Midterm, CS 581, Fall 2020

Instructions:

• The exam is due on October 11, 2020, by 10 PM, uploaded as a single PDF document into Moodle.

• You are not allowed to discuss this exam with anyone but the TA (Vlad Smirnov) or professor (Tandy Warnow).

• This is an open book exam in that you are welcome to consult the textbook. However, you should not copy text from the book – you should put answers in your own words. Also, you should not refer to any other material (i.e., other textbooks, websites, etc.). If you have questions about the problems in this exam (meaning if you do not understand what is asked), please contact the T.A. (Vladimir Smirnov); or me (Tandy Warnow) no later than October 8 at 5 PM; we will not be available to answer questions after that time.

• Note that for several questions, no justification is requested (and hence please don’t provide it, as no partial credit is given for those questions). For the other questions, it is not desirable that you go into tremendous detail; brief answers are most likely sufficient.
Part I: True/False (20 points) Instructions: Do not show your work. Circle the numbers that correspond to true statements. (1 point for each correct answer)

1. Maximum Parsimony and Maximum Likelihood under the No Common Mechanism Model have the same optimal trees.
2. CFN distances computed for finite length sequences always satisfy the triangle inequality.
3. The wheel graph on \( n \) vertices, with \( n \geq 5 \), is triangulated. (Note that a wheel graph is composed of \( n - 1 \) vertices in a cycle, each of them adjacent to the same node \( x \) that is not in the cycle.)
4. UPGMA applied to Hamming distances is statistically consistent under the CFN model when the strict molecular clock holds.
5. Let \( X \) be a set of unrooted binary trees, and we encode each tree \( T \) in \( X \) by its set \( Q(T) \) of quartet trees. Then \( X \) is compatible if and only if \( \bigcup_{T \in X} Q(T) \) is a compatible set of quartet trees.
6. Maximum likelihood under the Generalized Time Reversible model is statistically consistent, even when sequences are generated under the Jukes-Cantor model.
7. The problem of computing the probability that a given profile HMM generates a given sequence is NP-hard.
8. Suppose that the true tree \( T \) is fully resolved (i.e., binary) and \( \hat{T} \) is an estimated tree that may not be fully resolved. It is possible for the number of false positives of \( \hat{T} \) with respect to \( T \) to be larger than the number of false negatives.
9. Suppose we have a set \( X \) of DNA sequences generated by some GTR model tree. Now let \( Y \) be the set of sequences obtained by replacing every sequence in \( X \) by a sequence obtained by:
   - Changing every \( A \) into a \( C \), and every \( C \) into an \( A \)
   - Changing every \( T \) into a \( G \) and every \( G \) into a \( T \)
   Then the set of optimal maximum parsimony trees for \( X \) is the same as the set of optimal maximum parsimony trees for \( Y \).
10. Suppose we have a set \( X \) of DNA sequences generated by some GTR model tree. Now let \( Y \) be the set of sequences obtained by replacing every sequence in \( X \) by a sequence obtained by:
    - Changing every \( A \) into a \( C \), and every \( C \) into an \( A \)
    - Changing every \( T \) into a \( G \) and every \( G \) into a \( T \)
    Then the set of optimal maximum likelihood trees for \( X \) is the same as the set of optimal maximum likelihood trees for \( Y \) (where the likelihood is optimized under the GTR model).
11. If \( A \) and \( B \) are both \( n \times n \) additive matrices, then \( A + B \) is also additive.
12. Suppose \( \mathcal{T} \) is a compatible set of trees, each leaf-labelled by a subset of \( S \). Then every solution to MRP applied to \( \mathcal{T} \) is a compatibility tree for \( \mathcal{T} \).
13. If sequences evolve under a CFN model and also under strict molecular clock, then the rooted model tree topology is identifiable.
14. Every SPR move is a TBR move.
15. SVDquartets is a method for estimating species trees from unrooted gene trees.
16. Although statistical consistency cannot be established using a simulation study, statistical inconsistency can be established using simulations.
17. Saying that a phylogeny estimation method is statistically inconsistent means that it will converge to the wrong tree as the amount of data increases.

18. Maximum likelihood under the No Common Mechanism Model is statistically consistent.

19. A majority consensus tree of a set of trees that have the same set of leaves always exists and is unique.

20. The majority consensus tree of a set of two different trees on the same leafset is never fully resolved.
Part II: Multiple choice (9 points)  For each problem, select the best single answer. Do not show your work. (3 points for each correct answer)

1. Let $d$ be an $n \times n$ dissimilarity matrix. Finding an $n \times n$ additive matrix $D$ that minimizes
\[ \sum_{i,j} |D_{i,j} - d_{i,j}|^2 \]
is
(a) NP-hard
(b) solvable in polynomial time
(c) of unknown computational complexity

2. Suppose that every indel and every substitution has unit cost. Suppose that you are given a binary tree $T$ and an assignment of sequences to the leaves of $T$. Then the problem of finding optimal sequences at the internal nodes of $T$ to minimize the total edit distance across the tree (defined by summing over all the edges the edit distances between sequences labelling the endpoints of the edges) is
(a) NP-hard
(b) solvable in polynomial time
(c) of unknown computational complexity

3. The problem of computing the minimum edit distance between two sequences when every substitution has cost 1 and a gap of length $L > 0$ has cost $L^2$ is
(a) NP-hard
(b) solvable in polynomial time
(c) of unknown computational complexity
Part III: Short Calculations (31 points) Do the calculations requested, but unless otherwise asked, do not justify your steps.

