Math 2300 **DISCRETE MATH** Fall 2014

Dr. Duval **Project 3**

due Wednesday, December 3

Phylogenetic trees

In this project, we'll look at an algorithm to reconstruct a phylogenetic tree for a set of species, using data about the pairwise similarities between the species. One example of such data are the alignment scores we computed in Project 1. "Phylogenetic" refers to how the rooted tree is attempting to reconstruct the part of the "Tree of Life" for the set of species we are looking at, showing their evolutionary history. Every vertex on this rooted tree will represent a species, including some common ancestors we will hypothesize for our given species.

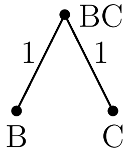
There are many ways to evaluate how similar two species are. One way would be to look at a specific region of their DNA, and evaluate the alignment, as we did in Project 1. We might work with the assumption that two species whose DNA is more similar have evolved away from each other more recently that two species whose DNA is less similar. We will try to put more similar species closer to each other on the evolutionary tree we are building. There may be different ways to measure how similar two species are to each other; each different way may produce a slightly different evolutionary tree. We will not attempt to deal with the problem of how to combine potentially different measures of similarity. We will work with just a single measure of similarity, which we will interpret as a kind of distance: Two species are more similar when this distance is smaller.

For instance, we might consider five species (A, B, C, D, E) whose pairwise distances are given by the following matrix:

	A	B	C	D	E
A	0	10	11	8	12
B	10	0	2	13	5
C	11	2	0	14	7
D	8	13	14	0	12
E	12	5	7	12	0

Since distance is symmetric (distance between A and B is the same as the distance between B and A), the matrix is also symmetric. We see that B and C are very similar (distance is only 2), while D and E much less similar (distance is 12, one of the larger values in the matrix).

We want to build a rooted tree where each vertex is a species, the leaves are the 5 species given above, and the distances between the leaves is at least approximately the distances given in the above matrix. (Unless we are very lucky it is impossible to make the distances exactly those given in the matrix.) There are many techniques for doing this. We will use a relatively simple one called Unweighted Pair Group Method with Arithmetic mean (UPGMA). The basic idea is take the two species most closely related, hypothesize a common ancestor to them, and then recursively build the tree using the common ancestor instead of the two species. For instance, in our example, B and C are most closely related (smallest distance in the matrix), so we will give them a common parent, a new vertex which we'll call BC. We'll let the distance between BC and each of B and C be half the distance between B and C, so that the total distance between the two vertices in the graph equals the distance given in the matrix:



Now we replace B and C in the matrix with the single entity BC, corresponding to the new vertex BC. We set the distance from each other vertices to the new vertex BC to be the average of the distances of the other vertices to B and to C. For instance, the distance between A and B is 10, and the distance between A and C is 11, so we'll set the distance between A and BC to be the average of 10 and 11, which is 10.5. Repeating this also to compute the distances from D and E to BC, we get the new matrix to be:

	A	BC	D	E
A	0	10.5	8	12
BC	10.5	0	13.5	6
D	8	13.5	0	12
E	12	6	12	0

Next, we repeat this process recursively, until the entire tree is built, though there is one additional wrinkle when we combine previously combined species (for instance, BC above) with other species. In our example, we see the new closest pair, with a distance of 6, is BC and E. In this case, when we compute the new distances to the newly combined (BC)E, we will take a weighted average of the distances to BC and to E, where the distance to BC will count twice as much, because there are two original species in BC, but only one in E. For instance, to compute the distance from A to (BC)E, the distance from A to BC is 10.5, and the distance from A to E is 12. We take the weighted average of 10.5 and 12, where 10.5 counts twice as much:

$$\frac{(2 \times 10.5) + (1 \times 12)}{2 + 1} = 11.$$

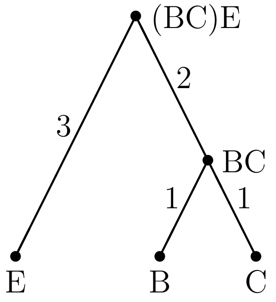
(Another way to interpret this weighted average is to take the usual average of the all the distances, in the original matrix, between A and each of B, C, and E: $(10 + 11 + 12)/3 = 11$. But that interpretation is less recursive, since we would have to keep going back to the original matrix to compute new distances.) Similarly, the distance from D to (BC)E is

