Spring 2020 CS123A Bioinformatics Project Proposal

Provide a ½ to 1 page description of your proposed project. Use this file/template below to provide the following information.

PROJECT TITLE: PSSM Replication

DESCRIPTION: [Describe the task or problem the group will address. Also motivate why you chose the problem.]

I want to be able to address how false positive matches may affect the PSSM cycles, and also be able to address how the E-value parameter cutoff may affect the results. I chose to work on the PSSM algorithm, since I want to see how it is implemented along with the BLOSUM62 before using the PSSM queries. I am interested to see how the PSSM is able to perform its cycles until no new sequences have been to the PSSM.

RESOURCES: [Describe the resources, DBs, Data, …etc. that you anticipate needing to complete the project]

The resources that I will need to complete the work will be using the PSI BLAST to determine whether the implementation of the PSSM hd similar results. I will also need to use the protein sequences from the NCBI DB to use to make the PSSM matrix when picking the highest similarity sequences with the initial sequence that will be used.

ASSISTANCE: [Describe any assistance that you anticipate needing]

As of right now, I think the main thing I'll need assistance on is understanding how the PSSM matrix works, and how to fill a PSSM matrix. Another thing I may need help on may be to choose the best scored sequences from the alignment of the initial sequence, in order to get a similar result that could potentially match with the results generated from the PSI BLAST.

STATEMENT OF WORK (SOW): [Give a rough breakdown of the major tasks to be completed. ← this can be refined later.]

1. Understand how the PSSM algorithm works, and the math behind it

2. Figuring out how to implement the PSSM algorithm with respect to using the BLOSUM62 alignment beforehand.
3. Trying to figure out how to code the algorithm in python ideally.
4. How to check if it gives similar results to the PSI BLAST

SCHEDULE: [Provide a rough schedule for completing the project]

3/19/20 - 3/31/20: Try to understand how the PSSM algorithm is used, and be able to fill one before starting the coding component.

4/1/20 - 4/20/20: Begin the coding process for the PSSM algorithm, and also test it with several protein sequences to see if it works as it is supposed to.

4/21/-4/26: Debugging of the code to make sure it provides similar results to the PSI BLAST using the same protein sequences that were gathered from the NCBI DB.

4/27/20-5/19/20: Final touches to the code, and the completion of the project report.

5/20/20: Project will be turned in.