

A Motion Planning Approach to Folding: From Paper Craft to Protein Folding*

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

Abstract

In this paper, we present a framework for studying folding problems from a motion planning perspective. Modeling foldable objects as tree-like multi-link objects allows us to apply motion planning techniques to folding problems. An important feature of this approach is that it not only allows us to study foldability questions, such as, can one object be folded (or unfolded) into another object, but also provides us with another tool for investigating the dynamic folding process itself. The framework proposed here has application to traditional motion planning areas such as automation and animation, and presents a novel approach for studying protein folding pathways. Preliminary experimental results with traditional paper crafts (e.g., box folding) and small proteins (approximately 60 residues) are quite encouraging.

1 Introduction

Folding is a very common process in our lives, ranging from the macroscopic level – paper folding or gift wrapping – to the microscopic level – protein folding. In most instances, while one desires a particular final state to be reached (e.g., the package is wrapped, or the protein’s structure is obtained), the knowledge of the dynamic folding process used to reach a particular state is of interest as well. For this reason, we believe motion planning has great potential to help us understand folding. In particular, while motion planning does have the ability to answer questions about the reachability of certain goal states from other states, its primary objective is to in fact determine the motions required to reach the goal.

The problem of folding (and unfolding) is an interesting research topic and has been studied in several application domains. Lu and Akella [23] consider a carton folding problem and its applications in packaging and assembly. In computational geometry, there are various paper folding problems as well [25]. In computational biology, one of the most important outstanding problems is protein folding, i.e., folding a

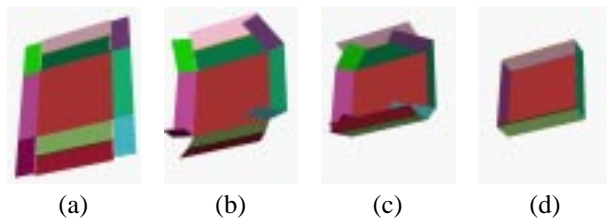


Figure 1: Snapshots of a carton folding.

one-dimensional amino acid chain into a three-dimensional protein structure.

There are large and ongoing research efforts whose goal is to determine the native folds of proteins (see, e.g., [15, 21]). 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. Although there have been some recent experimental advances [10], computational techniques for simulating this process are important because it is difficult to capture the folding process experimentally.

Our approach is based on the successful *probabilistic roadmap* (PRM) motion planning method [17]. We have selected the PRM paradigm due to its proven success in exploring high-dimensional configuration spaces. A major strength of PRMs is that they are quite simple to apply, requiring only the ability to randomly generate points in C-space, and then test them for feasibility. The protein folding problem has a complication in that the way in which the protein folds depends on factors other than the purely geometrical constraints which govern the polygonal problems. Nevertheless, we show that these additional factors can be dealt with in a reasonable fashion within the PRM framework.

Our preliminary experimental results with traditional paper crafts and small proteins of approximately 60 residues, or 120 degrees of freedom, are quite promising. See Figures 1 and 2 for some path snapshots. Further results on protein folding can be found in [27].

¹This 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.

2 Related work

Paper Folding. Many problems related to the folding and unfolding of polyhedral objects have recently attracted the attention of the computational geometry community [25]. For example, [8] shows that every polyhedron can be ‘wrapped’ by folding a strip of paper around it, which addresses a question arising in three-dimensional origami, e.g., [1]. In most cases, origami problems cannot be modeled as trees since the incident faces surrounding a given face form a cycle in the linkage structure. Such cycles, often called closed chains, impose additional constraints on the motion planning problem (see, e.g., [13, 19]). In this paper we are interested in problems with tree-like linkage structures. There are still many interesting problems involving folding of tree-like linkages. For example, not every tree-like linkage in the plane can be ‘straightened’ (called ‘locking’), that is, there are some pairs of configurations of the linkage which cannot be connected if the links are not allowed to cross [4]. In three dimensions, there exist open (and closed) chains that can lock [4, 5], while, in dimensions higher than three, neither open nor closed chains can lock [6].

