Course: COMPSCI 260
Name: Divya Koyyalagunta
NetID: dk160
Problem: 3
Problem Set: 5
Due: Fri 7 Nov 2016, 5pm
Using free extension (yes/no): no

Statement of collaboration and resources used (put None if you worked entirely without collaboration or resources; otherwise cite carefully):
My solutions and comments for this problem are below.

PROBLEM 3

1. D_1 is not ultrametric, and the triplet $1, 2, 3$ shows this:

   try $i, j, k = 1,2,3$

   $$d_{12} = d_{13} > d_{23}$$
   $$0.3 \neq 0.6$$

   try $i, j, k = 2,3,1$

   $$d_{23} = d_{12} > d_{13}$$
   $$0.5 \neq 0.3$$

   try $i, j, k = 3,1,2$

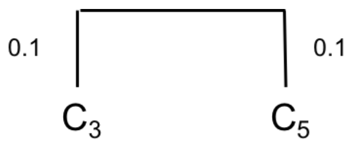
   $$d_{13} = d_{23} > d_{12}$$
   $$0.6 \neq 0.5$$

   so no shuffling of i, j, k will satisfy the ultrametricity criterion.

   the D_2 matrix is ultrametric, and so we can construct a UPGMA.

   first we put each sequence into its own cluster, so the set of all clusters starts as $\{C_1, C_2, C_3, C_4, C_5\}$

   the first minimum distance is between C_3 and C_5



   we will call the cluster above C_6, placed at height $\frac{d_{ij}}{2} = 0.1$. let's compute d_{6s} for all l

we will use the following formula to calculate the distance from C_6 and any other cluster C_s

$$d_{rs} = \frac{d_{ps}|C_p| + d_{qs}|C_q|}{|C_p| + |C_q|}$$

where $r = 6$

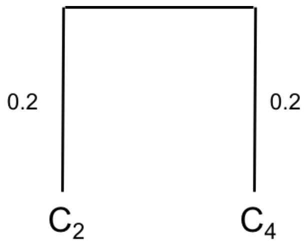
so for $s = 1$,

$$d_{16} = \frac{d_{13}(1) + d_{15}(1)}{2} = \frac{0.6 + 0.6}{2} = 0.6$$

continuing this process for all other clusters, our distance matrix becomes

D_2	C_1	C_2	C_4	C_6
C_1	0	0.8	0.8	0.6
C_2		0	0.4	0.8
C_4			0	0.8
C_6				0

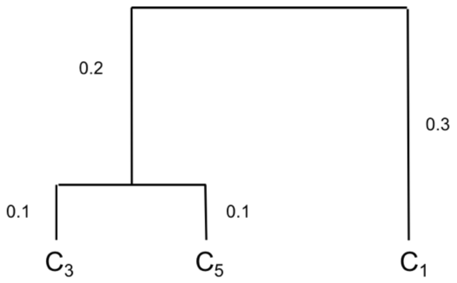
The next possible minimum distance is between C_2 and C_4



we will call the cluster above C_7, placed at height $\frac{d_{ij}}{2} = 0.2$. Using the same equations from above, we can update our distance matrix

D_2	C_1	C_6	C_7
C_1	0	0.6	0.8
C_6		0	0.8
C_7			0

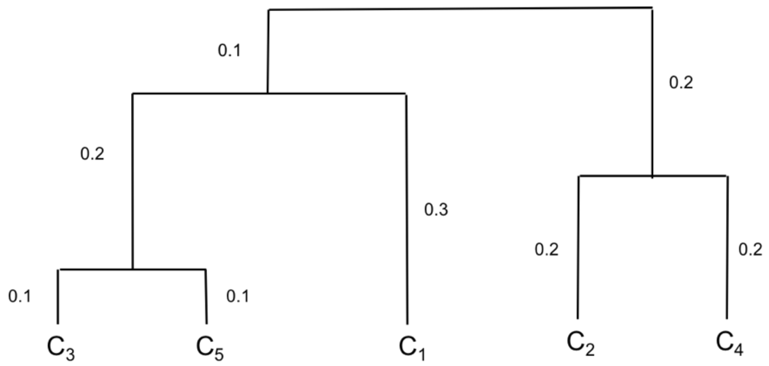
The next possible minimum distance is between C_6 and C_1



we will call the cluster above C_8, placed at height $\frac{d_{ij}}{2} = 0.3$. Using the equations from above, we update our distance matrix again

D_2	C_7	C_8
C_7	0	0.8
C_8		0

The next possible minimum distance is between C_7 and C_8, with the new node placed at height $\frac{d_{ij}}{2} = 0.4$.



We now have our final tree.

