A PATH PLANNING-BASED STUDY OF PROTEIN FOLDING WITH A CASE STUDY OF HAIRPIN FORMATION IN PROTEIN G AND L

GUANG SONG,[†] SHAWNA THOMAS,[†] KEN A. DILL,[‡] J. MARTIN SCHOLTZ,[§]
NANCY M. AMATO[†]

We investigate a novel approach for studying protein folding that has evolved from robotics motion planning techniques called *probabilistic roadmap* methods (PRMs). Our focus is to study issues related to the folding process, such as the formation of secondary and tertiary structure, *assuming* we know the native fold. A feature of our PRM-based framework is that the large sets of folding pathways in the roadmaps it produces, in a few hours on a desktop PC, provide global information about the protein's energy landscape. This is an advantage over other simulation methods such as molecular dynamics or Monte Carlo methods which require more computation and produce only a single trajectory in each run. In our initial studies, we obtained encouraging results for several small proteins. In this paper, we investigate more sophisticated techniques for analyzing the folding pathways in our roadmaps. In addition to more formally revalidating our previous results, we present a case study showing our technique captures known folding differences between the structurally similar proteins G and L.

1 Introduction

There are large and ongoing research efforts whose goal is to determine the native structure of a protein from its amino acid sequence.[1,2] A protein's 3D structure is important because it affects the protein's function. In this work, we assume the native structure is known, and our focus is on the study of protein folding mechanisms. 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. Such questions have taken on increased practical significance with the realization that mis-folded or only partially folded proteins are associated with many devastating diseases.[3] Moreover, increased knowledge of folding mechanisms may provide insight for protein structure prediction. Despite intensive efforts by experimentalists and theorists, there are major gaps in our understanding of the behavior and mechanism of the folding process.

[*] To appear in Proc. of the Pacific Symposium on Biocomputing (PSB 2003).

[†] To appear in Proc. of the Pacific Symposium on Biocomputing (PSB 2003). {gsong,sthomas,amato}@cs.tamu.edu Department of Computer Science, Texas A&M University, College Station, TX 77843-3112, USA.

[‡] dill@maxwell.ucsf.edu. Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-1204.

[§] jm-scholtz@tamu.edu. Department of Medical Biochemistry & Genetics, Texas A&M University, College Station, TX 77843-1114, USA.

Table 1: A comparison of protein folding models.

Comparison of Models for Protein Folding					
Approach	**Folding Landscape**	**#Paths Produced**	**Path Quality**	**Compute Time**	**Need Native**
Molecular Dynamics	No	1	Good	Long	No
Monte Carlo	No	1	Good	Long	No
Statistical Model	Yes	0	N/A	Fast	Yes
PRM-Based	Yes	Many	Approx	Fast	Yes
Lattice Model	Not used on real proteins				

In previous work,[4] we proposed a technique for computing protein folding pathways that is based on the successful *probabilistic roadmap* (PRM)[5] method for robotics motion planning. We were inspired to apply this technique to protein folding based on our success in applying it to folding problems such as carton folding and paper crafts.[6] We obtained promising results for small proteins (∼60 amino acids) and validated our pathways by comparing the secondary structure formation order with known experimental results.[7]

A major feature of our PRM-based framework is that in a few hours on a desktop PC it produces roadmaps containing large sets of unrelated folding pathways that provide global information about the protein's energy landscape. In this paper, we investigate more sophisticated techniques for analyzing the pathways in our roadmaps. In addition to more formally revalidating our previous results, we present a case study showing our technique captures known folding differences between the structurally similar proteins G and L.

1.1 Comparison to Related Work

Table 1 provides a summary comparison of various models for protein folding. Both Monte Carlo simulation and molecular dynamics provide a single, usually high quality, folding trajectory. Each run is computationally intensive because they attempt to simulate complex kinetics and thermodynamics. Statistical mechanical models, while computationally efficient, assume extremely simplified molecular interactions and are limited to studying global averages of folding kinetics. They also cannot detect multiple kinetics behavior such as the two-state and three-state kinetics exhibited by hen egg-white Lysozyme.[8,9] Lattice models[10] have been well studied and possess great theoretical value but cannot be applied to real proteins.

