Study Of Protein Folding

Monit Kanwat
Nitesh Vijayvargiya
Advisor : Dr. Amitabha Mukerjee
Department of Computer Science and Engineering
IIT Kanpur
{monit,nvijay,amit}@cse.iitk.ac.in

April 14, 2012

Abstract

Study of protein folding has been a highly studied problem for quite a while. Past years have seen a lot of work from the Amato Group from the Texas A & M University. We present an attempt to implement their work on finding the sequence of intermediate protein conformations a protein molecule may adapt to fold into the most stable state also known as its native state. We implement the calculation of energy and other useful parameters for a protein conformation as well as the sampling technique used to generate nodes for a PRM based RoadMap. We also verify the work of N. S. Bogatyreva and D. N. Ivankov by observing the behaviour of protein energy with variation in its native contact number. Future work involves creating the map and extracting the path of folding from a conformation to the native state.

1 Introduction

Proteins play a vital role in our metabolic systems. They perform a lot of functions especially for our Immmune System. To be functionally active, a protein molecule needs to get itself to a specific $3-D$ conformation known as its native state. It's the least energetic state of the molecule. This process of transitting from a conformation to the native conformation with many intermediate conformations, is known as the folding process of the protein molecule. Sometimes, proteins are unable to fold properly, which results into malfunctioning leading to diseases like Mad Cow's and Alzheimers'. So, a knowledge of the sequence of intermediate conformations might help medical diagnosis and cure of these diseases. Thus, protein folding is a very important problem to study.

Protein have a large number of degrees of freedom [refer Section 2.1]. Between any 2 conformations of a protein molecule, there are a large number of them possible as intermediates. Thus, prediction of the path which a protein molecule might take to go from one conformation to another involves search through a very high dimensional space involving an exponentially large number of points. As mentioned in a talk [Bis11] by Dr. Somenath Biswas, IIT Kanpur, the problem of predicting the complete $3-D$ structure of a protein's most stable state given the sequence of amino acids, is an NP-Hard problem. But, it is widely observed that even large protein molecules fold into their native state within a matter of microseconds. This is a somewhat paradoxical situation [Levinthal's Paradox].
So, nature has a time-efficient algorithm which we haven't been able to find out yet. This poses a challenge for theoretical Computer Science to come up with an efficient algorithm which even a small protein molecule knows and executes.

As we can now say that protein folding study is computationally very intensive. To reduce the time consumed, we must reduce our search space. Various approaches (like ab-initio) involve a large search space and hence are

inefficient. Approaches like Genetic Programming [LYYB08], Probabilistic Road Map [Tap09] are promising as they are efficient. They reduce their search space by using randomised search and applying certain heuristics. Nevertheless, they aren't deterministic. So, till date we don't have any deterministic polynomial time algorithm to solve the protein folding puzzle.

In this scenario, the Amato Group from the Texas University has done a lot of work on this problem. They have developed efficient techniques like Map Based Monte Carlo simulation, to find out the path from the huge search space of conformations. They use a PRM-based approach [Tap09] [TTA10] to study the folding kinetics of protein and RNA molecules. In this report, we present an attempt to implement their ideas as a course project and explore more about protein folding. We implement the node generation and sampling phases of their approach. Meanwhile, we observe that the results of [BI08] about the relationship between the contact number [Section 2.2] and the potential energy of a protein molecule are verified by our work for 1BMR protein molecule.

2 Preliminaries

2.1 Protein Molecule and Its Degrees of Freedom

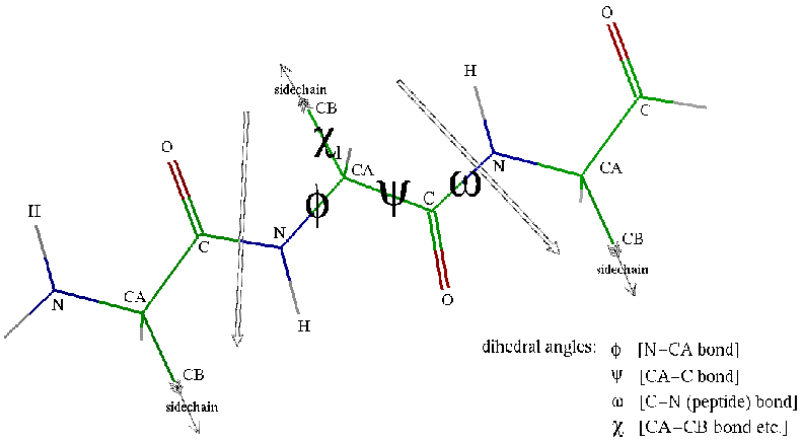


Figure 1: Amino acid model each amino acid has 2 degrees of freedom, the phi and psi torsional angles. [Uni]

