Using Motion Planning to Study Protein Folding Pathways*

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

Nancy M. Amato
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
amato@cs.tamu.edu

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. In particular, our work uses Probabilistic Roadmap (PRM) motion planning techniques which have proven to be very successful for problems involving high-dimensional configuration spaces. Our results applying PRM techniques to several small proteins (60 residues) are very encouraging. The framework enables one to easily and efficiently compute folding pathways from any denatured starting state to the native fold. This aspect makes our approach ideal for studying global properties of the protein's potential landscape. For example, our results show that folding pathways from different starting denatured states sometimes share some common 'gullies', mainly when they are close to the native fold. Such global issues are difficult to simulate and study with other methods.

Our focus in this work is to study the protein folding mechanism *assuming* we know the native fold. Therefore, 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 on the initial conformation. Our results indicate that for some proteins, secondary structure clearly forms first while for others the tertiary structure is obtained more directly, and moreover, these situations seem to be differentiated in the distributions of the conformations sampled by our technique. We also find that the formation order is independent of the starting denatured conformation. We validate our results by comparing the secondary structure formation order on our paths to known pulse-labeling experimental results. This indicates the promise of our approach for studying proteins for which experimental results are not available.

*This research supported in part by NSF CAREER Award CCR-9624315, NSF Grants IIS-9619850, ACI-9872126, EIA-9975018, EIA-9805823, and EIA-9810937, and by the Texas Higher Education Coordinating Board under grant ARP-036327-017.

To appear in the *5th International Conference on Computational Molecular Biology (RECOMB 2001)*, April 2001.

1. INTRODUCTION

The goal of motion planning is to compute a sequence of valid intermediate states to 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* which 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.

In this work, we concentrate on the application of the successful *probabilistic roadmap (PRM)* motion planning method [21] 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 [36], 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 [37] in applying this methodology to folding problems such as carton folding (with applications in packaging and assembly [30]), and paper crafts (studied in computational geometry [33]), provided some evidence of the feasibility of this approach for determining protein folding sequences. For example, note the parallels between periscope and protein folding in the path snapshots shown in Figures 1 and 2, respectively [37]. Our experimental results with small proteins of approximately 60 residues (modeled using approximately 120 degrees of freedom) are quite promising.

There are large and ongoing research efforts whose goal is to determine the native folds of proteins (see, e.g., [19, 27]). 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 [17, 35]. Although there have been some recent experimental advances [13], computational techniques for simulating this process are important because it is difficult to capture the folding process experimentally. 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

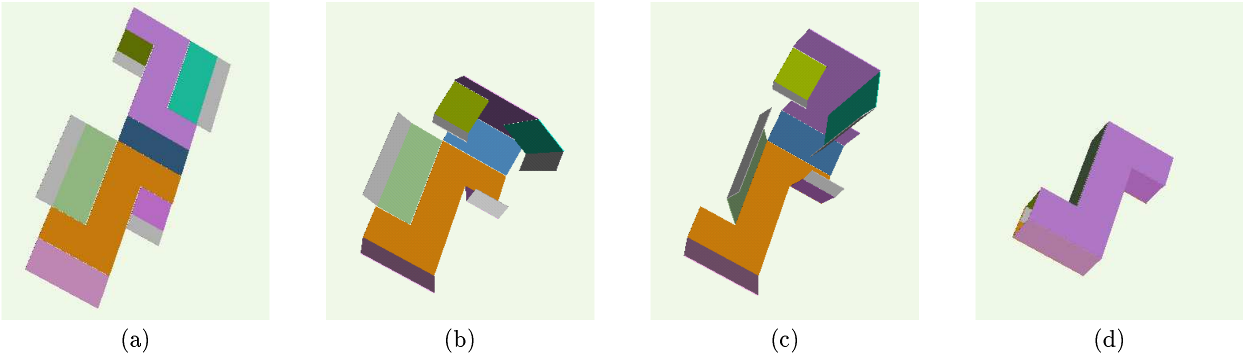


Figure 1: Snapshots of a carton folding.

