C3020 – Molecular Evolution

Exercises #3: Phylogenetics

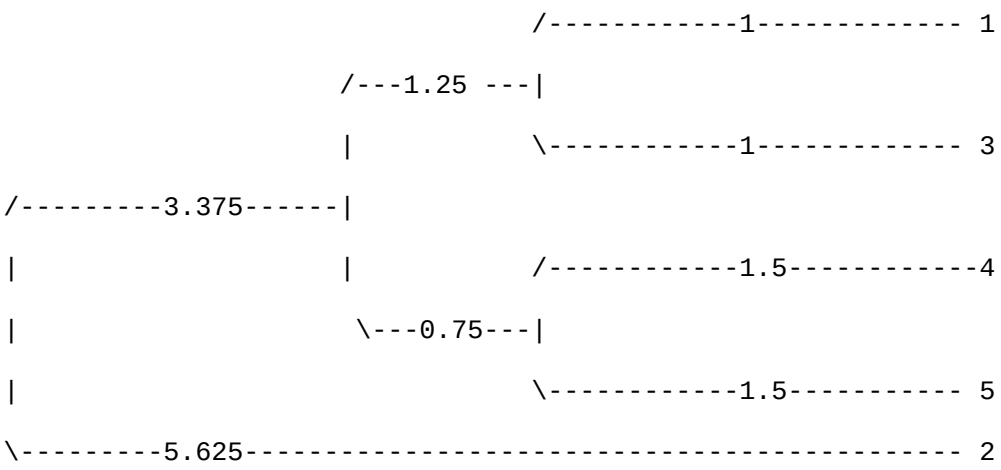
Consider the following sequences for five taxa 1-5 and the known outgroup O, which has the ancestral states (note that sequence 3 has changed from the earlier version to make computation easier)

```
1     ACAAACAGTT CGATCGATTT GCAGTCTGGG

2     ACAAACAGTT TCTAGCGATT GCAGTCAGGG

3     ACAGACAGTT CGATCGATTT GCAGTCTCGG

4     ACTGACAGTT CGATCGATTT GCAGTCAGAG

5     ATTGACAGTT CGATCGATTT GCAGTCAGGA

O     TTTGACAGTT CGATCGATTT GCAGTCAGGG
```

1. Make a distance matrix using raw distances (number of differences) for the five ingroup sequences.

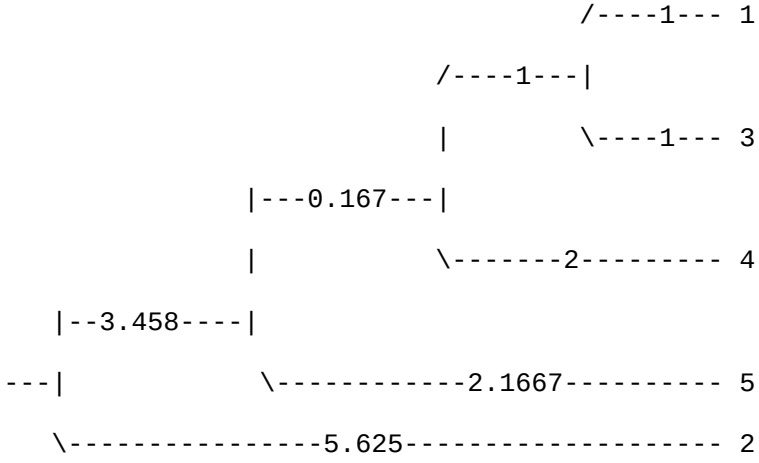
	1	2	3	4	5
1	-				
2	9	-			
3	2	11	-		
4	4	12	4	-	
5	5	13	5	3	-

2. Infer the UPGMA tree for these sequences from your matrix. Label the branches with their lengths.



To derive this tree, start by clustering the two species with the lowest pairwise difference -- 1 and 3 -- and apportion the distance between them equally on the two branches leading from their common ancestor to the two taxa. Treat them as a single composite taxon, with distances from this taxon to any other species equal to the mean of the distances from that species to each of the species that make up the composite (the use of arithmetic means explains why some branch lengths are fractions). Then group the next most similar pair of taxa -- here* it is 4 and 5 -- and apportion the distance equally. Repeat until you have the whole tree and its lengths. 2 is most distant from all other taxa and composite taxa, so it must be the sister to the clade of the other four taxa.