Our PRM approach, by constructing a roadmap that approximates the folding landscape, computes multiple folding pathways in a single run and provides a natural way to study protein folding kinetics at the pathway level. What we sacrifice is path quality, which can be improved through bigger roadmaps, oversampling, or other techniques.

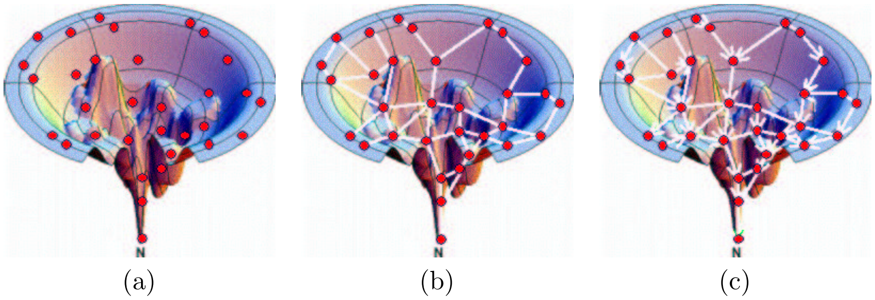

Figure 1: A PRM roadmap for protein folding shown imposed on a visualization of the potential energy landscape: (a) after node generation (note sampling is denser around **N**, the known native structure), (b) after the connection phase, and (c) using it to extract folding paths to the known native structure.

2 A Probabilistic Roadmap Method for Protein Folding

Our approach to protein folding is based on the *probabilistic roadmap* (PRM) approach for motion planning.[5] A detailed description of how the PRM framework can be applied to protein folding is presented in our previous work.[4] The basic idea is illustrated in Figure 1. We first sample some points in the protein's conformation space (Figure 1(a)); generally, our sampling is biased to increase density near the known native state. Then, these points are connected to form a graph, or roadmap (Figure 1(b)). Weights are assigned to directed edges to reflect the energetic feasibility of transition between the conformations corresponding to the two end points. Finally, folding pathways are extracted from the roadmap using standard graph search techniques (Figure 1(c)).

2.1 Modeling Proteins (C-Space)

The amino acid sequence is modeled as a tree-like linkage. Using a standard modeling assumption for proteins,[11] the only degrees of freedom (dof) in our model of the protein are the backbone's phi and psi torsional angles, which we model as revolute (rotational) joints taking values in $[0, 2\pi)$. Moreover, side chains are modeled as spheres and have zero dof.

Since we are not concerned with the absolute position and orientation of the protein, a *conformation* of an $n+1$ amino acid protein can be specified by a vector of $2n$ phi and psi angles, each in the range $[0, 2\pi)$, with the angle 2π equated to 0, which is naturally associated with a unit circle in the plane, denoted by S^1. That is, the conformation space (C-space) of interest for a protein with $n+1$ amino acids can be expressed as:

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

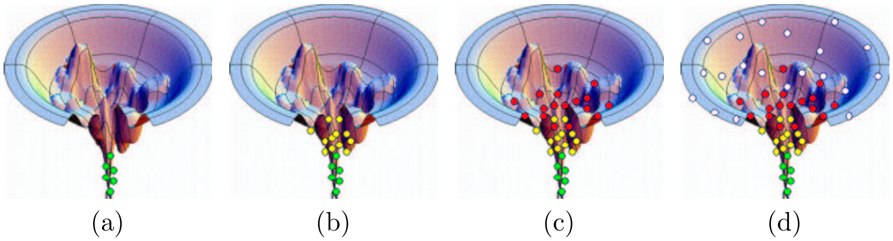

Figure 2: An illustration of our iterative perturbation sampling strategy shown imposed on a visualization of the potential energy landscape.

Note that $\mathcal{C}$ simply denotes the set of all possible conformations. The feasibility of a point in $\mathcal{C}$ will be determined by potential energy computations.

2.2 Node Generation

Recall that we begin with the known native structure and our goal is to map the protein-folding landscape leading to the native fold. The objective of the node generation phase is to generate a representative sample of conformations of the protein. Due to the high dimensionality of the conformation space, simple uniform sampling would take too long to provide sufficiently dense coverage of the region surrounding the native structure.

