Improved Sampling of Protein Folding Landscapes using Molecular Dynamics Data[†]

Manasi Vartak, Lydia Tapia, Shawna Thomas, and Nancy M. Amato
Parasol Lab, Dept. of Computer Science, Texas A&M University, College Station, TX 77843
{mvartak}@wpi.edu {ltapia, sthomas, amato}@cs.tamu.edu

The function of proteins depends upon their three dimensional structure or native state. Folding of a protein into an incorrect conformation has been shown to be the cause of diseases like Alzhiemers disease and bovine spongiform encephalopathy (Mad Cow disease). Current simulations using the Probabilistic Roadmap Method (PRM) can be used to predict folding pathways, but these simulations generate conformations based solely on the native state. This limits the proteins that can be studied by our method. We present an alternate way to generate conformations that can be used by the PRM method. The advantage of this approach to sampling is that it reduces dependence on the native state by using molecular dynamics (MD) simulations.

MD simulations are expensive for all but the smallest of proteins. Our approach overcomes this limitation by using MD simulation data for small overlapping fragments of the protein. This data is processed and used to bias our sampling to explore a larger conformational space. The PRM methods used here have been presented and validated previously. Compared to other methods like MD simulations, our PRM method combined with our new sampling method is fast and can be used to find conformations of full length proteins. We validate our results by comparing them to existing experimental data and our previous simulation results.

[†]This research supported in part by NSF grants ACI-9872126, IA-9975018, EIA-0103742, EIA-9805823, ACR-0081510, ACR-0113971, CR-0113974, EIA-0079874, IIS-0118097, by the Texas Higher Education Coordinating Board grant ATP-000512-0261-2001, and by the DOE.Vartak supported by the Computing Research Association CRA-W Distributed Mentor Program.Thomas supported in part by an NSF Graduate Research Fellowship, a PEO Scholarship, a Department of Education GAANN Fellowship, and an IBM TJ Watson PhD Fellowship. Tapia supported in part by a NIH Molecular Biophysics Training Grant (T32GM065088) and previously supported by a Department of Education GAANN Fellowship.

1 Introduction

As proteins fold to their native, functional state, they undergo critical conformational changes that affect their functionality. Some conformational changes are detrimental. For example, diseases such as Mad Cow disease or Alzheimer's disease are caused by misfolded proteins [3]. Insight into the kinetics and detailed mechanics of the folding process will help explain critical information about the protein such as its function and why it misfolds.

Simulating protein folding kinetics has been a difficult task performed on small structures through computationally expensive methods such as molecular dynamics [6] or Monte Carlo simulations[4, 10]. Studies on larger proteins have recently been accomplished, but these simulations have only been done on limited proteins represented with coarse models.

In our previous work [2, 18, 17], we studied protein folding through the application of a method that builds an approximate map of a protein's potential energy landscape. This map contains thousands of feasible folding pathways to the known native state enabling the study of global landscape properties. We obtained promising results for several proteins [18]. The pathways were validated by comparing secondary structure formation order with known experimental results. However, in this method, conformations are generated based solely on the native state.

In this work, we present a new way to sample the protein's potential landscape that reduces dependence on the native state. MD simulations for overlapping fragments of the protein are used to generate possible conformations for the whole protein.

Our method has been tested on protein G as a representative protein. Our studies indicate that our new method is fast and can be used for full length and detailed protein models. We validate our method by comparing the potential landscape and secondary structure formation orders generated by our method to those generated by our previous work.

2 Related Work

2.1 PRMs for Protein Folding

In previous work [1], we introduced an approach to protein folding that is based on the Probabilistic Roadmap (PRM) approach for motion planning [9]. We applied our method to a large number of structures and were able to identify subtle differences in the known experimental secondary structure formation order for proteins with very similar structures [14, 18].

Our method is simple and consists of two main steps: (1) sampling conformations in the landscape and (2) making transitions between sampled conformations. Sampling is the process by which conformations or nodes in the roadmap are generated. The set of all possible conformations of a protein is called the conformational space or C-space. Our methods bias sampling to increase density of nodes near the native state.

Sampling Methods. Sampling of a landscape decides how the nodes on the roadmap will be distributed and whether we can construct a complete roadmap. The quality and distribution of nodes ultimately decides the ability of our roadmaps to capture the important features of the energy landscape. In previous methods, nodes were generated based solely on the native state. In this section we briefly introduce these methods and discuss their strengths and weaknesses.

- TopRMSD: The TopRMSD method randomly perturbs the native state of the protein by changing the torisional angles (phi and psi) slightly.[1] This process is repeated till the required number of nodes are generated. TopRMSD is successful for small proteins typically less than 10 residues.,
- Layers: The Layers method is similar to TopRMSD in that it perturbs the native state by varying the torisional angles. But this method perturbs angles iteratively by using the idea of bins to group the nodes.[1] 10 bins are set up in descending order to hold conformations having the number of native contacts in a particular range. Native contacts are pairs of C_α atoms that are at a distance of less than 7Å in the native state as well as the said conformation. Randomly chosen conformations from the a bin are used as seeds to generate nodes for the upper bins.