* In fact there is a tie at this point, because the distance between 4 and 5 and the distance between 4 and the composite taxon (1,3) are equal (4 changes each). It is arbitrary which one you choose. If you choose to cluster 1, 3, and 4, the tree would look like this. Either answer is correct.



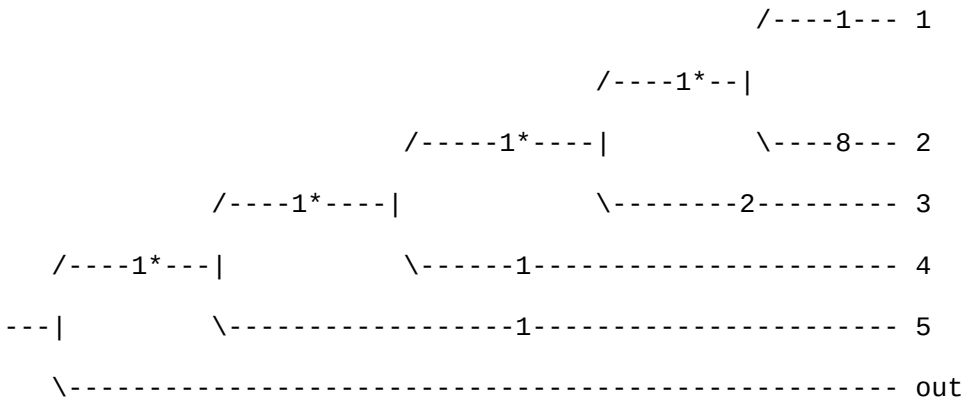
3. Using the parsimony criterion,

a. Which characters are phylogenetically informative?

Characters 2, 3, 4, and 27. Of the others, 14 are constant and 12 are autapomorphies (unique to one taxon).

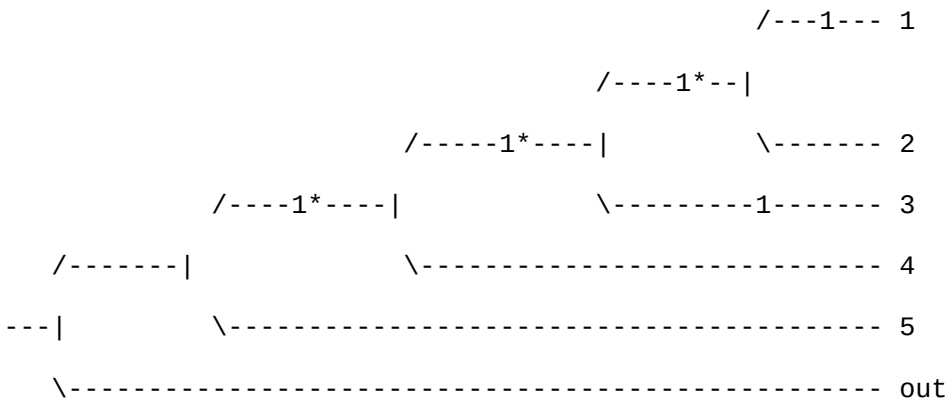
b. How many nucleotide changes (steps) does the tree rooted tree (((((1,2),3),4),5),0) require? How many homoplasious changes does it require? Mark the number of changes that occur on each branch.

Total length = 17. It has one homoplasious character (#27), which must be explained by two changes -- either a reversal to the ancestral state in the branch leading to species 2 or parallel changes in the branches leading to species 1 and 3 (The tree below assumes the latter). Many changes (12, if you count them) are autapomorphies. Changes that create synapomorphies are marked with asterisks and involve, moving from the root to the tips, characters 1, 2, 3 and 4.