Protein Folding. The protein folding problem is to predict a protein’s three-dimensional conformation based solely on its amino acid sequence. Many different approaches for predicting protein structure have been explored. In folding simulations, several computational approaches have been applied to this exponential-time problem, including energy minimization, molecular dynamics simulation, Monte Carlo methods, and genetic algorithms (see [15] and references therein). Among these, molecular dynamics is most closely related to our approach. Much work had been carried out in this area [7, 9, 12, 20], 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. Molecular dynamics simulations help us understand how proteins fold in nature, and provide a means to study the underlying folding mechanism, to investigate folding pathways, and can provide intermediate folding states.

Most of the proposed techniques have tremendous computational requirements because they attempt to simulate complex kinetics and thermodynamics. In this paper, we present an alternative approach that finds approximations to the folding pathways while avoiding detailed simulations. Our motion planning approach is based on the successful *probabilistic roadmap* (PRM) method [17] which has been used to study the related problem of ligand binding [3, 16, 26], which is of interest in drug design. The results were quite promising. Advantages of the PRM approach are that it efficiently covers a large portion of the planning space, in this case, the conformation space, and that it also provides an effective way to incorporate and study various initial conformations.

3 Preliminaries

C-spaces of folding objects. Both the paper polygon and the amino acid sequence are modeled as multi-link tree-like articulated ‘robots’, where fold positions (polygon edges or atomic bonds) correspond to joints and areas that cannot fold (polygon faces or atoms) correspond to links. The fold positions of the paper polygon are modeled as revolute joints. For the amino acid sequence of the protein, we consider all atomic bond lengths and bond angles to be constants, and consider only torsional angles (phi and psi angles), which we also model as two revolute joints (2 dof). Thus, in both cases, our models will consist of $n + 1$ links and n 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 . Therefore, the configuration space of interest for our multi-link objects can be expressed as:

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

Note that \mathcal{C} simply denotes the set of all possible configurations, but says nothing about their feasibility. The validity of a point in \mathcal{C} will be determined by collision detection for the polygon problems and by potential energy computations for the proteins.

Potential Function. The way in which the protein folds depends on the potential energy of the configurations. We start with:

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

which is similar to the potential used in [20]. 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 [20].

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 because it 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 is retained to capture the main features of the interaction.

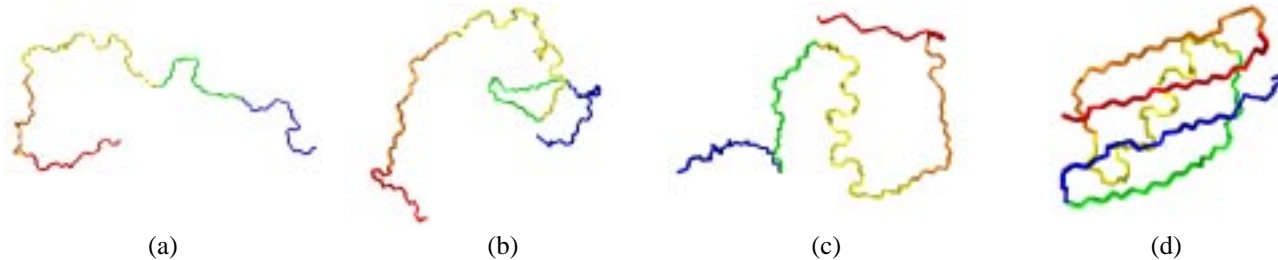


Figure 2: Snapshots of protein GB1 folding.

4 PRMs for Folding Problems

As mentioned in Section 1, our approach to the folding problem is based on the PRM approach to motion planning [17].

The folding problems, especially protein folding, have a few notable differences from usual PRM applications. First, as our problems are not posed in an environment containing external obstacles, the only collision constraint we impose is that our configurations be self-collision free, and, for the protein folding problem, our preference for low energy conformations leads to an additional constraint on the feasible conformations. Second, in PRM applications, it is usually considered sufficient to find *any* feasible path connecting the start and goal. For our folding problems, however, we are interested not only in whether there exists a path, but we are also interested in the *quality* of the path. For example, for the paper folding problems, one is interested in a path which makes a minimal number of folds, and for the protein folding we are interested in low energy paths.

I. Node generation. As described in Section 3, 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. Once all the joint angles are set, the object’s three-dimensional structure is fully determined.

For the paper folding, the configuration of each link is then calculated and self-collision among the links is checked. The node is discarded if any collision occurs.

