CS 490/590 Bioinformatics Project V: UPGMA for Phylogenetic Trees

Description: You are to implement the UPGMA clustering algorithm on input sets of multi-aligned sequences to construct phylogenetic trees.

Specifications**:** Your input (as file **sequences1.txt**) will be a set of multiply-aligned DNA sequences, such as the alignment.fasta sequences of the previous project with added labels in each sequence's header. As in the previous project, each sequence is a representative of its organism, and you will compute a distance matrix of those sequences based on the Jukes-Cantor distance calculations. You will run the centroid-linkage based UPGMA algorithm to cluster the organisms based on the distance metrics. Your final output, to the console, should be in the parentheses-based Newick format.

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 Thursday, October 31. Late submissions carry a -33% per day penalty.

The assignment is worth 10% of your total grade.