seem to reflect the fact that all alpha proteins tend to fold quite differently from all beta proteins. In particular, all alpha proteins tend to form the helices first, 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

[2]The results shown in Figure 11 use a variant of the method described in this paper. It still focuses sampling around the native state, but instead of using a set of normal distributions we generate new conformations by iteratively applying small perturbations to existing conformations. This version appears to produce smoother distributions and is much faster.

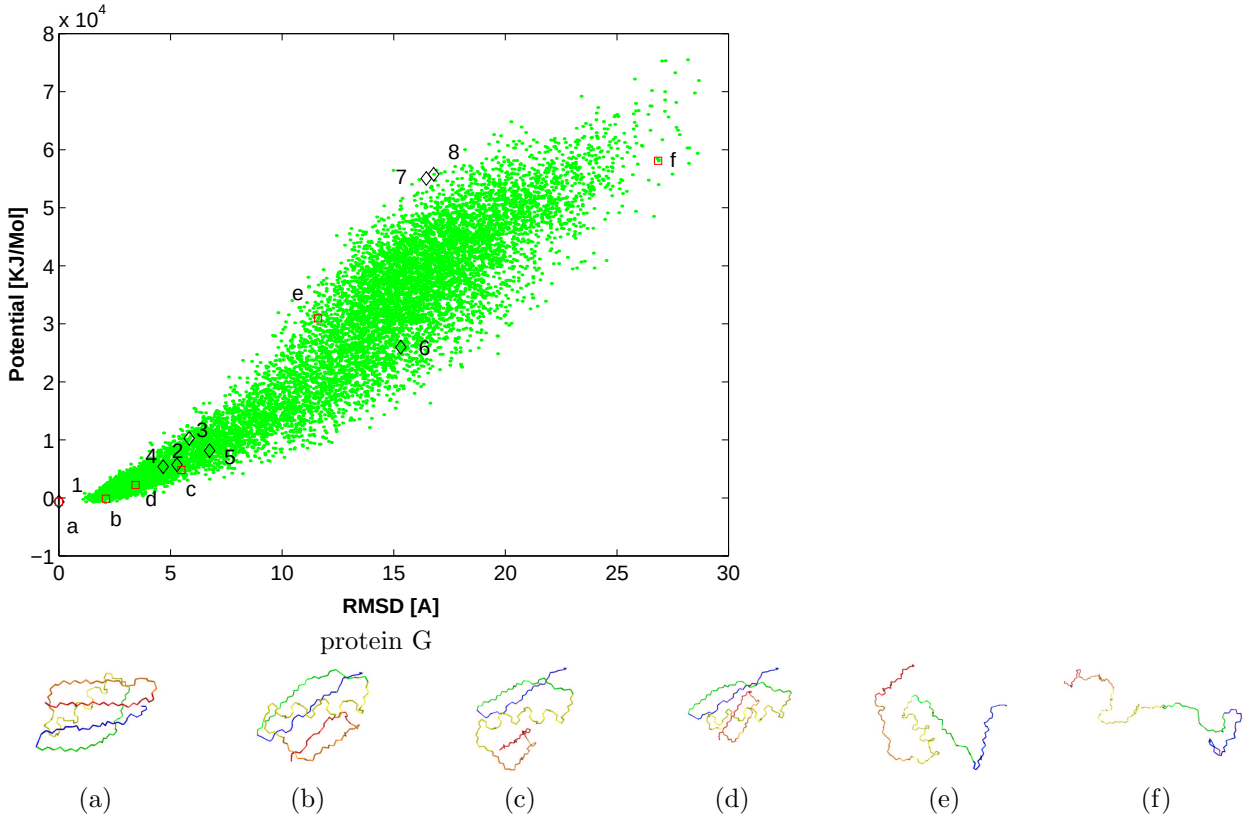


Figure 8: Folding from different starting conformations for protein GB1. Folding snapshots from the extended conformation are also shown in reverse order, labels running from a to f. The folding process from a random conformation to the goal is shown with numerical labels in the Potential vs RMSD plot (snapshots not shown).