For the protein molecular model, 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 defined in Section 3. The node is accepted and added to the roadmap based on its potential energy E with the following probability:

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

This acceptance criterion was also used when building PRM roadmaps for ligand binding in [26]. This filtering helps us to generate more nodes in low energy regions, which is desirable since we are interested in finding the pathways that are most energetically favorable (low energy). In our case, we set $E_{\min} = 50000$ KJouls/mol and $E_{\max} = 89000$

KJouls/mol, which favors conformations with well separated side chain spheres.

II. 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 (using Euclidean distance in C-space), for some small constant k , and then try to connect it to them using some simple local planner. For both the paper folding and protein folding models, each connection attempt performs feasibility checks for N intermediate configurations between the two corresponding nodes as determined by the chosen local planner. If there are still multiple connected components in the roadmap after this stage other techniques will be applied to try to connect different connected components (see [2] for details).

When two nodes are connected, the corresponding edge is added to the roadmap. We associate a weight with each edge. For the paper folding, the weight is simply N , the number of intermediate configurations on the edge. For the protein folding, the weight is:

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

where the probability P_i 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. By assigning the weights in this manner, we can find the shortest or most energetically feasible path when performing subsequent queries. A similar weight function, with different probabilities, was used in [26].

III. ‘Querying’ the roadmap. The resulting roadmap can be used to find a feasible path between given start and goal configurations. For our folding problems, we first 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. For the protein folding, if the potential of some intermediate node is too large (as compared to some predetermined maximum), a failure is reported, otherwise the path is returned.

5 Validating Folding Pathways

For the protein folding pathways found by our PRM framework to be useful, we must find some way to validate them 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 (i) the intermediate or transition states on the pathway, and the order in which they are obtained, or (ii) the formation order of secondary structures.

Folding intermediates have been an active research area over the last few years. It is thought that some, but not all, proteins go through intermediate states to reach the native conformation, see, e.g., [24]. Therefore, one possibility is to compare our folding pathways with experimental results known about folding intermediates.

The formation order of secondary structures 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 with results for some small proteins that have been determined by pulse labeling and native state out-exchange experiments [22].

6 Results and Discussion

We now describe results on paper folding and protein folding problems obtained using our PRM-based approach. For the paper folding problems we used the obstacle-based PRM called OBPRM [2], which generates nodes near constraint surfaces (C-obstacle surfaces). For the protein folding, the results presented follow the basic PRM approach [17] of uniform sampling in C-space. We used the RAPID [11] package for 3D collision detection. The experiments were performed on an SGI Octane R10000. In this paper we can only show path snapshots; movies can be found at <http://www.cs.tamu.edu/faculty/amato/dsmft>.

6.1 Models studied

We study two paper folding models: a *box* and a *periscope*. The periscope has 11 degrees of freedom (11 joints) and the box has 12. However, for the box, the number of dof can be reduced to five using symmetry arguments. Both foldings are non-trivial, and in fact, correspond to what are known as ‘narrow passage’ problems [14].

We present results for two small 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/>.

Paper Folding Roadmap Construction Statistics					
Model	dof	Gen	Con	#CC	#Nodes
Box	12(5)	38.7	201	1	1035
Periscope	11	13	177	1	883

Table 1: Roadmap construction statistics for the Box and Periscope models. The Box has 12 links, but its dof becomes 5 after symmetry is exploited. ‘Gen’ and ‘Con’ represent node generation and connection times in seconds, resp. #Nodes and #CC are the number of nodes and connected components, resp., in the resulting roadmap.

6.2 Paper folding results

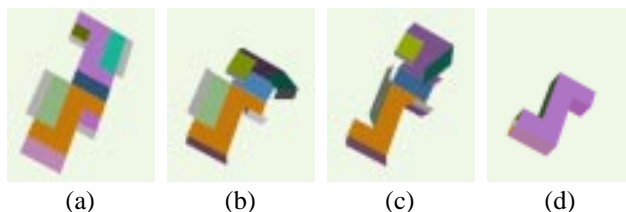


Figure 3: Snapshots of the periscope folding.

Some statistics regarding the roadmaps constructed for the paper folding problems are shown in Table 1. As can be seen, in both cases the problems were solved rather quickly with relatively small roadmaps. These results are really quite remarkable as the problems are actually considered to be quite challenging motion planning problems. Nevertheless, we see that just a few minutes were needed to construct roadmaps containing solution paths. We believe our success with these problems can be attributed to the tendency of the OBPRM roadmaps to contain nodes near the constraint surfaces (i.e., near self-collision configurations) which include configurations necessary for successful paths. For example, configurations in which the flaps of the box fold over other flaps. Snapshots of the folding paths found are shown in Figures 1 and 3 for the box and the periscope, respectively.

