CS195Z: Problem Set 2
Solution Key

April 2, 2009

1. The compatibility graph and tree are:

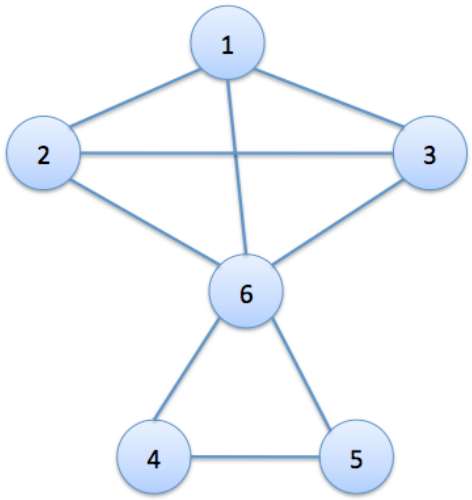


Figure 1: Compatibility graph.

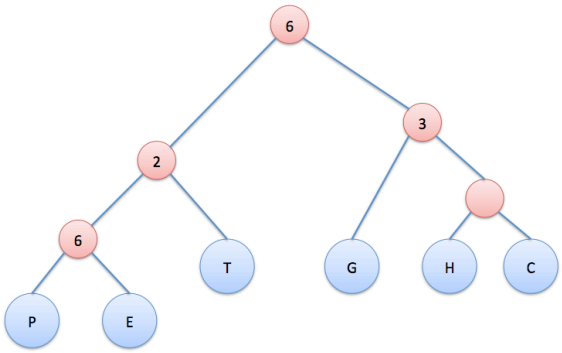


Figure 2: Compatibility tree.

2. For $n \leq 3$, the compatibility and parsimony trees are always the same. Parsimony trees minimize homoplasy. Recall that a character is incompatible with a tree if it requires exactly 2 changes of state,

whereas compatible characters require 0 or 1 changes of state. With $n = 4$, each character cannot evolve with homoplasy. So parsimony tree must equal compatibility tree. With 5 species, there is exactly one tree topology, and there are exactly 3 internal nodes in a compatibility tree. Consider the edge e shown in Figure 3 below. For every subtree with at most 3 leaves, we know the three leaves are compatible (because they must pass the four gametes test). This is true, for example, of the subtrees containing leaves $\{1, 3, 8\}$ and $\{4, 6, 8\}$ in Figure 4. The number of mutations on e, therefore, must equal the number of mutations on edges $e1$ and $e2$ below in Figure 4. So, the compatibility tree minimizes homoplasy, which implies it is equal to the parsimony tree. Parsimony trees with $n \geq 6$ need not equal compatibility trees because changing a character twice along a path may give a better parsimony score (even though this is homoplasy).

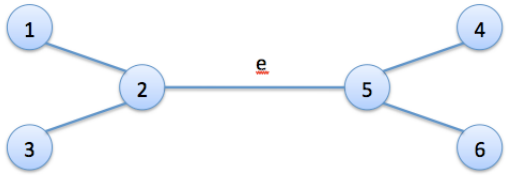


Figure 3: (a)

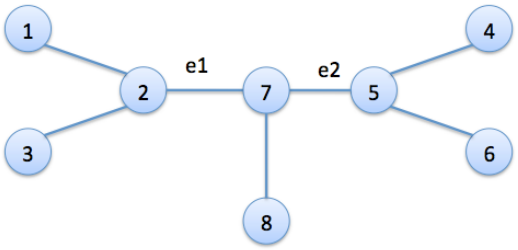


Figure 4: (b)

3. (a) Tree-derived distances are defined by positive weights on edges. Therefore, since the path between any two nodes contains either 0 or some positive number of edges, the total distance must be nonnegative.

   (b) $\rightarrow$: Because weights on edges are positive, a distance $d(x, y) = 0$ implies the path from x to y contains 0 edges, so x and y must be the same node.

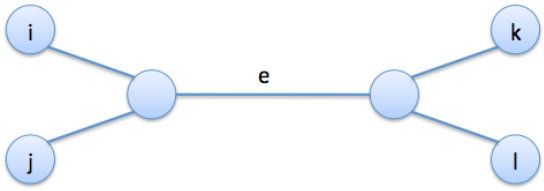
   $\leftarrow$: If 2 nodes are the same, then the path between them contains 0 edges. The sum of 0 positive numbers is 0.

   (c) By the definition of a tree, there is a unique path from x to y. Directions of edges don't affect their weights. So, the sum of weights on the x-to-y path is the same as the sum of weights on the y-to-x path.

