

Take home midterm, CS395T, Spring 2008

Instructions: This is an open-book exam, but you are not allowed to discuss the problems with anyone else (except me). 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. Show all your work!

Important: You are welcome to answer as many questions as you like. Each problem is worth 20 points. I will keep the top 5 problems in computing your grade.

However, if you wish to attempt the extra credit, please see the policy on that page for additional instructions.

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 K2P 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 K2P 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. Treat each of the model trees in the “Hobgoblin of Phylogenetics” paper as Cavender-Farris trees (in other words, just look at the substitution probabilities). For each of the following methods, indicate whether it is statistically consistent – and provide a proof either way.
 - Maximum Parsimony
 - Neighbor joining using Cavender-Farris model distances
 - UPGMA using Hamming distances
 - Maximum likelihood optimizing Cavender-Farris parameters
 - Maximum compatibility
 - The Four Point Method, using Cavender-Farris distances
 - The Four Point Method, using Hamming distances

3. For each of the CF model trees described in the previous problem, compute the probability of producing the pattern $A = B = C = D = 0$.
4. Prove or disprove: a character is *parsimony uninformative* if and only if it is compatible on every tree. (A character is said to be “parsimony uninformative” means that if you perform a maximum parsimony analysis without the character, you get the identical output as if you had kept the character. This is the same as saying that the character has the same score on every tree.)
5. Prove or disprove: maximum compatibility and maximum parsimony return the exact same set of optimal trees on every input.
6. 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 $[D_{ij}]$ satisfying $L_\infty(d, D)$ minimum over all $n \times n$ additive matrices $[D_{ij}]$.

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 .)
7. Write down the Dynamic Programming algorithm for computing the cost of the optimal global pairwise alignment when all indels and mismatches have cost 1, and then apply it to the following two sequences:
X = ACTA
Y = ATATACA

How many optimal global pairwise alignments can you find? Show two of them. (For this problem, just present the properly filled in DP matrix, the number of optimal global pairwise alignments you found, and two of them - no need to show the calculations for filling in the DP matrix.)

8. 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.

9. 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?

Extra Credit Instructions: If you wish to attempt the extra credit, please do *all* the problems on the exam as well.

Consider the following stochastic model of character evolution down a tree. (This is a strange model I will present!) The model tree is a rooted and binary tree, with node set $V = \{v_1, v_2, \dots, v_{2n-1}\}$, 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 .

- Prove this model is identifiable!
- Describe a polynomial time algorithm for this problem, prove it statistically consistent under this model, and determine its computational complexity.