The way to reconstruct the number of changes is to trace each of the characters on the cladogram individually, and add up the number of steps required by each character to get the total length for all characters.

You will notice that this procedure is rather irritating, given the number of characters in the matrix. It can be simplified if we look only at the four informative characters, which are the only ones that affect the relative lengths of the possible cladograms. Then we get a tree with length 5, which has the following number of changes per branch (assuming parallel changes for character 27):



c. How many nucleotide changes does the tree ((((1,4),(2,3)),5),0) require? How many of these are homoplasious?

Including only uninformative characters, length = 7. Total length = 19. (The autapomorphies in the first tree will be autapomorphies in any tree, so the difference between lengths with uninformative characters excluded and included will always equal 12. This means you can estimate for any tree the total length with uninformative characters included by calculating its length with only informative characters included.) On this tree, three characters (3, 4,and 27) are homoplasious, requiring two changes each, so a total of 6 of the 7 changes in informative characters are homoplasious.

d. Which one of these two trees is "better"? Why?

The first tree is better under the parsimony criterion, because it requires fewer homoplasious changes (parallelisms/reversals). This means it explains more shared character states as the result of inheritance from a common ancestor than tree in c does. Tree 2 needs ad hoc hypotheses of parallelism/reversal to explain why unrelated taxa share identical character states.

e. For what clade of taxa does an A in the third nucleotide position represent a synapomorphy? (note change in question).

Clade (1,2,3). This character has the same derived state in these three taxa and only these three taxa.

f. Name two nucleotide positions that contain a symplesiomorphic state for the group (1,2,3).

A in position 1 and C in position 2 are both symplesiomoprhies for this group -- they are the ancestral state and give no information about relationships within the group.

4. Using neighbor-joining,

a. Estimate S (the sum of branch length) on the unrooted partially

resolved tree in which only taxa 1 and 2 are joined as neighbors. Leave the outgroup out of the analysis.

As we learned in class (see also Graur and Li p. 189), the sum of the branch lengths can be calculated for any situation like this from the pairwise distances alone. Here,

S = 1/6(d13+d14+d15+d23+d24+d25) + 1/2(d12) + 1/3(d34+d35+d45) = 16.33.

b. On this tree, what are the branch lengths for the branches leading from the last common ancestor (LCA) of 1 and 2 to taxon 1 and to taxon 2?

To calculate branch lengths, consider this a tree with three taxa: species 1, species 2, and a composite taxon of species 3, 4, and 5, which we will call X. As we saw in class, the length of the branch from the internal node on a three taxon tree to any terminal can be calculated from the pairwise distances between the taxa. Thus, the length of the branch from the internal node to species 1 = (d12 + d1X - d2X)/2; the length of the branch leading to species 2 = (d12+d2X - d1X). The distance between any species and the composite taxon X is the average of the distances between that species and each of the species that make up X. Thus,

d1X = (2+4+5)/3 = 11/3, and

d2X = (11+12+13)/3 = 12

Using these figures, we can calculate the length of the branches from the equation above.

Length of branch leading to species 1 = (9 + 3.667 - 12)/2 = 0.333

Length of branch leading to species 2 = (9 + 12 - 3.667)/2 = 8.667

Note that these branch lengths add up to 9, the pairwise distance observed between species 1 and 2, which shows that this part of the tree has a good fit to the sequence data.

c. Estimate S on the unrooted tree in which only taxa 1 and 3 are joined as neighbors. What are the lengths for the branches leading to 1 and to 3 from their LCA?

S= 1/6(9+4+5+11+4+5) + 1/2(2)+ 1/3 (12+13+3) = 16.66.

d. Which of these two trees is better? What do you do next with the better one?