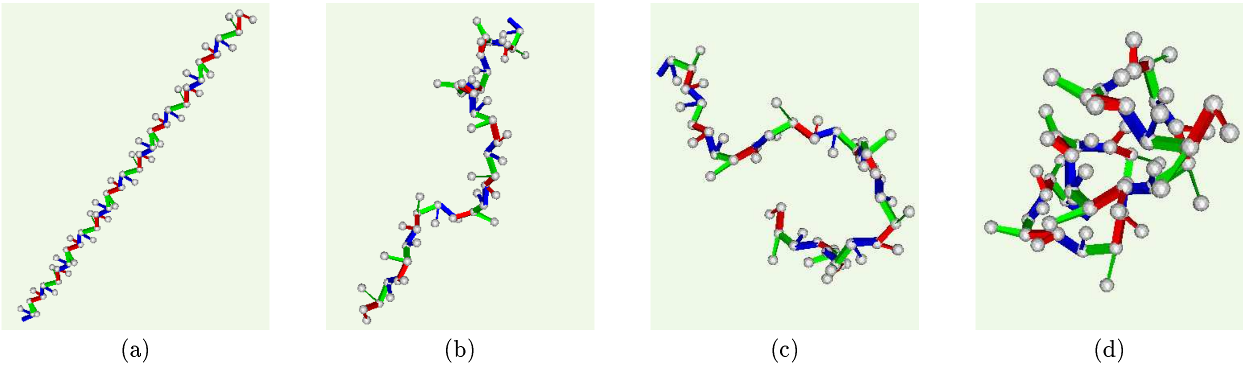


Figure 2: Snapshots of a 10 ALA chain folding.

studying global properties, such as secondary structure formation and funnel behaviors of the protein's potential landscape. For example, our results show that folding pathways from different starting denatured states sometimes share some common 'gullies', mainly when they are close to the native fold. 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 [3]. Since then, many different approaches for predicting protein structure have been explored (see [38] for a review). We note that our work presented here 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 [28, 40], molecular dynamics simulation [26], Monte Carlo methods [10, 22], and genetic algorithms [8, 39]. Among these, molecular dynamics is most closely related to our approach. Much work has been carried out in this area [11, 12, 15, 26], 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. The simulation result also depends heavily on the start conformation and can easily result in local minima. Most of the 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.

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 how we validated the pathways computed by our method. Our results for a 10-ALA polypeptide chain and two small proteins (protein A and protein GB1) 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 [24]. 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 [24], attention has been focussed on randomized or probabilistic methods.

As mentioned in Section 1, our approach to the folding problem is based on the probabilistic roadmap (PRM) approach to motion planning [21]. Briefly, PRMs work by sampling points 'randomly' from C-space, and retaining those that satisfy certain feasibility requirements (e.g., they must correspond to collision-free configurations of the movable object). Then, these points are connected to form a graph, or roadmap, using some simple planning method to connect 'nearby' points. During query processing, paths connecting the start and goal configurations are extracted from the roadmap using standard graph search techniques. (See Figure 3.)

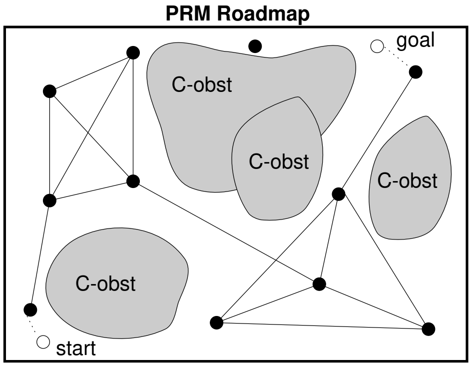


Figure 3: Querying a PRM roadmap (C-space).

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 effected using multiple applications of the feasibility test).

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. 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 fold positions (atomic bonds) correspond to joints and areas that cannot fold (atoms) correspond to links. 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). Side chains are modeled as spheres and are given no degree of freedom. Thus, for an amino acid sequence with k residues, our model will consist of $2k$ links and $2k$ revolute joints.

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 our proteins can be expressed as:

$$\mathcal{C} = \{q \mid q \in S^n\}. \tag{1}$$

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

As described in Section 2.1, since all joints are revolute, 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, as described in Section 3. The node is accepted and added to the roadmap based on its potential energy E with the following probability:

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

This acceptance criterion was also used when building PRM roadmaps for ligand binding in [36]. 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 chain spheres, for example, has higher potential and is thus more likely to be rejected during node generation.

