NEIGHBOR JOINING

Consider this tree:

(((A:0.01, B:0.08):0.01, C:0.01):0.01, D:0.01). This corresponds to the unrooted tree ((A:0.01, B:0.08):0.01, (C:0.01,D:0.02)). See also the sophisticated representation in Fig. 1.

Note the highly elevated rate in the branch leading to B.

If the observed differences exactly match the expectation from the tree, the distance matrix will be:

	A	B	C	D
A	-			
B	0.09	-		
C	0.03	0.1	-	
D	0.04	0.11	0.03	-

Now let's do a neighbor joining analysis on these data.

1. We must first try joining some pair of neighbors. Let's try (A,B).
2. Find the branch lengths. The tree has five branches: the terminal branches a, b, c, and d, and the internal branch i.
    a. To find these, we first estimate the terminal lengths by reducing the tree to three-taxon form A:a, B:b, CD:x (Fig. 2). To estimate branch lengths, we derive a new matrix where CD is one taxon, and the distances to it are the means of the distances to its component taxa.

	A	B	C
A	-		
B	0.09	-	
CD	0.035	0.105	-

    b. From the elements of this matrix, we can now calculate the branch lengths a, b, and x.

$$a = [(A.B) + (A.CD) - (B.CD)] / 2 = (0.09 + 0.035 - 0.105)/2 = 0.01$$
$$b = [(A.B) + (B.CD) - (A.CD)] / 2 = (0.09 + 0.105 - 0.035)/2 = 0.08$$
$$x = [(B.CD) + (A.CD) - (A.B)] / 2 = (0.105 + 0\ .035 - 0.09)/2 = 0.02$$

    c. This gives us two of our five lengths -- a and b, both of which you'll note match the lengths in the true tree. But we still need to decompose x into the internal branch i and the branches to the terminals c and d. To do that, use the three-taxon tree approach but

this time recongifure the original matrix making AB one taxon (Fig. 3). This yields the matrix:

	AB	C	D
AB	-		
C	0.065	-	
D	0.075	0.03	-

We can calculate the branch lengths on the three taxon tree as follows, where y is the branch leading to AB.

c=[(AB.C)+(C.D)-(AB.D)] / 2= (0.065 + 0.03 - 0.075)/2=0.01
d=[(AB.D)+(C.D)-(AB.C)] / 2 = (0.075 + 0.03 - 0.065)/2=0.01
y=[(B.CD)+(A.CD)-(A.B)] / 2 = (0.065 + 0.075 - 0.03) / 2 = 0.055

Now we have all the terminal branches; they are all correct.

d. Our last job is to find the internal branch length i. For this purpose, use the original matrix and the fully resolved tree, holding the terminal branches at the lengths inferred above (Fig. 4). The internal branch length i can be found this way:

   i = (AD + AC + BD + BC – 2AB – 2BC)/4. (See Fig. 5)

   We get i = [0.03 + 0.04 + 0.10 + 0.11 - 2(0.09) - 2(0.03)] / 4 = 0.01.

   We are done. Note that all the branch lengths were inferred correctly, although there is no clock.

e. Calculate the sum of the branch lengths on the tree by adding up the five lengths we found above. It is 0.13.

3. Now do this for each of the other possible neighbor pairs. If we pick pair (A,C), we evaluate tree ((A,C), (B,D)). We end up with the following branch lengths. (You should do this yourself to verify).

((A:0.015, C:0.015):-0.005, (B:0.085,D:0.025)). (See Fig. 6)

Note the negative internal branch length. That's a weird pathology of neighbor joining. It's hard to say what a negative branch length actually means. Statistically, however, this *is* the best fit to the observed data given the topology. There is no other set of branch lengths on this tree that could yield a smaller sum of squared deviations between the predicted distances from the branch lengths and the observed distances from the matrix.

The sum of the branch lengths on this tree is 0.135. This is longer than the length of the (A,B) tree.