the potential changes very little as the RMSD approaches zero. In contrast, the distributions for the all beta proteins are much smoother, indicating that the secondary and the tertiary structure are formed simultaneously. For the mixed alpha and beta proteins, the plots share some features of the plots for the all alpha proteins and for the all beta proteins. And moreover, the degree of similary seems to be related to the proportion of the protein composed of a particular secondary structure. For example, hen egg-white Lysozyme, whose secondary structure is mainly alpha, has a similar distribution to the all alpha Cytochrome C, and the distribution for protein GB1, which is more beta than alpha, is similar to that of protein CTX III, an all beta protein.

5.7 Folding from different start conformations

For PRMs, after building the roadmap, searching for the folding pathway from any denatured state is just another query (which are handled relively quickly as opposed to building a new roadmap). That is, there is nothing special about the extended conformation as a starting conformation. This is one of the key features of the PRM approach that distinguishes it from other simulation methods that compute a single folding trajectory.

Figure 7 and Figure 8 show the folding pathways imposed on the potential vs. RMSD plots for different starting conformations. One can see that different pathways tend to come together and appear to share some common portions ('gullies') as they approach the native fold. They also reflect some common behavior regarding the formation of secondary structure. The formation of the secondary structure can be appreciated in the (reverse) path snapshots shown beneath the plots, where the labels a-f correspond to the same labels in the plot above. One may also note that the folding paths for both protein A and protein GB1 cross themselves when ordered by RMS distance, i.e., for protein A we have *ab<u>dc</u>efgh*

