Lab - Protein-Protein Interfaces

Many biological functions involve the formation of protein-protein complexes. A complex is a structure consisting of two or more interacting proteins. The interface of interacting proteins is defined as the set of amino acids of the two proteins within a certain threshold distance. The characterization of protein-protein interfaces is important because it will enable the prediction of protein interactions providing insight into their function.

This project is about the determination of the amino acids which are at the interface of two interacting proteins. In this project, given two interacting chains A and B of a protein complex, you compute the Euclidean distance between every amino acid of chain A and every amino acid of chain B. If the distance of two amino acids is less than a certain threshold, both amino acids are registered as the interfaced amino acids. Generally, he threshold is selected arbitrarily by trial and error.

You have to write a program that computes protein-protein interfaces and analyzes them. It takes in input the name of a pdb file of a protein, the identifiers of two chains (for instance, A and B), and a threshold value. In a pdb file the interacting proteins in a complex are listed as separate chains.

The project consists of three parts.

Part 1. (40 points) In three dimensional space the Euclidean distance *d* of two points P1(x_1, y_1, z_1), and P2 (x_2, y_2, z_2) is defined as:

$$d = \sqrt[2]{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

You compute the Euclidean distance of atoms in different chains. To simplify the computation, we only consider C_alpha atoms, denoted by CA in the pdb files. Then for each CA atom of the first chain A you compute its distance from each CA atom of chain B. If such a distance is less than the input threshold then the pair of CA atoms is considered at the interface of the two chains and you output both the CA atom IDs and the corresponding amino acids. (see implementation section)

Once you have determined the interface, you do the following:

Part2. (20 points) For each chain, you compute the fraction of the interface amino acids lying on the secondary structures alpha helices and beta sheets. In other words, you determine the number of interface CA atoms lying on secondary structures over all interfaced CA atoms.

Part3. (20 points) For each interface amino acid of a chain, determine the closest interface atom of the same chain to the right in the primary sequence. Then determine the difference in position of the two amino acids in the sequence. For example, assume the sequence below is a fragment of the primary sequence of a chain starting at VAL numbered 33, and assume that you have determined that PRO, ALA and ALA are all interface amino acids,

33 34 35 36 37 38 39 40 41 42 43 44 45 .....
VAL LEU SER PRO ALA ASP LYS THR ALA VAL LYS ALA ALA .....

Then the closest interface amino acid of PRO is ALA and they are at distance 1, the closest of ALA is LYS at distance 2, and of LYS is ALA at distance 6.

Part3. (20 points) Visualization. Use JMOL or any other tool to visualize the two chains and their interface. You may do so by displaying the entire protein and then selecting a set of amino acids (those at the interface) and color them in a different color or display them in a different style. If you use JMOL you select the amino acids by using a console command, as explained at:

http://people.virginia.edu/~dta4n/biochem503/console.html

Implementation and Submission.

Submit a file with name "your_last_name_interface.ipynb".

Test your program on protein 1atp, chains E and I, or any other protein of your choice. Chain I is small so that your program will run fast. But try also with 4HHB chains A and B. As a threshold for the distance between two amino acids to be considered at the interface use 7 or 6 Angstrom, or choose your own threshold. You should specify the name of the pdb file, chains and threshold in your script.

Output.

Part1:

For output, you will report each pair of interfaced amino acids from two different chains in the following format:

<CHAIN_ID>:<AA CODE>(<AA NUMBER>) interacts with <CHAIN_ID>:<AA CODE>(<AA NUMBER>) as for instance in the line below

A:LYS(255) interacts with B:LEU(353) (This is an example not a real correct part of the output).

……

Part2:

Print the fraction to console:

<CHAIN_ID>

<Fraction> of the interface amino acids lying on alpha helices.

<Fraction> of the interface amino acids lying on beta sheets.

Part3:

Then for each chain and each interface amino acid you print:

<CHAIN_ID>

PRO: closest ALA at distance 1.

ALA: closest LYS at distance 2.

LYS: closest ALA at distance 6.