Homework #5

Selected solutions

5.7 Exercises

1. For the tree in Figure 5.3 constructed by UPGMA, compute a table of distances between taxa along the tree. How does this compare to the original dissimilarities of Table 5.1?

   Solution 1. For the tree in Figure 5.3 constructed by UPGMA, a table of distances between taxa along the tree is:

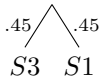
	S1	S2	S3	S4
S1		.425	.27	.55
S2			.425	.55
S3				.55

   It is different from the original one except for the closest two taxa which are $S1$ and $S3$. For some pairs, the new distance is larger than the original one and for some of them the new distance is smaller.

2. Suppose four sequences S1, S2, S3, and S4 of DNA are separated by dissimilarities as in Table 5.9. Construct a rooted tree showing the relationships between S1, S2, S3, and S4 by UPGMA.

	S1	S2	S3	S4
S1		1.2	.9	1.7
S2			1.1	1.9
S3				1.6

   Solution 2. From the table, we see that $S1$ and $S3$ are closest in the tree. Since $.9/2 = .45$, then

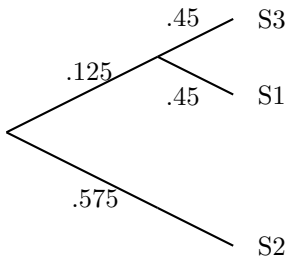


   Since $S1$ and $S3$ have been joined, we now collapse them into a single combined group $S1 - S3$ and find the distance between them and another taxa. Then we have

$$d(S1 - S3, S2) = \frac{d(S1, S2) + d(S3, S2)}{2} = \frac{1.2 + 1.1}{2} = 1.15$$
$$d(S1 - S3, S4) = \frac{d(S1, S4) + d(S3, S4)}{2} = \frac{1.7 + 1.6}{2} = 1.65.$$

	$S1 - S3$	S2	S4
$S1 - S3$		1.15	1.65
S2			1.9

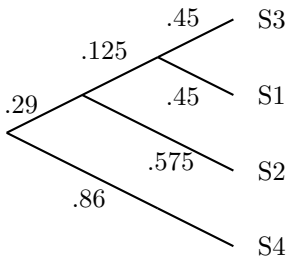
   Now we repeat the process, using the distances in the collapsed table. So $1.15/2 = .575$.



For the last step we find the distance between $S1 - S2 - S3$ with $S4$ that is

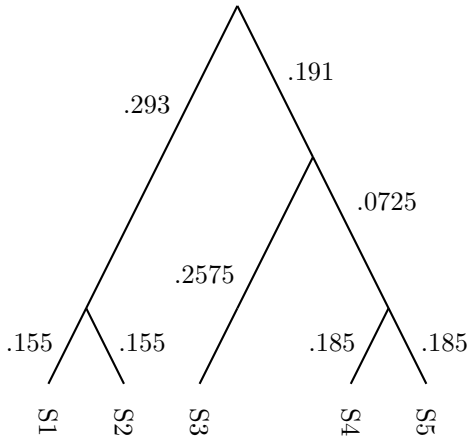
$$d(S1 - S2 - S3, S4) = \frac{d(S1, S4) + d(S2, S4) + d(S3, S4)}{3} = \frac{1.7 + 1.9 + 1.6}{3} = 1.73.$$

Then the UPGMA tree is



3. Perform UPGMA on the data in Table 5.3 that was used in the text in the example of the FM algorithm. Does UPGMA produce the same tree as the FM algorithm topologically? metrically?

Solution 3.



The topology of FM tree is the unrooted version of UPGMA tree, but they are different metrically.

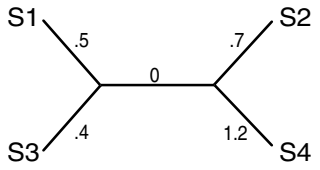
4 b. (plus a.) The fact that dissimilarity data relating three taxa can be exactly fit by appropriate edge lengths on the single unrooted topological 3-taxon tree is used in the Neighbor Joining algorithm.

a. There are several algebraic approaches: Either *ad hoc* algebra or methodical elimination of variables can be used, or matrix algebra. The nicest solution (since it makes the formulas memorable) is a geometric one: $d_{AB} + d_{AC}$ includes the edge x twice, and the edges y and z once, so subtracting d_{BC} gives $2x$, etc.

b. $x = .555$, $y = .079$, $z = .772$

5. Use the FM algorithm to construct an unrooted tree for the data in Table 5.9 that were also used in Problem 2. How different is the result?

Solution 5.



In problem 2, we get a rooted binary tree, but here we get an unrooted non-binary tree. In performing the algorithm, we join taxa in the same order, but for FM here it turns out we get a 0 branch length.

6. A desirable feature of a dissimilarity map on sequences is that it be *additive*, in the sense that if S0 is an ancestor of S1, which is in turn an ancestor of S2, then

$$d(\mathrm{S0}, \mathrm{S2}) = d(\mathrm{S0}, \mathrm{S1}) + d(\mathrm{S1}, \mathrm{S2}).$$

a. Explain why an additive dissimilarity map is desirable if we are trying to use dissimilarities to construct metric trees.

b. Give an example of sequences to illustrate that the Hamming dissimilarity might not be additive.

c. If mutations are rare, why might the Hamming dissimilarity be approximately additive?

