CS 490/590 Bioinformatics Project I: Alignments and Substitution Matrices

Description: This project has three components (i), (ii), (iii) for both undergraduate and graduate students, with additional part (iii).(a) for graduate students only:

(i) Allow the user to select nucleotide (DNA or RNA) or peptide (protein, amino acid sequence) sequence files in **FASTA format**, prompting for the FASTA format filename. If a nucleotide sequence is selected, **translate** that sequence into the corresponding amino acid sequence (assuming coding, 5'-3'). **Output your translation to the console.**

(ii) Allow the user to select either a matrix of amino acid **substitution scores** (e.g. BLOSUM, PAM, hydrophobicity) or a PAM(n) **mutation probability** matrix with given **n units of evolutionary divergence**. In the latter case, further prompt the user for the given units of evolutionary divergence, say n: You will **calculate the PAM-n substitution scores matrix** based on the mutation probability matrix as in class, and **output both your PAM-n mutation probability matrix and your PAM-n substitution scores matrix to the console**.

(iii) Finally, allow the user to select what kind of alignment is desired: global, local, or semiglobal. Have the user input gap penalties. You will use the substitution matrix (either input or calculated) of part (ii) to perform a protein alignment of the two amino acid sequences (either input or translated) of part (i).

a. **Graduate students only:** Additionally allow user to select **affine alignment** as one of the alignment types. If affine is selected, then the user should separately input the **gap start** penalty and the **gap extension** penalty.

Output both your OPT matrix (in the case of affine that includes 3 matrices) and your optimal alignment in a readable form to the console.

Note that all outputs are to be to the console (I will keep log files while running anyway).

I have and will upload various input files with explanations to Moodle. In particular, note PAM substitution matrix PAM250.txt, the BLOSUM substitution matrix BLOSUM62.txt, and the hydrophobicity based substitution matrix HP.txt. I have also uploaded the PAM-1 mutation probability matrix already. Notice the amino acid order in all matrices:

A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V

What to turn in: You must turn in a single zipped file containing your source code, a Makefile if needed for compilation, and a README file indicating how to compile/execute your program in addition to any other commentary concerning parts of your program that do or do not work.

Your program must be written in C/C++, Java, or Python and compile using an open source compiler on our home server home.cs.siue.edu. You must test your program on the home machine.

This assignment is due by MIDNIGHT of Sunday, October 11.

Late submissions carry a -33% per day penalty.