PROTEIN STRUCTURE PREDICTION

**FAST National University of Computer and Emerging Sciences**

**11/10/2018**

K173810- Aiman Siddiqui K173881- Saba Pervez K1738ab- Muneeb Khan

A Data Structure project proposal in the field of Bioinformatics.

**Description:**

Protein structure prediction strategies endeavor to decide the local structure of a given amino acid arrangement. A single protein consists of at least 20 amino acids. Each amino acid is represented by different alphabets, such as, Alanine(A), Arginine (R), Aspartic Acid(D), Glycine (G), Lysine(K), Methonine(M),etc. Protein sequence consists of different combinations of these amino acids’ symbols. Each protein differs from another by a single change in these sequences. For example, “MVMSELRWHTASPEDNKNSLKRDLLKSTPTSAREAAVHIM” is a protein sequence. In our project, we intend to take 3-4 known protein structures and an unknown structure and have to find the percentage of similarity between the unknown structure and each of the known structures.

In this problem we will be having the protein sequences saved in some text files. We will take input through these files. We will align the unknown sequence with the known ones one by one and simultaneously calculate the score. We will be using Global Alignment for this purpose. Then we will compare the score of matching with each of the known alignments. The sequence with the largest matching score will be the most similar sequence of the unknown sequence.

The dataset that we are using is from <https://www.uniprot.org> and we are using following species:

* Drosophila simulans
* Drosophila sechellia
* UPI0007E6510F
* Drosophila melanogaster
* An unknown sequence