The results presented in this paper use a biased sampling strategy that focuses sampling around the native state by iteratively applying small perturbations to existing conformations.[4] The process is illustrated in Figure 2. A node q is accepted and added to the roadmap based on its potential energy $E(q)$ with the following probability:

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

We set $E_{\text{min}} = 50000$ kJ/mol and $E_{\text{max}} = 70000$ kJ/mol which favors configurations with well separated side chain spheres. This acceptance test, which retains more nodes in low energy regions, was also used in PRM-based methods for ligand binding[12,13] and in our previous work on protein folding.[14]

2.3 Connecting the Roadmap

Connection is the second phase of roadmap construction. The objective is to obtain a roadmap encoding representative, low energy paths. For each roadmap node, we first find its k nearest neighbors, for some small constant k, and then try to connect it to them using local planning method. This yields a connectivity roadmap that can be viewed as a net laid down on the energy landscape (see Figure 1(b)).

When two nodes q_1 and q_2 are connected, the directed edge (q_1, q_2) is added to the roadmap. Each edge (q_1, q_2) is assigned a weight that depends on the sequence of conformations $\{q_1 = c_0, c_1, c_2, \ldots, c_{n-1}, c_n = q_2\}$ on the straight line in $\mathcal{C}$ 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 $\Delta E_i = E(c_{i+1}) - E(c_i)$.

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

This keeps the detailed balance between two adjacent states, and enables the weight of an edge to be computed by summing the logarithms of the probabilities for consecutive pairs of conformations in the sequence.

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

In this way, we encode the energetic feasibility of transiting from one conformation to another in the edge connecting them.

2.4 Extracting Folding Pathways

The roadmap is a map of the protein-folding landscape of the protein. One way to study this landscape is to inspect and analyze the pathways it contains.

An important feature of our approach is that the roadmap contains *many* folding pathways, which together represent the folding landscape. We can extract many such paths by computing the single-source shortest-path (SSSP) tree from the native structure (see Figure 1(c)).

3 Potential Energy Calculations

The way in which a protein folds depends critically on the potential energy. Our PRM framework incorporates this bias by accepting conformations based on their potential energy (Section 2.2) and by weighting roadmap edges according to their energetic feasibility (Section 2.3).

While our framework is flexible enough to use any method for computing potential energies, our current work uses a very simplistic potential.[4] Briefly, we use a step function approximation of the van der Waals potential component. Our approximation considers only the contribution from the side chains. Additionally, in our model of each amino acid, we treat the side chain as a single large 'atom' R located at the C_β atom. 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. If all the distances between all R 'atoms' are larger than 2.4 Å, then we proceed to calculate the potential as follows:

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + E_{hp}, \qquad (4)$$

The first term represents constraints which favor the known secondary structure through main-chain hydrogen bonds and disulphide bonds and the second term is the hydrophobic effect.

Finally, we note that in our case, the minimum potential is not necessarily achieved by the native structure, and thus our energy model does not yield true funnel landscapes as are shown in the figures.

4 Timed Contact Analysis

Contact analysis provides us with a formal method of validation and allows for detailed analysis of the folding pathways. We first identify the *native contacts* by finding all pairs of C_α atoms in the native state that are at most 7 Å apart. If desired, attention can be restricted to *hydrophobic contacts* between hydrophobic residues. To analyze a particular pathway, we examine each conformation on the path and determine the time step on the path at which each native contact appears. Although these time steps cannot be associated with any real time, they do give a temporal ordering and produce a *timed contact map* for the given pathway, see Figures 3 and 4.

The timed contact map provides a formal basis for determining secondary structure formation order along a pathway. Here, structure formation order is based on the formation order of the native contacts.[15] We have looked at several metrics to determine when a secondary structure appears: average appearance time of native contacts within the structure, average appearance of the first $x\%$ of the contacts, average appearance ignoring outliers, etc. We can also focus our analysis on smaller pieces of secondary structure such as β-turns (instead of the entire β-sheet). This is especially helpful when looking for fine details in a folding pathway.

