

Box 11.1: An Example of Phylogenetic Tree Construction Using the UPGMA Method.

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

1. Create a distance matrix involving four taxa, A, B, C, and D, the UPGMA method first joins two closest taxa together which are A and C (0.35 in grey), because all taxa are equidistant from the root. The branch length for A to the root is $AC/2 = 0.35/2 = 0.175$.

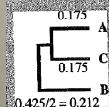


2. Because A and C were joined first, they are treated as one new composite taxon, which is used to create a reduced matrix. The distance of A-C cluster to every other taxon is the average of A to taxon B and C. Respectively, that means that the distance of B to A-C is $(AB + BC)/2$, and that of D to A-C is $(AD + CD)/2$.

	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70

	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70

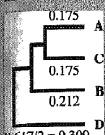
3. In the newly reduced distance matrix, the smallest distance is between B and A-C (in grey), which allows the grouping of B and A-C to create a three-taxon cluster. The branch length for the B is one half of B to the A-C cluster.



4. When B and A-C are grouped and treated as a single taxon, this allows the matrix to reduce further into only two taxa, D and B-A-C. The distance of D to the composite taxon is the average of D to every single component which is $(BD + AD + AC)/3$.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

5. D is the last branch to add to the tree, whose branch length is the half of D to B-A-C.



6. Because distance trees allow branches to be additive, the resulting distances between taxa from the tree path can be used to create a distance matrix; obviously, the estimated distances do not match the actual evolutionary distances shown, which illustrates the failure of UPGMA to precisely reflect the experimental observation.

	A	B	C
B	0.42		
C	0.35	0.42	
D	0.62	0.62	0.62

Box 11.2: Phylogenetic Tree Construction Using the Neighbor Joining Method

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

1. The NJ method is similar to UPGMA, but uses an evolutionary rate correction step (locus tree building). Using the same distance matrix as in the UPGMA tree building (see box 11.1), the first step of the NJ method is r -value and Q -value calculation. According to Eq. 11.1 and 11.2, r_i and Q_i for each taxon are calculated as follows:

$$r_A = AB + AC + AD = 0.4 + 0.35 + 0.6 = 1.35$$

$$r_B = r_A/(4-2) = 1.35/2 = 0.675$$

$$r_C = BA + BC + BD = 0.4 + 0.45 + 0.7 = 1.55$$

$$r_D = r_C/(4-2) = 1.55/2 = 0.775$$

$$Q_A = AB + CD + AD = 0.35 + 0.65 + 0.55 = 1.55$$

$$Q_B = r_A/(4-2) = 1.35/2 = 0.675$$

$$Q_C = DA + DB + DC = 0.6 + 0.7 + 0.55 = 1.85$$

$$r_D = r_C/(4-2) = 1.55/2 = 0.775$$

2. Based on Eq. 11.4 and the above r -values, the corrected distances are obtained as follows:

$$d'_{AB} = d_{AB} - 1/2 * (r_A + r_B) = 0.4 - (1.35 + 1.55)/2 = -1.05$$

$$d'_{AC} = d_{AC} - 1/2 * (r_A + r_C) = 0.35 - (1.35 + 1.55)/2 = -1$$

$$d'_{AD} = d_{AD} - 1/2 * (r_A + r_D) = 0.6 - (1.35 + 1.75)/2 = -1$$

$$d'_{BC} = d_{BC} - 1/2 * (r_B + r_C) = 0.45 - (1.55 + 1.35)/2 = -1$$

$$d'_{BD} = d_{BD} - 1/2 * (r_B + r_D) = 0.7 - (1.55 + 1.75)/2 = -1$$

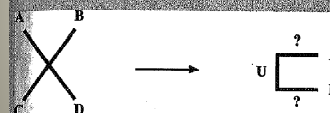
$$d'_{CD} = d_{CD} - 1/2 * (r_C + r_D) = 0.55 - (1.55 + 1.75)/2 = -1.05$$

3. The rate-corrected distances allow the construction of a new distance matrix.

	A	B	C
B	-1.05		
C	-1	-1	
D	-1	-1	-1.05

4. Before tree construction, all possible nodes are collapsed into a star tree. The pair of taxa with the shortest distances in the new

matrix are separated from the star tree first, according to the corrected distances. In this case, A and B as well as C and D are the shortest (-1.05, in grey). Therefore, the first node to be built can be either A-B or C-D. Choosing either pair first will give the same result. Let's choose A and B first and name the node U.