- Rigidity Layers: Rigidity Layers combines the Layers method with a more refined way to perturb a conformation.[18] This is done by identifying the rigid and flexible parts of a protein using the pebble game algorithm. These parts are then perturbed according to their flexibility, thus providing a physically realistic way to perturb conformations. It has been shown that this method works better than the other methods and requires a smaller roadmap to compute the correct formation orders.
- Our Contribution: The sampling technique called Fragment-MD introduced in this paper is used to sample the conformational space like the above methods. The advantage of this approach is that it reduces dependence on native state while benefitting from the accuracy of the MD simulations. Due to the computational expenses associated with MD, we use simulations for overlapping fragments of the protein instead of the whole protein. The MD data is then processed and used to bias the conformations. For the case study of protein G, we use data from 8 and 12 residue fragments.

Connecting the Roadmap. In the second step, connections (edges) are made between sampled conformations with similar structure. Weights are assigned to directed edges to reflect the energetic feasibility of transitioning between the two endpoint conformations. This combination of nodes and weighted edges forms a roadmap that approximates the energy landscape. This roadmap encodes thousands of folding pathways. The most energetically feasible pathways in the roadmap can be extracted using these weights.

Connections between two nodes, q_1 and q_2, are labeled with edge weights that reflect the energetic feasibility of transitioning between them. This is done by first identifying all the intermediate nodes, $q_1 = c_0, c_1, ..., c_{n-1}, c_n = q_2$, that connect q_1 to q_2. For each pair of consecutive conformations c_i and c_{i+1}, the probability P_i of transitioning from c_i to c_{i+1} depends on the difference between their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$:

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

This keeps the detailed balance between two adjacent states and enables the edge weight to be computed by summing the logarithms of the probabilities for all pairs of consecutive conformations in the sequence. With this edge weight definition, we can use simple graph search algorithms to extract the most energetically feasible pathways in the roadmap between two given states (e.g. from the unfolded state to the folded state).

Protein Model. We model the protein as an articulated linkage. Using a standard modeling assumption for proteins that bond angles and bond lengths are fixed [15], the only degrees of freedom in our model are the backbone's phi and psi torsional angles which are modeled as revolute joints with values in the range $[0, 2\pi)$.

Potential Energy Calculation. Our method is flexible and allows any potential function to be used. In this paper, we use a coarse potential function similar to [12]. We use a step function approximation of the van der Waals potential component and model side chains as spheres with zero dof. If any two spheres are too close (i.e., less than 2.4Å during sampling and 1.0Å during connection), a very high potential is returned. Otherwise, the potential is:

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

where K_d is 100 kcal/mol and $d_0 = d_c = 2$ Å as in [12]. The first term represents constraints favoring known secondary structure through main-chain hydrogen bonds and disulphide bonds, and the second term is the hydrophobic effect. The hydrophobic effect is computed as follows: if two hydrophobic residues are within 6 Å of each other, then the potential is decreased by 100 kJ/mol.

In our previous work [1, 18], we provided methods for building an approximate map of a protein's potential energy landscape [1, 18] and an RNA's folding landscape [16]. We have published results from our approximate maps for proteins up to 148 residues easily built on a desktop PC [18]. Our roadmaps give an approximate view of the protein folding landscape. In the past, we have successfully extracted low-energy pathways, validated secondary structure formation order, and seen general and consistent trends in reaction coordinates such as native contacts present and RMSD.

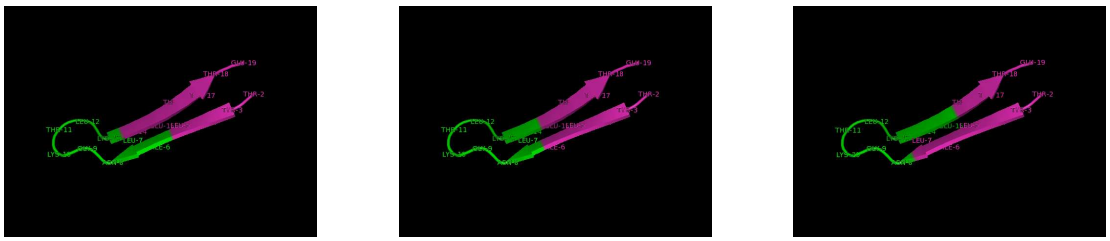
2.2 Molecular Dynamics

Molecular Dynamics (MD) is a versatile tool for studying trajectories of particles in any system. [11] MD simulations consist of the following main steps:

- Choosing a force field and potential energy function
- Finding the force on each particle
- Finding the equations of motion for each particle
- Integrating the equations of motion to give particle trajectories

These simulations are accurate and when applied to proteins, can give a single folding pathway. But they are expensive for all but the smallest of proteins.

