Genome 570 Winter, 2016
Phylogenetic Inference J. Felsenstein

Homework no. 3

Due Friday, February 19

- Using the same nucleotide or protein sequence data set that you used for Homework #2, infer a maximum likelihood phylogeny. (If that data set turned out to have big problems, you may choose another one).

- If you have been using protein sequences you should use a program that has a 20×20 amino acid model such as Jones-Taylor-Thornton (JTT) or other some reasonable one like WAG. For DNA sequences use one of the usual DNA models.

- The ML analysis should include allowing gamma-distributed variation of rates from site to site. If you need to choose the parameter value for the amount of rate variation, do so by maximum likelihood. That is, either have the program optimize the parameter (if it can) or do so yourself by trying different values of the parameter and finding the one that leads to the highest likelihood. The parameter will either be the Gamma distribution's "shape parameter" α or its Coefficient of Variation which is $1/\sqrt{\alpha}$.

- Since for some programs, trying different values of the rate variation parameter may be very slow, one way that works almost as well would be to get a tree topology for one value of the parameter, and then feed it in as a user-defined tree so that the program only uses that topology but does infer new branch lengths for it. Then try different parameter values in different runs, which should be much quicker than doing a full search, and find the parameter value with the highest likelihood.

- If we then start with that parameter value and do a search among tree topologies does the topology found differ? If so, and we then use that new topology to again re-estimate the parameter, Does this process converge?

- Report also on which sites show the smallest rate of change, and which show the most, and whether that pattern makes biological sense (third positions? active sites?).

- Also (after the rate variation parameter is determined) use it to do a bootstrap analysis with at least 100 bootstrap replicates. The bootstrap analysis includes making a majority-rule consensus tree of the 100 (or more) results.

- Some program packages do not do ML. I am not insisting that you use PHYLIP. In fact if you used it before, I would suggest trying another package, to broaden your experience. You can find packages in my Phylogeny Programs list (use that phrase in a search engine to find the pages). PAUP*, Phyml, RAxML, and MEGA are popular choices.

- Make sure in your report to show the ML estimate of the tree, and report on the rate variation parameters, which sites seem to have high and low rates (not so much the detailed site numbers as the description of which functional regions in the molecule have high and low rates). For the bootstrap analysis report the bootstrap P values. Does the amount of bootstrap support make sense in terms of which parts of the tree seem biologically sensible?

- Comment on the programs you used, how well they worked and how easy they were to use.

e-mail me `(joe (at) gs.washington.edu)` with a report in PDF or MS Word (.docx) form on the results. Mercy will once again be granted if you need more time, but a new assignment is coming on February 26 which *must* be turned in a week later as otherwise I will out of time in the quarter to grade them..