The tree with (1,2) is better. The fact that it has shorter total branch length means that the pairwise distances can be forced to fit on this tree more easily than on the tree with 1 and 3 as neighbors, because clustering taxa that are in fact more distantly related requires adding length into the internal branches to explain the observed distances.

The next thing to do is to resolve the unresolved node while holding (1,2)

constant. To do this, try all possible pairs of neighbors on this partially resolved tree and choose the tree with the shortest length.

e. Estimate S for the tree ((1,2),3),4,5) -- a possible "next step" tree from the tree in question 4a.

If you draw this unrooted tree, you will see that it treats 4 and 5 as neighbors. Using the method above, we find the length of this tree S = 1/6(4+12+4+5+3+5) + 1/2(3) + 1/3((9+2+11)=16.

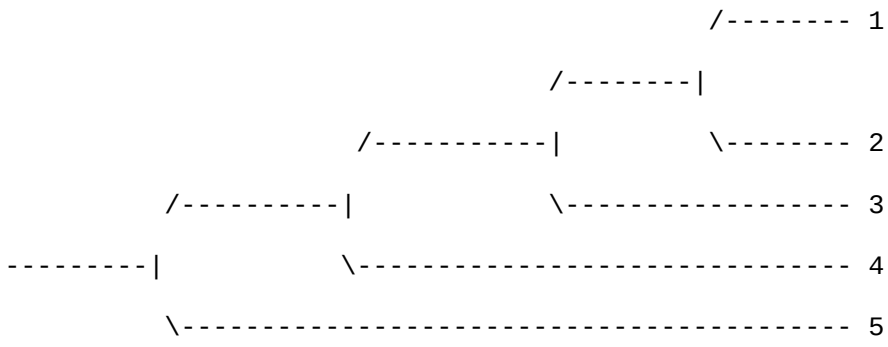
f. Estimate S for the tree ((1,2),4),5,3) -- another possibility for the next step.

This tree treats 5 and 3 as neighbors, so S = 1/6(2+11+4+5+13+3) + 1/2(5) + (1/3(9+4+12)= 17.167.

g. Which of these last two trees is better?

The first one, **((1,2),3),4,5)** is better for the reasons explained above.

h. Draw the rooted tree for the tree you chose as better, assuming that the outgroup attaches to the branch leading to taxon 5 and only taxon 5.



5. You now have 3 trees -- a UPGMA tree (problem #2), a parsimony tree (#3d), and a NJ tree (#4h). (Although you have not examined all possible phylogenies for either parsimony or NJ, assume that the best tree from question 3 is the most parsimonious tree and the best tree from question 4 is the best NJ tree. These are, in fact, the optimal tree for each method.)

a. Do the three methods agree on the phylogenetic relationships among 1-5?

The parsimony and NJ trees agree, but they are different from the UPGMA tree. Specifically, the former methods cluster species 1 and 2 together, but UPGMA puts 2 at a distant position, sister to the clade of species 1, 3, 4, and 5. The UPGMA tree also shows species 3 and 4 as a clade, while NJ and parsimony put 3 and 4 in a paraphyletic relationship to each other.

b. If there are differences, explain why one method might have given different results from the others.

The likely reason is unequal rates of evolution. By clustering sequences together based on raw similarity, the UPGMA method assumes a molecular clock. But the total pairwise distances on which this method is based include unique derived characters (autapomorphies -- character states that have changed in the branch leading to a terminal taxon) that in fact tell nothing about phylogenetic relationships. If one species has more autapomorphies than others, that species will have higher distances from the other taxa and will be erroneously assumed to have diverged very anciently. This is precisely the case for these data, because species 2 has many autapomorphies (and therefore a high rate of sequence evolution in this gene), as shown by the parsimony tree.