A protein molecule is a long sequence of amino acids (residues) joined together by peptide linkages. Consider the figure 1. Each amino acid has N, C_α and C atoms, which in sequence form the main backbone chain for the protein molecule. During folding, the bond angles and bond lengths in the molecule remain unchanged. The dihedral angles ϕ and ψ are the ones which change. Thus, to articulate it, we can say that a protein molecule is a linkage where each amino acid has 2 degrees of freedom. The side chains (attached to C_β atom) are assumed to be spheres with 0 dof. We can think it in terms of a robotic arm with so many degrees of freedom, where we want to go from some conformation to another. Thus, some similarities between robot motion and protein folding can be realised.

2.2 Contact Number

As defined in [ADS02], a native contact is a pair of C_α atoms in the native state of a protein molecule, which are in the range of about 7 $\mathring{A}$. The contact number for a conformation is defined as the number of native contacts in it, or in other words, number of common native contacts between the conformation and the native state for the protein molecule corresponding to the conformation.

A protein molecule folds from an open chain structure to a highly folded native state in order to minimise its potential energy. Hydrogen bond interactions among the H and O atoms and the hydrophobic nature of long hydrocarbon chains attached to the main chain of protein serve as the driving force for it to fold. As the protein folds, the hydrocarbon chains come near to each other trying to shield each other from the polar solvent around. Thus, we can observe that the C_α atoms would come closer as the protein folds. As aptly described in [BI08], the surface area of the molecule accessible to the solvent decreases gradually and the molecule stabilises resulting in an overall decrease in its potential energy. A simultaneous increase in contact number may be noted.

2.3 Energy

The potential energy of a protein molecule can be defined as

$$U = \sum_{restraints} K_d\{[(d_i - d_0)^2 + {d_c}^2]^{\frac{1}{2}} - d_c\} + E_{hp}$$

where $K_d = 100KJ/Mol$, $d_0 = d_c = 2\mathring{A}$ as described in [TTA10].

The first term represents the energy decrement due to the hydrogen bond interactions, while the second term represents that due to the Van der Waals' interaction where the hydrophilic hydrocarbon chains shield each other from the polar solvent. d_i represents the distance between the atoms involved in hydrogen bonding. $E_{hp} = 20KJ/Mol$, if 2 hydrophobic hydrocarbon chains (assumed to be spheres located at C_β atoms for computational reasons) are within the range of $6\mathring{A}$ from each other. If any pair of C_β atoms are at a distance less than $2.4\mathring{A}$, then that molecule is given a very high potential energy owing to high Van der Waals repulsion.

We see that calculation of the potential energy requires the coordinates of the atoms of the molecule conformation. Suppose, all information we have is the values of the parameters which are the degrees of freedom for the molecule; in other words, we only know the dihedral angles on the main chain of the protein conformation. We need to compute the atom coordinates from that data. We implemented these calculations which involve simple geometry concepts first worked upon by [PHR+05].

2.4 Energy Landscape

A conformation for a protein molecule can be represented as an n dimensional (say) vector (n is the number of degrees of freedom). Consider a function $E : \mathbb{R}^n \to \mathbb{R}$. Consider a conformation $c \in \mathbb{R}^n$ of the molecule, $E(c)$ is the potential energy of the conformation c. An energy landscape can be thought of as a map of protein conformations ($n+1$ dimensional), with each point mapped to its potential energy. A $3-D$ map with protein conformations of 2 degrees of freedom, would look like figure 2. This landscape is funnel shaped with the native state at the bottom.

