Thomas Muamar

Final Project Proposal

The tool I have chosen is a sequence similarity search tool. This involves a list or matches between the query sequence and the reference sequences in the database. The output will include additional information, such as alignment scores, e-values, and sequence identities, which will help match the quality of the matches. The use of this tool is important because it allows a researcher to gain insight into the structure, function, and evolution of biological molecules such as DNA, RNA, and proteins. It can help identify homologous sequences, predict protein structure and function. This tool will be built using various technologies such as python, CGI programming, HTML5, MySQL, JavaScript, and jQuery for client-side interactions.

In this tool one of the things being used is the python CGI programming. In this portion of it I will be writing a python script that will involve processing the user input, executing the database queries and will work with the html content. The user input in this case will most likely be a protein sequence that is in FASTA format. This user input can be inserted in as a text in a specified box saying to enter your protein sequence. This will then be processed the user input and go through a sequence alignment tool such as BLAST, FASTA, or HMMER which each of these algorithms will have their different strengths and weaknesses. I will most likely be looking at HMMER as the algorithm of choose for the sequence alignments. So once the user inputs their protein sequence the server-side code will use a sequence similarity search algorithm to search for similar sequences in the database. The results will be displayed to the user in a tabular format with information such as alignment scores, e-values, and sequence identities.

This algorithm will reference sequences from a relational database scheme that I will be creating using a source database from uniport kb. I will most likely be using python in order to parse the information into MySQL database. So, MySQL will be implemented to store and manage the reference sequences and their associated metadata. Using MySQL connector with python to connect to this database and allow me to retrieve results. I will most likely create tables that contain the protein sequences, annotations, taxonomy, and cross-references. The HTML template in this case will contain things like the placeholders for dynamic content, such as search forms, and search results. The HTML template will provide the structure and formatting of the web page and will work together with the python CGI script to produce the information such as the scores, e-value, and sequence identities. This will allow the user to depict which reference sequence is most similar to the query sequence. CSS styling will most likely be kept minimal to the point of having a clean layout and defining the fonts and making sure the information is in a desirable output. Most likely have a box with text above saying to input protein sequence code for similarity analysis. In this box users can input their sequence in FASTA format, or they can upload a file that is in FASTA format. Once they have done that there is a submit button which will allow the algorithm to depict similarities. The display for the output results will show the protein name and the information I talked about earlier and the sequences highlighting the similarities between the query sequence.

JavaScript and jQuery for the client-0side interactions will be used to make the user experience more interactive. Trying to do more client-side sorting and filtering as well as visualization of search results. Allowing the user to sort by the different values I specified. In the end the sequence similarity search tool will help users for analyzing protein sequences. It will be built using the combination of the technologies described.