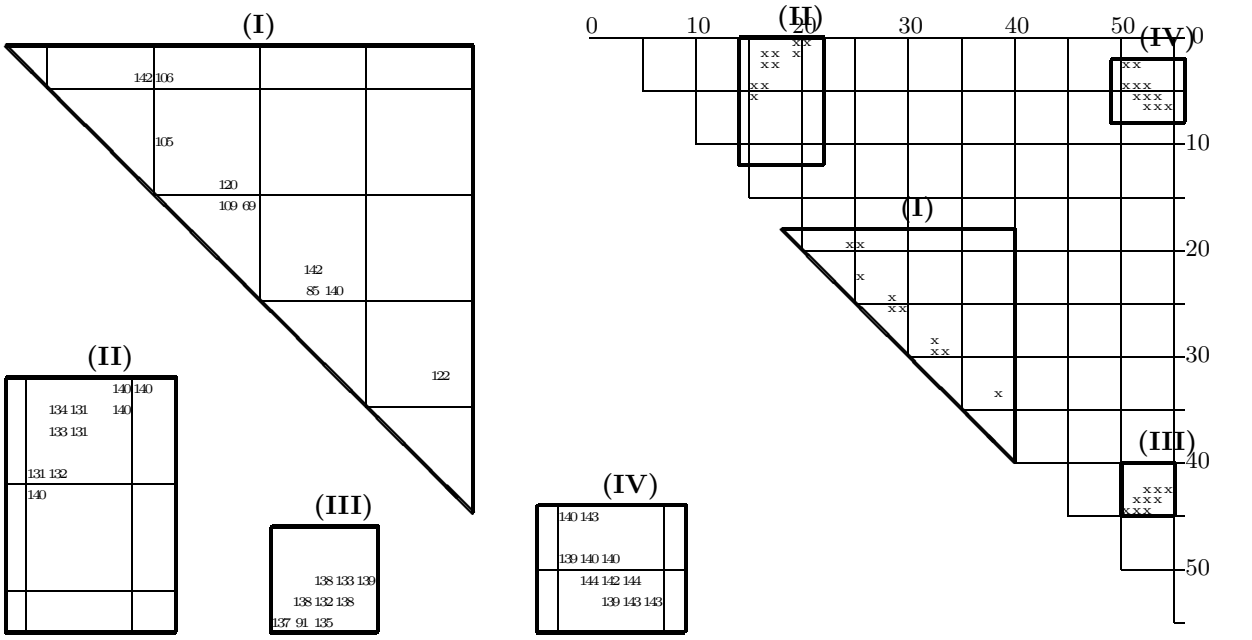


Figure 9: Protein GB1 (see Figure 12). The full contact matrix (right) and blow-ups (left) showing the time steps when the contacts appear on our path. Blow-ups I, II, III, and IV correspond to the alpha helix contacts, the beta 1-2 contacts, the beta 3-4 contacts, and the beta 1-4 contacts, respectively.

and for protein GB1 *ab*$\underline{dc}$*ef*. This is because on the way to the native fold, proteins may pass through states, such as the "molten globule" state, which are similar to the native fold and thus have small RMSD values. However, these are not necessarily intermediate states.

It is a simple matter to extract the shortest, or lowest weight, path between any two nodes in our roadmap using standard graph search techniques. In our case, since we are interested in analyzing multiple paths to a common goal (the native state), it is convenient to construct a single-source shortest path (SSSP) tree rooted at the native state [13]. This shortest path tree provides in some sense a description of the global folding environment and the potential landscape of the protein. Given the shortest path tree encoding folding pathways to the native state, we can analyze overall folding pathway behavior (e.g., secondary structure formation order) by studying the folding pathways from the *near–random conformations* in the tree, i.e., from the conformations with very few native contacts. Given a folding pathway starting from a near–random conformation, we determine the formation order of the native state contacts for the various secondary structures, and the contacts between secondary structures in the tertiary structure. As in our validation studies (Section 5.4.2), a contact is present when the participating atoms are ≤ 7 Å apart, and the formation time of a secondary structure or tertiary contact is the average of the appearance times for all contacts in that structure.

The results of such a study are shown in Figure 12 and Figure 13 for proteins GB1 and A, respectively. In the plots, the x-axis is labeled with the formation order of the structure for the largest percentage of paths, and the y-axis represents formation order. Each graph in the plot represents one permutation of the formation order that was present in the roadmap; the legend shows the percentage of paths that follow that order.

For protein GBI, Figure 12 shows that in about 60% of the pathways, the first structure formed is the alpha helix, which is followed by the C-terminal harpin (beta strands 3 and 4), and then the N-terminal harpin (beta strands 1 and 2), and finally the two harpins come together forming the contacts between the beta 1 and beta 4 strands. The remaining 40% of the pathways invert the formation order of the N-termial harpin and the C-terminal hairpin. This agrees with hydrogen exchange experiments and reflects the fact that the N-terminal harpin and C-terminal harpin are in some coarse sense symmetric.

Figure 13 shows that the secondary structure formation order (as well as the contacts between them) for protein A agrees with the experimental results, i.e., all three alpha helices form at about the same

