Efficient Sampling of Protein Folding Pathways using HMMSTR and Probabilistic Roadmaps

*Yogesh A. Girdhar, ‡Christopher Bystroff, *Srinivas Akella

Dept. of *Computer Science, ‡Biology, Rensselaer Polytechnic Institute, Troy, NY 12180
{girdhy,sakella}@cs.rpi.edu, bystrc@rpi.edu

Abstract

We present a method for constructing thousands of compact protein conformations from fragments and then connecting these structures to form a network of physically plausible folding pathways. This is the first attempt to merge the previous successes in fragment assembly methods with probabilistic roadmap (PRM) [2] methods. Previous PRM methods have used the knowledge of the true structure to sample conformational space. Our method uses only the amino acid sequence to bias the conformational sampling. Conformational sampling is done using HMMSTR [1], a hidden Markov model for local sequence-structure correlations. We then build a PRM graph and find paths that have the the lowest energy climb. We find that favored folding pathways exist, corresponding to deep valleys in the energy landscape. We describe the pathways for three small proteins with different secondary structure content in the context of a folding funnel model.

1. Introduction

Protein folding is a heirarchical process where small fragments are assembled into compact larger fragments subject to the conformational preferences of the small fragments and subject to the topological constraints that arise when differently shaped fragments are linked end-to-end. The allowable conformations of the chain are dictated not just by the energies of the conformations themselves but also by the accessibility of these conformations to a folding pathway that begins with a fully extended chain. We have explored the energy landscape of the conformational space of small proteins by random sampling of protein-like conformations, followed by linking of these conformations into a folding pathway. No knowledge of the true structure was used in building the conformational samples and pathways.

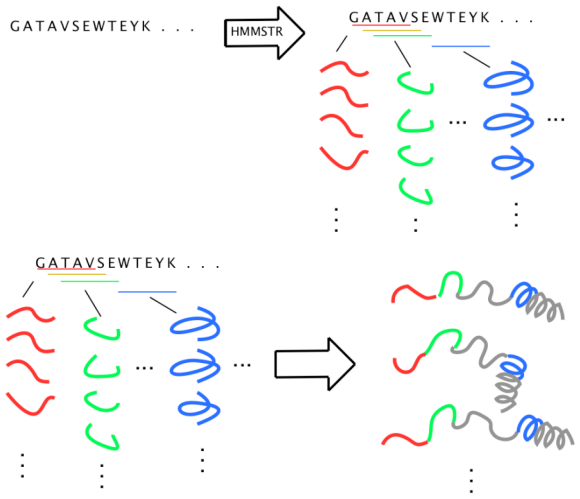

Figure 1. Generation of short fragments based on the sequence, followed by linking of these fragments to form samples.

2. Sample Generation using HMMSTR

HMMSTR is a hidden Markov model for local sequence/structure correlation. It uses knowledge of preferred orientations of amino acid sequences from data in the Protein Data Bank (PDB) to predict the Phi and Psi torsion angles of local sequences. Given the amino acid sequence code of a protein and a window size, HMMSTR generates a set of likely Phi and Psi angles for each overlapping window for the entire sequence. Once we have these local structures, we then proceed to building a complete configuration out of these local structures. We start with a random anchor window in the protein, and then choose a fragment randomly, with a probability proportional to its score given by HMMSTR. We then walk towards both the ends of the protein from the anchor window, assigning angles for all remaining windows.

Although HMMSTR predicts the local structure of an amino acid sequence, it does not give us any information about non-local interactions. Hence we

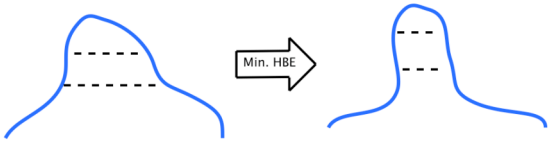

Figure 2. Monte Carlo energy minimization

apply Monte Carlo minimization on each of the samples to minimize van der Waals collisions (VDW), radius of gyration (RG) and hydrogen bond energy (HBE). As a result of this, we get samples which are compact and show non-local hydrogen bonding.

3. PRM Graph

PRM graph is a nearest neighbor graph which represents the protein folding funnel landscape. Each node of this directed graph represents a valid low energy (no VDW collisions) conformation of the protein, and each edge (u, v) represents a possible transition from u to v. We use Dijkstra's shortest path algorithm to find the folding pathways on the PRM graph. Weight of each edge is set to the difference in the energy of the conformations. All the negative edge weights are set to zero.

4. Results & Discussion

We evaluate our approach with three different proteins: a 16-residue long β-hairpin from Protein G – 2GB1(16), and a 28 residue long Fbp28Ww Domain from Mus Musculus – 1E0L(28). After generating pathways, we built a highway graph, which is a subgraph of the PRM graph, with only the 100 most popular configurations. Popularity is simply measured as the number of times a configuration appears for all the pathways.

Our new technique for sampling the energy landscape of a folding funnel does not use any information about the native conformation; hence this method is unbiased by the native state. Using HMMSTR results in samples that are more biologically feasible as compared to random sampling.

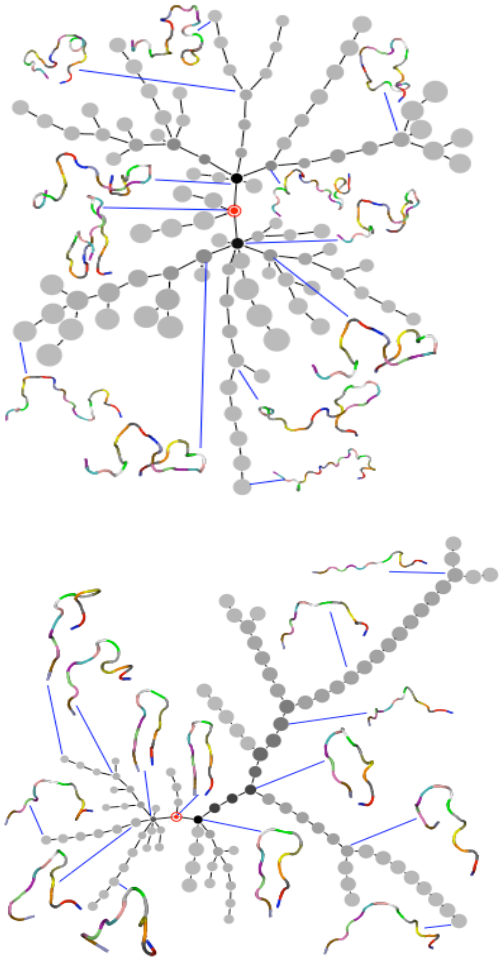

Figure 3. Highways for 28 residue long 1E0L (top) and 16 residue long β-hairpin of 2GB1 (bottom). Bigger circles correspond to high energy configurations, darker circles correspond to more popular configurations in all pathways.

6. References

[1] C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. Journal of Molecular Biology, 301(1):173–190, 2000.

[2] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transactions on Robotics and Automation, 12(4):566–580, 1996.