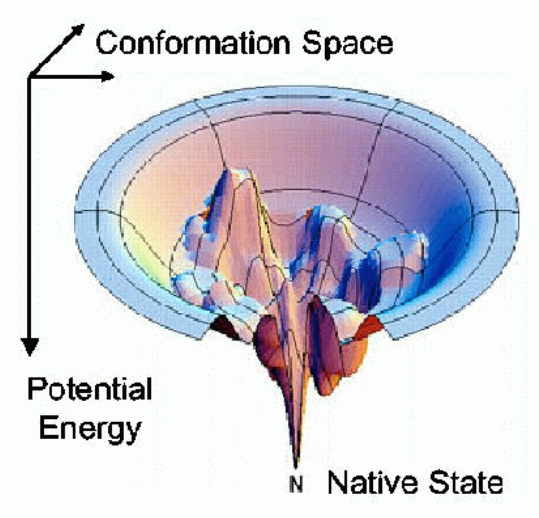


Figure 2: Visualization of the energy landscape. [Uni]

2.5 Distance Between 2 Conformations

Each protein conformation (with n degrees of freedom) is a vector of the form $c \in \mathbb{R}^n$. We can define the distance between 2 conformations as the $\mathbb{L}_2$ norm of their difference vector.

$$dist(c_1, c_2) = || \, c_1 - c_2 \, ||_2.$$

3 Methodology

3.1 Overview

We try to implement the PRM based approach discussed in Lydia Tapia's paper [TTA10]. This approach was originally developed for robot motion planning i.e., to find out the series of feasible conformations (or intermediate states) a robot may adapt to go from one conformation (or orientation) state to another. On similar basis, this can be used for protein motion planning (predicting the folding pathway). In this approach, we know the native state of the protein molecule under consideration. A large number of conformations are sampled from the molecule's energy landscape. Then, for each conformation (sample point), transitions from it to its neighbours are considered. This sample space can be represented as a directed graph $G = (V, E)$, where each $v \in V$ represents a sample point and each $(u, v) \in E$ represents the transition from conformation u to v. The edge weights represent the energetic feasibility of the transition. This graph (connected) is the Road Map for the protein molecule, where we have a large number of paths between any 2 points. Next, starting with a conformation, the native state can be reached by traversing the graph.

3.2 Sampling of Nodes

Node generation for the map is started by giving Gaussian perturbations to the native state. Next, the new nodes are used as seeds for the next random perturbation. Firstly, a large number of nodes are generated by giving Gaussian perturbance to the native state. The perturbation is chosen from Gaussian distribution centred about the seed conformation with standard deviation as $1°$. Next, with the generated nodes as seeds, perturbation is chosen cyclically from the set $\{3°, 5°, 10°, 20°, 40°\}$ of standard deviations. This method restricts the set of nodes from being completely random. Instead, it looks as if the protein molecule is being opened gradually with increasing perturbations.

3.2.1 Biased Sampling

The conformation map is desired to have enough nodes around the native state. Nodes far off (having very high energy) from the native state are discouraged to join the map. The graph should be dense around the minimum potential energy region so that the sequence of transitions leading to the native state can be observed. It is ensured by assigning probability to each sampled node, of being accepted in the map. For a conformation c with potential energy $E(c)$ the probability of acceptance to the map is defined as

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

where $E_{min} = 50000$ KJ/Mol and $E_{max} = 89000$ KJ/Mol as mentioned in [ADS02].

3.2.2 Variety in Sampled Nodes

The sampling is biased around the native state with some thresholds of energy mentioned above. It is also desirable that the map covers states spread widely around the native state. As shown in [BI08], the degree of folding of a protein molecule can be characterised by its contact number. The more the contact number, the more the protein has folded. Thus, by keeping enough number of sampled nodes for each contact number, it can be ensured that the conformation map generated will have enough width around the native state. A suitable strategy discussed in [ADS02] has been implemented with results showing a decrease in potential energy with increase in contact number.

The strategy involves keeping bins to store the generated nodes. The number of bins is proportional to the contact number of the native state (which is the maximum, by definition). Each bin can take a maximum number of sampled nodes. During each iteration of sampling, a few nodes are randomly selected from a bin as seeds for Gaussian perturbations. The sampled and selected nodes are put into the bins according to their contact number. The bin with contact number 1 less is selected for the next round of seeds. This process goes on until all (or as per user needs) bins are full. A bin is full if it has atleast N nodes, where $N \leq capacity(Bin)$ is the user choice.
The strategy discussed by [ADS02] is not an algorithm as it doesn't provide complete details about the parameters chosen. It's left upto the researcher. There is a possibility that this method won't terminate or might take a long long time (because we need to wait until the bins get full), which restricted us to be able to generate only about 1000 sample nodes, see results section 5.