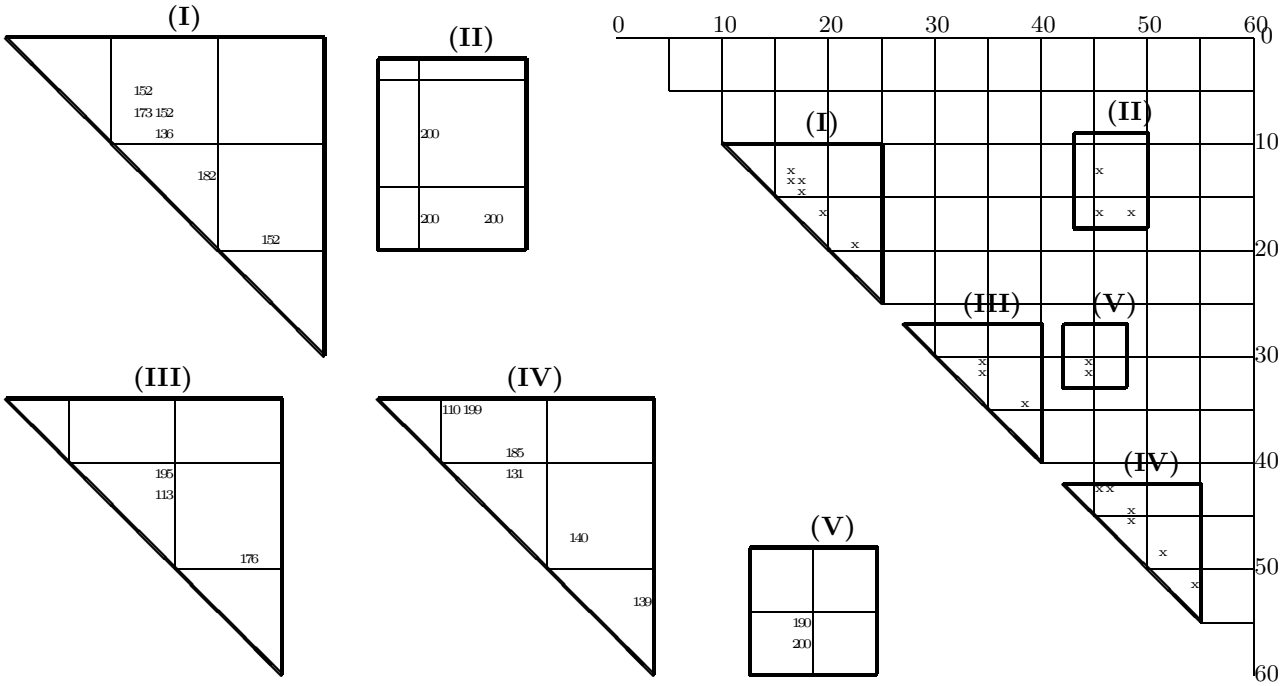


Figure 10: Protein A (see Figure 13). The full contact matrix (right) and blow-ups (left) showing the time steps when the contacts appear on our path. Blow-ups I, II, III, IV, and V correspond to the alpha helix 1 contacts, the 1-3 helix contacts, the alpha helix 2 contacts, the alpha helix 3 contacts, and the 2-3 helix contacts, respectively.

time. In particular, about 60% of the paths first formed alpha helix 3, then alpha helix 1, then alpha helix 2 (the center one), then the contacts are formed between helices 2 and 3, and then finally the contacts are formed between helices 1 and 3. There were three other permutations present, representing 20%, 15% and 5% of the paths. In all of them, the helices came together in the same order, and they differed only in the order in which the helices were formed. Our results also indicate that the two helices at the ends have a strong tendancy to form earlier than the helix in the middle – in only 15% of the paths was the central helix (2) the first to form.

6 Conclusion and Future Work

In this paper, we present a framework for studying protein folding using motion planning techniques. Our approach, which is based on the PRM motion planning method, was seen to produce interesting results for representative small proteins. One of the most important benefits of this approach to folding problems is that it enables one to study the dynamic folding process itself. Unfortunately, it is difficult to appreciate this from the few path snapshots we are able to display in a paper. (Movies can be viewed on our webpage [4]). Nevertheless, we believe that our results establish that this is a promising approach which deserves further investigation.

In current work, we are using a variant of the method described in this paper. It still focuses sampling around the native state, but instead of using a set of normal distributions we generate new conformations by iteratively applying small perturbations to existing conformations. This version appears to produce better results, and connection is much faster, especially for larger proteins. We are also further refining our potential energy approximation, investigating more sophisticated sampling techniques to concentrate more nodes near the local maxima observed in the path profiles, and are exploring additional validation mechanisms (e.g., comparison to other simulation approaches). We are also performing more extensive analysis of our paths and are studying more proteins.

