Analysis of Protein Folding Pathways

Prof. Luis Garreta

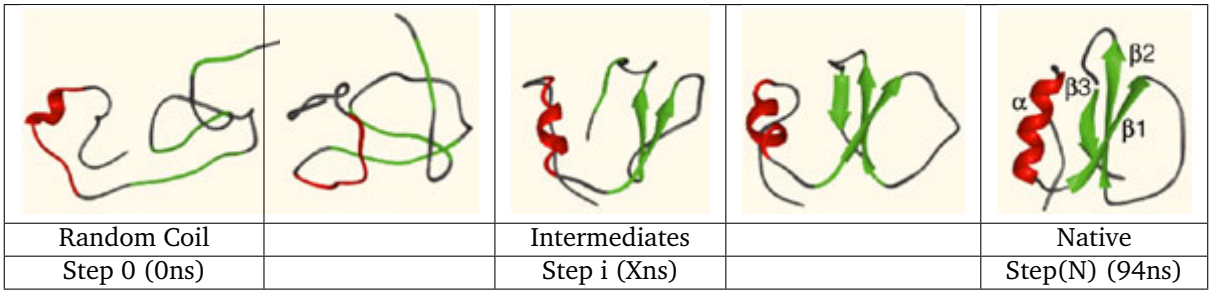
April 12, 2017

Doctorado en Ingeniería
Pontificia Universidad Javeriana
Cali - Colombia

1 Introduction

1.1 Protein Folding Pathways

Proteins fold into three dimensional (3D) structures directly related with the function they do. The folding starts with the protein in an unfolded state or *random coil* and after continuous biophysical interactions, the protein goes through different states until it reaches a stable 3D structure known as *the native state*. The sequence of states that the protein goes through is known as its *protein folding pathway.* The rules or algorithm that determines the protein folding still is unknown and it is one of the most important problems to resolve for diverse fields of knowledge, including the computational biology.



Random Coil		Intermediates		Native
Step 0 (0ns)		Step i (Xns)		Step(N) (94ns)

Figure 1. The folding pathway of chymotrypsin inhibitor 2.
(http://www.nature.com/nrm/journal/v4/n6/fig_tab/nrm1126_F1.html)

Although different theories have been formulated to explain the protein folding, not single theory explains the full process. For example, pathways theory refers to the protein following a sequence of preestablished states, from the unfolded to the folded state. This theory was relevant during the 70's and 80's but it lost importance as experiments did not show specific pathways in most proteins.

1.2 Simulated Protein Folding Trajectories

The most used technique for simulating the protein folding process is the Molecular Dynamics (MD). This kind of simulation are highly demanding on computational resources, so there are only a few complete simulations for small to short proteins. MD simulates the folding of a protein by physical movements of atoms and molecules giving detailed information of them. Each trajectory is composed by thousands of conformational structures depending of the time step of the simulation and which correspond to possible conformational states the protein passes through.

Furthermore, there are other techniques used for generating protein folding trajectories. Although they are less detailed than MD, they can generate simulations for larger proteins. For example, the Probabilistic Roadmap Method (PRM) constructs the folding pathway of a protein by destabilizing its native structure trying to create the set of possible states or structures that are part of its folding pathway. The reverse of this pathway will be the folding pathway of the protein.

2 General Objective

Establish the relation between conformational and biophysical changes of a proein during folding.

More specific objectives are:

- Run a cluster analysis on the protein conformations of various PRM simulated pathways using structural information.
- Evaluate a set of biophysical properties on the protein conformations of simulated pathways.
- Reduce the set of properties to a more workable set of variables or components.
- Run a cluster analysis on the protein conformations using the new components.
- Establish the relation of the resulting cluster between the two cluster analysis.
- Reduce the dataset of one MD simulated pathway
- Establish the relation between conformational changes and biophysical properties in the above MD pathway

3 Motivation

At present, experiments using improved devices have shown that although there is no a specific pathway, there are some "obligated" points or folding states that the protein pass through. In this sense, if we can determine these "obligated" states, we can characterize and extract knowledge of them which can be used to improve algorithms or discover new rules of protein folding.

4 Hipothesis

The folding simulation of a protein shows a pathway composed by many conformations in which the protein is changing due to the interaction of its atoms. We can cluster these conformations using the distance between pairs of conformations (using the RMSD distance) and expect that they form few clusters representing main global states. The bipopysical properties of these global states can be used to establish the characteristics of the "obligated" states.

We hypothesize that conformational changes during protein folding are directly related with the changes of biophysical properties of the protein.

5 Methodology

Our approach will be to cluster the protein conformations of many folding pathways, firstly by using only the structural information, and secondly using the information given by the biophysical properties evaluated on the protein conformations. As a result we expect that the formed clusters in both analysis look similar according to a similarity index.

We will use two datasets of protein folding simulations: the one given by PRM simulations from the Amato group; and the other given by MD simulations from the David Shaw group. It will be necessary to apply data reduction techniques for both the set of properties and the dataset of protein conformations.

Two analysis will be performed, first on the PRM simulations, characterized by short pathways (50∼300 conformations); and second on one the MD simulations, characterized by large pathways (∼1000000 conformations).

5.1 Analysis on PRM simulations

The following procedures will be performed using the PRM simulations:

1. To search for groups of conformations according to their structural information, run a clustering analysis on the whole set of protein conformations using only the structural information. Run both a hierarchical and k-means clustering with the following parameters:

   - As a distance measure use the RMSD between two conformations.
   - As the number of clusters, define four clusters

2. To characterize protein conformations, evaluate the set of fourteen biophysical properties on all protein conformations using a own toolkit for protein analysis.

3. To reduce the number of properties and search for new variables (hidden properties), run a Principal Component Analysis on all evaluated conformations.

4. To search for groups of protein conformations characterized by their biophysical properties, run a clustering analysis using the new components. Run both a hierarchical and k-means clustering with the following parameters:

   - As a distance measure use the Euclidean distance between components.
   - As a number of clusters use four clusters.

5. To compare both clustering results: structural information and biophysical properties, compute the Jaccard Index for each pair of clusters and take the best value as the corresponding cluster:

Hierarchical	**c1'**	**c2'**	**c3'**	**c4'**	**Best**
c1					
c2					
c3					
c4					

and

k-means	**c1'**	**c2'**	**c3'**	**c4'**	**Best**
c1					
c2					
c3					
c4					

   where c1,c2,c3,c4 are the clusters formed using the structural information, and c1',c2',c3',c4' are the cluster formed using the components reducing the biophysical properties.

5.2 Analysis on the MD simulation

This analysis uses one MD simulations and it follows the same steps of the previous PRM analysis but it has a preliminary step o step0 that significantly reduces the dataset ($\sim$1000000 conformations). The steps to reduce the dataset are:

1. Split the full dataset into K subdatasets (K is a input parameter, e.g. 1000). If the full dataset has N conformations, then it will be created K datasets of N/K conformations.

2. Reduce each subdataset by calculating the medoid among conformations. The medoid is as the centroid but it correspond to a member of the data set.

3. Create the new dataset with all the medoids.

4. Continue with the steps as the above PRM analysis but using as input dataset the previously created with the medoids.