Due Date: December 29, 2015 (23:55)

CNG 465
Introduction to Bioinformatics

Fall 2015-2016

Assignment #3

Programming Assignment on Protein Structures

Structural Alignment by Extending a Seed Alignment

In this assignment your goal is to implement a new structural alignment algorithm described below. Your program will take two PDB (Protein Data Bank) files as input protein structures and report the discovered local alignment and how similar the aligned parts are.

Here is a step by step description of the algorithm:

1) Get two protein structures as input from the user as PDB (Protein Data Bank) files.

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. In other words parse only the lines that start with the "ATOM" keyword and has "CA" as the atom name. Put all these coordinates of CA atoms in an array in the order they appear in the PDB file.

3) Find a seed structural alignment between the two protein structures by sliding a window of size 10 over each protein structure. In other words, you should check for each possible stretch of 10 amino acids in both structures and superimpose them to see how similar these two local structures are. You may use the code at the following address or you may find other sources on the Internet for superimposition of these windows of size 10:

http://www.ceng.metu.edu.tr/~tcan/ceng465_f1314/Assignments/superimpose.zip

The superimposition algorithm need the Singular Value Decompostion of a 3x3 matrix. This is accomplished by using the SVD function in the Jama library provided in the zip file above. You may use other implementatins, in C, Matlab, etc. if you want.

4) You will select the best pair of local structures (both of which size 10) which are most similar to each other, i.e.,with minimum RMSD, as the seed structural alignment.

5) You will extend the seed alignment to the left and to the right simultaneously (one amino acid to the left and one amino acid to the right and then iterate, another amino acid to the left and another amino acid to the right, etc) and check the RMSD of this extension. In other words, the seed alignment will grow 2 amino acids at a time, length 10,12,14… etc. You will run the superimposition method for the extended part to find the RMSD again. You will continue this extension process as long as the structural alignment score (defined below) does not decrease (in other words, keep extending as long as the score increases or stays the same). The score of the extension should be checked two amino acids at a time. For example, add one amino acid to the left and one amino acid to the right and then check the score, if it is same or larger than the previous score, continue adding another pair of amino accids to the left and to the rigt and check the score, and so on.

Structural alignment score = Length of alignment / RMSD

So the initial score of the alignment will be equal to *10 / RMSD of the best pair*, and the final alignment you report will not have a lower score than this initial score.

6) Report the final alignment similar to the example output given below:

```
Alignment results:
=====================
Alignment length:  45
Aligned amino acids:
Prot1: 1ABC    86-130
Prot2: 2XYZ    23-67
RMSD: 2.3
Alignment Score: 19.565
```

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 one type of record in the PDB files: the ATOM record. 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. Also note, that the amino acid numbers given in the ATOM record are 217, 218, 219 which may not match the indices in your CA coordinate array. In your alignment result, you may report your array indices as the aligned parts, i.e. the order of amino acids in the PDB file as you read the ATOM records (starting from 0). Input protein structures will be single-chain proteins. Just read all the ATOM records in the PDB file and construct a single vector/array of coordinates as your read the ATOM records for the CA atoms and use the order in this vector/array when reporting your alignment.

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 provide some example outputs in ODTU-Class. You are also free to share your outputs with your friends.

Submission

Submit your program (source code and executable) with a README file, which describes how to run your program, as a zip bundle via ODTU-Class before the deadline. Late submission is -20 pts per day.