Bioinformatics Algorithms Lab
Lab 2

Progressive Sequence Alignment and UPGMA tree

Breakdown of parts

For this lab, you will need to build the final assignment in three stages. Be sure to complete each stage before starting the next. Note that you have multiple weeks to complete this assignment. It will be helpful to attempt at least one stage per week, possibly leaving you a week of buffer in case you encounter an unexpected problem.

Stage 1: Pairwise alignment

Using the dynamic programming method discussed in lecture, you will need to conduct a global sequence alignment with affine gap penalties between two sequences stored in a single FASTA file. Sample files will be available off of myCourses. Note that your program will need to ask the user if nucleotide or protein sequences are in the file. If protein, use the PAM100 matrix (in myCourses) for match/mismatch scores. Your output should include the pairwise sequence alignment in a format easily read by a naïve biologist.

Be sure to ask the user for any scoring parameters, as well as the name of the input file. Try to avoid hardcoding as much as possible.

Stage 2: Progressive Multiple Sequence Alignment

Read in a single FASTA file with multiple sequences and perform the simple progressive alignment as given in lecture. Again, ask the user for parameters as appropriate. In the final alignment, be sure to include a consensus. The consensus may include an "*" where all sequences agree for a particular residue. Where there is disagreement at a residue, think about the best way to represent it in the consensus...this is good material for online lab discussions. ☺

Stage 3: UPGMA tree generation

Extend your progressive alignment to also output a Newick format tree for the sequences. Be sure to include transformed distances in the tree, as described in lecture. Limit distances to two decimals.