3.3 Map Generation: Connecting the nodes with edges

An edge from node i to nodes j denotes the transition from conformation represented by i to that of j. The edge weights reflect the feasibility of the transition involved. Nature always favours the transition from high energy to lower energy. Thus, we keep edge weights for such transitions low. But, there are times when a protein molecule has to take a high energy conformation (intermediate) only to go into a very low energy state. Such transitions happen because of a tradeoff among the Hydrogen Bonding, Van der Waals' interaction and solute solvent interaction. Sometimes a more favourable hydrogen bond (low energy state) can only be formed by going through a highly repulsive (high energy state) Van der Waals' interaction state. Such interactions can be taken care of by allowing transitions from a lower energy state to a higher energy state; but their feasibility is low and hence the edge weights are kept high.

For each roadmap node, possible transitions are selected by considering its k-nearest neighbours. The distance function used is discussed in Section 2.5. Consider, 2 conformations c_1 and c_2 of a protein molecule. Consider the vector joining them in the conformation space ($n-D$, where n is the number of degrees of freedom). Let $a_1, a_2, a_3, \ldots, a_k$ are some intermediate conformations lying on the vector joining c_1 and c_2. The feasibility of the transition from c_1 to c_2 depends upon the feasibility of the sequence of transitions $(a_0 = c_1) \rightarrow a_1 \rightarrow a_2 \rightarrow \ldots a_k \rightarrow (a_{k+1} = c_2)$. This is considered because the conformations c_1 and c_2 may be quite far on the map generated. The nodes $a_1, a_2, a_3, \ldots, a_k$ allow us to bridge this gap (if any) and observe the transition feasibility correctly. For each pair of consecutive states (say a_i and a_{i+1}), the transition feasibility is a function of $\Delta E_i = E(a_{i+1}) - E(a_i)$. The probability of transition [TTA10] can be seen as

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

Thus, the edge (c_1, c_2) in the graph will get weight $\sum_{i=0}^{k} -\log P_i$. Thus, energetically feasible transitions get 0 weight while others get a positive weight.

Now, a graph is generated. What remains is given 2 conformations (one is preferably the native state), finding the sequence of intermediate conformation states that the protein molecule may adapt during the folding process.

3.4 A Random Walk

When the roadmap (graph) is constructed, for each node, probabilities can be assigned to all the outgoing edges based on their weights. This determines the probability with which a conformation can transit to another. According to the technique discussed in [TTA10], from the initial conformation by choosing the subsequent state by the edge probabilities assigned, the native state can be reached.

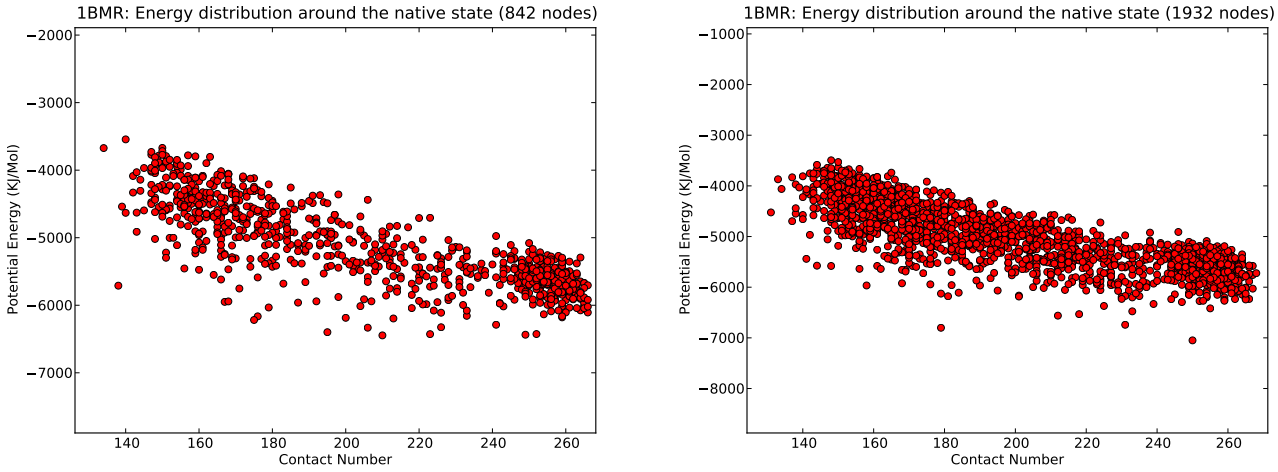
4 Data And Implementation

Data for the native state for a large number of protein molecules is publicly available on various Protein Data Bank websites in the form of .pdb files, including the WorldWide Protein Data Bank ([BHN03]), RCSB Protein Data Bank ([RU03]). We took data (.pdb file) for 1BMR protein from the Folding Results Database of the Protein Folding Server, Texas A & M University [Gro].