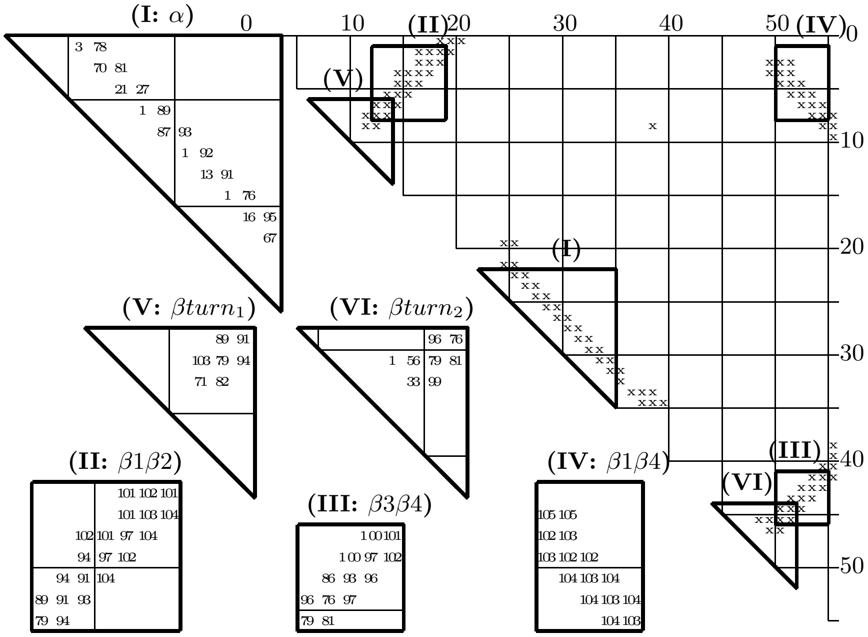

Figure 3: Timed Contact Map for Protein G. The full contact matrix (right) and blow-ups (left) showing the time steps when contacts appear on a path. The blow-ups: I - alpha helix contacts, II - beta 1-2 contacts, III - beta 3-4 contacts, IV - beta 1-4 contacts, V - turn 1 (beta 1-2) contacts, and VI - turn 2 (beta 3-4) contacts.

Because the roadmap contains multiple pathways, we can estimate the probability of a particular secondary structure formation order occurring. If the roadmap maps the potential energy landscape well, then the percentage of pathways in the roadmap that contain a particular formation order reflects the probability of that order occurring.

5 Experimental Validation and Discussion

In this section, we present results obtained using our PRM-based approach. For each protein studied, we construct a roadmap, extract the folding pathways as described in Section 2.4, and analyze the pathways as described in Section 4.

We study several small proteins[a] in detail, see Table 2. The structures for all the proteins were obtained from the Protein Data Bank.[17] Protein A is an

[a] Abbreviations of proteins: G, B1 immunoglobulin-binding domain of streptococcal protein G; L, a 62-residue variant[16] of B1 immunoglobulin-binding domain of peptostreptococcal protein L; A, B domain of staphylococcal protein A; CTXIII, Cardiotoxin analogue III.

