Exploring Bioinformatics: A Project-Based Approach

Key for Test Your Understanding Questions, Chapter 7

1. Given aligned sequences for four species with distances W-X = 1.8, W-Y = 0.8, W-Z = 2.4, X-Y = 1.8, X-Z = 2.4 and Y-Z = 2.4, cluster the sequences using single linkage and show the result in Newick format.

	W	X	Y	Z
W	0			
X	1.8	0		
Y	0.8	1.8	0	
Z	2.4	2.4	2.4	0

	WY	X	Z
WY	0		
X	1.8	0	
Z	2.4	2.4	0

	WYX	Z
WYX	0	
Z	2.4	0

As shown by the matrices above, W would first merge with Y, then X with the W-Y cluster, leaving Z as the last merge. In Newick format, this is ((W,Y),X),Z)

2. Using the distance data that you used to sketch your phylogenetic tree for whales and their relatives (Chapter 6, Web Exploration exercise #8), apply the clustering algorithm to those data. Do you get the same groupings as in the tree you drew?

Using the Kimura data (see the last page of this document for the matrices and how they are used to cluster the species), the grouping is ((((((whale,porpoise),hippo),giraffe),camel),dog),rat). This is exactly as phylogeny.fr drew the tree. The distances will change with the other distance methods, but the grouping should not change.

3. Try the UPGMA linkage method instead of the single linkage method for our sample dataset presented in the Understanding the Algorithm section. Do you get the same groupings? The same distances?

	A	B	C	D	E	F
A	0					
B	1	0				
C	3	2	0			
D	7	6	4	0		
E	17	16	14	10	0	
F	19	18	16	12	2	0

	AB	C	D	E	F
AB	0				
C	2.5	0			
D	6.5	4	0		
E	16.5	14	10	0	
F	18.5	16	12	2	0

	AB	C	D	EF
AB	0			
C	2.5	0		
D	6.5	4	0	
EF	17.5	15	11	0

	ABC	D	EF
ABC	0		
D	5.3	0	
EF	16.3	11	0

	ABCD	EF
ABCD	0	
EF	13.6	0

As shown by the matrices above, the grouping is the same: ((((A,B),C),D),(E,F)). But the distances are somewhat different. Interestingly, this method removes some of the ambiguity in merging the clusters: with single linkage, the second merge could be either AB with C or E with F, both having a distance of 2. But UPGMA makes this unambiguous, with the EF merge occurring before the ABC merge.

4. In the sample dataset used in the Understanding the Algorithm section above, at the second merge, we had a choice of either merging cluster (AB) with C (which we chose to do) or merging clusters E and F (which we ignored); both choices had a distance value of 2. Use the clustering algorithm to determine how the tree would have come out if we'd chosen E and F instead; would it have been different? Do you think this

would always be the case? In other words, does the arbitrary choice of one grouping when there are two possibilities have the potential to affect our view of the evolutionary relationships?

In the case of these sample data, it would not have made any difference which option was chosen: E and F are on a distinct branch of the tree from all the others either way. However, one can imagine a dataset in which this would make a difference: if species C, for example, was similar in its distance from AB and from EF, the choice made at this point might determine which of the other two C was grouped with. So, it's clearly desirable to develop methods that avoid these ambiguities where possible.

5. The tree in Figure 7.6 is drawn as a cladogram, not a phylogram: that is, the branch lengths are not strictly proportional, although the evolutionary pathways are shown correctly. Try putting branch lengths onto the tree, using the data in Figure 7.4A. What problem do you encounter? How would you explain this difficulty, biologically? (Hint: what assumption are we implicitly making when we calculate distances between clusters?) In the On-Your-Own Project, you'll see how the neighbor-joining algorithm deals with this important complication by changing the way the distances between clusters are calculated.

   We know that the distance from A to B is 1, from A to C is 3 and from B to C is 2. However, if we make the branches from *y* to A and B 0.5 each, then it becomes impossible to correctly show the distance from *y* to C: based on the A-C distance, it would be 2.5 (3 – 0.5), but based on the B-C distance, it would be 1.5 (2 – 0.5). This problem arises because both the linkage methods we've looked at here assume a constant rate of evolution (mutation): that the branches from *y* to A and *y* to B should be the same lengths. In the real world, this is sometimes approximately true, but we see many examples where more mutations have occurred in one species than another since they diverged from a common ancestor.

Matrices for clustering of the Kimura data for mammals:

	rat	horse	dog	camel	giraffe	hippo	whale	porpoise
rat	0							
horse	0.580	0						
dog	0.515	0.265	0					
camel	0.517	0.276	0.240	0				
giraffe	0.551	0.294	0.266	0.217	0			
hippo	0.519	0.298	0.274	0.209	0.166	0		
whale	0.495	0.294	0.239	0.181	0.162	0.111	0	
porpoise	0.502	0.276	0.219	0.181	0.147	0.093	0.050	0

	rat	horse	dog	camel	giraffe	hippo	WP
rat	0						
horse	0.580	0					
dog	0.515	0.265	0				
camel	0.517	0.276	0.240	0			
giraffe	0.551	0.294	0.266	0.217	0		
hippo	0.519	0.298	0.274	0.209	0.166	0	
WP	0.495	0.276	0.219	0.181	0.147	0.093	0

	rat	horse	dog	camel	giraffe	HWP
rat	0					
horse	0.580	0				
dog	0.515	0.265	0			
camel	0.517	0.276	0.240	0		
giraffe	0.551	0.294	0.266	0.217	0	
HWP	0.495	0.276	0.219	0.181	0.147	0

	rat	horse	dog	camel	GHWP
rat	0				
horse	0.580	0			
dog	0.515	0.265	0		
camel	0.517	0.276	0.240	0	
GHWP	0.495	0.276	0.219	0.181	0

	rat	horse	dog	CGHWP
rat	0			
horse	0.580	0		
dog	0.515	0.265	0	
CGHWP	0.495	0.276	0.219	0

	rat	horse	DCGHWP
rat	0		
horse	0.580	0	
DCGHWP	0.495	0.265	0

	rat	others
rat	0	
others	0.580	0