Final Report
Marina Kogan's DMP Project 2004

Motivation

One of the principles of biochemistry is the relationship between protein structure and function. Since it is generally believed that structure determines function and that proteins are not rigid macromolecules, the scientific community is exploring the structural flexibility of proteins. Modeling protein flexibility is crucial to the automation of drug design since many diseases such as Alzheimer's and Mad Cow disease are attributed to protein misfolding. Adding flexibility to proteins is computationally challenging because the space of the spatial arrangements of all atoms in a protein, i.e. the space of possible protein conformations, remains large and intractable for Turing machines. However, exploration of the conformational space of a protein is critical for folding and for drug design.

There are continuing attempts to develop algorithms that produce conformational trajectories for folding or for moving the molecule from one state to another by searching the conformational space in an efficient manner. With such a deluge of conformational search techniques, it is necessary to develop quantitative methods that can rank these search techniques by comparing their respective protein trajectories. Our ultimate goal is to develop trajectory analysis techniques that can detect critical events and differences/similarities among protein folding trajectories.

Problem Statement

Our project has several layers of goals, ranked by their conceptual difficulty:

- Analyze available techniques of detecting critical events in motion and comparison of motion signatures in fields such as computer graphics, video surveillance, information theory, statistics, and bioinformatics.
- Develop new methods and test their success on toy models.

Due to a lack of available techniques for trajectory analysis, we develop a new method that can detect critical events such as bends and loops and compare among different trajectories of a simple protein model of 60 beads.

Prior Work

Needham and Boyle [3] propose some standard methods for analyzing and comparing trajectories, such as: displacement computed as the difference between respective x and y coordinates of two interest points (for equal lifetime trajectories the comparison is point to point); Euclidean distance between two points; standard statistical measures such as mean, median, standard deviation, and minimum/maximum; average displacement defined as the average value of the displacement between each pair of the compared points for spatially separated trajectories; temporal translation as the time shift that would optimally align two temporally separated trajectories; area between trajectories normalized by the average length of the paths. All theses measures are well known and applicable to many simple cases, but they fail to depict similarity/difference of trajectories for such a complex event as protein folding.

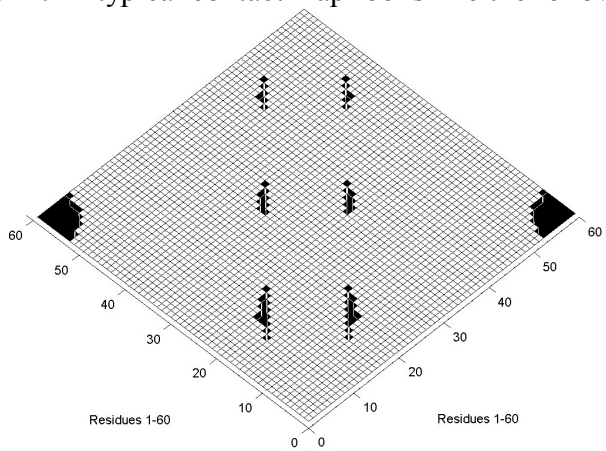
Fatih Porikli [4] proposes to use Hidden Markov Model for trajectory analysis – a probabilistic model representation of trajectories that allows comparing trajectories of different lifetimes and sampling rates, though only after optimally determining the number of states and events within those states that would properly model trajectories. The problem with this method is adequately defining the states and events emitted by each state.

Chandler et al [2] suggest using Transition Path Sampling for derivation of ensembles of trajectories for protein folding, while Vlugt and Smit [5] offer using parallel tampering to speed up the process of transition path sampling for difficult cases. Their work suggests that a transition from unfolded to folded state doesn't happen along a single trajectory, but rather an ensemble of similar trajectories. Distinguishing between ensembles of trajectories, rather than single paths, therefore, is much more applicable to the protein folding problem.

Arriving to such a conclusion after analysis of Chandler et al [2] and Vlugt and Smit [5], we resolved to develop a method of comparison between ensembles of protein folding trajectories. This method had to classify trajectories in the same ensemble as "similar", and trajectories belonging to different ensembles as "different." The method we decided to implement was inspired by the work of Best and Hege [1], which uses interatomic distances as a metric for measuring the extent to which the protein is folded. Change in the interatomic distances over time is an acceptable quantitative measure to illustrate protein's change of shape while folding into a specific conformation.

