Genome 541 Spring, 2010
Computational Molecular Biology J. Felsenstein

Homework no. 4
Due Sunday, May 2

Write a program (most conveniently using code from the previous homeworks) to read in a data set of DNA sequences (including the possibility of gap symbols).

1. Compute a table of distances between the DNA sequences using the Jukes-Cantor formula. Print it out. The Jukes-Cantor distance is given in the projection for lecture 4 of my block of lectures as the first formula in the first panel titled "Approximate variances for distances", where D in that formula is the fraction of sites differing between the pair of sequences and $\hat{\mathtt{t}}$ is the distance we want to compute. Note that a site that has a "?" or "-" (gap) state is simply to be dropped from any pair of species where one of them has that state (it is not dropped from consideration in other pairs of species).

2. Then construct a UPGMA tree from these distances. Either write out a Newick tree for it, or print a table of some sort that shows the topology and the branch lengths. The UPGMA algorithm is described in Ewens and Grant, and also in the lecture projection for lecture 4 of my block of lectures, available from the course web page.

3. Then (the most fun part) print it out in some interesting and relatively readable pictorial form (you can just use "character graphics" if you want). i.e., not just the Newick form but something more readable by people.

The data set is the same primate one that you used in the previous assignment. You can also try some other data set, but at least you should include the primate data set results too.

Show me (by email) the output of your program, and, as usual, attach the source code for use in case something appears to be wrong.