Our approach overcomes this limitation by using MD simulations for small overlapping fragments of the proteins rather than the whole protein at once. The fragments resemble a sliding window with two consecutive fragments differing by one residue. For protein G, we used MD data for 8 and 12 residue fragments of the protein. Figure 1 shows a set of such overlapping fragments for hairpin 1 of Protein G. The MD data used in this paper is courtesy of the Dill group at UCSF.



(a) Residues 6 - 13 (b) Residues 7 - 14 (c) Residues 8 - 15

Figure 1: The overlapping fragments (shown in green) for hairpin1 of protein G

3 Metrics Used

The following is a listing of the metrics used to assess the quality of the nodes produced by our methods. We use several metrics to evaluate our nodes since no single method is effective for all conformations. Certain metrics like Euclidean distance and RMSD are effective only when the structures are close to the native state.

3.1 Potential energy

Potential energy is one of the main metrics used in our method as it is one of the few metrics that doesn't depend on the native state. Details of the potential energy function used in this paper have been described in Section 2.1.

3.2 Euclidean distance

The Euclidean distance metric captures the amount of physical movement (around the torsional angles) that a start configuration, c_a, would undertake to move to a goal conformation, c_b.[2] This distance is approximated by measuring the difference in the phi/psi angle pairs of the two conformations. The Euclidean distance metric, $d_E(c_a, c_b)$, between two conformations $c_a(\phi_1^a, \psi_1^a, \ldots, \phi_n^a, \psi_n^a)$ and $c_b(\phi_1^b, \psi_1^b, \ldots, \phi_n^b, \psi_n^b)$ is:

$$\sqrt{\frac{(\phi_1^a - \phi_1^b)^2 + (\psi_1^a - \psi_1^b)^2 + \cdots + (\phi_n^a - \phi_n^b)^2 + (\psi_n^a - \psi_n^b)^2}{2n}}$$

3.3 RMSD

Root Mean Square Distance (RMSD) is measured between the atoms of the protein in the two conformations. [8, 7] Our model consists of six atoms for each amino acid, C, C_α, R, O, N, and H. Therefore, a protein with n amino acids, will have $6n$ atoms in our protein model. The coordinates of these atoms can be specified as x_1 to x_{6n}. For two conformations $c_a(x_1^a, x_2^a, \ldots, x_{6n}^a)$ and $c_b(x_1^b, x_2^b, \ldots, x_{6n}^b)$, the distance, $d_R(c_a, c_b)$, is:

$$\sqrt{\frac{\|x_1^a - x_1^b\|^2 + \|x_2^a - x_2^b\|^2 + \cdots + \|x_{6n}^a - x_{6n}^b\|^2}{6n}}$$

However, the above equation only considers the distance for a single translation and orientation of the two configurations. The RMSD must be the minimization of $d_R(c_a, c_b)$ over all possible translations and orientations.

C_α RMSD: Another variant of RMSD is to compare only the C_α atom for each amino acid in the RMSD calculation. This speeds up the RMSD calculation.

3.4 Native contacts present

A native contact is a pair of C_α atoms that are at a distance of less than 7Å in the native state as well as the given conformation. The number of native contacts present in any conformation can be used as an indication of how close it is to the native state. These contacts are used to define the formation of secondary structures in a conformation and to calculate the formation orders in a roadmap.

3.5 Hydrogen bonds present

Repetitive hydrogen bonds (H-bonds) in the protein generate secondary structures like helices and sheets. Hence, the number of corresponding H-bonds present in the native state and a given conformation can be used a metric. This metric is more specific than native contacts and takes longer to compute but it maybe a better indication of native-like structure.

3.6 Other metrics

Pymol[5] was used to compare our lowest energy conformations to the native state. It was also used to spot the discrepancy between the metrics of native contacts and H-bonds as described in Section 6. Ramachandran plots were used to compare the angles in the lowest energy conformations from Method 1 to those in the native state.

4 Sampling by Fragment-MD method

The Fragment-MD method uses MD simulations of overlapping fragments of the protein to bias the generation of new conformations. For using the MD data, it is processed to give a probability distribution for each residue of the protein. One distribution is obtained per fragment the residue is present in. This probability distribution is then used in various ways to generate new conformations.

We studied two main methods to generate conformations based on the MD data: (i) Method 1 (without the native state) and (ii) Method 2 (using the native state). In this section, we discuss some of the approaches used to implement these methods and provide comparisons between them. We validate our results by comparing the roadmaps generated by our methods with those from previous sampling methods in Section 5.

4.1 Method 1

Method 1 generates conformations without using the native information. Nodes are constructed by picking phi and psi angles for each residue based on the probability distributions from MD data. Some approaches

used to implement this method are listed below.

Approach A: For our first implementation, we chose the probability distribution for each residue from the fragment in which that residue was third in sequence. Phi and psi angles were then chosen based on this probability distribution.