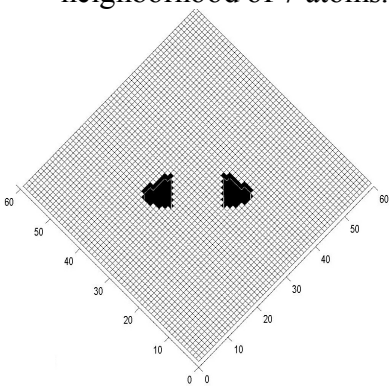
Approach

Since we wanted to detect when two different parts of the protein would come into contact with each other, we decided to abstract on interatomic distances and simplify the information of protein structure into a binary contact map. After establishing an empirical threshold (cutoff), we designate two atoms as not being in contact with each other if their interatomic distance lies above the threshold. Therefore, the only contacts resulting from a protein conformation come from pairs of atoms that are closer than the threshold to each other. As a result, a protein conformation can be modeled as a binary symmetric matrix with entries 0 or 1. A typical contact map looks like the following:

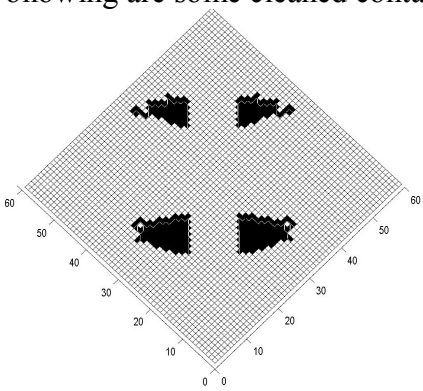


Graphical representations for our contact vectors were generated using Matlab graphing utilities

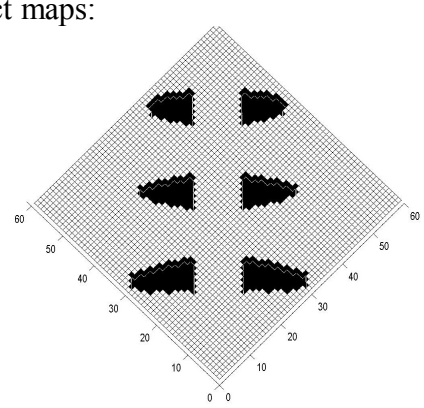
Since we compare trajectories through the contact maps for every conformation, it becomes important to filter out anything that classifies as noise from the contact maps. Considering our bond length of 0.1 angstrom, we found it appropriate to experiment with different thresholds in the range 0.05 – 2.00 angstroms until we visually determined that the contact maps revealed depressions that we associated with the presence of bends in the toy examples that we created. We decided to work with a threshold of 0.6 and filtered out local structure that was an artifact of atoms bonded with each other by ignoring a neighborhood of 7 atoms. Following are some cleaned contact maps:



1bend



2bends



3bends

After filtering out a neighborhood of 7 atoms, a discontinuity in the presence of contacts signals the forming of a bend because we attribute the presence of contacts between neighbor atoms to the local bonded structure. Therefore, contacts among atoms that are far apart in the protein sequence encode the presence of a bend. Our detection of the actual number of bends depends on counting the frequency of such discontinuities. Detecting the presence of a loop in a protein conformation involves finding contacts between pair atoms that are arbitrarily far apart in the protein sequence.

As we count the number of discontinuities in a trajectory of protein conformations, we notice that certain bends appear and disappear quickly. To get rid of unstable structure that starts to form in a protein, we only count discontinuities that persist over a window of time, which we set empirically. This makes our method robust and less dependent on the whimsical nature of bend forming and disappearing.

Experimental Setup:

The simple protein example is a string of 60 beads/aminoacids with bond length of 0.1 angstroms. We were provided with simple trajectories where the protein goes from an unfolded state to five different folded states: symmetric 1 bend, asymmetric 1 bend, 2 bends, 3 bends, and a loop.

VDW representation of our simple model through VMD

Results

We were able to produce clean contact maps that encoded the folding for all different five shapes and were able to distinguish among these contact maps using our clustering method. Detection of discontinuities in the list of contacts helps us identify the presence of a bend. The frequency of such discontinuities encodes the number of bends in a structure. We were able to correctly find the number of bends for the protein conformations we had, detect the event of a new bend forming or disappearing, and distinguish between trajectories for the simple protein of 60 beads.

Discussion and Future Work

Although our method was applied to a simple protein model of 60 beads, it is capable of detecting bends in any general contact map. This method helped us to get started on the general problem of trajectory analysis and understand some of the technical difficulties with the filtering out of unnecessary information from contact maps. Trajectory analysis is an important problem to extend to real protein ensembles and our method provides the first insights into the difficulties and ways to detect and classify the kinds of structure in proteins. This DMP project, though limited to a few weeks, will provide the starting ground for students in Computer Science and Bioinformatics in the comparison of any trajectory ensembles.

References:

[1] Best, C. and Hege, H.-C. *Visualizing and identifying conformational ensembles in molecular dynamics trajectories* .Comp. in Science and Engineering 4 (3), 68 (2002).
[2] Chandler, D., Bolhuis, P., Dellago, C., Geissler, P. *Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark.* Annu. Rev. Phys. Chem., 59, 291-318 (2002).
[3] Needham, C J; Boyle, R D. *Performance evaluation metrics and statistics for positional tracker evaluation.* in: Crowley, J L, Paiter, J H, Vincze, M & Paletta, L

(editors) Computer Vision Systems Third International Conference, ICVS 2003, pp. 278-289 Springer-Verlag. 2003.
[4] Porikli, Fatih. *Trajectory Distance Metric Using Hidden Markov Model Based Recognition.* 2003
[5] Vlugt, Thijs J. H. and Smit, Berend. *On the efficient sampling of pathways on the transition path ensemble.* PhysChemComm, 2001, 2.