Thomas Muamar

**Advanced Practical Computer Concepts for Bioinformatics**

**Final Project Narrative**

**Background:**

The reason I choose to do a sequence similarity tool for proteins is because to me it seems like an important tool for scientists when they come across an unknown protein sequence. When researchers come across an unknown sequence it is easy to go to a website and input their sequence and allow them to see if there are any similarities between their sequences and ones that have been identified. This program allows an individual to input their unknown protein sequence and display other proteins in the uniport kb database with different values that allows a user to depict similarities. The local alignment between the protein sequences can allow for an individual to see regions of similarity or similar sequence motifs within their larger sequence. These short, conserved regions that get identified can be related by evolution and can be responsible for a specific function in their specific unknown sequence.

**Issues in Implementation:**

During the implementation of this tool one of the problems I had come across was trying to figure out a way to implement one of the sequence similarity tools into my code. At first, I was trying to get hmmer tool to be implemented into my code but one thing that seemed to be a big problem with it was that it required to create a hmmer profile and then it ends up using hmmsearch command to search through the database. From doing this it seemed like it would take longer than normal. With blastp as an implementation it allowed for a faster process for searching for protein similarities. The only thing that I was unable to figure out was I had to have it go through my database create a separate file and then search through that file to grab the similarities.

Another implementation that I found hard to do was trying to get the local alignments to be like they would in a blastp output. Every time I tried having it produced in a way that would look like the ncbi alignments I ended up breaking my table. The solution for me was to have the aligned query sequence and subject sequence be side by side on the table and use JavaScript and CSS to highlight the similarities when the user hovers over. This allows for the user interaction and allows for the user to see what is similar between the two sequences. Something I believe I could have done that would have made it easier was supply another webpage that would show the given alignments in tabular format with the rest of the data displayed.

One of the issues that caused a lot of time consumption was getting a lot better at using JavaScript. Going through a lot of different online to try and figure out how to get things to work. Especially with the highlighting since I wanted to have both of them highlighted when the user was hovering over the text. CSS styling seemed like an easy way of doing it at first but for some odd reason I couldn’t get it to highlight when the subject sequence was being hovered over. I tried using the sibling selectors of ‘+’ and ‘~’ but still couldn’t get it to select them both when user hovers over either one.

**Project Methods and technologies:**

In this project the technologies I used are python, CGI, javascript, HTML5, page styling with CSS, and a relational database schema. Python in this project was used in order to create my database as will as the CGI script to conduct the analysis. The relational database was created as a table called proteins with three columns which were id, seq\_id and sequence. The CGI was created using python within this code it used MySQL connector in order to allow for the blastP algorithm to search through the database.

The HTML template is connected to the CGI script so once the user submits a file or post through the protein sequence submitted it will go through the CGI script which will allow for the sequence submitted to be utilized in a sequence similarity search. JavaScript was used to allow sorting of the table and highlighting the aligned query and subject sequence. CSS styling was used to style the page and provide a better display of the table for the aligned query sequences and subject sequences.

**Discussion/Validation of tool:**

To make sure that Blastp was working correctly and producing the correctly results with the database provide I used this website to <https://www.ebi.ac.uk/Tools/sss/ncbiblast/> in order to make sure that it was right. This tool allowed me to select the same search parameters and database I was using. The display of the data was similar for the percent identity and the bit score, and e-value also produced the same results when searching for similar protein sequences.

Once this tool was working with my web page, I was able to parse the results into a tabular format displaying the query id, sequence id, percent identity, alignment length, mismatches, gap opens, e-value, bit score, and the aligned sequences. I find this tool to be super useful to use because it allows a user to search their unknown protein and identify if it has any similarities with others. If there is one that is similar user is able to check similarities between the sequence by hovering over them and can also click on the reference sequences ID to redirect them to information about it.

**Conclusion:**

In the end this project was useful for developing my skills in HTML, python, CSS, and JavaScript. It allowed me to get a better perspective of the different coding tools we can use and how much more flexibility we have with other coding tools. Especially when making the website more interactive with JavaScript it allowed me to get better understanding of how it worked. Once you are able to figure out how to solve or find another solution to your issues it makes it satisfying to see your results.