Due Date: December 25, 2016 (23:55)

CENG 465
Introduction to Bioinformatics

Fall 2016-2017

Assignment #3

Programming Assignment on Protein Structures

Finding a seed structural alignment using intra-molecular distance matrices

In this assignment your goal is to implement a method to find the most similar 20 amino-acid long similar substructures of two protein structures using their intra molecular distance matrices. Your program will input two input protein structures and report two integers which are the starting amino-acid indices of the length-20 seed match in the respective proteins structures. You may use a helper program, which will be provided, to inspect the RMSD of this seed alignment, too.

Here is a step by step description of the algorithm you are expected to implement:

1) Get two protein structures as input from user as PDB (Protein Data Bank) files. You may assume that the PDB files will contain single chain protein structures with amino acid indices starting at 1.

2) Read the PDB files and represent each protein as a sequence of 3D points with (x,y,z) coordinates, one coordinate for each amino acid, which is the coordinate of the alpha carbon of that amino acid.

3) Compute two intra-molecular distance matrices for these two structures by computing the Euclidean distance between all amino acids of a protein structure.

4) Find the most similar 20x20 sub-matrices between the two intra-molecular distance matrices, by sliding 20x20 windows over both of the intra-molecular distance matrices. You may use the L1-distance metric to find the distance between two 20x20 matrices. In other words, in L1-distance, the distance between two 20x20 matrices A and B is given as:

$$d(A,B) = \sum_{i=1}^{20} \sum_{j=1}^{20} |A_{ij} - B_{ij}|$$

5) Your program should find the 20x20 sub-matrix pair with the minimum L1-distance and report the starting indices of the amino-acids for these sub-matrices.

6) For example, given two imaginary protein structures as 1ABC.pdb and 2XYZ.pdb, your output should look like:

Seed Alignment found at:
==========
Prot1: 1ABC 86
Prot2: 2XYZ 23

7) If you want to check how good the seed alignment is, you may use the Java program given at the link below with the following parameters:

java superimpose 1ABC.pdb 2XYZ.pdb 86 23

http://www.ceng.metu.edu.tr/~tcan/ceng465_f1617/Assignments/superimpose.zip

Additional Information:

The details of the PDB format can be found at:

http://www.wwpdb.org/documentation/format33/v3.3.html

However, you will only need to read the ATOM record of PDB files. The ATOM record contains the coordinates of the atoms that make up the structure. For each amino acid, you are only going to use the CA atom (alpha-Carbon) coordinates. The atom records look like below:

```
ATOM      2  CA  SER A 217       9.923  23.155  -3.178  1.00 40.91           C
ATOM      8  CA  SER A 218       8.001  22.803   0.087  1.00 38.93           C
ATOM     14  CA  GLY A 219       4.872  20.798  -0.806  1.00 30.77           C
```

The atom type of alpha-Carbon is indicated as CA in the third column. The (x,y,z) coordinates are the first triplet of floating point numbers. For examle for the first SER amino acid the CA coordinates are (9.923, 23.155, -3.178). All you need to read for each amino acid are these CA coordinates. You may find example PDB files at the Protein Data Bank web site:

http://www.rcsb.org/pdb/home/home.do

You are free to use any programming language to develop the required program. You are also free to use any online resource that you can find on the Internet.

We will not provide any example outputs. However, you are free to share your outputs with your friends in the ODTU-Class forum.

Submission

Submit your program (source code only) via ODTU-Class before the deadline. Late submission is -20 pts per day.