Approach B: Our next approach was to combine the existing Layers method with Method 1 to get a better distribution of nodes. We generated part of the total nodes using Approach A and then used the Layers method to iteratively generate new nodes from these. We tried several variations in this approach by altering the percentages of nodes generated by Approach A. It was found empirically that constructing 30% of the total nodes by the Layers method gave the best node distribution.

An important variation in this method was to sort the nodes generated by Approach A and then use the best nodes as seeds for the Layers method. We sorted the nodes based on various metrics like potential energy, RMSD, euclidean distance and native contacts. To minimize dependence on the native state, our final choice was to sort the nodes by potential energy and then pick the top 50 percent of the nodes as seeds for the Layers method. This approach produced a more extensive node distribution as compared to Approach A.

Approach C: To accomodate the dependence between angles of consecutive residues, we changed our way of choosing probability distributions. Distributions for consecutive residues were now picked from the same fragment instead of using different fragments for each residue. The objective was to conserve conformations got from the MD simulations. Adjustments had to be made for the first and last residues of each fragment because the phi value for the first residue and psi value for the last residue were assumed to be 0. The method produced a marked improvement in the quality of nodes generated.

Approach D: The Layers method perturbs conformations uniformly without regard to the rigidity of residues. So, to obtain better conformations, we modified our method to include the Ridigity Layers method along with the Layers one. This combination slightly improved the quality of nodes generated and as expected it gave better results than Approach A, B and C.

4.2 Method 2

Our results indicate that one of the major drawbacks of Method 1 was that it did not generate native-like nodes. The values of metrics like RMSD and potential energy were also not close to the expected ones. But Method 1 provided the framework to implement a method that would make use of native information as well as MD data to construct conformations. Method 2 begins with the native state and then biases its angles gradually toward the MD conformations. Some of the approaches used to implement it were:

Approach A: The initial approach consisted of using the native state and then changing its angles to bias them towards the MD data. For better results, we used rigidity analysis (like the Ridigty Layers method) to determine the flexible parts of the protein and then biased angles in those parts. The probability distribution towards which angles in a residue would be biased was chosen randomly. This method used the same iterative approach as used in Layers and Rigity Layers.

We encountered problems with this implementation because the MD data corresponded only to partially folded states of the protein. With no unfolded states, the roadmap could not be analyzed for folding pathways or formation orders. So the method was modified to randomly perturb partially folded conformations to generate unfolded ones.

Approach B: In the previous approach, we randomly picked distributions towards which we biased our angles. As with Approach A of Method 1, this did not capture the dependence between angles of consecutive residues. Hence, we modified the method so that consecutive residues were biased towards distributions from the same fragment. The objective as before, was to conserve the conformations got from MD simulations. This produced an improvement in the formation orders for Protein G.

Some of the variations in this approach were to bias the native state towards a particular probability

distribution set picked from the MD data i.e. instead of using all the overlapping fragments, we used a set of non-overlapping fragments that spanned the protein. Adjustments had to be made for residues at the beginning and end of the protein.

5 Experimental Results

We tested our method on protein G (Figure 2 (a)) which has been used as a benchmark protein for various studies. protein G is a 56 residue protein with 1 α helix and two β sheets.

(a) (b)

Figure 2: Comparison of conformations (a) Protein G native state (b) Top conformation after sorting nodes by native contacts, RMSD and potential energy.

5.1 Performance of Method 1

Since we were unable to get native-like nodes from Method 1, its performance was judged based on the node distributions obtained. We evaluated our results by comparing the node distributions generated by our method to the ones obtained from previously validated methods. This included the comparison of graphs of the various properties like RMSD, Euclidean distance, native contacts present and potential energy (Figure 3). We also compared the distribution of nodes with histograms. The top few conformations were compared to the native conformation using Pymol (Figure 2) and Ramachandran plots.

