Genome 570 Winter, 2016
Phylogenetic Inference J. Felsenstein

Homework no. 2
Due Friday, February 5 at the end of the evening

- Use a data set of sequences of the same gene, over the same set of species, with DNA data, and also one with protein sequences. I have provided 27 such data sets available on the course webpage, and you can use those, or you can download your own from Genbank. Note that the DNA data sets on the webpage are a bit different from the ones used for the previous homework, as some regions of the gene are now omitted as dubious in quality or quality oif alignment. So if you use those, download them again. Make sure to tell me which gene you used.
- As before, the data sets should have about 40 sequences (or species).
- Use a program (or feature of a program) that computes an appropriate matrix of distances for the species from DNA sequences. Also use a program or feature that computes a matrix of distances for the same species from the protein sequences. Be sure to report what program and what distances you used, and why. Does the distance you used cope with possible differences of rate of evolution among sites. (By the way, if there are deletions in the sequences, most distance programs will ignore those sites for any pair of sequences where one or both have the gap – but will *not* drop them from the whole data set.)
- Use a program or programs that compute a trees by a distance matrix method from these two distance matrices. These can either use a least squares method (such as the Fitch-Margoliash method), or the neighbor-joining method. Or if you are enthusiastic, both. Show the trees.
- Are these trees rooted or unrooted? If rooted, is that by outgroup?
- Comment on the speed of the methods (exact timings not needed).
- In addition, run the UPGMA method for each of the two distance matrices. Are these trees rooted? If so, is that by outgroup?
- How different are the DNA and protein trees in the non-UPGMA method and in the UPGMA method?
- How reasonable are these trees?
- In which parts of the tree are the DNA trees most accurate (recent divergences or ancient divergences? In which parts of the tree are the protein trees most accurate?
- Comment on how well the programs functioned and how easy or hard they were to use.
- Report the results to me in a short (2-5 pages or so) report. Show some results if needed.

There are many distance matrix programs. Aside PHYLIP, MEGA, and PAUP*, you will find many others listed in my webpages on "Phylogeny Programs".

Hints for those using PHYLIP programs: After using Dnadist or Protdist to compute a distance matrix, you have to rename `outfile` to `infile` (or else when you run the distance program you will overwrite your distances and crash).