7 Acknowledgements

We would like to thank Jean-Claude Latombe for pointing out to us the connection between box fold-

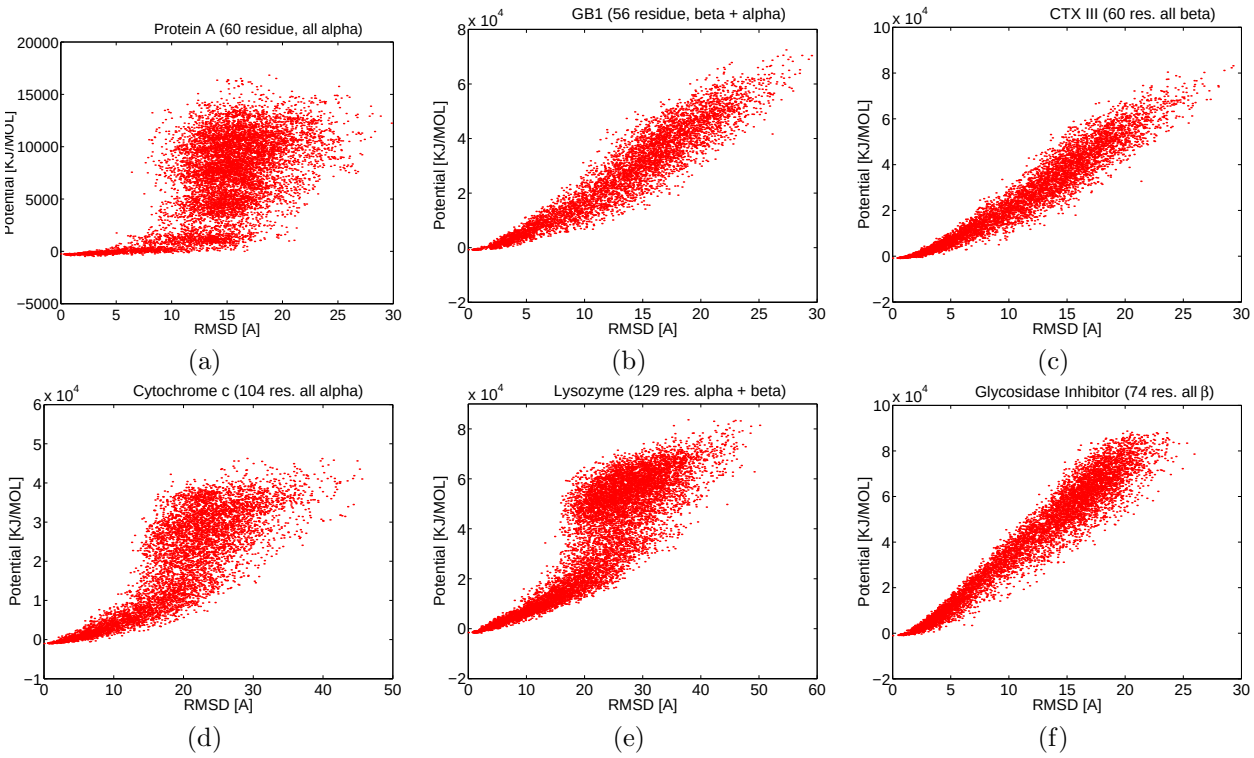


Figure 11: The potential vs RMSD distribution, or the 'landscape', for protein (a) A, (b) GB1, (c) CTX III, (d) Cytochrome c, (e) hen egg white Lysozyme, and (f) α-Amylase Inhibitor. The two proteins in the first (left) column are all alpha proteins, the middle column contains mixed alpha and beta proteins, and the third (right) column contains all beta proteins.

ing and protein folding. We would also like to thank Marty Scholtz for suggesting validation using the hydrogen exchange experimental results, and Ken Dill, Michael Levitt, and Vijay Pande for useful suggestions.

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*, 96:11305–11310, 1999.