2.2.1 *Gaussian Sampling*

Due to the high dimensionality of the conformation space, simple uniform sampling would have to be very dense to reliably characterize the important features of the potential energy landscape. One idea is to use the Ramachandran plot [34] 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 are $\{5°, 10°, 20°, 40°, 80°, 160°\}$. The small STDs capture the detail around the goal, and the larger STDs ensure adequate roadmap coverage of the conformation space. Similar biased sampling strategies have been applied successfully in robotics applications [2, 7, 16, 18, 20, 25, 41], where oversampling in and near narrow

[1] We would like to remind the readers 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.

passages in C-space is crucial for some problems. Our simulation results shown in Section 5 show that this Gaussian sampling strategy is very useful. Baker and coworkers [4, 1] and Muñoz and Eaton [32] recently also take advantage of the topology of the native state to predict the folding rates and mechanisms of some proteins.

2.3 Constructing 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 configurations between the two corresponding nodes as determined by the chosen local planner (the number of such configurations is, e.g., the resolution used for collision detection, 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. If there are still multiple connected components in the roadmap after this stage (which is generally the case, and in fact is sometimes unavoidable, see, e.g., [6, 9]), other techniques will be applied to try to connect different connected components (see [2] for details).

When two nodes are connected by the local planner, the corresponding edge is added to the roadmap. We associate a weight with each edge. For two consecutive intermediate conformations i and $i+1$ on the edge, we first calculate their potential energies, i.e., E_i and E_{i+1}, and then the probability of moving from conformation i to $i+1$ is determined by:

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

which keeps the detailed balance between two adjacent states. And the total weight of the edge is:

$$Weight = \sum_{i=0}^{N-1} -log(P_i),$$

By assigning the weights in this manner, we can find the most energetically feasible path when performing subsequent queries. A similar weight function, with different probabilities, was used in [36].

2.4 'Querying' the roadmap

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

For our folding problems, it is convenient to actually connect the start and the goal into the roadmap, just as was done for the other roadmap nodes in the connection phase. Dijkstra's algorithm is then used to find the smallest weight path between the start and goal configurations. 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. The advantage of adding the start and the goal to the roadmap is two-fold. First, the roadmap is augmented after each query is performed (this has been noted as a possible optimization for regular PRM applications as well). Second, this facilitates our search for the lowest weight path.

2.4.1 Path optimization: resample near peaks

Since the nodes are generated randomly and connected using straight–line connections, the path returned by the query will not be optimal. In this case, we will attempt to 'push,' or deform, the given path into a better path. We used this strategy successfully in CAD applications to transform invalid user-collected paths into valid paths [5].

In this case, we will resample around all the nodes on the query path that have higher potential than some user specified threshold. In particular, 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 iterate for some fixed number of iterations, connect all new nodes into the roadmap, and then the query is redone.

3. POTENTIAL ENERGY COMPUTATIONS

For potential energy calculations, we start with:

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

which is similar to the potential used in [26]. The first term represents constraints which favor the known secondary structure through main-chain hydrogen bonds and disulphide bonds. The van der Waals interaction among atoms is considered in the second term. All parameters can be found in [26].

We can use this potential for small problems, such as the 10 ALA polypeptide we study, when the computational requirements are relatively mild. However, even for relatively small proteins (around 60 residues), there will be 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. This is computed by considering only the contribution from the side chains and modeling each side chain with a fixed-size rigid sphere (a further approximation). The side chain was 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 still capture the main features of the interaction.

We have also included the hydrophobic effect in a simplistic

Protein Folding Roadmap Construction Statistics							
Model	**dof**	**Gen**	**Con**	**#N sam**	**#N BigCC**	**#edges**	**#N path**
Protein 10-ALA	20	47	896	500	495	6099	4
		189	3749	2000	1990	24504	4
		1005	25229	10000	9978	127763	5
Protein GB1	112	111	1811	500	500	6395	7
		443	7303	2000	2000	25854	5
		2265	48069	10000	9996	132104	6
Protein A	120	188	1905	500	498	5982	4
		734	8008	2000	1994	24824	6
		3737	52517	10000	9976	127282	10

Table 1: Roadmap construction statistics for Protein GB1 and Protein A. 'Gen' and 'Con' represent the node generation and connection 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.

way for the proteins we study. We assign a hydrophobicity value of 1 to all non-polar amino acids, and 0 to the rest. When the sidechains 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 $E_h = 20KJ/Mol$.

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.
- The formation order of secondary structures.

