Protein Classification

Course Teacher: **Sir Rafi**

Lab Instructor: **Miss Safia**

Group Members: **Hatim(17K-3626)**

Anas(17K-3782)

Sheryar (17k-3620)

Objective:

The main idea behind the project Protein Classification was to align the input string with various classes of Proteins, and by aligning, determine the class which has the highest score. This allows us to classify unknowns with known protein classification which helps determine the properties of unknown sequences of proteins.

Algorithms Used:

- **NeedleMan-Wunsch for Global Alignment**
- **Smith-Waterman for Local Alignment**

Approaches Used:

- **K-nearest neighbor approach (ML Algorithm)**
- **Average Score**
- **Top Score**

Protein Classes Used:

For the sake of project demo, out of several protein classes only 10 classes with 10 subclasses were used. The source code allows expanding the classes and also the subclasses to the desired number.

1. TRANSCRIPTION (Class 0)
2. HYDROLASE (Class 1)
3. TRANSFERASE (Class 2)
4. OXIDOREDUCTASE (Class 3)
5. HYDROLASE/HYDROLASE INHIBITOR (Class 4)
6. VIRAL PROTEIN (Class 5)
7. LYASE (Class 6)
8. VIRUS (Class 7)
9. IMMUNE SYSTEM (Class 8)
10. TRANSPORT PROTEIN (Class 9)

Models of the Program:

7 models are developed of the same program on basis of their accuracy and approaches. The description of these models is given in the excel sheet attached along with the parameter values used in the alignment algorithms.

Program Functionality:

To avoid Exponential time during the alignment, Dynamic Programming was used to decrease the order of time to Polynomial time.

Phase 1:

The input string each time is selected from a subclass from one out of ten classes.

Phase 2:

The input string is then compared with all the classes to find the alignment score.

Phase 3:

The aligned string is formed and the score is calculated based on the algorithm. The alignments done are global and local both.

Phase 4:

Best score or average score is selected based on the model, which determines the best class similar to the input string.

Program Logic:

The program is based on the nearest neighbor approach where it is assumed that a class subclasses share some similarities when compared. Based on that logic, it computes the score using Needleman-Wunsch and Smith-Waterman Algorithm to find the best similarities with respect to input string.

Conclusion:

Protein classification is from bio-informatics field. It can help identify the characteristics and the purpose of an unknown protein by comparing it with the known classes of the proteins. This can help classify many unknown proteins.