Among various other details, the data file contains the coordinates of all the atoms in the protein molecule in the native state. The atom coordinates facilitated the calculation of the dihedral angles of the protein molecule which are to be perturbed. Next, from the new set of dihedral angles, atom coordinates are recalculated in order to compute the potential energy of the conformation. All the distances used are in Angstroms and all the angles are in Degrees.

5 Results So Far

The sampling technique discussed, allowed us to generate data points for 1BMR protein molecule. The *Energy* vs. *Contact Number* plots indicate that with increase in the *contact number* of a protein molecule, the associated potential energy decreases. We find them more or less in accordance with the idea of [BI08]. They argue that during folding, the surface area of the protein accessible by the polar solvent around decreases which must be characterised by the contact number of the protein conformation. As a protein molecule folds, its potential energy decreases and the contact number increases.
We sampled around 900 and 2000 nodes and observed the change in contact numbers with energies. The contact number for the native state was found to be 269. See figure 3.



(a) 842 nodes

(b) 1932 nodes

Figure 3: 1BMR : Energy vs. Contact Number Plots

6 Conclusion

We conclude that the potential energy of the conformations for the 1BMR molecule around the native state decreases as the native contact number of the conformation increases. This was a result of [BI08], where they

characterise the degree of protein folding by the solvent accessible surface area which can be seen in terms of the native contact number.

7 Future Work

We could implement calculating the details for a protein conformation given the parameters; i.e. given the values of all the dihedral angles, we can calculate the potential energy and the native contact number for the molecule. We are also able to generate around 1000 sample nodes, which needs to be increased to of the order of say 10^5 nodes for the PRM approach to work properly. The main work left can be seen as:-

1. Increasing the capacity to generate a large number of sample nodes.
2. Creating the roadmap by assigning weights and transition probabilities to the edges.
3. Given an intermediate conformation, finding the sequence of states it may take to fold into the native state.

References

[ADS02] Nancy M. Amato, Ken A. Dill, and Guang Song. Using Motion Planning to Map Protein Folding Landscapes and Analyze Folding Kinetics of Known Native Structures. In *6th International Conference on Computational Molecular Biology (RECOMB 2002)*, 2002.

[BHN03] H. M. Berman, K. Henrick, and H. Nakamura. World Wide Protein Data Bank, 2003.

[BI08] N. S. Bogatyreva and D. N. Ivankov. The Relationship Between the Solvent-Accessible Surface Area of a Protein and the Number of Native Contacts in its Structure. *Molecular Biology, 2008*, 42(6):932–938, 2008.

[Bis11] Dr. Somenath Biswas. Protein Folding Challenge and Theoretical Computer Science. September 2011.

[Gro] Amato Group. The Protein Folding Server.

[HC07] Mike Harms and Marcin Cielik. pdb-tools, 2007.

[HDD11] John Hunter, Darren Dale, and Michael Droettboom. matplotlib, October 2011.

[LYYB08] Rufei Lu, Lauren Yarholar, Warren Yates, and Dr. Miguel Bagajewicz. Protein Folding Prediction. March 2008.

[PHR+05] Jerod Parsons, J. Bradley Holmes, J. Maurice Rojas, Jerry Tsai, and Charlie E. M. Strauss. Practical Conversion from Torsion Space to Cartesian Space for in-silico Protein Synthesis. *Journal of Computational Chemistry*, 26(10):1063–8, 2005.

[RU03] Rutgers and UCSD. An Information Portal to Biological Macromolecular Structures, 2003.

[Tap09] L. Tapia. *Intelligent Motion Planning and Analysis With Probabilistic Roadmap Methods for the Study of Complex and High-Dimensional Motions.* PhD thesis, Texas AM, 2009.

[TTA10] Lydia Tapia, Shawna Thomas, and Nancy M. Amato. A Motion Planning Approach To Studying Molecular Motions. *Communications In Information And Systems*, 10(1):53–68, 2010.

[Uni] Texas AM University. The Protein Folding Server.