Due Date: Sunday, June 13, 2021 (23:59)

CENG 465
Introduction to Bioinformatics
Spring 2020-2021

Assignment #3

Programming Assignment on Protein Structures

Finding the diameter of a protein structure

In this assignment, your goal is to write a program (in a language of your choice) which will process any given protein structure provided as a PDB file and find the *diameter* of the protein structure as the longest distance between any two alpha Carbon atoms in its backbone.

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 only consider the alpha carbon atom (indicated as CA in the "Atom name" field of the ATOM record) coordinates. The protein structure may contain multiple protein chains and you will read all alpha carbon atom coordinates from all the chains into a single list of coordinates. You will then find the farthest pair of alpha carbon atoms and compute the *diameter* of the protein structure as the Euclidean distance between these farthest alpha carbon atoms and report that distance in Angstroms. You will also be required to report additional information on these amino acids that define the diameter, such as their "Residue name", "Chain identifier", and "Residue sequence number" as indicated in the 4th, 5th, and the 6th fields of their ATOM record. See the PDB file format description referred to in the above link for more information about the ATOM record.

Example outputs

Report the amino acid pair in the order that they appear in the PDB file: the first one that appears in the PDB as the first amino acid in the information line, for example, like "Between <first aa> and <second aa>". Use three digits after the decimal point when reporting the diameter.

Input:
7dwb.pdb
Example run on commmand line:

```
$hw3 7dwb.pdb
Diameter = 137.409 Angstroms
Between A:SER(265) and F:ASP(387)
```

Input:
2wfi.pdb
Example run on commmand line:

```
$hw3 2wfi.pdb
Diameter = 41.852 Angstroms
Between A:GLN(6) and A:ALA(161)
```

Test cases

Run your program on the 10 protein structures provided in the following link:

http://www.ceng.metu.edu.tr/~tcan/ceng465_s2021/hw3_proteins.zip

and report the outputs of your program in a single TXT file in the following format:

#<test case number>
<pdb file name>
<output of test case>
#<test case number>
<pdb file name>
<output of test case>
…..
…..

Please follow this format strictly and do not forget to report theinformation on amino acid pairs in the order that they appear in the PDB files. Your output txt for the two examples given above should look like:

Content of output.txt:

```
#1
7dwb.pdb
Diameter = 137.409 Angstroms
Between A:SER(265) and F:ASP(387)
#2
2wfi.pdb
Diameter = 41.852 Angstroms
Between A:GLN(6) and A:ALA(161)
```

It is guaranteed that the pair of amino acids that define the diameter is unique in each of the test cases.

Submission

Submit your source code and your output txt file as a single ZIP bundle via ODTU-Class before the deadline. The deadline is not subject to postponement. Late submission is -15 points per day.