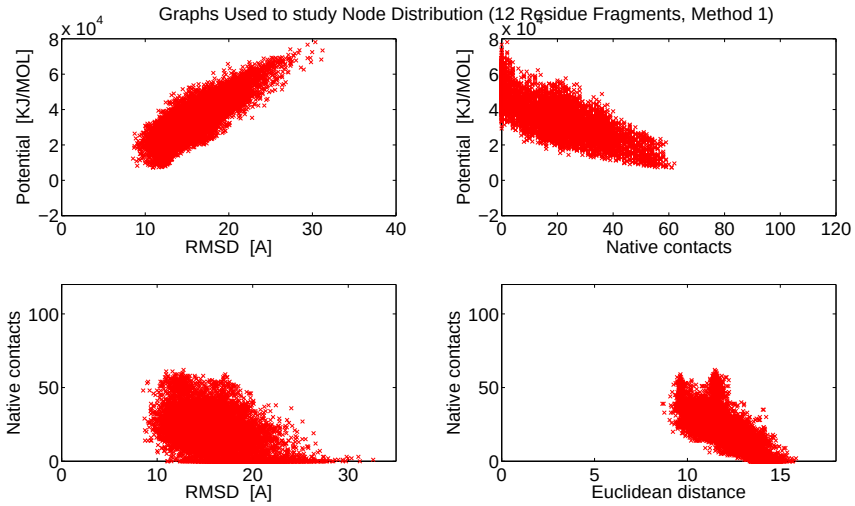
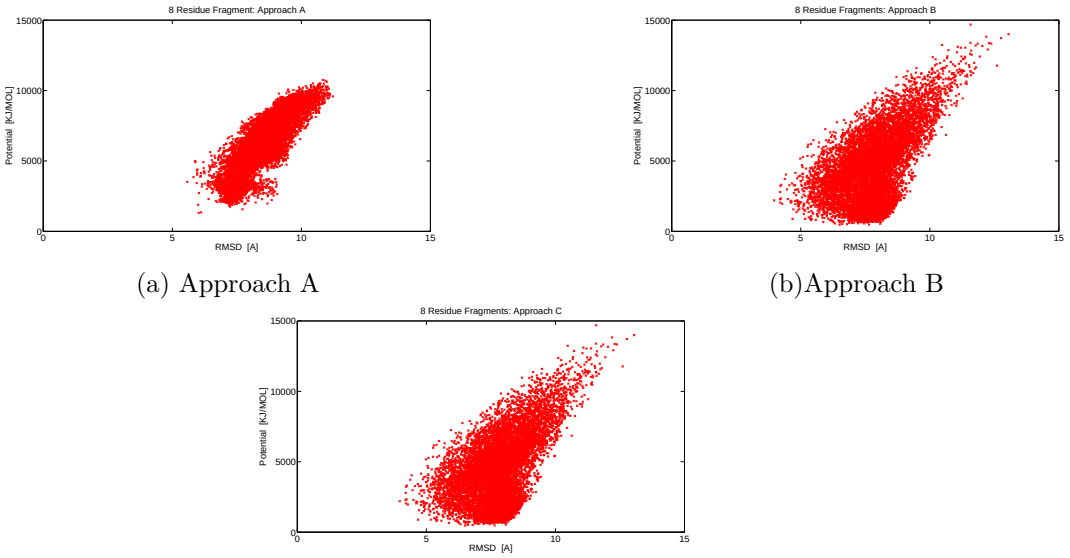


Figure 3: Graphs used to study node distribution

Comparison of Approaches A, B, C and D. Figure 4 shows the Potential energy vs.RMSD plots obtained for 10000 nodes generated by Approach A, B and C for hairpin1 of protein G, spanning residues 1 - 20. Approach D was not run on hairpin 1. We can see that the distribution of nodes expands from A to B. Approach C improves the quality of conformations though it does not change the node distribution.

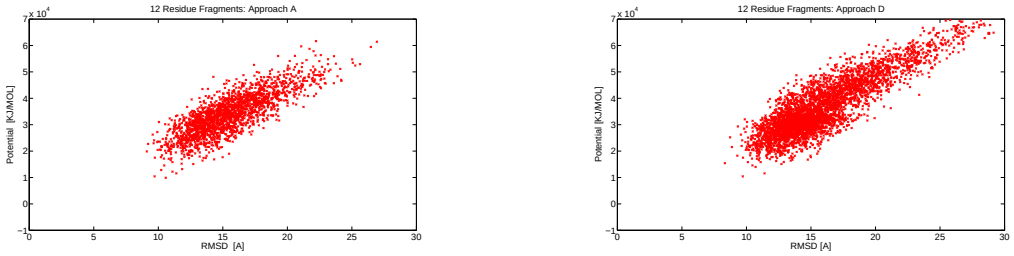


(a) Approach A (b)Approach B

(c)Approach C

Figure 4: Potential vs. RMSD plots for conformations of hairpin 1 of protein G (8 residue fragments).

Figure 5 compares the distributions obtained from Approach A and D for protein G. We can see that Approach D produces a marked improvement in the node distribution.

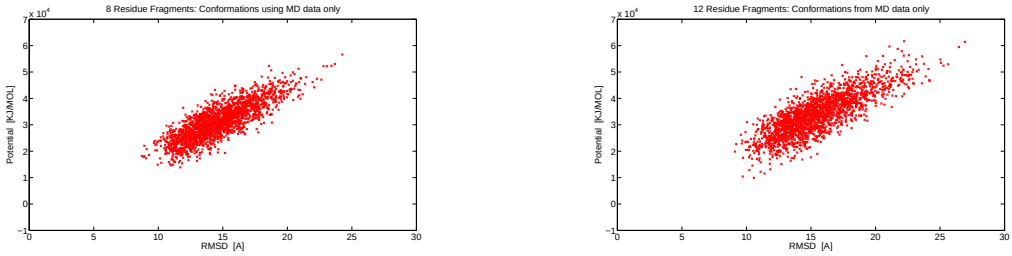


(a) Approach A (b) Approach D

Figure 5: Potential vs. RMSD plots for Approach A and D applied to protein G (12 residue fragments).