6.3 Protein folding results

The results for the protein folding examples are also very interesting. Some statistics regarding the roadmaps constructed for the protein folding problems are shown in Table 2. We provided the goal 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 2 and Figure 4, respectively.

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 [18, 22] indicate that the alpha helix and beta strand 4 form first and are protected during

Protein Folding Roadmap Construction Statistics								
Model	dof	Gen	Con	#N sam	#N ret	#N BigCC	#edges	#N path
Protein GB1	112	130	1500	5000	594	559	898	1
		500	5300	20000	2508	2381	3890	2
		2600	42300	100000	12392	11865	20433	3
Protein A	120	400	1300	5000	555	508	767	5
		1600	5800	20000	2308	2140	3352	4
		9500	50100	100000	11715	11057	17719	6

Table 2: 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 and ‘#N ret’ is the number of nodes retained after rejecting nodes with high potentials. ‘#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.

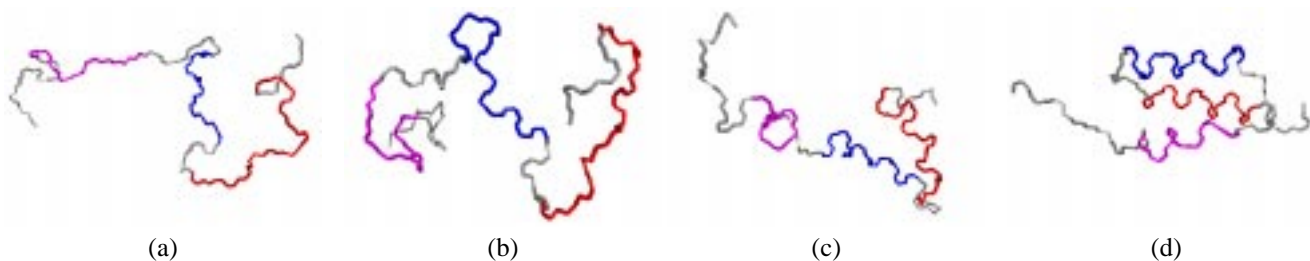


Figure 4: Snapshots of protein A folding.

hydrogen-deuterium exchanges. This was consistent with the path found by our method. For example, from the snapshots shown in Figure 2, one can see that the alpha helix in the middle forms first.

Protein A has 60 residues (120 dofs), and consists of three alpha helices. The pulse labeling results [22] show that the three alpha helices form at about the same time. As seen in the path snapshots in Figure 4, our paths seem to be consistent with these results.

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.

Analyzing folding pathways. 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 for proteins GB1 and A in Figure 5(a) and 5(b), respectively. 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 are contributed by the van der Waals interaction, which is consistent with the tight packing of atoms in the native fold. The similarity among these paths 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.

7 Conclusion and Future Work

In this paper, we present a framework for studying folding problems from a motion planning perspective. Our approach, which is based on the PRM motion planning method, was seen to produce interesting results for representative problems in paper folding and protein folding. One of the most important benefits of this approach to folding problems is that it enables one to study the dynamic folding process itself. We believe that our results establish that this is a promising approach which deserves further investigation.