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 intermediate states to reach the native conformation, see, e.g., [31]. (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 a very important feature since it is related to a fundamental question in protein folding. Do secondary structures always form before the tertiary structure, or is tertiary structure 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 pulse labeling and native state out-exchange experiments [29].

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 at `http://www.cs.tamu.edu/faculty/amato/dsmft`.

5.1 Implementation details

The PRM code we used was our OBPRM software library [2], which implements a variety of PRM variants. The experiments were performed on an SGI Octane R10000 machine.

5.2 Models studied

To prepare ourselves to study the folding of real proteins, 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 at `http://www.imb-jena.de/`.

We next present results for two small real proteins. Protein GB1 has 56 residues (112 dof) and consists of one alpha helix and two beta-sheets. Its structure has been determined by both NMR and crystallography. Protein A has 60 residues (120 dof) and consists of three alpha helices. The pdb files used for the proteins were 1GB1.pdb and 1BDD.pdb, respectively, from the Protein Data Bank at `http://www.rcsb.org/pdb/`.

5.3 Roadmap construction: sampling and efficiency

Table 1 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 15 hours on an SGI octane R10000. 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 quite small (around 30 minutes for our largest problems).

Since the roadmap nodes reflect somewhat the conformation space for a given protein, it is illustrative to examine the distribution of the sampled nodes. Figure 4(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 4(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