Effect of Length of Residues. Our method was tested on MD data from 2 types of fragments of protein G: 8 and 12 residue fragments. The node distribution from these two types of fragments is shown in Figure 6. As seen from the figures, it was found that the 12 residue fragments gave a better node distribution than the 8 residue fragments. The Euclidean distance and potential energy limits were lower while the number of hydrogen bonds was higher for the 12 residue fragments nodes compared to the 8 residue ones. The other metrics had comparable values.

Performance of Method 1. Landscapes from Method 1 were computed and compared to those obtained from the Rigidity Layers method. The landscapes show that native-like conformations are not obtained from Method 1. But the placement of the node distribution in the graph (Figure 7 (b)) is very similar to what is expected (Figure 7 (a)). From this we can conclude that the Method 1 performs well for nodes that are

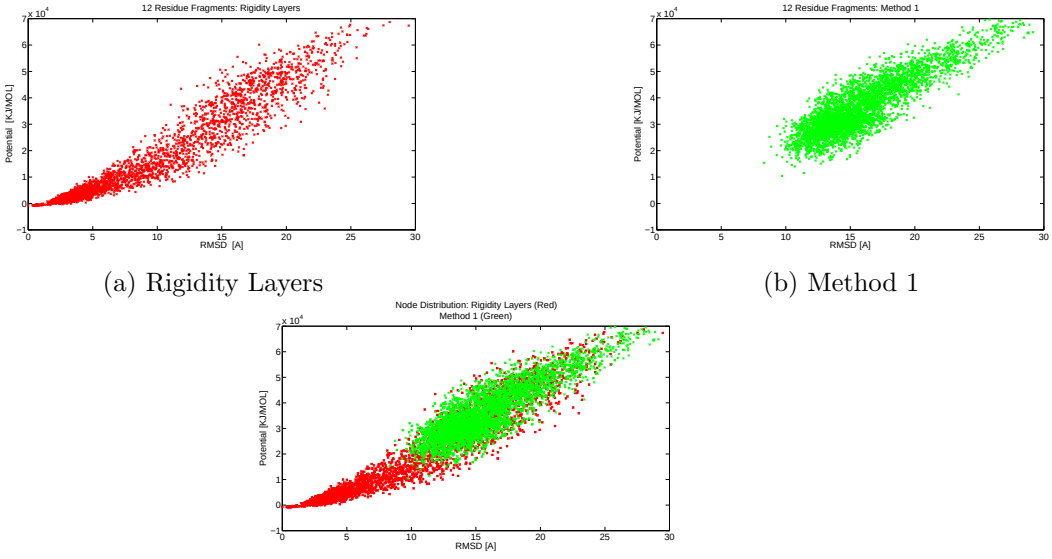


(a) 12 residue fragments (b) 8 residues fragments

Figure 6: Potential vs. RMSD plots.

away from the native state.

Figure 2 shows protein G and the top conformation generated by Method 1 after sorting the nodes by native contacts, RMSD and potential energy. As seen from these conformations, the α helix of protein G is correctly formed but we have an additional helix as well. The Ramachandran plot for the top conformation indicates that more phi-psi pairs are in the helix region of the plot than expected. We believe that this may be due to the nature of the MD data.



(a) Rigidity Layers (b) Method 1

(c)

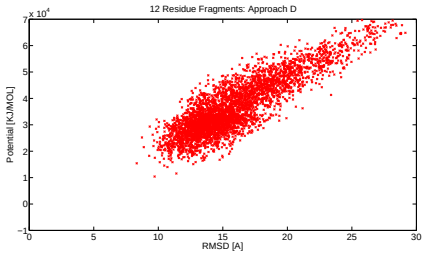
Figure 7: Comparison of landscapes through Potential vs. RMSD plots. (c) shows rigidity layers (red) and Method 1(green) plotted on the same scale.

5.2 Method 2: Expansion of Method 1 to use the native state

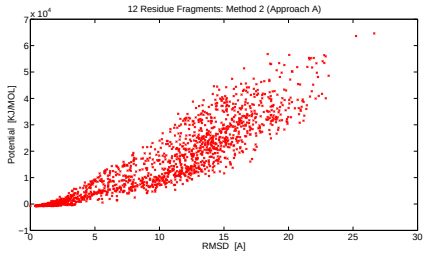
Method 2 begins with the native state and then perturbs its angles to bias them towards the MD data. As opposed to this, Method 1 uses no native state knowledge. This can be seen from the potential energy vs. RMSD plots for the two method as shown in Figure 8. Near native nodes are only obtained by Method 2.

