CMSC 12200 Project Proposal

Name of the Group and Members

Group name: lyz
Members: Jason Zhang, Tuoyuan Li, Ruilin Yang

Description of Projects and Goals

In protein biochemistry research, evolutionary conservation of amino acid residues in a protein offers valuable insight to researchers. Typically, more evolutionarily conserved residues are more important to the structure and function of a protein, and are hence more relevant to research. However, generating a conservation table and visualizing the data is a laborious process involving data selection, processing and manual manipulation of protein visualization, often spanning multiple softwares and databases. In this project, our team aims to design a user-friendly software using python that combines these tasks into one simple input and output. The program will automatically collect data, generate a Multiple Sequence Alignment, calculate the conservation, and map the conservation to a three-dimensional protein structure.

Background

Proteins are large biomolecules that serve diverse functions in all living organisms. They are made of amino acids, which form long chains that fold into complex three-dimensional structures. Each individual amino acid is called a residue. Each amino acid chain has an N-terminus and a C-terminus, and residues can be numbered from one starting from the N-terminus.

Throughout evolution, random mutations occur in the DNA, which can alter the sequence of the protein. When this happens,there can be one of two outcomes. Either the organism dies, which means the mutation will not pass on to its offspring, or the organism survives, which means the change in protein sequence will remain in all its descendants. Intuitively, if a residue important to the function of the protein is mutated, the first outcome is more likely. Conversely, the second outcome is more likely for a mutation in a less important residue. Over time, the residues that see fewer changes

(more outcome one) throughout evolution are said to be evolutionarily conserved, and this quality can be quantified given a large dataset. Consequently, evolutionary conservation is used as a proxy to identify residues of research value.

There are two important databases used for protein research. Protein Data Bank (PDB) hosts proteins whose three-dimensional structure is determined by X-ray crystallography or cryo-EM. UniProt is a large database with curated information on many genes and proteins, including those whose structure is not known. Each protein entry in PDB would have a link to its page on UniProt, but not vice versa. We will be using both databases for our project.

The sequence of a protein is often coded in a FASTA file. FASTA files contain the name of the protein, the species from which the protein comes from (if available) and a unique identifier code, in addition to the amino acid sequence. A Multiple Sequence Alignment (MSA) can be generated from many FASTA sequences of similar proteins. The best practice for the filtering of FASTA sequences and the construction of MSAs using various new algorithms is a topic of active research in protein bioinformatics, and is beyond the scope of this project. Instead, we will generate MSA using all the available data and the classical Clustal Omega algorithm.

An important component of our project is visualization, since we have limited our protein to only those with known structures. We will be visualizing our protein in a program called PyMol, a Python-based visualization software with excellence compatibility with data from PDB as well as Python.

Data Source

- We will select proteins from the Protein Data Bank (https://www.rcsb.org/) which hosts protein structure
- We will obtain information about amino acid sequences as well as other relevant information from UniProt (https://www.uniprot.org).

Task List and Timeline

* Texts colored in red indicate challenging part of the project

Week4:

- Select a protein from the PDB (protein database) (https://www.rcsb.org/)

- Find the UniProt identification code (e.g. P19812, unique for each protein entry)
- Open the UniProt page of that protein entry

Week 5:

- Find the FASTA sequence of that protein entry
- Find the name of that protein and the organism it came from
- Search on the UniProt database of similar proteins

Week 6:

- Narrow down the search result list using user's parameters
- Download the FASTA sequence for all of the proteins that satisfy the user requirement
- Store and process the FASTA sequence in a Pandas data structure

Week7:

- Importing Python packages and uses them to generate a Multiple Sequence Alignment (MSA)
- Using the multiple sequence alignment, calculate a useful parameter about the protein
    - We will calculate the Conservation of the amino acid residue by default
    - We may calculate more advanced parameters (Statistical Coupling Analysis) if time permits

Week 8:

- Using Python packages, we will map the conservation results (or other parameters we calculated) onto a PyMol protein structure obtained from PDB

Week 9:

- Review and testing

Week 10:

Final presentation