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

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

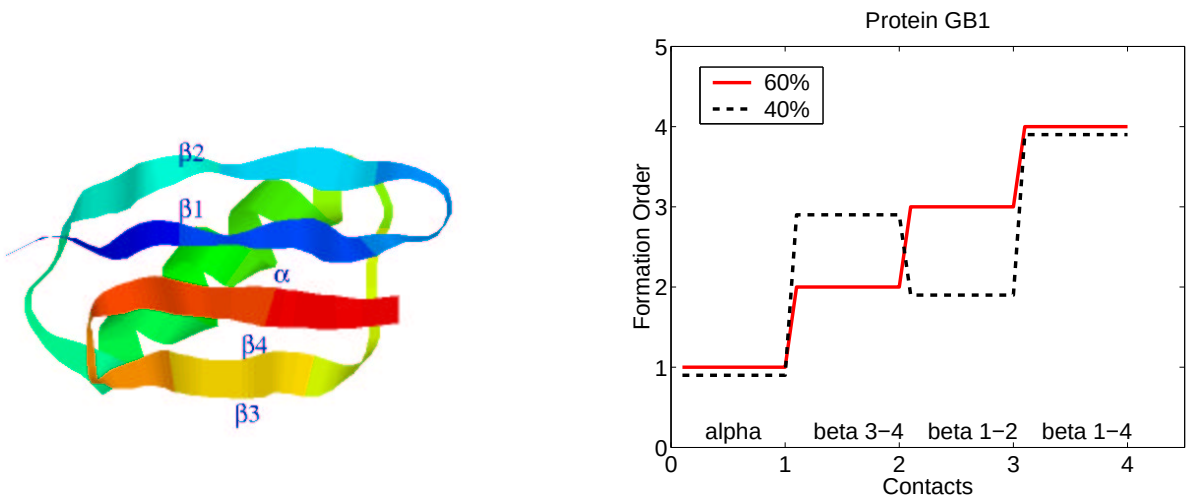
[4] Nancy M. Amato. Motion planning group webpage. `http://parasol-www.cs.tamu.edu/people/amato/`.

[5] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

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

[7] O. B. Bayazit, G. Song, and N. M. Amato. Enhancing randomized motion planners: Exploring with haptic hints. *Autonomous Robots, Special Issue on Personal Robotics*, 10(2):163–174, 2001. Preliminary version appeared in ICRA 2000, pp. 529–536.

[8] O. 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.*



(a) (b)

Figure 12: Protein G and its secondary structure formation order. The x-axis is labeled with the formation order of the structures (contacts) for the largest percentage of paths, and the y-axis represents formation order. Each graph in the plot represents one permutation of the formation order that was present in the roadmap; the legend shows the percentage of paths that follow that order.

Autom. (ICRA), pages 954–959, 2001. This work was also presented as a poster at RECOMB'01.

[9] T. Biedl, E. Demaine, M. Demaine, S. Lazard, A. Lubiw, J. O'Rourke, M. Overmars, S. Robbins, I. Streinu, G. Toussaint, and S. Whitesides. Locked and unlocked polygonal chains in 3D. In *Proc. 10th ACM-SIAM Sympos. Discrete Algorithms*, pages 866–867, January 1999.

[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.U. Bowie and D. Eisenberg. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA*, 91:4436–4440, 1994.

[12] J. Cantarella and H. Johnston. Nontrivial embeddings of polygonal intervals and unknots in 3-space. *J. Knot Theory Ramifications*, 7:1027–1039, 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] D.G. Covell. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.*, 14:409–420, 1992.

[15] V. Daggett and M. Levitt. Realistic simulation of naive-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.*, 22:353–380, 1993.

