ECS 129
Assignment: Option3 (Programming)
Protein Structure Prediction

Due: Wednesday, March 2nd, 2022

Protein geometry

Predicting the structure of a protein remains a formidable task. However, in the last three years, AlphaFold and its current version, AlphaFold2, have proven to be quite successful in solving this challenge, as demonstrated in successive protein structure prediction challenges (CASP). In this assignment, you will:
- Predict the structure of two protein sequences, using AlphaFold
- Write a program that allows you to compare the results of AlphaFold with the corresponding ground truth structures available in the PDB
- Perform those comparisons using your program, and discuss them.

The two protein sequences

**Sequence 1:**
> Fimbrial adhesin|Proteus mirabilis (strain HI4320) (529507)
> SIFSYITESTGTPSNATYTYVIERWPETSGILNPCYGWPVCYVTNKHHTVNGTGGNPA
> FQIARIEKLRLLAEVDVVLKNRSFPIEGQTTHRGPSLNSNQECVGLFYQPNSSGISPRGK
> LLPGSLCGIAPP

**Sequence 2:**
> CST complex subunit CTC1|Homo sapiens (9606)
> AISQAIIRLLLVEDGTAEAVVTCRNHHVAALGLCPREWASLLD

Computing the RMSD to compare two protein structures

You will write a program that computes the Root Mean Square Deviation (RMSD) between the CA atoms of two protein structures. While this will be studied in class, your implementation will follow the paper by Coutsias, Seok, and Dill, “Using quaternions to compute RMSD” available on the web page.

Notes:
- You can use the computer language you want (C, C++, Java, Python, R, Matlab among others)
- You may need a library to compute the eigenvalues / eigenvectors of a real symmetric
matrix. Such libraries are readily available in most computer languages.

- You may, or may not implement the computation of the rotation matrix that corresponds to the optimal RMSD; your choice!
- Your program will basically read two files (one for each structure to be compared), isolate the CA for each structure, compute the RMSD using the algorithm on page 1855 of the paper, and outputs the RMSD.

**Predicting the structures of the two protein sequences.**

You will use either AlphaFold2, available at:

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

(see https://www.youtube.com/watch?v=mTjYvI3KCY for how to use it)

or RosettaFold, available at


or both!

**Comparing with the gold standard**

Sequence 1 corresponds to the protein structure identified as 6YAF chain A from the PDB (www.rcsb.org)
Sequence 2 corresponds to a fraction of the protein structure identified as 1w6w chain B from the PDB.

For convenience, I have provided the corresponding PDB file on the web page.

*Please provide both the source code of the program you wrote, and a report describing the results. There is no need to send a lengthy write-up, but it should definitely include an introduction, results and analysis, a conclusion, and references to published work, if needed.*

**Good Luck!**