MATH 469/769 ASSIGNMENT #4

Assignment Due: April 4, 2025, 5:30pm

The total number of marks possible marks for the assignment is 50. All students should attempt all questions. Make sure you show all your work, and make sure that your work is your own. Your reasoning and work is more important than your answer.

---

1. (a) Why is a trivial split always compatible with any other split?

   (b) Show that the following multiple sequence alignment does not admit a perfect phylogeny:

Taxon	1	2	3	4	5	6	7	8	9	10
a	A	T	G	C	C	A	T	G	A	A
b	A	T	G	C	C	T	T	G	T	G
c	C	T	A	A	C	A	C	G	T	G
d	A	A	A	A	T	A	T	A	T	G
e	A	T	A	C	C	A	C	G	A	A

[8]

2. (a) Use tree popping for this MSA, moving from left to right to create a phylogenetic tree.

Taxon	1	2	3	4	5	6	7	8	9	10
a	A	A	A	C	C	A	T	G	T	G
b	A	A	A	C	C	A	T	G	T	G
c	C	T	A	A	C	T	T	G	T	A
d	A	T	A	A	T	T	T	A	T	A
e	A	T	G	A	C	A	C	G	A	A
f	A	T	A	C	C	A	T	G	T	G
g	A	T	A	A	C	A	T	G	A	A

   Try another order and verify that you arrive at the same tree.

   (b) Show that a tree obtained by tree popping, using a compatible set of splits, is unique.

[12]

3. Consider the Jukes-Cantor distance between two sequences,

$$d_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}d_f)$$

where d_f is the fraction of sites that differ between the two sequences.

   (a) What is d_{JC} for a pair of identical sequences?

(b) What happens when $d_f > 3/4$? Intuitively, why would such a high value of d_f not make sense in the Jukes-Cantor model?

(c) What happens to the JC distance as $d_f \to 3/4$ from below (i.e. d_f is just under 3/4)? What is you interpretation of this?

[8]

4. A matrix D on a set of taxa X satisfies the triangle inequality if for any (a, b, c) in X,

$$D(a, b) \leq D(a, c) + D(b, c)$$

(aside: this is a requirement for a metric, and if D meets it, we can call D a distance matrix. Otherwise, we refer to D as a dissimilarity matrix).

(a) Draw the (only) unrooted tree that is possible for 3 taxa (a, b, c). Let x, y and z denote the lengths of the branches.

(b) For a set of distances on the three taxa, given by D, show that if D obeys the triangle inequality, the tree's path lengths can match the distances perfectly, and solve for the branch lengths.

(c) In class, we discussed the 4-point condition, namely that for any 4 taxa in X, we can label them i, j, k, ℓ such that

$$D(i, k) + D(j, \ell) = D(i, \ell) + D(j, k) = L$$

and

$$L \geq D(i, j) + D(k, \ell).$$

Interpret the 4-point condition in terms of splits and the 4-gamete theorem.

(d) Write a set of distances on 4 taxa that does not correpond to the path lengths in any tree.

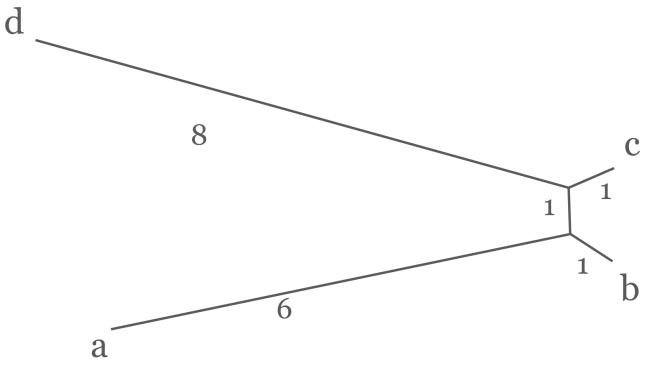
[12]

5. We talked about the neighbour-joining method in class. The choice of a pair that minimizes Q, where

$$Q_{ij} = (n - 2)D_{ij} - \sum_k D_{ik} - \sum_k D_{jk},$$

is somewhat mysterious. (Why not just choose the closest pair of tips to join, at each stage?)

Consider this tree, with branch lengths indicated:



(a) If you used D to choose which nodes to join, in the first iteration of neighbour joining, which nodes would be joined?

(b) Perform neighbour joining to recover the tree shown (use a computer if you like). Which two nodes are joined in the first iteration of the algorithm?

[10]