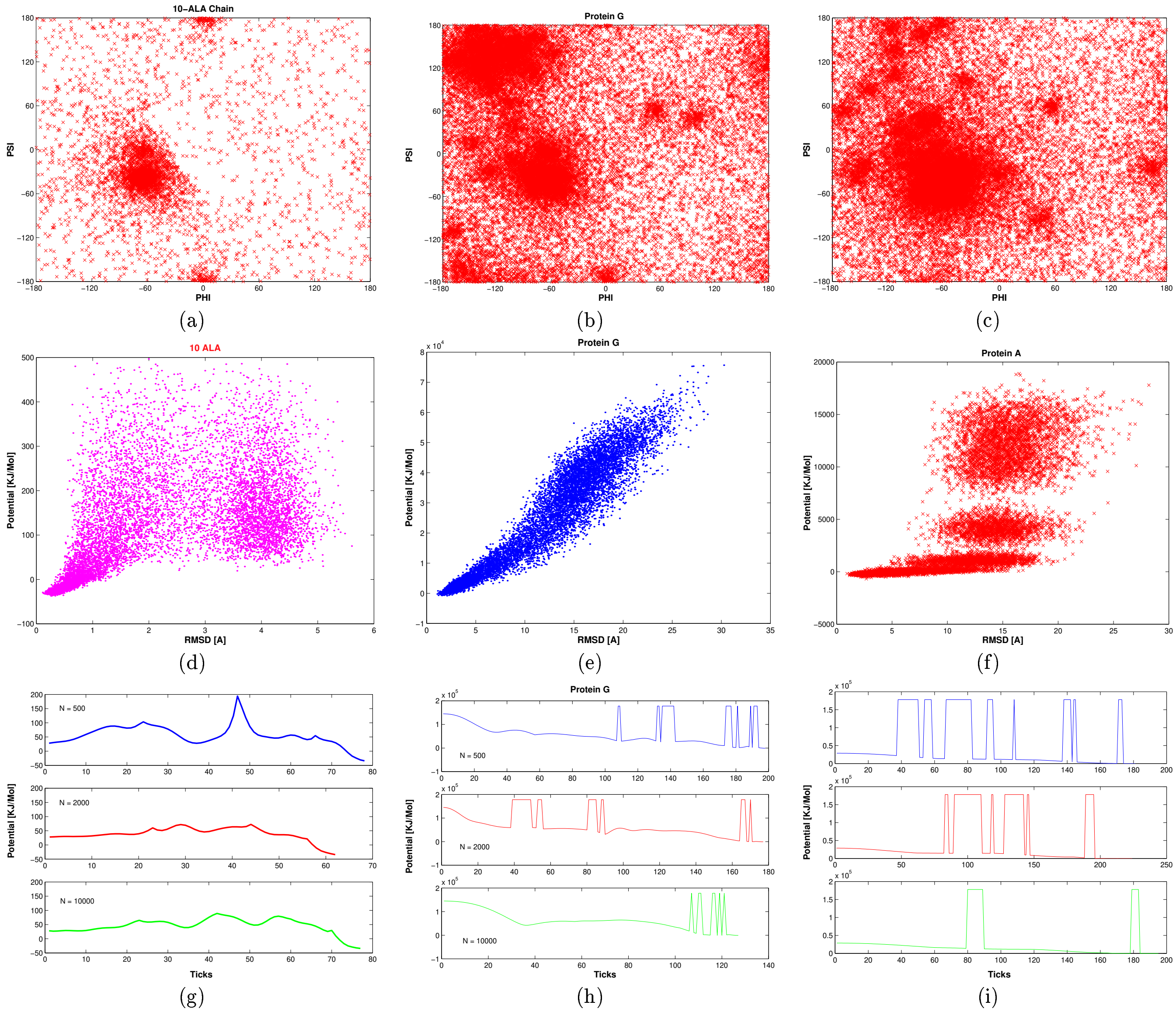


Figure 4: (a-c) Phi/Psi plots, (d-f) Potential/RMSD plots, and (g-i) path profiles for the 10-ALA chain (a,d,g), protein GB1 (b,e,h), and protein A (c,f,i).

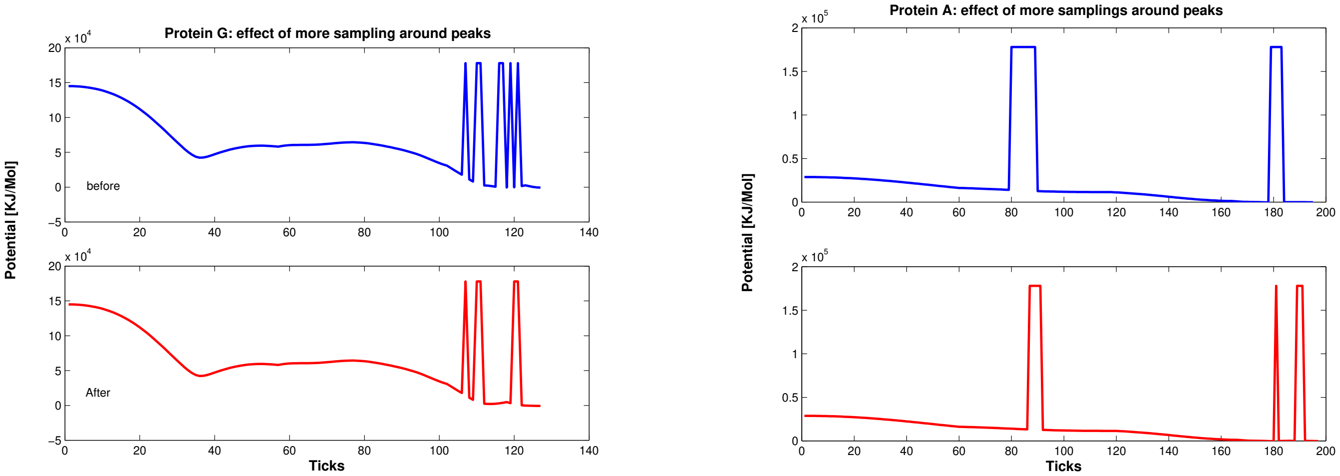
intriguing to note the differences in the plots for protein GB1 and protein A, which suggest different folding behaviors.

5.3.1 Validation of folding pathways

Protein GB1 has 56 residues (112 dofs), and consists of a central alpha helix and two beta-sheets, each composed of two beta strands. Pulse labeling experimental results [23, 29] indicate that the alpha helix and beta strand 4 form first and are protected during hydrogen-deuterium exchanges. This was consistent with the path found by our method. For example, from the snapshots shown in Figure 6, one can clearly see that alpha helix in the middle of the polypeptide forms first.

Protein A has 60 residues (120 dofs), and consists of three alpha helices. The pulse labeling results [29] show that the three alpha helices form at about the same time. Our paths seem to be consistent with these results, as seen in the path snapshots in Figure 6.

In general, these results are very encouraging. In both cases, the formation order of the secondary structures seems to agree with the results of the pulse labeling experiments. 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. 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 contacts are formed between different residues, e.g. as in [14]. The results of this analysis have further validated our paths, and moreover, provide us with a tool for performing more detailed analysis of the folding pathways.



(a) (b)

Figure 5: Potential energy profiles for paths before (top) and after (bottom) resampling for (a) protein GB1 and (b) protein A.

