Project Name: Protein Dynamics

Group Member: Haotian Tang, Xinyao Yin, Yuxiao Tian

Group Number: 23

Supervisor Name, Email, Organization: N/A, N/A, N/A

Git version control: https://github.com/XxTxXTxX/CapstoneProject/tree/main/meeting

Contributors: Yuxiao Tian, Xinyao Yin, Haotian Tang

Contributions: Xinyao Yin contributed to the project's concept and provided biological background information, Haotian Tang confirmed the information, and Yuxiao Tian contributed the descriptions.

Project Description:

Our goal is utilizing machine learning to **predict a protein's three dimensional shape** given its **amino acid sequence** and various **environmental factors**, such as the pH level surrounding the amino acid when protein is formed, temperature, and cellular location. There environmental factors play a critical role in determining how a protein folds and functions in biological systems. By incorporating these variables, we seek to enhance the accuracy of protein structure predictions beyond what is achieved by sequence information alone, like in the state-of the art software, AlphaFold. More importantly, our approach will provide researchers with the possibility of environmental interactions to enable them to observe how those environmental factors can influence protein generation, which is not present in AlphaFold.

In addition to the predictive model, we will develop and deploy a **web-based user interface** that allows researchers to generate protein structures by simply inputting an amino acid sequence. This tool will provide a streamlined platform for scientists to explore protein folding in various environments, offering valuable insights into protein behavior in different physiological conditions.

When a protein fails to fold correctly due to environmental factors that lead to gene mutations, it can lose its functionality, potentially leading to diseases. Disorders such as **Alzheimer's** and **Parkinson's** are closely associated with **protein misfolding and aggregation**. In the field of **drug design**, researchers must often engineer proteins with precise functions, which requires a deep and accurate understanding of the protein folding process, thus the **origin of this project**.

The success of this project will provide researchers with accurate protein structure predictions, which can significantly accelerate **drug discovery**, and also allows for the development for more effective and targeted disease treatments. Additionally, by incorporating **environmental factors** such as pH, temperature, and cellular location, our model will offer deeper insights into how proteins behave under different physiological conditions. This comprehensive approach will not only aid in **drug design** but also in **understanding protein misfolding** diseases like **Alzheimer's** and **Parkinson's**, potentially opening new avenues for therapeutic interventions. The web-based platform we develop will serve as a valuable tool for scientists, enabling them to generate and explore protein structures efficiently, thereby advancing both **basic research** and **applied biomedical science**.

We will be using Python and PyTorch to build a machine learning model, and javascript, HTML, CSS, Django to build the web-based interface that allows users to input amino acid sequences and retrieve protein structure predictions, which also containing database system to store user input and output. We will also use Biopython for parsing protein data (PDB format) and handling bioinformatics operations. Finally, PyMOL will allow us to visualize protein structures and 3D folding and NGL viewer allows for the same functionality in the web interface. RCSB Protein Data Bank (PDB) is the primary source for 3D protein structure data. PDB entries often provide crystallization conditions, including pH and temperature, which are essential for training the model. UniProt provides detailed amino acid sequence and functional data. We will cross-reference UniProt data with PDB entries to ensure accurate sequence-to-structure mapping.