1. (6 pts) Consider the following pair of unrooted trees $T_1 = ((a, b, c), (d, e, (f, g)))$ and $T_2 = (a, (c, (b, (d, (e, (f, g))))))$.
   - Draw $T_1$ and $T_2$ (2 pts)
   - Draw three different trees $A, B, C$ so that $T_1$ is their strict consensus and $T_2$ is their majority consensus (4 pts)
2. (10 pts) Consider the following set of four sequences:

- \( u = (b, a, b, c, c, b) \)
- \( v = (b, b, a, c, a, b) \)
- \( w = (c, b, a, d, a, b) \)
- \( x = (c, b, d, d, c, c) \)

(a) What is the best maximum parsimony score for this input? (3 pts)

(b) Draw the maximum parsimony tree(s) on this input, and explain how you got the result. (3 pts)

(c) Draw the unrooted version of the tree produced by UPGMA on this input, given the matrix of normalized Hamming distances (and explain how you got the result). (4 pt)
3. (15 pts) Find the pairwise alignment corresponding to the minimum edit distance between sequences $s = ACATTA$ and $s' = CCAT$ under the assumption that each single letter indel costs 2 (and an indel of length $L$ costs $2L$) and each substitution costs 3. Do this by running the modification to Needleman-Wunsch to allow for these cost, and recall that the entries of the matrix are the cost – and so are always non-negative.

   (a) Write down the recursive formula for Needleman-Wunsch for this particular problem (specified by indel cost and substitution cost) (4 pts)

   (b) Show the matrix you compute (6 pts)

   (c) Write down all the pairwise alignments with the best score (5 pts).
4. (6 pts) Consider the gap-free alignment on sequences \( u = CCCCC \) and \( v = TTTTT \). For the following questions (each is worth two points), just give your answers—do not show how you derived them.

(a) Write down the associated unadjusted profile \( P_{u,v} \) for this alignment (i.e., it will not contain any insertion or deletion states).

(b) How many sequences have non-zero probability of being generated by this profile?

(c) Suppose \( x \) is a third sequence of length 5 and we let \( P_{u,x} \) be the associated unadjusted profile for the gap-free alignment of \( u \) and \( x \). Give an example of a sequence \( x \) for which \( P_{u,x} \) can generate exactly two sequences.
Part IV: Longer Questions (40 points) Instructions: Show all your work for these questions.

1. (10 points) Suppose that a set of binary characters evolve down a tree $T$ without any homoplasy. Prove or disprove: the Hamming distance matrix between the sequences at the leaves is additive.
2. (10 pts) Consider the following optimization problem. The input is an $n \times n$ dissimilarity matrix $d$ and the output is an $n \times n$ additive matrix $D$ that minimizes the $L_\infty$-distance to $d$ (i.e., $L_\infty(D, d) \leq L_\infty(D', d)$ for all other $n \times n$ additive matrices $D'$). Suppose $\Phi$ is an algorithm that exactly solves this optimization problem.

- Describe a statistically consistent pipeline for estimating CFN model trees from input binary sequences that uses $\Phi$ (3 pts)

- Sketch a proof that the pipeline is statistically consistent (7 pts)
3. (10 pts) Consider the following algorithms for species tree estimation. The assumption is that the input is a set of true unrooted gene trees sampled from the distribution on unrooted gene trees under the multi-species coalescent (MSC) model. Hence, every tree in the input has the same set of species.

- Algorithm A computes the strict consensus of the input gene trees, and returns that as its estimate of the species tree.
- Algorithm B returns the most frequently observed gene tree (if there is a tie, it returns any of the most frequently observed gene trees).
- Algorithm C throws out half the gene trees (randomly selected) and then runs ASTRAL (in default mode) on the remaining trees.
- Algorithm D adds a random tree to the input set of $k$ gene trees, and then runs ASTRAL (in default mode) on the set of $k + 1$ gene trees it obtains.

For each of these algorithms, answer whether it is statistically consistent under the MSC, and briefly justify your answer. Note that you can use results already established in the textbook. (Points: 2 points for answering about algorithms A and B, 3 points each for answering about algorithms C and D.) Use an extra page if you need it.
4. (10 pts) 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 (i.e., for different edges $e, e'$ we may have $p(e) \neq p(e')$). 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.

Answer (a)–(c) on the following pages, but do not provide your reasoning. For (d), you should provide your reasoning. Also, if you wish to answer (e), you must provide full details, as only a fully correct answer to (e) will yield extra credit.
(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. Each question below asks you to write the character states at the leaves under different scenarios. Recall that the root is always labelled by 0. (4 pts)

i. Suppose that a character evolves down this model tree but *never changes its state*. What are the character states at the leaves for this character? (1 pt)

   $a =$
   $b =$
   $c =$

ii. 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? (1 pt)

   $a =$
   $b =$
   $c =$

iii. 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? (1 pt)

   $a =$
   $b =$
   $c =$

iv. 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? (1 pt)

   $a =$
   $b =$
   $c =$
(b) Suppose the following four sequences evolve down some unknown model tree of this type, and recall we already know that the root label is 0:

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

What is the tree topology, and what are the labels at the nodes of the tree? (2 pts)

(c) Suppose the following six sequences evolve down some unknown model tree of this type, and recall that we know that 0 is the label at the root:

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

What is the tree topology, and what are the labels at the nodes of the tree? (2 pts)
(d) Suppose the following three sequences are given to you. Recall that 0 is the label at the root. 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. (2 pts)
(e) (Extra credit, up to 5 additional points.) 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?