$$\frac{(2 \times 13.5) + (1 \times 12)}{2 + 1} = 13.$$

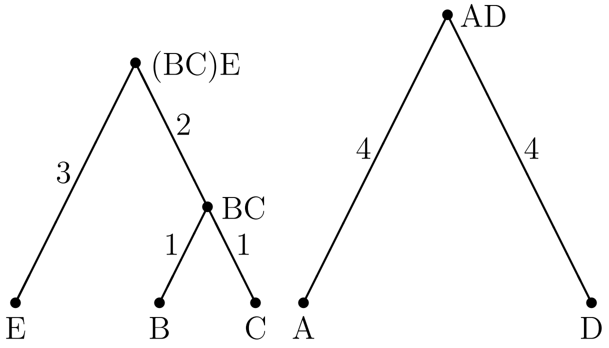
This makes our new matrix:

	A	(BC)E	D
A	0	11	8
(BC)E	11	0	13
D	8	13	0

Meanwhile, we also put in a new vertex (BC)E to be the common parent of BC and E, setting the distance from (BC)E to E to be 3, half the distance from E to the average of B and C. The distance from (BC)E to BC will only be 2, so that the total distance from (BC)E to each of B and C is also 3. Notice how this makes the distance in our tree between E and B and between E and C each 6, which is what we'd put as the average distance between E and each of B and C in our second matrix.



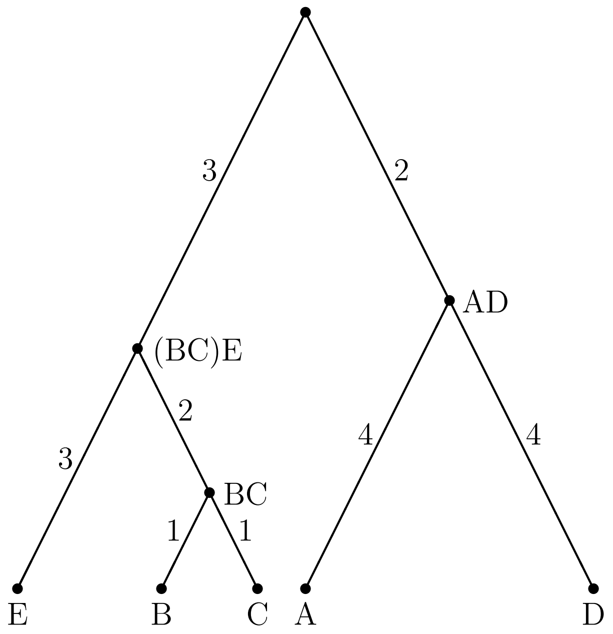
The next step is to give a common parent to A and D (since they are closest in the latest matrix), and call it AD. Since the distance between A and D is 8, we'll set the distance from AD to each of A and D to be 4, giving the following diagram



(note that our graph is not a tree right now, since it is not connected, but we will connect it all up shortly), and also the following matrix

	AD	(BC)E
AD	0	12
(BC)E	12	0

Finally, we will give AD and (BC)E a common parent, making sure its distance to each of its descendant leaves is 6, half of 12:



This is our phylogenetic tree from the original matrix.

One final note about computing new distances when there are even more vertices involved. When combining two vertices each of which corresponds to previously combined species, both parts get weighted by the number of original species in each group. For instance, when combining L(MN), which is a combination of 3 original species, with PQ, which is a combination of 2 species, to make (L(MN))(PQ), and computing the distance from between vertex X and this new vertex, the distance from X to L(MN) is weighted by a factor of 3, and the distance from X to PQ is weighted by a factor of 2. If the distance from X to L(MN) is 10 and the distance from from X to PQ is 20, then we set the distance of X to the new vertex (L(MN))(PQ) to be

$$\frac{(3 \times 10) + (2 \times 20)}{3 + 2} = 14.$$

Now it's your turn! Use UPGMA to construct a phylogenetic tree for each of the following distance matrices. In each case, be sure to show all your steps.

1.

	A	B	C
A	0	4	8
B	4	0	12
C	8	12	0

2.

	A	B	C	D	E
A	0	9	5	8	9
B	9	0	12	15	16
C	5	12	0	5	6
D	8	15	5	0	3
E	9	16	6	3	0

3.

	A	B	C	D	E	F	G
A	0	4	5	6	11	9	7
B	4	0	1	3	7	12	10
C	5	1	0	2	8	13	11
D	6	3	2	0	6	11	15
E	11	7	8	6	0	5	7
E	9	12	13	11	5	0	2
E	7	10	11	15	7	2	0