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

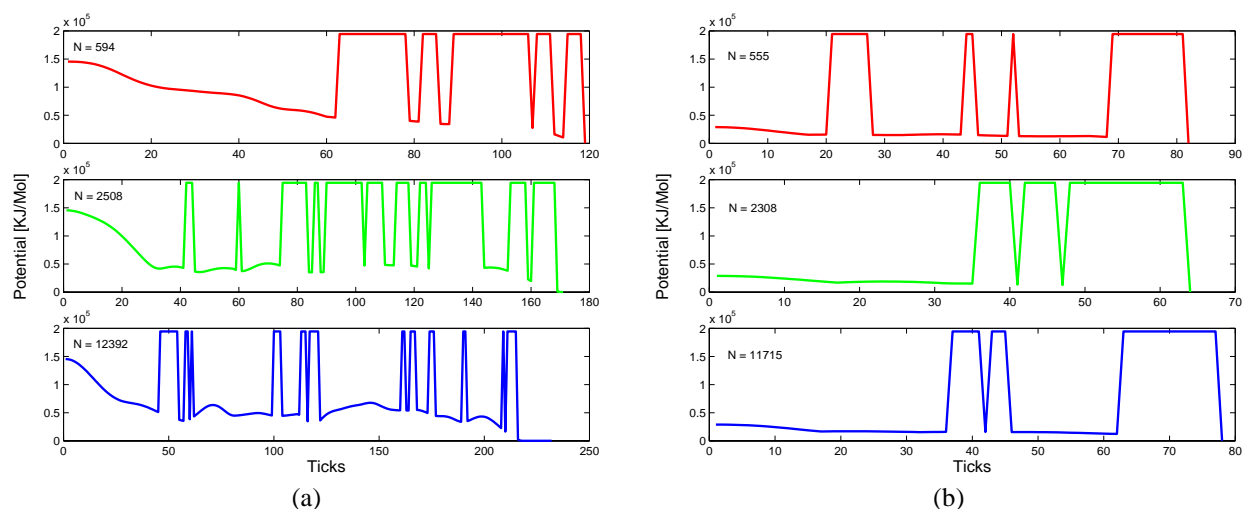


Figure 5: Potential along the folding path shown for each intermediate configuration on the path ('tick') for different sized roadmaps. (a) Protein GB1, roadmaps with $N = 594, 2508, 12392$ nodes (top to bottom), (b) Protein A, roadmaps with $N = 555, 2308, 11715$ nodes (top to bottom).

suggesting validation using the pulse labeling results, and Michael Levitt and Vijay Pande for useful suggestions.

References

- [1] Jin Akiyama. Why Taro can do geometry. In *Proc. 9th Canad. Conf. Comput. Geom.*, page 112, 1997.
- [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] 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. Autom. (ICRA)*, 2001. To appear. This work will be presented as a poster at RECOMB'01.
- [4] 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.
- [5] J. Cantarella and H. Johnston. Nontrivial embeddings of polygonal intervals and unknots in 3-space. *J. Knot Theory Ramifications*, 7:1027–1039, 1998.
- [6] R. Cocan and J. O'Rourke. Polygonal chains cannot lock in 4D. In *Proc. 11th Canad. Conf. Comput. Geom.*, pages 5–8, 1999.
- [7] V. Daggett and M. Levitt. Realistic simulation of naive-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.*, 22:353–380, 1993.
- [8] E. D. Demaine, M. L. Demaine, and J. S. B. Mitchell. Folding flat silhouettes and wrapping polyhedral packages: New results in computational origami. In *Proc. 15th Annu. ACM Sympos. Comput. Geom.*, pages 105–114, June 1999.
- [9] 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.
- [10] 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.
- [11] S. Gottschalk, M.C. Lin, and D. Manocha. Obb-tree: A hierarchical structure for rapid interference detection. Technical Report TR96-013, University of N. Carolina, Chapel Hill, CA, 1996.
- [12] J.M. Haile. *Molecular Dynamics Simulation: elementary methods*. Wiley, New York, 1992.
- [13] 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.
- [14] 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.
- [15] G. N. Reeke Jr. Protein folding: Computational approaches to an exponential-time problem. *Ann. Rev. Comput. Sci.*, 3:59–84, 1988.
- [16] L. Kavraki. Geometry and the discovery of new ligands. In *Proc. Int. Workshop on Algorithmic Foundations of Robotics (WAFR)*, pages 435–448, 1996.
- [17] 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.
- [18] J. Kuszewski, G.M. Clore, and A.M. Gronenborn. Fast folding of a prototypic polypeptide: The immunoglobulin binding domain of streptococcal protein g. *Protein Science*, 3:1945–1952, 1994.
- [19] 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.
- [20] M. Levitt. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Bio.*, 170:723–764, 1983.
- [21] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai. Protein folding: the endgame. *Annu. Rev. Biochem.*, 66:549–579, 1997.
- [22] R. Li and C. Woodward. The hydrogen exchange core and protein folding. *Protein Sci.*, 8:1571–1591, 1999.
- [23] 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.
- [24] C.R. Matthews. Pathways of protein folding. *Annu. Rev. Biochem.*, 62:653–683, 1993.
- [25] 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.
- [26] 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.
- [27] G. Song and N. M. Amato. Using motion planning to study protein folding pathways. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 278–287, 2001.