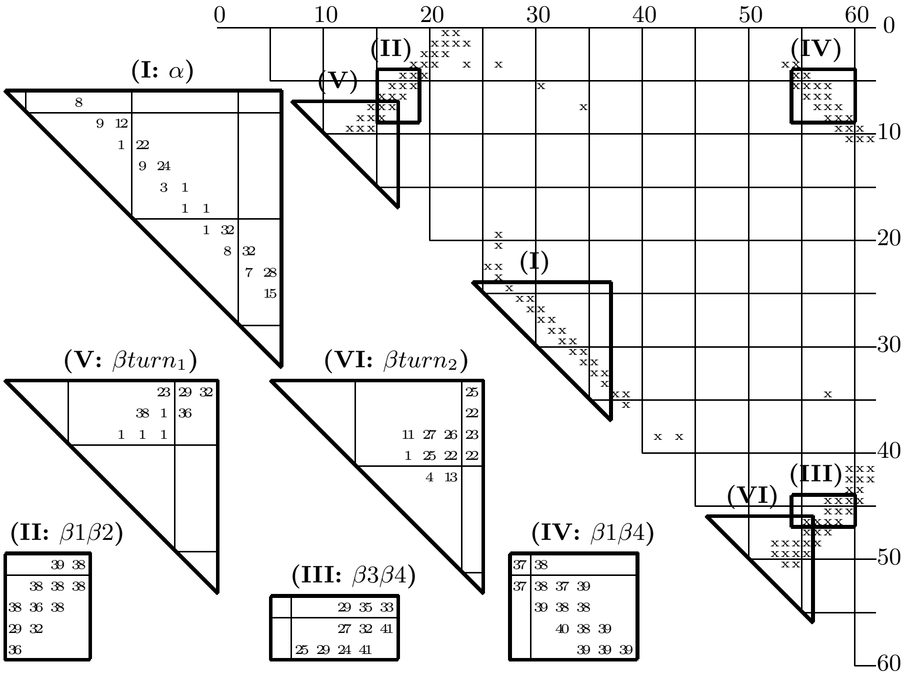

Figure 4: Timed Contact Map for Protein L. The full contact matrix (right) and blow-ups (left) showing the time steps when contacts appear on a path. The blow-ups: I - alpha helix contacts, II - beta 1-2 contacts, III - beta 3-4 contacts, IV - beta 1-4 contacts, V - turn 1 (beta 1-2) contacts, and VI - turn 2 (beta 3-4) contacts.

all alpha protein composed of three alpha helices. Protein G and protein L are mixed proteins which are both composed of one α-helix and a 4-stranded β-sheet. CTXIII is an all beta protein composed of a 2-stranded β-sheet and a 3-stranded β-sheet.

As discussed in Section 1.1, our PRM-based method sacrifices accuracy in favor of rapid coverage. As can be seen from the running time and roadmap statistics shown in Table 2, our roadmaps containing thousands of folding paths are computed in just a few hours on a desktop PC. In contrast, traditional methods such as molecular dynamics, compute a single trajectory, have tremendous computational requirements, and are subject to local minima.

Contact analysis was performed on the pathways for proteins A, G, L and CTXIII. Timed contact maps for proteins L and G are shown in Figure 3 and Figure 4. The dominant formation order found for each protein is shown in Table 3. It is clearly seen that our results are in good agreement with the hydrogen-exchange experimental results described by Li and Woodward.[7]

Table 2: Proteins studied. Shown are number of residues (#res), number of α helices and β strands ($\alpha + \beta$), roadmap size (#nodes), and construction time.

Proteins and Roadmap Statistics						
name	**pdb**	**brief description**	**#res**	**SS**	**#nodes**	**time (hr)**
G	1gb1	Protein G, B1 domain	56	$1\alpha + 4\beta$	16407	6.985
A	1bdd	Protein A, B domain	60	3α	21917	11.325
CTXIII	2crt	Cardiotoxin III	60	5β	14532	6.386
L	2ptl	Protein L, B1 domain	62	$1\alpha + 4\beta$	17407	9.152

Table 3: The secondary structure formation order on dominant pathways in our roadmaps and some validations. The brackets indicate there was no clear order. The last column compares our results with those from hydrogen-exchange experiments.

Secondary Structure Formation Order and Validation				
pdb	Out-Exchange[7]	Pulse-Labeling[7]	Our SS Formation Order	Comp.
1gb1	[α,β1,β3,β4], β2	[α,β4], [β1,β2,β3]	α, β3-β4, β1-β2, β1-β4	Agreed
1bdd	[α2,α3], α1	[α1,α2,α3]	[α1,α2,α3], α2-α3, α1-α3	Agreed
2crt	[β3,β4,β5], [β1,β2]	β5, β3, β4, [β1,β2]	β3-β4, [β1-β2,β3-β5]	Similar
2ptl	[α,β1,β2,β4], β3	[α,β1], [β2,β3,β4]	α, β1-β2, β3-β4, β1-β4	Agreed

All proteins seem to form local contacts first, and then those with increasing sequence contact order, like a zipper process.[18,15] Finally, we note that our results for CTXIII may be affected by the four disulphide bonds which we model as hydrogen bonds.

5.1 Protein G and Protein L: A Detailed Study

Proteins G and L present a good test case for our technique because they are known to fold differently although they are structurally similar. In particular, although they have only 15% sequence identity,[19] they are both composed of an α-helix and a 4-stranded β-sheet. β strands 1 and 2 form the N-terminal hairpin (hairpin 1) and β strands 3 and 4 form the C-terminal hairpin (hairpin 2). Experimental results show that β-hairpin 1 forms first in protein L, and β-hairpin 2 forms first in protein G.