2. Both matrices are additive, so we don't have an example that doesn't satisfy the additivity criterion.

   The D_1 matrix is not ultrametric but it is additive, so we can create an NJ tree.

   we first start with the original distance matrix (which we will call d):

d	X_1	X_2	X_3	X_4
X_1	0	0.3	0.6	0.5
X_2		0	0.5	0.6
X_3			0	0.9
X_4				0

from this we will calculate an adjusted distance matrix called D, which is calculated by the following formulas:

$$D_{ij} = d_{ij} - (r_i + r_j)$$

$$r_i = \frac{\sum_{k \in L} d_{ik}}{|L| - 2}$$

where L is the set of current leaf nodes so now we produce a D matrix like the one below

D	X_1	X_2	X_3	X_4
X_1	0	-1.1	-1.1	-1.2
X_2		0	-1.2	-1.1
X_3			0	-1.1
X_4				0

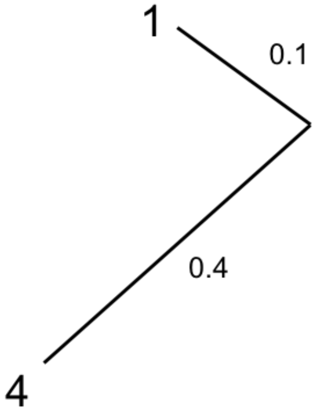
from here we choose an $i,\ j$ based on the minimum D_{ij}. we choose $i, j = 1, 4$.

now we create a new node $k = 5$, we now set:

$$d_{ik} = \frac{1}{2}\left(d_{ij} + \left(r_i - r_j\right)\right) = \frac{1}{2}\left(0.5 + (0.7 - 1.0)\right) = 0.1$$

$$d_{jk} = \frac{1}{2}\left(d_{ij} + \left(r_j - r_i\right)\right) = \frac{1}{2}\left(0.5 + (1.0 - 0.7)\right) = 0.4$$

now we can join k to i and j:



and using the following equation,

$$d_{km} = \frac{d_{im} + d_{jm} - d_{ij}}{2}\ , \forall\ m$$

we can now update our d matrix and D matrix

d	X_1	X_2	X_3	X_4	X_5
X_1	0	0.3	0.6	0.5	
X_2		0	0.5	0.6	0.2
X_3			0	0.9	0.5
X_4				0	
X_5					0

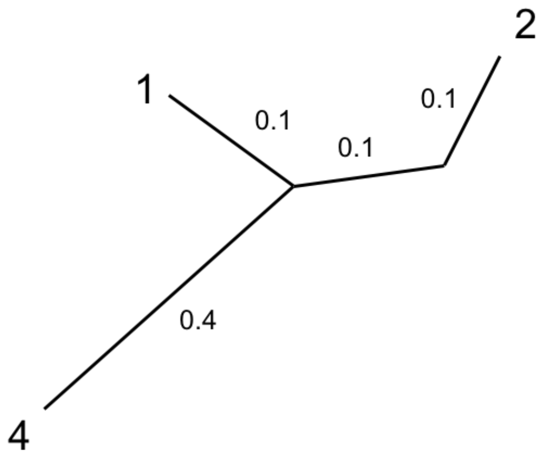
D	X_1	X_2	X_3	X_4	X_5
X_1	0				
X_2		0	-1.2	-1.2	
X_3			0	-1.2	
X_4				0	
X_5					0

from here we choose an $i,\ j$ based on the minimum D_{ij} again. we choose $i,j = 2,5$ (in the case of a tie, the topology of the tree will be the same regardless of which we choose), and create a new node $k = 6$

$$d_{ik} = \frac{1}{2}\left(d_{ij} + \left(r_i - r_j\right)\right) = \frac{1}{2}\left(0.2 + (0.699 - 0.7)\right) \approx 0.1$$

$$d_{jk} = \frac{1}{2}\left(d_{ij} + \left(r_j - r_i\right)\right) = \frac{1}{2}\left(0.2 + (0.7 - 0.699)\right) \approx 0.1$$

