Genome 570
Phylogenetic Inference

Winter, 2016
J. Felsenstein

Homework no. 1

Due Wednesday, January 20

- Get a computer program or programs that can do parsimony on nucleotide or protein sequences, or on discrete characters. You should also get one that is able to do branch-and-bound search as well as heuristic search. For a list of programs, many of them free, see
  `http://evolution.gs.washington.edu/phylip/software.html`
  or type "Phylogeny Programs" into Google. Some possibilities include `PAUP*`, `PHYLIP`, `MEGA`, `TNT`. Some of the other excellent general-purpose programs there may not have both heuristic search and branch-and-bound. It is OK to get several programs that together have these capabilities – they need not all be in one program.

- Get a data set, either molecular sequences or discrete characters. It should have at least 20 sequences (or species) and of course these should be aligned. More than 20 (say about 40) is better. For molecular data sets there should be at least 200 sites. (500 or more is better). If you do not have your own data sets see the course web site where there both some individual data sets and some database web sites where aligned sequences can be downloaded. (Some of the datasets available there have too few species for this assignment).

- Run heuristic searches for the most parsimonious trees on data sets of size 10, 15, 20, ... species (or sequences).
  - Explain exactly what choice of rearrangement strategy was used: Does the program use nearest-neighbor interchanges? SPR? TBR? Sequential addition? Usually the general strategy is explained in the program documentation.

- How much time do these take? (For some data file formats you can do different size data sets just by deleting whole species). Plot the run time against number of species. It may be best to use log-log plot for that. Report on whether the run time seems to be proportional to a power of the number of species, and what that power is.

- Run branch and bound searches on some of those same data set sizes (start with the smallest ones and stop when it seems likely to take too long). Make sure that you set parameters so that each search (that you don't abandon as taking too long) finishes completely and is not broken off short of completion.
  - How much time do these take? Does the run become impossibly slow?
  - How do the trees found by branch-and-bound compare in number and parsimony score with heuristic searches of the same dataset?

- Are the trees found reasonable? Why or why not?
- Report the results to me in a short (2-4 pages or so) report. Show some results if needed. I prefer to receive the report in PDF format, but can read Microsoft Word (.DOC or .docx) format if needed.
- Comment on the program you used, how well it worked and how easy it was to use.

You should e-mail me `(joe (at) gs.washington.edu)` with the report on the results.