Due Date: Sunday, May 10, 2020 (23:59)

CENG 465
Introduction to Bioinformatics
Spring 2019-2020

Assignment #3

Programming Assignment on Protein Structures

Finding interaction hotspots of dimers

In this assignment, your goal is to process a PDB file of a protein structure that contains two amino acid chains and find which residue pairs from different chains physically interact in the quaternary structure of that protein.

Specifically, you will read and parse a given input PDB file of a protein structure and consider only the ATOM records. For more information on the PDB format you may refer to:

http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html#ATOM

The format uses fixed column (i.e., character) positions for different information provided in a single ATOM record (i.e., on a single line).

You will determine the position of each amino acid by its beta Carbon (CB) atom's coordinates only (You will use the alpha Carbon (CA) for Glycine, because it does not contain beta carbon). For every pair of amino acids from two different chains of the protein, you will compute the Euclidean distance between them and identify two amino acids as interacting if the distance between them is less than 8 Angstroms. Find all the pairs of interacting residues between the two chains. Note that the pairing may not be one to one. In the next step, you will group interacting pairs of amino acids to determine interaction hot spots. If the closest distance between two interacting residue pairs A-B and A'-B' (A and A' is from the same chain, B and B' is from the same chain) is less then 8 Angstroms, i.e., if either A-A', B-B', A-B', or A'-B is also less then 8 Angstroms, A-B and A'-B' interaction pairs will be grouped together. A single pair of interaction in a group will be sufficient to add a new interaction to that group based on this rule. Report the discovered interaction hot spots as groups of interaction pairs. Note that an interaction hot spot will not share any interacting amino acid pair with another interaction hot spot (if they did, they would be a single group).

Run your program on the 6 example protein structures provided at:

http://www.ceng.metu.edu.tr/~tcan/download/assignment3_proteins.zip

Report the hot spots (i.e., groups of interacting amino acid pairs) you have found for each protein structure in a short report. For example, like below for the protein 4uap.pdb:

```
There are 39 interacting pairs.
Group 1: THR(6)-SER(16)
Group 1: SER(8)-ASP(10)
Group 1: SER(8)-SER(11)
Group 1: SER(8)-SER(12)
Group 1: SER(8)-LEU(14)
Group 1: GLY(9)-GLY(9)
Group 1: GLY(9)-ASP(10)
Group 1: GLY(9)-SER(11)
Group 1: GLY(9)-SER(12)
Group 1: ASP(10)-SER(8)
Group 1: ASP(10)-GLY(9)
Group 1: ASP(10)-ASP(10)
Group 1: ASP(10)-SER(11)
Group 1: ASP(10)-SER(12)
Group 1: SER(11)-SER(8)
Group 1: SER(11)-GLY(9)
Group 1: SER(11)-ASP(10)
Group 1: SER(11)-SER(11)
Group 1: SER(11)-LEU(20)
Group 1: SER(12)-SER(8)
Group 1: SER(12)-GLY(9)
Group 1: SER(12)-ASP(10)
Group 1: SER(12)-GLU(55)
Group 1: SER(12)-THR(57)
Group 1: LEU(14)-SER(8)
Group 1: SER(16)-THR(6)
Group 1: LEU(20)-SER(11)
Group 1: GLU(56)-SER(12)
Group 1: THR(58)-SER(12)
Group 1: LYS(126)-ASP(48)
Group 1: TYR(95)-GLN(47)
Group 1: TYR(95)-ASP(48)
Group 1: GLY(98)-GLN(47)
Group 1: GLY(98)-ASP(48)
Group 1: GLU(100)-GLN(47)
Group 1: ASP(48)-LYS(124)
Group 1: GLN(47)-GLY(97)
Group 1: ASP(48)-TYR(94)
Group 1: ASP(48)-GLY(97)
Group 1: THR(6)-SER(16)
Number of groups = 1
```

The output should contain which group each interacting pair belongs to and the 3 letter amino acid codes and their order on the chain as shown above. The order of reported amino acid pairs is not important.

Submission

Submit your code and your report, which contains the results for the example 6 PDB files, as a single ZIP bundle via ODTU-Class before the deadline. The deadline is not subject to postponement. Late submission is -10 points per day.