5.3 Performance of Method 2

Since Method 2 starts from the native state, we could construct complete roadmaps using it and the results are analyzed based on the roadmap and secondary structure formation orders. Comparison of the potential



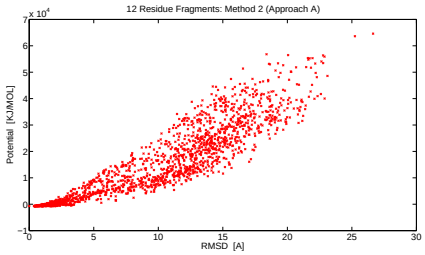
(a) Method 1, Approach D



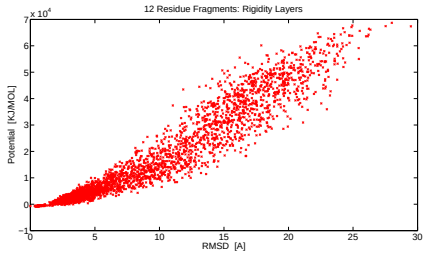
(b) Method 2, Approach A

Figure 8: Potential vs. RMSD plots

energy vs. RMSD plots from Method 1 with those from the existing Rigidity Layers method (Figure 9) shows that the node distributions are very similar with the landscapes having the same general trends.



(a) Method 2, Approach A



(b) Rigidity Layers

Figure 9: Potential vs. RMSD plots.

Table 1 shows a comparison between the performance of Rigidity Layers and Method 2. Method 2 (Approach A and B) requires fewer nodes and takes lesser time to build the roadmap than Rigidity Layers. Additionally, Approach B performs even better than approach A. However, there is a discrepancy in the secondary structure formation orders (SSFOs) got from Method 2. Experimental results indicate the α helix forms first, followed by $\beta 3$, $\beta 4$ called the β-hairpin 1 and lastly $\beta 1$, $\beta 2$ called the β-hairpin 2[13]. As seen from Table 1, the Rigidity Layers method finds the correct formation order as α, $\beta 3$-$\beta 4$,$\beta 1$-$\beta 2$ while the results from Method 2 indicate it to be α, $\beta 1$-$\beta 2$,$\beta 3$-$\beta 4$. We believe that this may be due to the MD simulations or our present method of picking fragments for processing.

There is a marked improvement in the SSFO got from Approach B as compared to Approach A. The last row of the table gives one of the top results obtained from variations in Approach B. In this case, only one set of non-overlapping fragments spanning the protein was used instead of using all the overlapping fragments. Though this variation needs more nodes and time to generate the map, the SSFO obtained is much better than the previous ones.

6 Discussion: Native contacts vs. Hydrogen bonds

When we used native contacts as a metric, inconsistent results were obtained for hairpin 1 of protein G. In spite of having close to 100 percent contacts, our conformations were not native-like. When the conformations were analyzed for the number of H-bonds, it was found that only about 40% of the native H-bonds had been formed though almost all native contacts were present. This seemed to indicate that the number of H-bonds may be a better metric than the number of native contacts. On the other hand, in the case of protein G, the results for native contacts and H-bonds seemed to agree. It is not clear as to which is a better metric in this case. Figure 10 shows the hydrogen bonds (yellow dotted lines) present in the native state of hairpin 1 and our lowest energy conformation. This conformation has 21 (of 24) native contacts but only 2 (of 7) H-bonds.

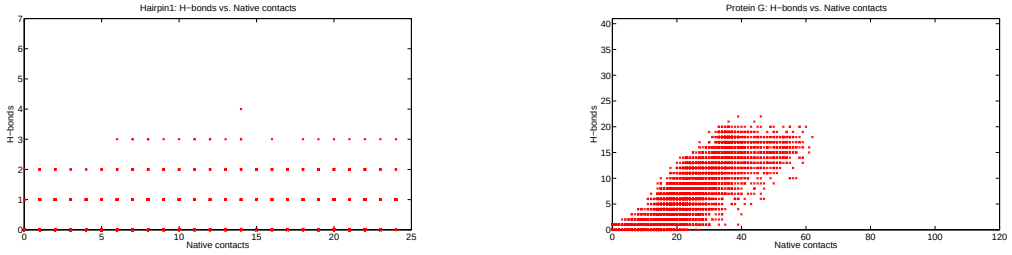
Index	Method	Nodes (N)	Edges (E)	Connectivity (E/N)	E+N	Generation time (min)	Seondary Structure Formation Order (SSFO)
1	Rigidity Layers	3001	109754	36.57	112755	106.73	α, β3-β4,β1-β2(95%)
2	Method 2 Approach A	1851	61996	33.49	63847	71.36	α, β1-β2,β3-β4 (100%)
3	Method 2 Approach B	1701	54872	32.26	56573	57.56	α, β1-β2,β3-β4 (75%) α, β3-β4,β1-β2 (25%)
4	Method 2 Approach B variation	2701	91474	33.86	94175	106.97	α, β1-β2,β3-β4 (53%) α, β3-β4,β1-β2(47%)

