**Introduction to Bioinformatics**

**Fall 2016**

**Programming Project**

**Reduced Representation of a Protein Structure**

Assigned 11/14/2016

Due date see individual task

**Cluster of atoms**

There are three purposes in this project. (1) To be able to work with actual protein structure data from the database; (2) To be able to perform simple calculation using the data; (3) To realize that complex data can be simplified for certain purposes.

**Task 1 – Parsing a protein structure extracted from the Protein Data Bank (PDB)**

**(due 11/22/16)**

The input is a protein structure file in the format of PDB. Your program will take the name of the PDB file given by a user and output a file of the alpha-carbon (CA) information of the first chain of the protein. The output is in the PDB file format so that you can load and see in Chimera. You may just extract the lines of CA atoms in the input file.

**Task 2 – Reduced representation of side chains, helices and beta-strands**

**(due 12/7/16)**

Once you are comfortable extracting information from a PDB file, you are ready to calculate reduce representation.

The reduced representation of a side chain is the geometric center of side chain atoms. Side chain atoms are those atoms other than backbone N, CA, C, O and H (if available) in an amino acid. A few examples will be given in classes.

The reduced representation of a helix is a set of points near the central axis of the helix. Utilizing the natural of the spiral in the helix, an axis point can be obtained from the geometric center of four consecutive points of the helix, where *CA* is an alpha-carbon.

The reduced representation of a beta-strand is a set of points near the central line of the beta-strand. A point of the central line is defined as the geometric center of four consecutive backbone atoms (without O). As an example, a point can be the geometric center of , , , and where , , , and the *i*-th nitrogen, alpha-carbon, carbon, and *i+1-*th nitrogen atoms respectively on the beta-strand. Alternatively, the geometric center of , , , and can also be a point on the central line.

Your programs are expected to perform the following tasks.

1. Take an input file from a user in the format of PDB.
2. Output the side chain centers in a PDB file using “S” as a label for the atom. This allows you to load the file for visualization in Chimera.
3. Output a PDB file of those points in axes of helices. You may also use “S” as the label for the point in the PDB file.
4. Output a PDB file of those points in the central lines of beta-strands. You may also use “S” as the label fo the point in the PDB file.

**Task 3 – Extra-credit**

**(due 12/7/16)**

Use Hermite spline to interpolate those points on the axis of each helix. Use Hermite spline to interpolate those points on the central line of a beta-strand. The output files are in PDB format.

**Report**

A brief report of how to run your program and what test you have performed should be included. I will test your program using 10 proteins randomly chosen from the PDB.

**Suggested Testing Data**

You may consider to use the following PDB files to test your programs.

1AKY

1AOP

2P8Y

1S04

1ATZ

2QTR

1D5T

1ELU

1A12

1B3A