**Due Date: June 5, 2022 (23:59)**

# CENG 465
# Introduction to Bioinformatics

## Spring 2021-2022

## Assignment #3

Programming Assignment on Protein Structures

## Finding Residues in Contact in a Protein Structure

In this assignment, your goal is to implement a method to find amino acid pairs that are in close proximity in the structure (i.e., in three dimensions) but far away in the sequence. In particular, your program will get input a protein structure (which may be composed of multiple amino acid chains) as a single PDB file and two numbers, one integer, S, to be used as the distance threshold in the sequence and a real number, D, to be used as a distance threshold in the stucture. You willl then report amino-acid pairs (for each chain of the structure) whose distance in 3-dimensions is less than or equal to D, and which are at least S amino acids apart in the sequence (i.e., if one amino acid is the $i^{th}$ amino acid, the other has to be $i+S^{th}$ or more, or vice versa).

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

**1)** Get a protein structure as input from the user as a PDB (Protein Data Bank) file.

**2)** Read the PDB file and get the sequence of each chain from the SEQRES records and the coordinates of each aminno acid from the ATOM records. Use the alpha carbon coordinate of amino acids as their coordinates.

**3)** For each chain, compute pairwise distances, as the the Euclidean distance, between all amino acid pairs in that chain. Output amino acid pairs whose distance in 3-dimensions is less than or equal to D, and which are at least S amino acids apart in the sequence.

**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 and SEQRES records 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:

https://www.rcsb.org/

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 Student Discussion forum.

## Submission

Your program should be run from the command line similar to the format below:

```
> hw3 <pdb file name> <D> <S>
Or
>python hw3 <pdb file name> <D> <S>
```

E.g.,

```
> hw3 1ABC.pdb 4.5 20
```

You may assume that the PDB file input will be in the same directory as your executable

Submit your program (source code only) via ODTU-Class before the deadline. Late submission is -15 pts per day.