Take home final, CS394C, Fall 2012

Instructions:

Your solutions are due by Monday, May 7, at noon (delivered in hardcopy to Laurie Alvarez in PAT 141 and by email as a pdf file to tandy@cs.utexas.edu).

This is an open-book exam, but you are not allowed to discuss the problems with anyone else. You should put your name on every page, and staple the pages together (just in case the pages come apart). Partial credit will be given, for example for work that has arithmetic mistakes but otherwise indicates that the concepts are understood. Except for problem 1, please give *reasons* for your answers rather than just stating your answer, and show calculations that you made to determine the answer.

1. For each of the following statements, say only whether it is true or false (do not give a reason or proof).
   - Neighbor Joining computed using Jukes-Cantor distances is statistically consistent under the GTR model.
   - Neighbor Joining computed using GTR distances is statistically consistent under the Jukes-Cantor model.
   - Maximum likelihood (optimizing parameters under Jukes-Cantor) is statistically consistent under the GTR model.
   - Maximum likelihood (optimizing parameters under the GTR model) is statistically consistent under the Jukes-Cantor model.
   - UPGMA based upon Hamming distances is statistically consistent under the Jukes-Cantor model.
   - UPGMA based upon Jukes-Cantor distances is statistically consistent under the Jukes-Cantor model.
2. Give a proof that Maximum Parsimony is not statistically consistent for some Cavender-Farris model tree.
3. Give a proof that Maximum Parsimony is statistically consistent for some Cavender-Farris model tree.
4. (a) Write down the Dynamic Programming (DP) algorithm for computing the cost of the optimal global pairwise alignment when all indels and mismatches have cost 1 (thus, a gap of length k is considered k single indels, and hence has cost k). Be sure to provide full information: the meaning of each subproblems, the order in which the subproblems are computed, how they are initialized, and where the solution is provided.

   (b) Apply the DP algorithm to the following two sequences:

X = ACTA
Y = ATATACA

(Just present the properly filled in DP matrix; no need to show how you obtained each value!)

(c) How many optimal global pairwise alignments can you find?

(d) Show two of the optimal pairwise alignments.

5. Consider the following three trees, each of which is supposed to be an estimate of the true tree.

   - Tree T_1 has one internal edge defining 12|3456
   - Tree T_2 has one internal edge defining 1234|56
   - Tree T_3 has one internal edge defining 15|2346

   Two of the three trees above have 0 FP (false positives) with respect to the true tree. Which two must these be? Give a possible "true tree" which proves your statement valid.

6. Henry performs a simulation study of sequence evolution, and calculates maximum parsimony trees for the datasets he obtains, producing several "optimal" solutions. He then calculates the majority and strict consensus trees, and compares them to the model tree (known to him because he performed the simulation study).

   (a) He obtains trees T_1 and T_2 with the following error rates:
   - T_1 has 3 false positives and 18 false negatives
   - T_2 has 4 false positives and 16 false negatives

   Assuming he made no mistakes in his calculations of error rates or consensus trees, which tree is the strict consensus and which is the majority consensus? (Why?)

   (b) Same set-up, but now T_1 and T_2 have the following error rates:
   - T_1 has 3 false positives and 18 false negatives
   - T_2 has 4 false positives and 20 false negatives

   What do you conclude now? Is it possible for one of these to be the majority consensus and the other the strict consensus? If not, why not?

   (c) Let $gt = ((a, b), (c, (d, e)))$ and $ST = ((a, (b, (c, (d, e)))))$. Compute the minimum number of duplications needed to reconcile gt with ST, and draw the embedding of the gene tree gt into the species tree ST.

7. Let $T_1 = ((b, (a, e)), (c, d)), T_2 = ((c, (e, (a, b))), d)$, and $T_3 = ((e, (a, b)), (c, d))$. Compute the duplication cost for all ordered pairs of these trees (i.e., for all i, j, if you treat T_i as the gene tree and T_j as the species tree, the number of duplications implied by that pair). Which of these trees would give the smallest total duplication cost if it were the species tree?

Extra Credit Problems

To obtain credit for these problems, you will need to provide proofs.

1. Suppose you have a circular ordering of the leaves of a tree, and the true quartet tree on every subsequent quartet of leaves. Give an algorithm to compute the tree from this set of n quartet trees, and prove it correct.

2. Let Φ be an exact algorithm for the L_∞-nearest tree problem, as follows:

   Input: $n \times n$ dissimilarity matrix $[d_{ij}]$

   Output: $n \times n$ additive matrix $[D_{ij}]$ such that $L_\infty(d, D)$ is minimum over all $n \times n$ additive matrices $[D_{ij}]$. Here, $L_\infty(d, D) = max_{ij}|d_{ij} - D_{ij}|$.

   For this method, answer the following questions in the context of the Cavender-Farris model:

   (a) Is Φ statistically consistent under the Cavender-Farris model if applied to Cavender-Farris distances? Why or why not?

   (b) Let $[D'_{ij}]$ be an $n \times n$ additive matrix corresponding to an edge-weighted tree (T, w). Find the biggest $\delta > 0$ for which $\Phi(d)$ is guaranteed to be an additive matrix for the same tree T whenever $L_\infty(d, D') < \delta$. (Prove that the statement holds for your choice of δ.)

3. Consider the following stochastic model of character evolution down a tree. The model tree is a rooted and binary tree, with node set $V = \{v_1, v_2, \ldots, v_{2n-1}\}$, where n is the number of leaves (i.e., the internal nodes and leaves are all labelled). The root is v_1. Every character that evolves down this tree begins with the state 1 (recall that the root is v_1). Each edge e of the tree has a substitution probability $p(e) > 0$ indicating the probability that the character will change its state on the edge. However, if a character changes its state on the edge (v_i, v_j) (with v_j below v_i), then the state of the character at v_j will be j.

   - Describe a polynomial time algorithm to estimate the tree (just the unrooted topology), prove it statistically consistent under this model, and determine its computational complexity.