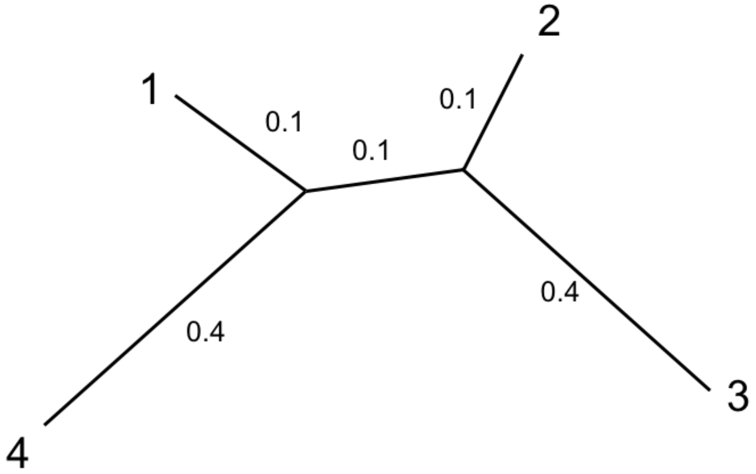
we can add this new node to the tree:



we update the d matrix

d	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0					
X_2		0	0.5	0.2		
X_3			0	0.5	0.4	
X_4				0		
X_5					0	
X_6						0

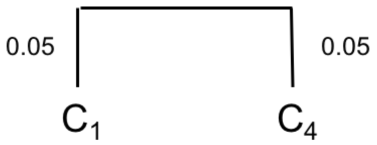
the node list currently only contains 2 leaf nodes, so we just merge the last two (X_3 and X_5), to get the final tree:



b) The SODM distance metric satisfies both the ultrametricity and additivity conditions, so we can construct a UPGMA. Let's use the same process from (b) part 1.

first we put each sequence into its own cluster, so the set of all clusters starts as $\{C_1, C_2, C_3, C_4\}$

the first minimum distance is between C_1 and C_4



we will call the cluster above C_5, where this cluster is placed at height $\frac{d_{ij}}{2} = 0.05$
we will use the following formula to calculate the distance from C_5 and any other cluster C_s

$$d_{rs} = \frac{d_{ps}|C_p| + d_{qs}|C_q|}{|C_p| + |C_q|}$$

where $r = 5$

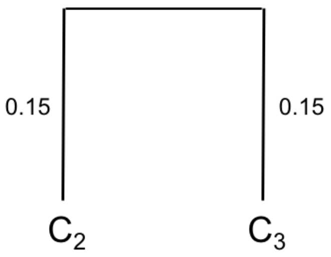
so for $s = 2$,

$$d_{25} = \frac{d_{12}(1) + d_{24}(1)}{2} = \frac{0.5 + 0.5}{2} = 0.5$$

continuing this process for all other clusters, our distance matrix becomes

	C_2	C_3	C_5
C_2	0	0.3	0.5
C_3		0	0.5
C_5			0

the next minimum distance is between C_2 and C_3

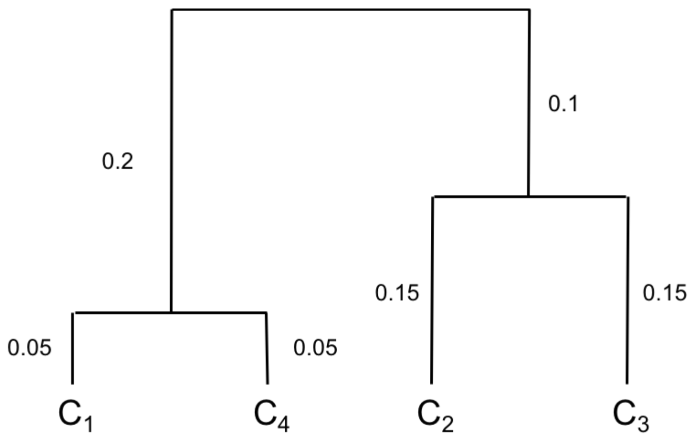


we will call the cluster above C_6, placed at height $\frac{d_{ij}}{2} = 0.15$

we again update our distance matrix using the same formulas

D_2	C_5	C_6
C_5	0	0.5
C_6		0

the next possible minimum distance is between C_5 and C_6, so our final tree becomes:



where the final node is placed at height $\frac{d_{ij}}{2} = 0.25$

Just as in problem set 4, the phylogeny above shows us that, if we assume the molecular clock hypothesis is true, the human and mouse SODM sequences are more similar (which follows that the UPGMA shows that these two share a more recent common ancestor), and *E.coli* and *Bacillus subtili* are also more similar to each other. Since the branch lengths in UPGMA give a sense of time, we can see that the two bacteria diverged before the human and mouse did, which is reasonable considering evolutionary history.

This phylogeny is also consistent with the fact that the distance metric was based on the alignments of the protein sequences themselves, so the distances between more similar protein sequences resulted in a lower distance metric between those two sequences, and consequently a phylogeny where the species with those two sequences share a more common ancestor. Using the UPGMA assumption of the molecular clock hypothesis, this means that having a more recent common ancestor in the visualized tree also corresponds to less mutations (less differences between the two sequences), and thus this result also confirms the alignments of the protein sequences in problem set 4.