CS 598 Algorithmic Genomic Biology
Midterm, 2016

Instructions: Do all the problems in Problem Set 1. Then, do either Problem Set 2 or one of the research projects in Problem Set 3. If you wish, you can do both Problems set 2 and one of the research projects in Problem Set 3, and I'll give you the best of the two grades.

This is an open-book exam. You are welcome to look online for help, read textbooks, papers, etc. You are also welcome to ask me (Tandy Warnow) for help. However, you should not discuss the problems with anyone else.

The midterm counts 40% of your course grade. Please write clearly so that you can receive full points. It is due at 11:10 AM on March 29, in class. If you will be late to class, please make sure I receive your solutions before then, for example by email or by giving your solutions to Elaine Wilson (my assistant).

Problem Set 1: 60 points

1. Define what it means for a matrix to be ultrametric, and draw the rooted tree with edge weights that proves that a 4×4 matrix in which all diagonal entries are 0 and all off-diagonal entries are 4 is ultrametric.

2. Consider the unrooted tree given by $(1, ((2, 3), (4, (8, 9)), (5, (6, 7))))$. Root this tree at leaf 5, draw this rooted tree, write the Newick string for the rooted tree you obtain.

3. Draw two binary unrooted trees on leafset $\{a, b, c, d, e, f\}$ that induce the same tree on $\{a, b, c, d, e\}$ but have no non-trivial bipartitions in common.

4. This problem has to do with the two matrices in Tables 1 and 2, below. One of the two matrices given below is additive.

   (a) Find the additive matrix, and describe the method you used to find it.

   (b) For the additive matrix, draw its corresponding unrooted tree topology.

   (c) Now compute the lengths of each of the internal branches of the tree.

	a	b	c	d	e
a	0.00	0.30	0.45	0.30	0.60
b	0.30	0.00	0.35	0.40	0.50
c	0.45	0.35	0.00	0.55	0.35
d	0.30	0.40	0.55	0.00	0.70
e	0.60	0.50	0.35	0.70	0.00

Table 1: T1

	a	b	c	d	e
a	0.00	0.15	0.20	0.15	0.30
b	0.15	0.00	0.20	0.20	0.25
c	0.20	0.20	0.00	0.25	0.30
d	0.15	0.20	0.25	0.00	0.40
e	0.30	0.25	0.30	0.40	0.00

Table 2: T2

5. Suppose $l_1, l_2, \ldots, l_k$ are non-negative integers, and M_{ij} is a matrix defined by $M_{ij} = 0$ if $i = j$ and otherwise $M_{ij} = l_i + l_j$. Is M additive? Prove or disprove.

6. Suppose you have a collection $\mathcal{T}$ of binary trees, each of them different, all on the same leafset $\{1, 2, 3, \ldots, n\}$. Suppose that the set $\mathcal{T}$ is compatible. Express the maximum size of $\mathcal{T}$ as a function of n.

7. Consider the following three unrooted trees:

   - $T_1 = (1, (3, (5, (6, 7))))$
   - $T_2 = (1, (2, ((4, 8), (3, 7))))$
   - $T_3 = (2, ((4, (3, 5)), 1))$

   Answer the following questions:

   (a) Are these unrooted trees compatible? Justify your answer.

   (b) Root all the three trees at leaf 1, and draw the rooted versions of these trees. Are these rooted trees compatible? Justify your answer.

8. Let $S = \{s_1, s_2, \ldots, s_n\}$ be a set of binary sequences of length k, and let (T, P) be a rooted CFN tree on the same leafset, where T is the model tree topology and P is the set of substitution probabilities on the edges. Which of the following is the correct running time of the dynamic programming algorithm for computing the $Pr(S|(T, P))$?

   - $\Theta(nk)$
   - $\Theta(2^n k)$
   - $\Theta(2^k n)$
   - $\Theta(n^2 k)$

9. Let $S = \{s_1, s_2, \ldots, s_n\}$ be a set of binary sequences of length k and let T be a binary tree on the same leafset. Which of the following is the correct running time of the dynamic programming algorithm for computing the parsimony score of T with this set of sequences at the leaves?

   - $\Theta(nk)$
   - $\Theta(2^n k)$
   - $\Theta(2^k n)$
   - $\Theta(n^2 k)$

10. Suppose every substitution costs 1 and a single letter indel has cost 2 (so an indel of length K costs $2K$). You want to compute the edit distance between two input strings, using dynamic programming.

    (a) Fill in the DP matrix for the edit distance between $S = AAGTAT$ and $S' = CAAGGAC$.

(b) What is the edit distance between S and S'?

(c) What is the minimum edit transformation achieving this edit distance?

(d) What is the pairwise alignment for this transformation?

Problem Set 2: 40 points

1. (10pts) Suppose you have a collection $\mathcal{T}$ of trees, not necessarily binary but each of them different, all on the same leafset $\{1, 2, 3, \ldots, n\}$. Suppose that the set $\mathcal{T}$ is compatible. What is the maximum size of $\mathcal{T}$ (expressed as a function of n).