5.4 Analyzing folding pathways

The results for the protein folding examples are very interesting. We provided the goal (native fold) conformations beforehand, and then searched in the roadmap for the minimum weight path connecting the extended amino acid chain to the final three-dimensional structure. Snapshots of folding paths found by our planner for protein GB1 and protein A are shown in Figure 6.

5.4.1 Effect of (re)sampling density

By analyzing the paths found, we may be able to gain some insight into the natural folding process. Towards this end, we analyzed the profiles of the potential energies of the intermediate conformations on the folding paths. This is shown in Figure 4(g-i) for the 10-ALA chain, protein GB1, and protein A, respectively, for three different sized roadmaps. We expect that as the 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. In particular, we can see in the plots that as the number of nodes, N, is increased, the paths seem to improve in quality, and have fewer and smaller peaks in their profiles.

Another interesting point is the similarity among the paths for all roadmap sizes. In particular, they all illustrate that there is a peak (or peaks) near the goal conformation. 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.

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 5, resampling around the peaks 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.

5.4.2 Potential landscapes & secondary structure formation

Do secondary structures form first, as sub-units, before they pack together to reach the tertiary structure? We believe our framework, which enables one to perform targeted sampling of the conformation space with variable resolution, may be a valuable tool for studying such questions.

Consider, for example, the potential vs. RMSD distribution plot for protein A shown in Figure 4(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 6, we see that even at conformations with RMSD 10 Å away from the native fold, the three helices have already formed, and thus the narrow corridor from 10 angstrom to the native fold is only the packing of the secondary structures to form the tertiary structure. This process can be thought of as being governed by hydrophobic forces. Since the hydrophobic effect has a relatively small magnitude, the energy varies little during the packing.

In contrast, the potential vs. RMSD distribution plot for protein G shown in Figure 4(e) has a rather different structure from that of protein A. This is consistent with our paths