In native state out-exchange experiments for protein G and L, numerous NHs out-exchange very slowly, which makes it difficult to unambiguously identify the slowest out-exchange residues. It is found that the slow exchanging NHs are in β_1, β_3, β_4 and the helix for G, and the α helix, β_1, β_2, and β_4 for L. On the other hand, pulse-labeling experiments identify that the first NHs to gain protection during folding are in α and β_4 for G and α and β_1 for L. (See Li and Woodward [7] and references therein.) In summary, out-exchange and pulse-labeling experiments strongly suggest that the α and β_4 form first for G and that the α and β_1 form first for L. Furthermore, this is consistent with Φ-value analysis on G [19] and L [20] which indicates that, in the folding transition state, β-hairpin 2 is more formed than the rest of the structure for

Table 4: Comparison of analysis techniques for proteins G and L using roadmaps computed with energy thresholds $E_{\min} = 50,000$ kJ/mol and $E_{\max} = 70,000$ kJ/mol. For each combination of contact type (all or hydrophobic) and number of contacts (first x% to form), we show the percentage of pathways with a particular secondary structure formation order. Recall that β-hairpin 2 (β3-β4) forms first in protein G and β-hairpin 1 (β1-β2) forms first in protein L.

Comparison of Analysis Techniques – Helix and Hairpins							
			analyze first x% contacts				
Name	Contacts	SS Formation Order	20	40	60	80	100
Protein G	all	α, β3-β4, β1-β2, β1-β4	76	66	77	55	58
		α, β1-β2, β3-β4, β1-β4	23	34	23	45	42
	hydrophobic	α, β3-β4, β1-β2, β1-β4	85	78	77	62	67
		α, β3-β4, β1-β4, β1-β2	11	11	9	8	8
		α, β1-β2, β3-β4, β1-β4	4	10	14	29	24
Protein L	all	α, β1-β2, β3-β4, β1-β4	67	76	78	78	92
		α, β1-β2, β1-β4, β3-β4	15	4	4	4	4
		α, β3-β4, β1-β2, β1-β4	19	20	18	18	4
	hydrophobic	α, β1-β2, β3-β4, β1-β4	54	65	74	73	86
		α, β1-β2, β1-β4, β3-β4	9	3	3	2	2
		α, β3-β4, β1-β2, β1-β4	36	32	23	26	13

G and β-hairpin 1 is similarly more formed for L.

For both protein G and L, we use the same definition of the beta strands as is contained in the protein Data Bank. These definitions include all the observed residues that are found in the slowest exchange core in the native state out-exchange experiments and that are among the first gaining protection in the pulse labeling experiments, see Figures 3 and 4. This enables us to have a fair comparison of our results with those from these experiments.

Table 4 shows our results. For each protein, one roadmap was constructed and then its (thousands of) pathways were studied using the different analysis methods described in Section 4. When only the specified contacts were considered, the percentage of paths that had the given secondary structure formation order is shown. For example, for all contacts, and limiting our consideration to only the first 60% of the contacts to form, in 77% of the pathways for protein G β-hairpin 2 (β_3-β_4 contacts) formed before β-hairpin 1 (β_1-β_2 contacts), while in 82% of the pathways for protein L β-hairpin 1 formed before β-hairpin 2. Thus, the helix and β-hairpin 2 form first by a significant percentage for protein G, while for protein L, the helix and β-hairpin 1 consistently form first by a significant percentage. Both results agree very well with experimental observations. We also performed the study considering only the hydrophobic contacts, and obtained similar results, further confirming our findings.

We also study the formation order of β turns (see Figures 3 and 4 for

Table 5: β turn formation: comparison of analysis techniques for proteins G and L using the same roadmaps as in Table 4. Recall that turn 2 forms first in protein G and turn 1 forms first in protein L.

