Questions for Chapters 1-9

The answers to most of these questions should be in the textbook. Some of them will require some sleuthing or calculations.

1. Consider the Cavender-Farris-Neyman (CFN) model. What are the parameters of a CFN model tree? What do these parameters mean?

2. What does it mean to say that a method is statistically consistent for estimating the CFN model tree topology?

3. If a method M is statistically consistent under the Jukes-Cantor model, is M consistent under the GTR model as well? What about vice-versa?

4. What is the CFN distance correction? Why is it used? What happens if you use NJ or some other distance-based method (e.g., the Naive Quartet Method) on sequences generated under the CFN model, but you don't correct the distances? Do you still have a statistically consistent method?

5. Is finding the best maximum likelihood (ML) tree topology and numeric parameters solvable in polynomial time?

6. What is the computational complexity of computing the probability of a set of sequences being generated on a given CFN model tree?

7. What is the computational complexity of computing the maximum parsimony (MP) score of a given set of sequences on a fixed tree tpoology?

8. What is the computational complexity of finding a best MP tree for a given set of sequences?

9. What is the computational complexity of determining if a set of unrooted trees is compatible?

10. What is the computational complexity of determining if a set of rooted trees is compatible?

11. What does it mean to say a matrix is additive?

12. What is the Four Point Condition?

13. What is a dissimilarity matrix? Why is it not properly speaking a distance matrix?

14. Suppose you have a method M that takes as input an $n \times n$ dissimilarity matrix d and returns an additive matrix A that minimizes $L_\infty(d, A)$. Is M a statistically consistent method for estimating CFN model tree topologies? (Assume that CFN distances are computed.)

15. Consider a set $\mathcal{T}$ of trees, all of them on the same leafset S. Let T_{sc} be the strict consensus tree of $\mathcal{T}$ and T_{maj} be the majority consensus tree of $\mathcal{T}$. If these two consensus trees are not identical, is it always the case that T_{maj} refines T_{sc}? Why?

16. Suppose T is a binary tree with n leaves. How many edges does it have?

17. Suppose T is a model CFN tree (and hence binary tree) on n leaves, and T^* is a star tree (no internal edges). What is the FN rate of T^*? What is the FP rate of T^*? What is the Robinson-Foulds error rate?

18. Find a Hamming distance matrix M and its CFN distance matrix D so that the Four Point Method applied to M produces a tree T that is different from the result of applying the Four Point Method to D.

19. Consider a CFN model tree with topology $((A, B), (C, D))$, and with substitution probabilities 0.01 on every edge incident with a leaf, and substitution probability 0.49 on the internal edge.

    - Compute the probability of $A = B = 1, C = D = 0$.
    - Compute the probability of $A = B = 0, C = D = 1$.
    - Compute the probability of $A = B = C = D = 1$.
    - Compute the probability of $A = C = 0, B = D = 1$.
    - Compute the probability of $A = D = 0, B = C = 1$.
    - Is MP statistically consistent on this model tree?

20. Consider the simple edit distance between two DNA sequences X and Y to be the minimum number of single nucleotide substitutions and single nucleotide indels needed to transform X into Y.

    - Is the simple edit distance symmetric?
    - Does the simple edit distance satisfy the triangle inequality?
    - What is the computational complexity of computing the simple edit distance between X and Y, if X has length L and Y has length L'?
    - What is the simple edit distance between AAC and $TCGA$?

21. Consider the edit distance cost function where each single letter indel has cost C and every substitution has cost C'. Show how to set C' so that for all X and Y, the minimum cost transformation of X into Y would never include any substitutions.

22. Under the simple edit distance, the cost of an indel of P consecutive letters is just P. Suppose you change the cost of an indel of P consecutive letters to be $C_0 + P$, where C_0 is some constant. How would you compute the minimum edit distance between two strings?

23. Suppose you have two strings X and Y and you want to find a pairwise alignment where you maximize the number of sites with identical letters. How would you solve this problem?

24. Suppose you have two DNA strings X and Y and a “match” value for every pair of nucleotides. Some match values can be negative, but when the two nucleotides are identical the match value is always positive. You want to find a pairwise alignment that has the maximum total match value. How would you solve this problem?

25. Suppose you have two DNA strings X and Y and a “match” value for every pair of nucleotides. Some match values can be negative, but when the two nucleotides are identical the match value is always positive. You want to find a substring Y' of Y so that the pairwise alignment of X and Y' has the maximum total match value. (A substring is a consecutive subsequence.) How would you solve this problem? (Hint: this is called “local alignment”.)