TEAM DOUBLE HELIX | PROJECT UPDATE

PROTEIN SUB-LOCALIZATION PREDICTION THROUGH PROTEIN SEQUENCE ANALYSIS

Protein sorting, the transport of protein after its synthesis, is a one of the most complex process in a cell. Large scale researches have been going on in predicting the sub-localization of the proteins based on primary, secondary and tertiary sequencing.

Our project aims at predicting the sub-localization of proteins in an organelle of cell using primary sequencing. The Python-based GUI application would be based on Tkinker and BioPython Library. This project gives us insight about the usage of Bioinformatics concepts like multiple sequence alignments, usage of PSI-BLAST as well protein sequence extraction from NCBI. We will be making our local protein database as well with the help of Entrez E-Utilities API services.

01

TAKING AN INPUT PROTEIN SEQUENCE

The user will be providing an input into GUI program as a string or he can input the sequence in the form of a FASTA file/ .txt file. The program would parse string according to appropriate delimiters and store it in a hashmap.

02

EXTRACTION OF PROTEIN SEQUENCE FROM NCBI

Since our main objective is to do to comparisons through primary sequencing, we need to have a local database of few proteins from each of the organelle which takes into account the homology among different species so that we could have most accurate results for a wider range of inputs. The local database of proteins is made through scripting using Entrez E-Utilities tool by NCBI. E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data fro.

03

RUNNING BLAST

With the help of the parsed fasta file, we will do a PSI-BLAST for the protein sequence using the BLAST Common URL API. The NCBI-BLAST Common URL API allows us to run searches remotely.

04

THE RESULTS

The program would finally process on the BLAST results to find the most appropriate organelle where the protein input sequence could be found. We will be evaluating against an appropriate threshold of E-Value score (Minimum e-value means best result) and Query Score (maximum percentage means best result). This way, we will find the organelle where the protein is most likely to be found.

STAGES ACHIEVED

RESEARCH PHASE

The research for our project has been done in the 2 meetings our group has had after the project approval. We have discussed among ourselves the resources and the tools we would be using (mentioned in the the flowchart above.) There are lot of resources available online but using the most appropriate our needs and then deciding the suitable libraries for it was the main task.

WORK DISTRIBUTION

The work has been divided among the group members and right now everyone is learning about the tools and how to implement those tools to the best of our needs. The next step includes local dataset formation, coding the program and testing it for bugs.