COMP 4980 Project Proposal(#5 bioinformatic tools)

Problem and proposal

PAM and BLOSUM are the powerful global sequence alignment tools that can help analyzing the protein sequence. However, in the comp 4980 lectures this semester, the two alignments were only introduced briefly, some core mechanics were not mentioned.
In order to understand these two alignment methods. Nobutaka Kim and I are proposing to build a website like EMBOSS Needleman-Wunsch tool webpage. Thus this will be a team project and this project will contain two part. The front end website that take the protein sequences from users and display the calculated scoring metrics to the users. The second part is back end engine that read the users input and do the PAM or BLOSUM scoring metrics calculations.
Since there is already PAM and BLOSUM tool online, so we can use these online tools to check the correctness of the implemented program.

Proposed schedule:

Nov.4 - Nov.6

Implement front end website.

Implement Needleman-Wunsch global sequence alignment tool.

Nov.6-Nov.14

implementing the PAM sequence alignment tool

Nov.15-20

Implementing the BLOSUM sequence alignment tool

Nov.21-Nov.23

Testing

Front End specs:

We will make a frontend using Plotly Dash which is fully asynchronous and pure Python. Dash Bio library specifically provides components for bioinformatics visualization, alignment, etc. We are also using Biopython to process fasta file inputs from user upload, manual input and searches by accession number on NCBI website using their APIs.

We considered hosting local copies of NCBI databases but they were 97 Gigabytes or larger so that is impossible at this time.
Also with regards to visualizations, Dash Bio actually implements algorithms behind the scene for its graphs so we have to find a way to output something using Pandas dataframe with Dash Table and perhaps some of their other visualization tools or even Matplotlib to provide a nice UI/UX for the end user, biologists, who may get tired of looking at black and white output from Emboss.

We also found an NCBI link to download all the PAM and BLOSUM matrices that can supplement the app. I am also going to have a in the frontend where NCBI search results are shown so that users can select other accession numbers based on their initial search.
Also included will be a to check any file uploads before running the algorithm since this is an expensive step for users with limited data such as mobile. We are going to try implement a progress bar for PAM but the BLOSUM algorithm is recursive so this may or may not be possible within the project time frame although I have implemented a counter for recursion once before.
It should be possible to save the results or to email them to a provided email.
Dr. Yan has been included in the Github repository via email invitation so please feel free to browse the code base divided into frontend and backend. There is a needleman.ipynb in /backend/ and in /frontend/project/ there is an .ipynb that shows some of the layout and use of NCBI APIs with outputs to confirm they worked.

The final app will be hosted on an AWS server and accessible for students during the presentation as we hope to walk them through at least one each of “search-and-run-Needleman” for PAM and BLOSUM. In addition, if time allows, a short explanation on the algorithms will be given and with any other remaining time, a short presentation on Plotly Dash for bioinformatics.

Although Fei is our algorithm expert, having walked through the pseudocode together, I feel comfortable that her implementation is accurate and robust.