Solution 6.

a. On a tree, we would definitely want distances to be additive so it's natural to want that for a dissimilarity map too.

b. Consider the single site:

$$\begin{array}{ll} S_0\text{:} & \texttt{A} \\ S_1\text{:} & \texttt{G} \\ S_2\text{:} & \texttt{A} \end{array}$$

Then $d(S0, S2) = 0$, $d(S0, S1) = 1$ and $d(S1, S2) = 1$. We can see that that the Hamming dissimilarity is not additive since 2 mutations are not seen in the computation of $d(S0, S2) = 0$.

c. As we can see in part (b), the problem is caused in Hamming dissimilarity because of repeated substitutions that are not picked up in comparing S_0 and S_2. If mutations are rare, then multiple substitutions are rare too and the Hamming distance will not 'miss' them (or miss only a few of them) when comparing S_0 and S_2.

7. While any dissimilarity values between 3 taxa can be fit to a metric tree (possibly with negative edge lengths), that's not the case if we want to fit an ultrametric tree.

a. If the three dissimilarities are 0.3, 0.4, and 0.5, find an unrooted tree that fits them, and explain why no choice of a root location can make this tree ultrametric.

b. If the three dissimilarities are 0.2, 0.3, 0.3 find an unrooted tree that fits them, and locate a root to make it ultrametric.

c. If the three dissimilarities are 0.3, 0.3, and 0.4, find an unrooted tree that fits them, and explain why no choice of a root location can make this tree ultrametric.

Solution 7.

a. In order for a molecular clock hypothesis to hold, all the terminal vertices would have to be equidistant from the root. This is impossible. The root cannot be placed at the internal node since the edge lengths are different. Moreover, the root cannot be placed on any of the three edges since no two of the edges have the same length.

b. Since the two shortest edge lengths are equal to .1, it is possible to assume a molecular clock. The root would have to be placed on the edge of length .2 at a distance of .05 from the internal node. Then all terminal vertices are .15 from the root.

c. Here two of the edge lengths are equal, but their length .2 is larger than the length of the third edge. This means it is not possible to locate the root on either of the longer edges nor the shorter edge and achieve equal distances from the root. Of course, the internal node could not serve as a root either, if a molecular clock is to be assumed.

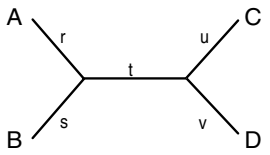
8. While distance data for 3 taxa can be exactly fit to an unrooted tree, if there are 4 (or more) taxa, this is usually not possible.

a. For the tree $((a, b), (c, d))$, denoting distances between taxa with notation like d_{ab}, write down equations for each of the 6 such distances in terms of the 5 edge lengths. Explain why if you use dissimilarity values in place of the distances these equations are not likely to have an exact solution.

b. Give a concrete example of 6 dissimilarity values so that the equations in part (a) cannot be solved exactly. Give another example of values where the equations can be solved.

Solution 8.

a.

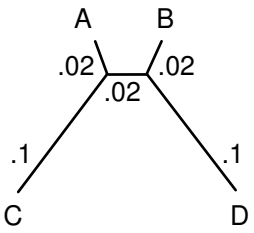


b. $d_{AB} = r + s$, $d_{AC} = r + t + u$, $d_{AD} = r + t + v$, $d_{BC} = s + t + u$, $d_{BD} = s + t + v$, $d_{CD} = u + v$; As this is a system of six equations in only five unknowns, in general there will not be a solution.

c. Answers may vary; one possibility follows. For the distances $d_{AB} = .2$, $d_{AC} = .3$, $d_{AD} = 1.33$, $d_{BC} = .29$, $d_{BD} = 1.3$, $d_{CD} = 1.19$, the system does not have a solution, whereas for the distances $d_{AB} = .17$, $d_{AC} = .32$, $d_{AD} = 1.33$, $d_{BC} = .29$, $d_{BD} = 1.3$, $d_{CD} = 1.19$, the system has a solution.

A good way to generate such examples is to start with a 4-taxon tree and write down the pairwise distances. These values will have a solution. Why? Then perturb just one entry of the table and there should be no solution Why?

11. Suppose the unrooted metric tree in Figure 5.12 correctly describes the evolution of taxa A, B, C, and D.



a. Explain why, regardless of the location of the root, a molecular clock could not have operated.

b. Give a dissimilarity table by calculating tree metric distances between each pair of the four taxa. Perform UPGMA on that data.

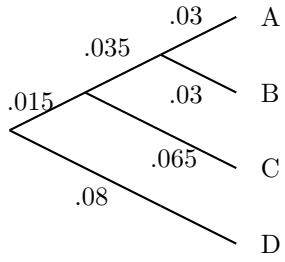
c. UPGMA did not reconstruct the correct tree. Where did it go wrong? What was it about this metric tree that led it astray?

d. Explain why the FM algorithm will also not reconstruct the correct tree.

Solution 11.

The closest taxa metrically are not *sister* on the correct tree. Both UPGMA and FM will fail in the very first step.

By UPGMA, we get



13. For the quartet tree $((a, d), (b, c))$ that was not explicitly treated in Section 5.4 , write the inequalities and equalities that hold expressing the four-point condition.

Solution 13. The inequalities and equalities that hold expressing the four-point condition for this quartet are:

$$d(a, d) + d(b, c) < d(a, c) + d(b, d) = d(a, b) + d(b, c).$$