Comparison of Analysis Techniques – Helix and Turns							
			analyze first x% contacts				
Name	Contacts	SS Formation Order	20	40	60	80	100
Protein G	all	α, turn 2, turn 1	53	52	52	50	50
		turn 2, α, turn 1	15	9	17	22	22
		α, turn 1, turn 2	25	33	26	23	24
	hydrophobic	α, turn 2, turn 1	96	96	85	96	87
		α, turn 1, turn 2	4	4	12	2	11
Protein L	all	α, turn 1, turn 2	24	30	37	38	41
		1st turn, α, 2nd	3	4	4	4	6
		α, turn 2, turn 1	73	63	60	48	39
	hydrophobic	α, turn 1, turn 2	72	68	72	70	69
		turn 1, α, turn 2	5	9	5	7	15
		α, turn 2, turn 1	23	22	22	23	15

our definition). Remarkably, the results (see Table 5) are in good agreement with those obtained using the beta strands. For protein G, the second β turn forms consistently earlier than the first β turn, which confirms our results that the second hairpin forms first. For protein L, our results show that the second β turn forms first when considering all contacts. However, when only hydrophobic contacts are considered, then the first β turn forms first by a significant percentage. This indicates that some hydrophobic contacts form earlier in the first turn than in the second.

6 Conclusion and Future Work

We have shown that our PRM-based approach for studying protein folding pathways is able to correctly reproduce known folding differences between the structurally similar proteins G and L. This result gives confidence in our approach and implies it could be valuable for analyzing proteins whose structure is known but for which we lack experimental data on the folding pathway.

Acknowledgments

This research supported in part by NSF Grants ACI-9872126, EIA-9975018, EIA-0103742, EIA-9805823, ACR-0081510, ACR-0113971, CCR-0113974, EIA-9810937, and EIA-0079874. Song supported in part by an IBM TJ Watson PhD Fellowship and Thomas supported in part by an NSF Graduate Research Fellowship. We would like to thank Xinyu Tang for helping with the experiments, Jean-Claude Latombe for pointing out the connection between box folding and protein folding, and Michael Levitt and Vijay Pande for useful suggestions.

References

1. Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. *Annu. Rev. Biochem.* **66**, 549–579 (1997).
2. Reeke, Jr., G. N. *Ann. Rev. Comput. Sci.* **3**, 59–84 (1988).
3. Lansbury, P. *Proc. Natl. Acad. Sci. USA* **96**, 3342–3344 (1999).
4. Amato, N. M. and Song, G. *J. Comput. Biol.* **9**(2), 149–168 (2002). RECOMB 2001 special issue.
5. Kavraki, L., Svestka, P., Latombe, J. C., and Overmars, M. *IEEE Trans. Robot. Automat.* **12**(4), 566–580 August (1996).
6. Song, G. and Amato, N. M. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 948–953, (2001).
7. Li, R. and Woodward, C. *Protein Sci.* **8**, 1571–1591 (1999).
8. C.M.Dobson, Sali, A., and Karplus, M. *Angew. Chem. Int. Ed.* **37**, 868–893 (1998).
9. Radford, S. E. and Dobson, C. M. *Phil. Trans. R. Soc. Lond.* **B348**, 17 (1995).
10. Bryngelson, J., Onuchic, J., Socci, N., and Wolynes, P. *Protein Struct. Funct. Genet* **21**, 167–195 (1995).
11. Sternberg, M. J. *Protein Structure Prediction.* OIRL Press at Oxford University Press, (1996).
12. Bayazit, O. B., Song, G., and Amato, N. M. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 954–959, (2001). This work was also presented as a poster at RECOMB'01.
13. Singh, A., Latombe, J., and Brutlag, D. In *7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, 252–261, (1999).
14. Song, G. and Amato, N. M. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, 287–296, (2001).
15. Fiebig, K. M. and Dill, K. A. *J. Chem. Phys* **98**(4), 3475–3487 (1993).
16. Yi, Q. and Baker, D. *Protein Sci* **5**, 1060–1066 (1996).
17. `http://www.rcsb.org/pdb/`.
18. Dill, K. A., Fiebig, K. M., and Chan, H. S. *Proc. Natl. Acad. Sci. USA* **90**, 1942–6 (1993).
19. McCallister, E. L., Alm, E., and Baker, D. *Nat. Struct. Biol.* **7**(8), 669–673 (2000).
20. Kim, D. E., Fisher, C., and Baker, D. *J. Mol. Biol.* **298**, 971–984 (2000).