5. The branch lengths for A and B to the node U are calculated according to Eq. 11.4:

$$d_{AU} = (d_{AB} + (r'_A - r'_B))/2 = (0.4 + (0.675 - 0.775))/2 = 0.15$$

$$d_{BU} = (d_{AB} + (r'_B - r'_A))/2 = (0.4 + (0.775 - 0.675))/2 = 0.25$$



6. The new cluster allows the construction of a reduced matrix. This starts with actual distances. Unlike in UPGMA, the distance from a taxon to a node is the average of the original distances to each of the components of the comparable taxon, subtracted from the inferred branch lengths.

$$d_{CU} = 1/2(d_{CU} + d_{DU}) + (d_{CU} - d_{DU})/2 = [(0.35 + 0.15) + (0.45 - 0.25)]/2 = 0.2$$

$$d_{DU} = 1/2(d_{CU} + d_{DU}) + (d_{DU} - d_{CU})/2 = [(0.35 + 0.15) + (0.25 - 0.45)]/2 = 0.15$$

	U	B
C	0.20	
D	0.45	0.35

7. Based on the reduced distance matrix, a new set of r_i and r'_i values are calculated.

$$r_U = CU + CD = 0.2 + 0.35 = 0.55$$

$$r'_U = r_U / (3 - 2) = 0.55 / 1 = 0.55$$

$$r_D = DU + CD = 0.15 + 0.35 = 0.5$$

$$r'_D = r_D / (3 - 2) = 0.5 / 1 = 0.5$$

$$r_B = BU + BU = 0.2 + 0.45 = 0.65$$

$$r'_B = r_B / (3 - 2) = 0.65 / 1 = 0.65$$

8. The new r_i and r'_i values allow construction of the corrected distance matrix.

$$d'_{CU} = d_{CU} - 1/2 * (r_U + r'_U) = 0.2 - (0.55 + 0.55)/2 = -0.2$$

$$d'_{DU} = d_{DU} - 1/2 * (r_U + r'_U) = 0.15 - (0.55 + 0.55)/2 = -0.375$$

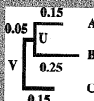
$$d'_{CB} = d_{CB} - 1/2 * (r'_U + r'_B) = 0.55 - (0.55 + 0.65)/2 = -0.325$$

	U	B
C	-0.5	
D	-0.375	-0.325

9. In the corrected distance matrix, C to node U has the shortest distance (-0.5, in grey). This allows creation of the second node named V. The branch length is calculated as in step 5.

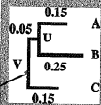
$$d_{UV} = (d_{CU} + (-r'_U)) / 2 = (0.2 + (-0.55)) / 2 = -0.175$$

$$d_{DV} = (d_{DU} + (-r'_U)) / 2 = (0.15 + (-0.55)) / 2 = -0.2$$



10. Because D is the last branch to be decomposed from the star tree, there is no need to convert to r and r' because r' is infinitely large when $n = 2$ (its branch length is calculated as one half of the sum of D to nodes V and D to C, subtracted from respective branch lengths).

$$d'_{DV} = 1/2(d_{DV} - r'_D) = (d_{DV} - d_{DV})/2 = [(0.45 - 0.05) + (0.25 - 0.15)]/2 = 0.2$$



11. When the overall branch lengths are compiled into a distance matrix, which is used to compare with the original distance matrix, it is clear that the estimated distances completely match the actual evolutionary distances, indicating that this treeing method is able to satisfy the constraint of the experimental observation in this case.

	U	B
C	-0.5	
D	-0.375	-0.325