   (d) Either z is on the path between x and y or it is not. If it is, then the x-to-y path is equal to the concatenation of the x-to-z and z-to-y paths, implying $d(x, y) = d(x, z) + d(z, y)$. If it is not, then let v be the node at which the x-to-y and x-to-z paths diverge. Then $d(x, y) = d(x, v) + d(v, y)$, $d(x, z) = d(x, v) + d(v, z)$, and $d(y, z) = d(y, v) + d(v, z)$. So, $d(x, y) = d(x, v) + d(v, y) < d(x, v) + d(v, y) + 2d(v, z) = d(x, z) + d(z, y)$.

4. To compute the trimming parameter δ, we must consider quartets of leaf nodes. For a quartet of leaves $1 \leq i, j, k, l \leq n$, such that i, j and k, l are neighbors, respectively, we compute the min-weight leaf

edge among the 4 of them. Let e be the edge (or path) between the LCAs of i, j and k, l, as in Figure 4.



The weight of e is:

$$d(e) = \frac{d(i,k) + d(j,l) - d(i,j) - d(k,l)}{2} \tag{1}$$

Readjust the pairwise distances between these four leaves by subtracting $d(e)$ from them:

$$d'(i,k) = d(i,k) - d(e) \tag{2}$$
$$d'(i,l) = d(i,l) - d(e) \tag{3}$$
$$d'(j,k) = d(j,k) - d(e) \tag{4}$$
$$d'(j,l) = d(j,l) - d(e) \tag{5}$$

Of these 4 leaves, then, we can decide which leaf has the shortest leaf edge by summing the distances from each leaf to all other nodes. The leaf that minimizes this sum is the one whose leaf edge is shortest. Without loss of generality, suppose i is this node. Then we compute the weight of i's leaf edge by:

$$\delta_i = \frac{d'(i,j) + d'(i,k) - d'(j,k)}{2} \tag{6}$$

Then we delete i, j, k, l from the graph and, for every other quartet of leaves, we compute the min leaf edge out of those 4 leaves. Once we have considered every node (as part of some quartet), we can take the min δ computed over all quartets as the trimming parameter.

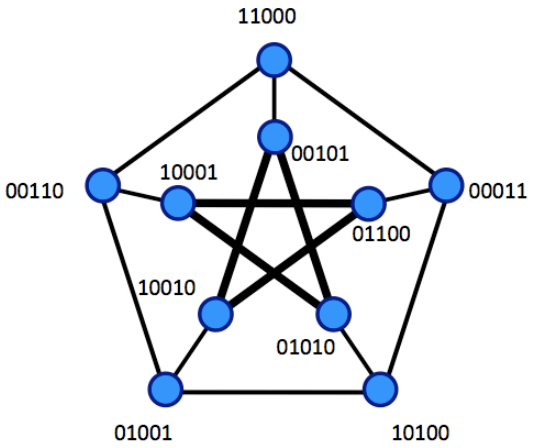
5. A distance is additive if and only if the four-point condition holds for every quartet of nodes. For an n x n matrix that is ultrametric, then for any 3 species i, j, k, it must be the case that two of the pairwise distances between them are equal and are at least as great as the third. Without loss of generality, let us suppose that, for leaves i, j, k, $d(i,k) = d(j,k) \geq d(i,j)$. Now consider another leaf l such that the triple i, j, l satisfies $d(i,l) = d(j,l) \geq d(i,j)$. Then because $d(i,l) = d(j,l)$ and $d(i,k) = d(j,k)$ and $d(k,i) \geq d(i,j)$ and $d(l,i) \geq d(i,j)$, it must be the case that $d(k,i) = d(l,i) \geq d(k,l)$ (by the ultrametric property). Therefore, we have that $d(i,k) = d(i,l) = d(j,k) = d(j,l)$. Therefore, we have that $d(i,j) + d(k,l) \leq d(i,k) + d(j,l) = d(i,l) + d(j,k)$, which means that the four-point condition holds.

6. (a) Let c be a binary character. Suppose that out of n species, i of them have state 0. (Note that for $n < 3$, there cannot be non-trivial characters.) If $2 \leq i \leq n-2$, then c is non-trivial. We can choose how to assign the 0 state to species in $\binom{n}{i}$ ways. Then the total number of non-trivial binary characters on n species is:

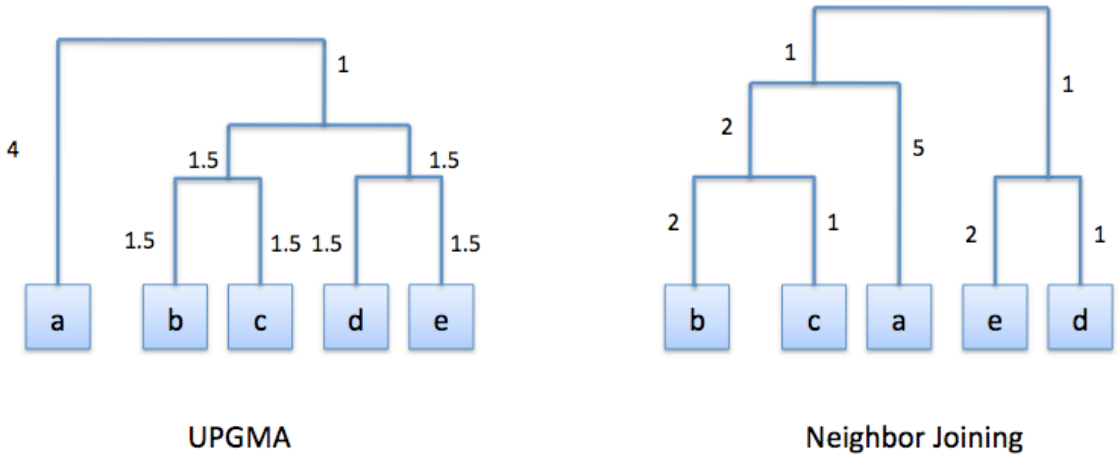
$$\sum_{i=2}^{n-2} \binom{n}{i} = s^n - \binom{n}{0} - \binom{n}{1} - \binom{n}{n-1} - \binom{n}{n} = 2^n - 1 - n - n - 1 = 2^n - 2n - 2 \tag{7}$$

Because we consider complementary state assignments to be the same, i.e. 00011 is the same as 11100, we must then divide this amount by 2: $2^{n-1} - n - 1$.

(b) For $n = 5$, there are 10 non-trivial characters. The compatibility graph of non-trivial characters on 5 species is given by the Peterson graph:



7. The trees for UPGMA and Neighbor-Joining are:



8. We express the likelihood of sequences S_1, S_2 given edge lengths t_1, t_2 as:

$$L(S_1, S_2|t_1, t_2) = (r_t)^{n_1}(s_t)^{n_2} = (\frac{1}{4}(1 + 3e^{-4\alpha t}))^{n_1}(\frac{1}{4}(1 - e^{-4\alpha t}))^{n_2} \tag{8}$$

Then the log-likelihood is:

$$\ln(L(S_1, S_2|t_1, t_2)) = n_1 \ln(\frac{1}{4}(1 + 3e^{-4\alpha t})) + n_2 \ln(\frac{1}{4}(1 - e^{-4\alpha t})) \tag{9}$$

To maximize the log-likelihood, we then find the value of t for which the derivative of $\ln(L)$ is equal to 0:

$$\frac{d\ln(L(S_1, S_2|t_1, t_2))}{dt} = n_1 \frac{-3\alpha e^{-4\alpha t}}{\frac{1}{4}(1 + 3e^{-4\alpha t})} + n_2 \frac{\alpha e^{-4\alpha t}}{\frac{1}{4}(1 - e^{-4\alpha t})} = \frac{(\frac{-1}{16}\alpha e^{-4\alpha t})(3n_1 - n_2) + (\frac{1}{16}\alpha e^{-8\alpha t})(3n_1 + 3n_2)}{(1 + 3e^{-4\alpha t})(1 - e^{-4\alpha t})} \tag{10}$$

If we set this value to 0, we get:

$$(\frac{1}{16}\alpha e^{-4\alpha t})(3n_1 - n_2) = (\frac{1}{16}\alpha e^{-8\alpha t})(3n_1 + 3n_2) \rightarrow \frac{3n_1 - n_2}{3n_1 + 3n_2} = \frac{e^{-8\alpha t}}{e^{-4\alpha t}} = e^{-4\alpha t} \tag{11}$$

$$\frac{\ln(3n_1 - n_2)}{\ln(3n_1 + 3n_2)} = \ln(e^{-4\alpha t}) = -4\alpha t \tag{12}$$

$$\frac{-1\ln(3n_1 - n_2)}{4\ln(3n_1 + 3n_2)} = \alpha t = t_1 + t_2 \tag{13}$$

9. Splits distance = 9 + 9 - 2*6 = 6.

   NNI = 4. From T_1, swap the following to get T_2: (6,4), (3,6), (1,6), (1,3).

10. The splits metric is:

$$\rho(T, S) = |\Sigma(T)| + |\Sigma(S)| - 2|\Sigma(T) \cap \Sigma(S)| \tag{14}$$

    Note that two trees on n species must share all leaf-edge splits. Note that, for $n \geq 3$, a full binary tree has $n - 3$ internal edges (i.e. edges between internal vertices). So, if $|\Sigma(T) \cap \Sigma(S)| = n$ then, in order for $\rho(T, S)$ to equal $2n - 6$, we can calculate:

$$2n - 6 = (2n - 3) + (2n - 3) - 2n. \tag{15}$$

    So, therefore, T and S must share all leaf-edge splits and no internal-edge splits.