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

Description: This project has three components:

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

(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 to a file named **PAM<n>.txt**. E.g. PAM250.txt for PAM250 substitution scores.
EXTRA CREDIT: Correctly implementing the logarithmic time matrix powering to compute the PAM-n mutation probabilities will result in extra credit of 6% towards course grade.

(iii) Finally, implement the **global alignment** algorithm using the substitution matrix. 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).

Additional details: Parts (i) and (iii) are primarily based on your two previous projects. Even if you didn't get them working perfectly the first time, you can retrospectively look at the perl implementations provided. Part (ii) is really the "new" stuff for this project, while all else is putting things together. Therefore, I recommend that you first get your project working for the case that the user wishes to select a **substitution scores matrix** instead of the mutation probability matrix. After you get that working, which you hopefully do quickly, work on deriving the PAM-n substitution scores matrix from the PAM-1 mutation probabilities.

I 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. 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

Your expected console output and its format are similar to your second project in the case that the input sequences are peptides and the selected matrices substitution scores instead of probabilities. For input nucleotide sequences, however, **additionally** output their translated amino acid sequences. And, for mutation probability matrices, **additionally** output to a file the corresponding substitution scores matrix.

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 execute your program.

Your program must be written in C/C++ or Java and compile using an open source compiler such as g++, gcc, or javac.

I will demo a Perl implementation of this project in class/lab so you will understand I/O etc..

This assignment is due by MIDNIGHT of Wednesday, October 9.

Late submissions carry a -33% per day penalty.

CS490 Students: The assignment is worth 12% of your total grade.

CS450 Students: The assignment is worth 10% of your total grade.