STAT 882 - STATISTICAL PHYLOGENETICS SPRING 2010

HOMEWORK #1 – DUE IN TWO WEEKS

For this problem set, we'll consider the following DNA sequence data:

Mouse	`ACCAAAAAAACATCC`
Human	`ACCCCACTCACCCAT`
Gorilla	`ACTATACCCACCCAA`
Bovine	`ACCAAACCTGTCCCC`
Rhesus macaque	`ACTTCACCCGTTCAC`

1. For the tree given below,

   (a) Find the length using Fitch parsimony (include all the sites).

   The length of each site, as well as the total tree length, is given in the table below:

Site	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*Total*
Length	*0*	*0*	*2*	*2*	*2*	*0*	*1*	*2*	*2*	*2*	*2*	*2*	*1*	*1*	*2*	*21*

   (b) Which of the sites in the alignment are not *parsimony informative* (i.e., have the same length on all possible phylogenies)?

   Sites 1, 2, 6, 7 and 13 have the same length on all possible phylogenies.

   In general, any site which is constant (all taxa have the same nucleotide) or for which there is only one taxon with a nucleotide that is different from the nucleotide shared by all other taxa will be parsimony uninformative.

   (c) Find the most parsimonious reconstruction at the internal nodes for site number 3.

   Several most parsimonious reconstructions are possible, and any of these may be chosen. One such reconstruction is to assign either a C or a T at the root node. The ancestral node to Mouse and Bovine should have a C. The ancestral node of Rhesus, Gorilla, and Human should have a T. The ancestral node to Human and Gorilla should have a T.

   Note to course participants: Many of you re-arranged the tree to obtain the short-

est length for this site. I graded the subsequent problem under your answer here, since I realized that "most parsimonious reconstruction" wasn't defined in lecture.

In general, a most parsimonious reconstruction is an assignment to states at the internal nodes on a fixed phylogeny that gives the minimal length to that site. The remainder of these solutions are written assuming that the most parsimonious reconstruction on the tree given in the assignment is used.

2. Note that for Kimura's two parameter model, the transition probabilities are:

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-\mu t} + \frac{1}{2}e^{-\mu t\frac{\kappa+1}{2}}, & i = j, \\ \frac{1}{4} + \frac{1}{4}e^{-\mu t} - \frac{1}{2}e^{-\mu t\frac{\kappa+1}{2}}, & i \neq j, \text{ transition} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t}, & i \neq j, \text{ transversion.} \end{cases}$$

Assuming that $\mu = 0.3$, and $\kappa = 1.5$, calculate the probability that over a time period of length $t = 0.1$,

(a) the base at a particular site remains the same

This probability is given by

$$P_{ii}(t) = \frac{1}{4} + \frac{1}{4}e^{-0.03} + \frac{1}{2}e^{-0.0375} = 0.9742$$

(b) a transition occurs at a particular site

This probability is given by

$$P_{ii}(t) = \frac{1}{4} + \frac{1}{4}e^{-0.03} - \frac{1}{2}e^{-0.0375} = 0.0110$$

(c) a transversion occurs at a particular site

This probability is given by

$$P_{ii}(t) = \frac{1}{4} - \frac{1}{4}e^{-0.03} = 0.0074$$

3. Next, we'll use our results from the previous two problems to think about the computation of the likelihood for the tree above.

   (a) For the tree shown above, consider the third site in the DNA sequence. Assuming that the lengths of the branches of the tree are all 0.1, that $\mu = 0.3$, and that $\kappa = 1.5$, find the likelihood for the data from site 3 only, assuming that the states of the internal nodes are as defined in the most parsimonious reconstruction that you found in Problem #1.

   Using the most parsimonious reconstruction from the tree in Problem #1, we see that there are 6 branches along which there is no change in nucleotide and 2 branches along which there is a transition. Since all branch lengths are assumed to be the same (and to be given by 0.1 so that the substitution probabilities from (a) can be used), we have that the likelihood is given by $(0.9742)^6(0.0110)^2 = 0.0001034$.

   (b) What you found in the previous exercise is the likelihood of that particular tree for site 3, *given specific states at the internal nodes*. How would you find the overall likelihood for the given tree at that site?

   If we were not given the nucleotides that occurred at the interval nodes, we would have to take the sum of the likelihoods computed for all possible nucleotides which could have occurred at the internal nodes. For example, the Mouse-Bovine ancestor could have had an A, C, G, or T. For each base at the Mouse-Bovine ancestor, the Human-Gorilla ancestor could also have had an A, C, G, or T, etc. We would enumerate all possibilities and then sum their probabilities. The basic calculation for each assignment of states to the internal nodes is as in (a), except that we need to weight the assignment of state to the root node by 0.25 (under the HKY model).

   (c) How would you find the likelihood of the tree given in the figure for *all* of the data above?

   To find the likelihood for the entire tree over all of the sites, we would repeat the procedure described in question (b) above for each site, and then multiply all of these site-wise likelihoods together.

   (d) How would you find the *maximum likelihood tree* for this problem?

   To find the maximum likelihood tree for this problem, we would repeat the procedure described in question (c) for all possible trees with 5 sequences, and pick the tree with the highest likelihood. For larger problems, it may not be possible to

consider all trees, and a strategy for efficiently searching the space of trees may be needed.

4. We'll now consider this data set using the programs we've discussed in class so far. A NEXUS-formatted file containing the data can be downloaded from the course website: http://www.stat.osu.edu/∼lkubatko/stat882/.

   (a) Carry out a parsimony analysis in PAUP*.

   (b) Use Modeltest to select the best-fit evolutionary model for these data.

   (c) Use both PAUP* and GARLI to carry out a maximum likelihood analysis for these data, using the evolutionary model you selected in part (b). To carry out the analysis in PAUP*, you will need to set the criterion to likelihood by issuing the following command after reading in the data: `set criterion=likelihood`. The other options (e.g., `hsearch`) are similar to the parsimony analysis.

   (d) Remove one of the non-informative sites you identified in Problem #1(b) from the data set. You can do this with the `exclude` command – e.g., issuing the command `exclude 1` will remove the first character from the data set. Recompute the score of the most parsimonious tree and the likelihood of the maximum likelihood tree. Have they changed? If so, how? Why?

   This problem was generally well-done by students, and individual results can vary. A couple comments are:

   (a) The parsimony tree found in part (a) was found by everyone. There do not appear to be local optima for this small problem, which is to be expected.

   (b) The model preferred by both hLRT and AIC in ModelTest is HKY.

   (c) There was variation in the ML trees found by GARLI and PAUP. This was due to two factors: differences in how the HKY model was specified in the programs, and differences in the thoroughness of the search. Still, it is surprising that one may obtain different estimates of the ML tree in such a small problem!*

   (d) The key observation here is that the parsimony estimate is not affected by removing uninformative sites (though the length may or may not change, depending on whether constant sites are excluded or sites with a single character change are excluded), while the ML estimate may change (and the likelihood of the tree will always change). The likelihood analysis uses all of the sites in determining which tree is optimal, while parsimony only uses parsimony informative sites.