Artificial Intelligence and Machine Learning (AIML) – Project

Names: **Sec: 2**

Mahalaya Talluri – 2320040027, Sec A

V. Divya Sri – 2320040141, Sec B

Problem Statement: Calculate protein aggregation in a given protein sequence.

Proteins, the complex molecular machinery that holds together the structures of life, rely on their three-dimensional structures to function. However, experimental protein systems are time-consuming and expensive. Our work uses reproduction models to generate 3D protein structures directly from amino acid sequences.

Dataset:

1. **Title:** Protein Secondary Structure
   Source: Kaggle
2. **Title:** Protein Dataset
   Source: Kaggle

Algorithm:

Step 1: Data Preparation: Find the Protein Dataset that has sequences and information about various protein structures and the extract sequences for those proteins.
Step 2: Feature Extraction: Compute crucial features from the protein sequences:

a) **Amino acid composition:** Find out how much of every type of amino acid (hydrophobic, polar, charged, etc.) it has.
b) **Secondary structure information:** Use the Protein Secondary Structure Dataset to detect secondary structure elements (alpha-helix, beta-sheet, random coil) corresponding to each residue.
c) **Sequence Patterns:** Look for aggregation-prone motifs (e.g., GxxxG, Q/N-rich regions).

Step 3: Models For Predicting Aggregation: Train models using machine learning algorithms based on the extracted attributes:

a) **Random Forests, Support Vector Machines (SVM), or Deep Learning Models** (like Recurrent Neural Networks or Convolutional Neural Networks).
b) We will use labeled data from the second structure dataset to predict the aggregation propensity score for every residue.

Step 4: Threshold determination: Define a threshold above which aggregation propensity scores indicate that residues are prone to aggregation.

Step 5: Visualization and Interpretation: The protein sequence can be visualized for aggregation-prone areas in this visualization step. Then find out which ones are more likely to aggregate. And how do they relate to secondary structure elements?

Step 6: Biological Implications: Use aggregation predictions to analyze protein function and disease contexts. Investigate how aggregation affects protein stability and cellular processes.

Step 7: Validation and Experimental Confirmation: Validate predictions using experimental assays (e.g., Thioflavin T binding, sedimentation assays). Confirm aggregation behavior in vitro or in vivo.

Step 8: Application in Protein Engineering: The predicted areas that will lead to aggregation can be used to guide the way forward in protein engineering strategies. Increase protein solubility and stability by decreasing its aggregation propensity

Step 9: Iterative Refinement: Continuously update and refine the algorithm based on new experimental data and insights.

Expected Outcome:

Our goal is to identify areas in protein sequences that exhibit a propensity for aggregation in this study. For instance, we gain insights into potential aggregation sites by analyzing specific stretches of hydrophobic amino acids and β-sheet-forming motifs. By visualizing these regions on the protein sequence, we can better understand their spatial distribution and subsequent consequences for protein structure. Notably, such knowledge has considerable biological implications, particularly with diseases like Alzheimer's and Parkinson's at stake. The predictions made need to be confirmed experimentally so that they can be trusted; drug design applications, on the other hand, help enhance stability as well as purification of proteins through biotechnology. In this manner, our work contributes to scientific advancement by encouraging synergies between computer-driven speculation and tangible life sciences.