Table 1: Comparison of performance of Rigidity Layers and Method 2 (Approach A and B)

(a) (b)

Figure 10: Conformations and H-bonds (yellow dotted lines): (a) hairpin 1 and (b) top conformation after sorting by native contacts, RMSD and potential energy

Figure 11 shows a plot of the the number of H-bonds vs. native contacts for hairpin 1 and protein G. The figure shows that in the case of hairpin 1, in spite of having upto 100% native contacts we can have only about 40% of the H-bonds. Figure 12 shows the RMSD distribution after sorting the nodes by H-bonds, RMSD, potential energy and native contacts, RMSD, potential energy for hairpin 1 and protein G. From the figure, we can see that for hairpin 1, initial sorting by native contacts results in very high RMSDs for the top sorted nodes. This is not the case for protein G.

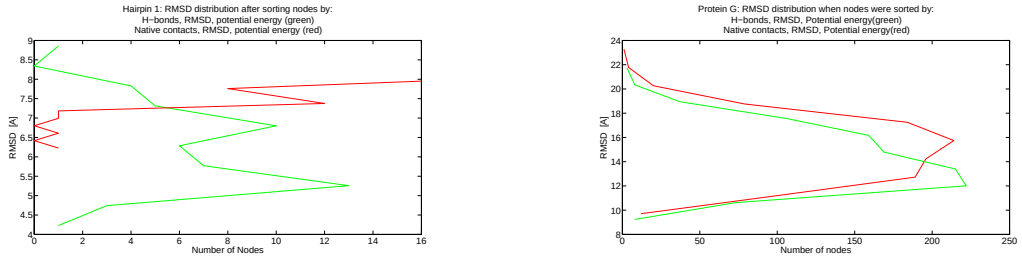


(a) hairpin 1 (b) protein G

Figure 11: H-bonds vs. native contacts.

7 Conclusion

In this paper, we presented a new way of generating conformations to be used in the PRM method called Fragment-MD. The advantage of our approach is that it reduces dependence on the native state by making use of Molecular Dynamics simulations. As explained in Section 7.1, the landscapes from Method 1 of the Fragment-MD approach have the same general characteristics as those from existing sampling methods.



(a) hairpin 1 (b) protein G

Figure 12: RMSD distribution after sorting nodes by H-bonds, RMSD, potential energy (green) and native contacts, RMSD, potential energy (red)

On the other hand, Method 2 not only produces landscapes very similar to the existing methods, but also needs fewer nodes and lesser time to construct the roadmap. The discrepancy in the secondary structure formation orders was partially resolved by using a single set of non-overlapping fragments to bias the angles. But further work in this area is necessary before using this method for other proteins.

Our findings also suggest that native contacts may not be a reliable metric for all molecules. Tallying the results from the native contacts with those from the H-bonds may help to determine the effectiveness of native contacts as a metric for the given molecule.

In all, the Fragment-MD approach to sampling is promising due to the reduced time and fewer nodes required to generate roadmaps. With optimization and further work, it can be expected to produce better results than the existing methods.

8 Acknowledgments

We would like the thank the Dill group at University of California San Fransisco for providing the MD simulation data for protein G.

References

[1] N. M. Amato, K. A. Dill, and G. Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, 10(3-4):239–256, 2003. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2002.

[2] N. M. Amato and G. Song. Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9(2):149–168, 2002. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2001.

[3] F. Chiti and C. Dobson. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333–366, 2006.

[4] D. Covell. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.*, 14(4):409–420, 1992.

[5] W. DeLano. The pymol molecular graphics system (2002). *DeLano Scientific, Palo Alto, CA, USA.*, 2002.

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

[7] B. Horn, H. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *J. Opt. Soc. Am.*, 5(7):1127–1135, 1988.

[8] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr*, A34:827–828, 1978.

[9] L. E. Kavraki, P. Svestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

[10] A. Kolinski and J. Skolnick. Monte Carlo simulations of protein folding. *Proteins Struct. Funct. Genet.*, 18(3):338–352, 1994.

[11] V. Lamberti, L. Fosdick, and E. Jessup. A hands-on introduction to molecular dynamics. *J. Chem. Edu.*, 79:601–606, 2002.

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

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

[14] G. Song, S. Thomas, K. Dill, J. Scholtz, and N. Amato. A path planning-based study of protein folding with a case study of hairpin formation in protein G and L. In *Proc. Pacific Symposium of Biocomputing (PSB)*, pages 240–251, 2003.

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

[16] X. Tang, S. Thomas, L. Tapia, and N. M. Amato. Tools for simulating and analyzing RNA folding kinetics. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, 2007.

[17] L. Tapia, X. Tang, S. Thomas, and N. M. Amato. Kinetics analysis methods for approximate folding landscapes. In *Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, page to appear, 2007.

[18] S. Thomas, X. Tang, L. Tapia, and N. M. Amato. Simulating protein motions with rigidity analysis. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 394–409, 2006.