Note that the terminals on the incorrect tree have become longer than the terminals on the true tree and the internal branch has become shorter. This is because in order to fit the observed distances onto the wrong topology, we need to explain the relatively large observed distance between the putatively "sister" taxa A and C as due to changes along the terminal branches leading to these taxa; same for the distance between B and D. Thus the terminals become longer. But now we have to fit the fairly small observed differences between A and D to a tree in which the terminal branches leading to these taxa have been lengthened, so the algorithm has to take length out of the internal branch. By forcing observed distances between taxa off the internal branch onto multiple terminals, the total length of the tree becomes longer.

If we fit branch lengths to the (AD) tree, we would also get a total length longer than 0.13. Thus we choose (AC) as the topology that produces branch lengths that best fit the data. We therefore join these two as neighbors. Because this is only a four-taxon tree, this one neighbor-joining step is sufficient to fully resolve the tree. If this were a larger tree, we would find the next best pair of neighbors by repeating this process, considering AC as one taxon and keeping this relationship fixed.

Fig 1.

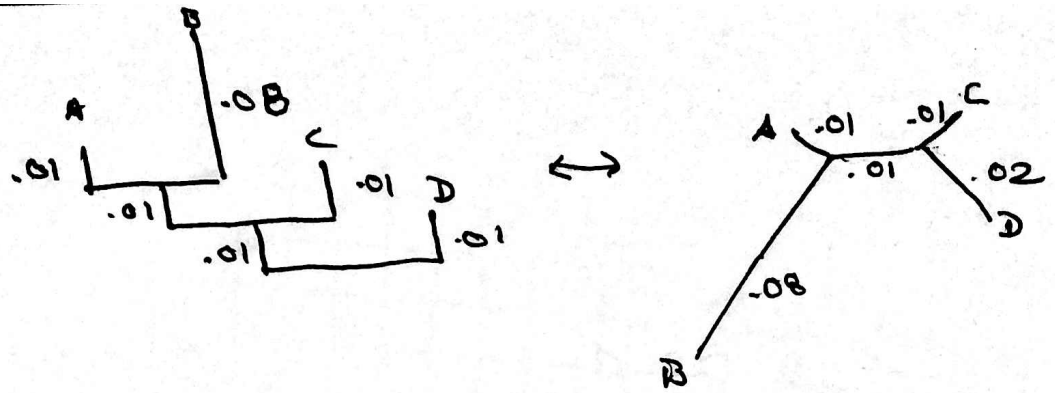


Fig 2

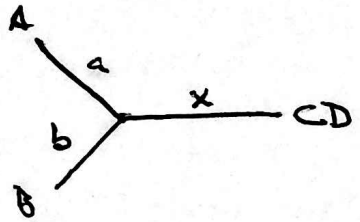


Fig 3.

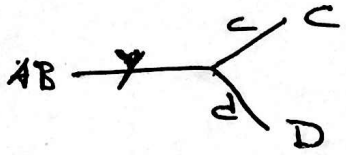


Fig 4.

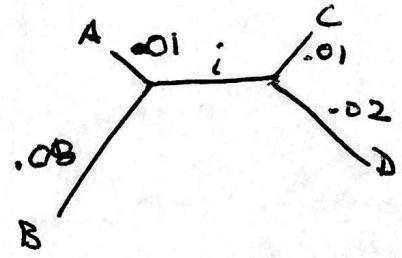
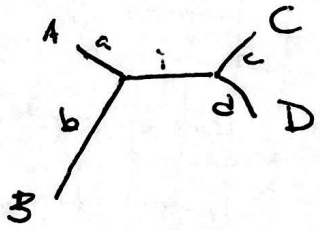


Fig 5



$$i = \frac{\text{⟋} + \text{⟍} + \text{⌐} + \text{¬} - 2(\rangle) - 2(\langle)}{4}$$

Fig 6