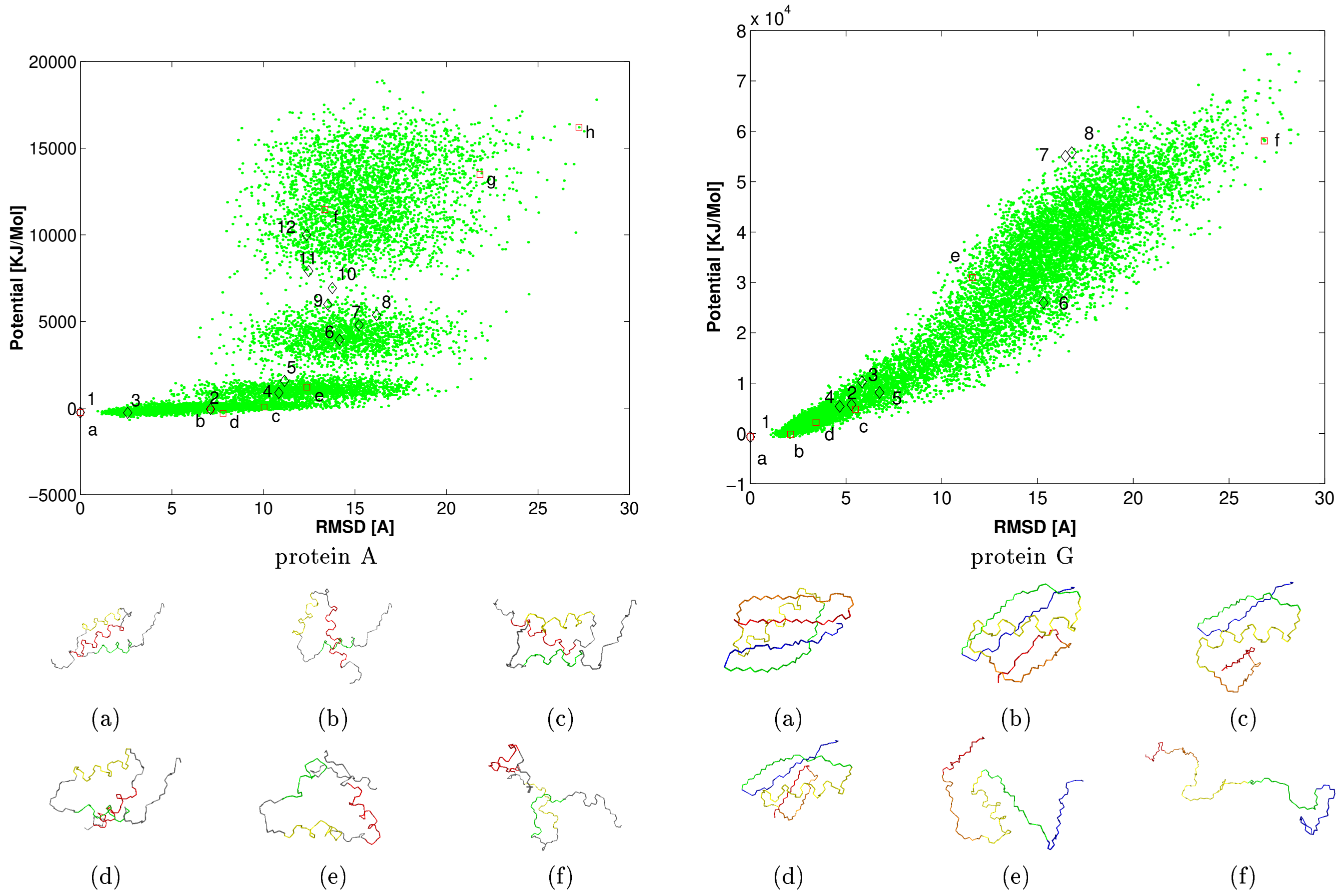


Figure 6: Folding from different starting conformations for protein A (on the left) and protein GB1 (on the right). Folding snapshots from the extended conformation are also shown for both proteins in reverse order, labels running from a to f. The folding process from a random conformation to the goal is shown with numerical labels in the Potential vs RMSD plot (snapshots not shown).

(and the pulse-labeling results [29, 23]), where the secondary structures form at different times (the alpha helix first, and then the beta sheets). Moreover, the beta sheets appear to reach their tertiary positions directly, and thus we do not have the clearly separable packing of secondary structure phase as was seen for protein A. This is reflected in the smooth funnel nature of the potential vs. RMSD plot.

In summary, we believe the shapes of the potential vs. RMSD plots of our roadmaps are related to the shape of the folding funnels which influence overall folding behavior. In particular, we believe the energy distributions suggest the formation time/order of secondary and tertiary structure, and changes in the distributions might indicate different stages of the folding process. In general, it is quite interesting to see the very different shapes observed for the different proteins when the same methodology was applied in both cases. This is an issue which we intend to investigate more thoroughly with a larger set of proteins.

5.4.3 *Folding from different start conformations*

For PRMs, after building the roadmap, searching for the folding pathway from any denatured state is just another query, i.e., there is nothing special about the extended conformation as a starting conformation. As previously noted, queries are generally answered quickly. Figure 6 shows 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 'gullies' as they approach the native fold. They also reflect some common behaviors 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.

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 at `http://www.cs.tamu.edu/faculty/amato/dsmft`.) Nevertheless, we believe that our results establish that this is a promising approach which deserves further investigation. In current work, we are further refining our potential energy approximation, investigating more sophisticated sampling techniques to concentrate more nodes near the peaks 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 folding and protein folding. We would also like to thank Marty Scholtz for suggesting validation using the pulse labeling results, and Michael Levitt and Vijay Pande for useful suggestions.

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

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

[5] O. B. Bayazit, G. Song, and N. M. Amato. Enhancing randomized motion planners: Exploring with haptic hints. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 529–536, 2000.

[6] 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.

[7] 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.

[8] 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.

[9] J. Cantarella and H. Johnston. Nontrivial embeddings of polygonal intervals and unknots in 3-space. *J. Knot Theory Ramifications*, 7:1027–1039, 1998.

[10] D.G. Covell. Folding protein α-carbon chains into compact forms by monte carlo methods. *Proteins: Struct. Funct. Genet.*, 14:409–420, 1992.

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

[12] 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.

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

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

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

[16] L. Han and N. M. Amato. A kinematics-based probabilistic roadmap method for closed chain systems. In *Proc. Int. Workshop on Algorithmic Foundations of Robotics (WAFR)*, 2000.

[17] B. Honig. Protein folding: From the levinthal paradox to structure prediction. *J. Mol. Bio.*, 293:283–293, 1999.

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

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

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

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

[25] 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.

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

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

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

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

[30] 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.

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

[32] 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.

[33] 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.

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

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

[36] 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.

[37] G. Song and N. M. Amato. 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.

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

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

[40] 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.

[41] 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.