Euclidean Metric for RMSD

COCS 174, Project Proposal
Milka Doktorova

April 12, 2012

1 Problem

In this project I propose to apply techniques from machine learning to design a Euclidean metric for the efficient computation of protein structure similarity.

Proteins are macromolecules composed of one or more chains of amino acids folded into compact 3-dimensional shapes. Their structure determines their function and has thus been extensively studied and used in the analysis of various functional characteristics of the macromolecules.

An important part in this analysis is the ability to determine how similar two structures are. A widely used metric is the root mean square deviation (RMSD) between the atomic coordinates of the two structures. However, RMSD computation requires initial alignment of the structures, which makes it expensive and cumbersome for use in large-scale studies.

In addition to the atomic coordinates however, a protein *conformation*, that is, the spacial arrangements of the atoms in the protein, can be represented in bond-angle-torsion (BAT) coordinates. The BAT coordinates consist of the bond distances and dihedral angles between the atoms in the molecule. As such, they fully determine the protein's structure but are independent of its particular embedding in $\mathbb{R}^3$. Therefore, the comparison of the BAT coordinates of two structures does not require the structures' initial alignment but simply computing the Euclidean distance in this representation does not yield an accurate measure of the similarity of the structures.

The goal of this project is to train a model to use the BAT coordinates of the structures in order to approximate the actual RMSD between them. In particular, I want to find a Euclidean metric based on the BAT coordinates that can be used for the accurate, i.e. RMSD-like, efficient computation of protein structure similarity. Such metric would be very useful in determining nearest neighbors in a large set of protein structures, in performing protein motion simulations and numerous other applications based on protein structure comparison.

2 Approach

I read the paper by Weinberger et al. published in NIPS in 2006 and called "Distance Metric Learning for Large Margin Nearest Neighbor Classification". The approach that they describe is based on semidefinite programming but I am not sure how applicable it will be to my problem since I want to approximate the actual RMSD values as opposed to learn a metric that correctly identifies nearest neighbors.

Another paper that describes a similar distance learning approach based on convex optimization is "Distance Metric Learning, with Application to Clustering with Side-Information" by Xing et al. published in NIPS in 2002. It shows how to learn a metric that respects the relationships between given data points that have been pair-wise classified as similar or dissimilar. Again however, this is more related to classification rather than learning how to compute exact values.

In a different study from 2007 Weinberger and Tesauro describe a metric learning for kernel regression that sounds more applicable to my problem.

3 Data

A protein consists of a backbone of carbon and nitrogen atoms, and side-chains of amino acids attached to it. In order to test the feasibility and applicability of the approach, I am going to start with segments with length 50 of just the backbone atoms, without considering the side-chains. All structures for this experiment will be taken from the Protein Data Bank, a main online repository of experimental protein structures.

For the training set I am planning to select 1000 segments. Each segment will be taken from a different structure and its particular location within the structure's backbone will be chosen at random. I will compute all pair-wise RMSD values for the 1000 segments, as well as each segment's BAT coordinates. Since the bond distances in the BAT coordinates in this case are constant, I am going to exclude them from the computation and instead, consider only the dihedral angles between the atoms in the segments. I will then train a model to use these BAT coordinates in order to approximate RMSD values. I will test the model on a different set of segments that does not intersect with the training set (and for which I will also compute the pair-wise RMSDs in order to be able to test the accuracy of my model).

When selecting the segments for the training set I will try to choose them in such a way that results in a nice distribution of the RMSD values.

4 Plan

By the project milestone due date I plan to have all my data collected (both for the training and testing stages), as well as found, and maybe implemented, the best approach for the problem.