2. (10pts) Let S and S' be two DNA sequences, with S of length L and S' of length L'. Give a polynomial time dynamic programming algorithm to determine the length of the longest common subsequence of S and S'. (Note that a common subsequence is not the same thing as a common substring; for example, AAA is a common subsequence of S=ATTGATA and S'=TAGGATCA, but AAA is not a substring of either S or S'.)

3. (20pts) Consider the following type of character evolution down a rooted binary tree T, in which every node is labelled by a unique integer (which may be positive, negative, or zero); note this means that in a tree with n leaves, there are $2n - 2$ distinct labels. We do not assume that the label of a node is larger or smaller than its parent node, but we do assume that the label at the root is 0. The state of the character at the root is always 0.

   Every edge e in the tree T has a substitution probability $p(e)$ with $0 < p(e) < 1$. On an edge $e = (x, y)$, with x the parent of y, the character changes its state with probability $p(e)$; if it changes state, then the new state is y. As with other models we've studied, if there are multiple sites that evolve down the same tree, we assume that the substitution probabilities $p(e)$ govern all the sites, but can differ between edges. We also assume that the labels at the nodes are part of the model tree, and so are the same for all characters that evolve down the tree.

   (a) Suppose the rooted model tree T has topology $(a, (b, c))$. Let the parent of b and c be labelled by 3, and let a be labelled by 5, b be labelled by 2 and c be labelled by 4. Recall that the root is always labelled by 0.

   - Suppose that a character evolves down this model tree but *never changes its state*. What are the character states at the leaves (a, b, c) for this character?
   - Suppose that the character evolves down this model tree and changes exactly once - on the edge from the root to a; what are the character states at the leaves for this character?
   - Suppose the character evolves down this model tree and changes exactly once - on the edge from the root to the parent of b and c. What are the character states at the leaves for this character?
   - Suppose the character evolves down this model tree and changes state on every edge of the tree. What are the character states at the leaves of the character?

(b) Suppose the following four sequences evolve down some unknown model tree of this type:

- $u = (3, 0, 1)$
- $v = (3, 0, 5)$
- $w = (0, 8, 2)$
- $x = (0, 8, 4)$

What is the tree topology, and what are the labels at the nodes of the tree? (Recall we already know that the root label is 0.)

(c) Suppose the following five sequences evolve down some unknown model tree of this type:

- $A = (4, 2, 0, 3, 1)$
- $B = (4, 2, 0, 3, 6)$
- $C = (0, 2, 0, 3, 7)$
- $D = (0, 0, 0, 3, 8)$
- $E = (0, 0, 5, 5, 9)$
- $F = (0, 0, 5, 5, 10)$

What is the tree topology, and what are the labels at the nodes of the tree?

(d) Suppose the following three sequences are given to you. Is it possible that they evolve down some unknown model tree of this type?

- $A = (4, 0)$
- $B = (4, 2)$
- $C = (0, 2)$

If so, present the tree; otherwise prove this cannot be the case.

(e) Describe a polynomial time statistically consistent method to infer the model tree topology from the site patterns. What is the running time of your algorithm? (Don't just say "polynomial".) What is your justification for saying it is statistically consistent under this model?

Problem Set 3: 40 points Do one of the following.

1. Write a paper in which you compare gene trees computed on a biological dataset with at least 50 unaligned sequences using at least two different techniques. You can use your own dataset or find a published dataset. Your paper should provide enough detail to be reproducible (e.g., software version numbers and commands, access to datasets), and should have some interesting discussion about what you observe, and if you were able to make comments about the methods you used. Your grade on this problem will be based on the content, writing, and scientific insight.

    - If you wish, you can use an "alignment-free" method (of your choice), in addition to a method that either co-estimates alignments and trees (e.g., PASTA) or a two-phase method. If you use two-phase methods, then use at least two different multiple sequence alignment methods, of which at least one must be from the following set – Clustal, MAFFT, Opal, Prank, PAGAN, PASTA, and UPP – and then compute a maximum likelihood tree (any software you like). If you compare ways of running PASTA, vary one of the following parameters: subset size, subset aligner, or decomposition strategy (longest branch vs. centroid).
    - Get bootstrap support on the branches of the tree you compute.
    - Compare the gene trees, taking bootstrap support into account. Where are they different? Are these differences interesting or important? What is your interpretation of these differences? If one method did particularly poorly, was there something about the data that was difficult for the method? What did you learn about the methods you used?

2. Write a paper in which you compare species trees computed on a biological dataset with at least 10 genes and between 10 and 100 species. It would be most interesting if you pick a dataset where gene tree heterogeneity has been observed or where it is expected. You can use your own dataset or find a published dataset. Your paper should provide enough detail to be reproducible (e.g., software version numbers and commands, access to datasets), and should have some interesting discussion about what you observe, and if you were able to make comments about the methods you used. Your grade on this problem will be based on the content, writing, and scientific insight.

    - Compute gene sequence alignments and gene trees using reasonable methods. (If you are using a dataset from a published study, these may already be computed for you!)
    - Compute species trees using at least two coalescent-based methods and one concatenation analysis. Reasonably fast coalescent-based methods include SVDquartets, MP-EST, ASTRAL, and ASTRID.

- Compare the species trees that you obtain using different species tree estimation methods. Where are they different? Are these differences interesting or important? What is your interpretation of these differences? What does this tell you about the methods you used?