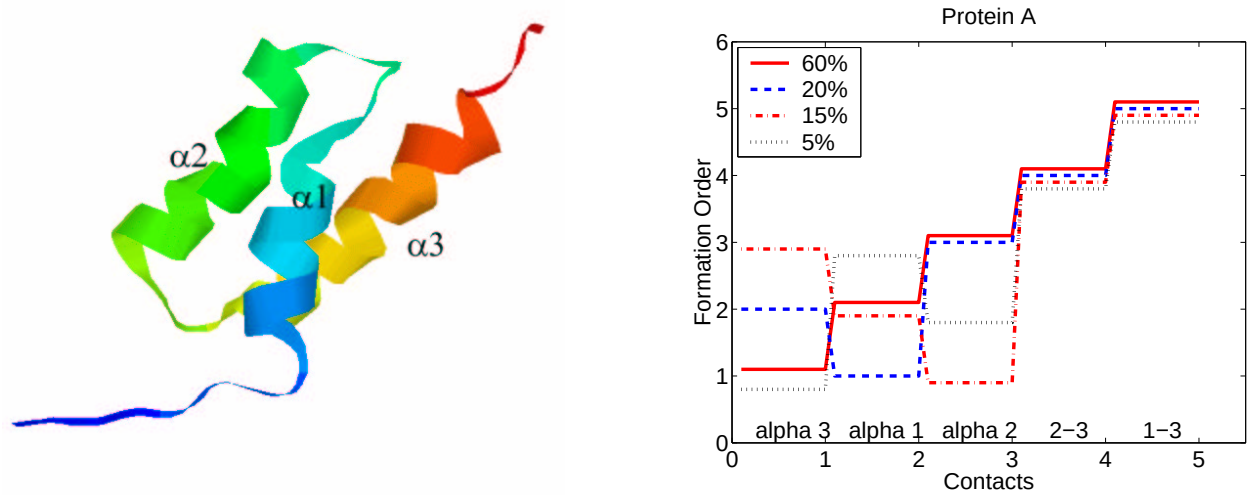
[16] Y. Duan and P.A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[17] W.A. Eaton, V. Muñoz, P.A.Thompson, C. Chan, and J. Hofrichter. Submillisecond kinetics of protein folding. *Curr. Op. Str. Biol.*, 7:10–14, 1997.

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

[19] J.M. Haile. *Molecular Dynamics Simulation: elementary methods*. Wiley, New York, 1992.

[20] 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.



(a) (b)

Figure 13: Protein A and its secondary structure formation order. The x-axis is labeled with the formation order of the structures (contacts) for the largest percentage of paths, and the y-axis represents formation order. Each graph in the plot represents one permutation of the formation order that was present in the roadmap; the legend shows the percentage of paths that follow that order.

[21] B. Honig. Protein folding: From the Levinthal Paradox to structure prediction. *J. Mol. Biol.*, 293:283–293, 1999.

[22] 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.

[23] The IMB Jena Image Library of Biological Macromolecules. http://www.imb-jena.de.

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

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

[26] 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.

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

[28] J. Kuszewski, G.M. Clore, and A.M. Gronenborn. Fasting folding of a prototypic polypeptide: The immunoglobulin binding domain of streptococcal protein G. *Protein Sci.*, 3:1945–1952, 1994.

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

[30] 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.

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

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

[33] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

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

[35] 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.

[36] C.R. Matthews. Pathways of protein folding. *Annu. Rev. Biochem.*, 62:653–683, 1993.

[37] 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.

[38] 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.

[39] The Protein Data Bank. `http://www.rcsb.org/pdb/`.

[40] G.N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Adv. Prot. Chem.*, 28:283–437, 1968.

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

[42] E.I. Shakhnovich. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Op. Str. Biol.*, 7:29–40, 1997.

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

[44] G. Song and N. M. Amato. How does it fold? searching for folding pathways using a motion planning approach. Technical Report TR00-014, Department of Computer Science, Texas A&M University, May 2000.

[45] G. Song and N. M. Amato. A motion planning approach to folding: From paper craft to protein structure prediction. Technical Report TR00-001, Department of Computer Science, Texas A&M University, January 2000.

[46] 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.

[47] 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.

[48] G. Song, S. L. Miller, and N. M. Amato. Customizing PRM roadmaps at query time. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 1500–1505, 2001.

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

[50] S. Sun. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.*, 2:762–785, 1993.

[51] S. Sun, P. D. Thomas, and K. A. Dill. A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng.*, 8:769–778, 1995.

[52] 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.