6. Morphological data do not resolve the relationships among the mammalian orders primates, artiodactyls, and rodents. You would like to use molecular data to establish which lineage diverged first from the others. In your laboratory, your research assistant obtains sequences from a cow, a human, and a mouse for three genes: psi-n-globin (a pseudogene of globin, the oxygen-transporting protein in blood), histone A1 (one of the proteins that packs DNA in chromatin), and 18S ribosomal RNA. You use parsimony for a phylogenetic analysis.

a. Which gene do you expect to be most useful for resolving the relationships among these taxa?

Phylogenetic analysis requires the use of genes that provide an appropriate balance between variability and conservation of sequence for the particular taxa at hand. As we have seen, the mammalian orders split from each other on the order of 60-120 million years ago. In this case, 18S is most likely to be useful. The three genes will have radically different rates of divergence. The pseudogene, subject to no selective constraints, will diverge rapidly. Over such a long period of time, it is unlikely to have any useful information for resolving these relationships; it would be appropriate for more recent divergences, such as for those between species in a genus. Histone we know is one of the most conserved proteins known -- its amino acid sequence is almost completely constant among all metazoans. The selective constraints for this protein are so strong that there is unlikely to be adequate variation to give any adequate information for resolving mammalian relationships. (There will be variation at synonymous positions, but since these are subject to virtually no selective constraints, they will behave like sites in a pseudogenes and will not be informative either). 18S is more conserved than a pseudogene but more variable than histone -- it is your best bet in this case.

b. What other information do you absolutely need to resolve the relationships among these taxa?

You need the sequence from at least one known outgroup to root the tree. Otherwise you can infer only an unrooted tree; in fact for three taxa there

is only one unrooted tree, so an analysis without a root would tell you absolutely nothing. An appropriate outgroup would be one or more taxa that diverged from your three mammalian taxa before they diverged from each other -- for instance, marsupials and/or monotremes would be good. You could also use a bird or reptile, but a more distantly related outgroup is less reliable than a more closely related one.

c. Suppose the gene you chose gives the phylogeny ((primate,rodent),artiodactyl), and the node supporting a primate-mouse clade has a bootstrap value of 55 and a Bremer support of 2.

- **What does a bootstrap value of 55 mean? -- explain in a sentence or two. What does a Bremer support of 2 mean?**

A bootstrap of 55 means that this node appeared in the most parsimonious tree for 55% of the bootstrapped data matrices created by randomly sampling nucleotide positions from the original alignment until the new matrix contains the same number of characters that the original one did. 55% implies that a considerable proportion of the sites in the sequence (an unknown fraction - not 55%!) support rodent-artiodactyl or primate-artiodactyl clades.

A Bremer support of 2 means that the best tree that does not contain the primate-mouse clade is 2 steps longer than the most parsimonious tree that does contain it. This means that the alternative phylogeny requires 2 extra parallelisms/reversals to explain shared character states as due to something other than common descent.

- **How confident are you that you have the true phylogeny? Justify your answer.**

I’m not confident, because these are both fairly low values. If this node appears in only about half of the most parsimonious trees for the bootstrapped matrices, there is considerable phylogenetic noise/homoplasy in the sequences, and the node in question may be the result of "sampling error" -- the favoring of one tree over another by chance alone. A Bremer support of two means that the phylogeny with this node is only a slightly more parsimonious explanation of the sequence data than the next best tree, requiring just two extra character states to be explained by homoplasy. So the data supports this node, but not very convincingly.

- **If this is not the true tree, what are some reasons you might have gotten this tree anyway? What should you do to address these problems?**

The sequence used might have considerable noise in it (lots of homoplasy), which could result in the true tree not being the shortest one because the signal is weak relative to all this noise. Another way of saying this is that the characters in this sequence support this tree solely due to chance/sampling error, but if you had more informative characters in your sample they would eventually support the "true" tree over the erroneous tree.

An incorrect phylogeny could also be due to long branch attraction: if two of the sequences diverged much more rapidly than the others, they would share a considerable number of character states due to chance alone, and they would cluster together in the most parsimonious tree even if they are not in fact from the most closely related taxa. This problem would be exacerbated if the two taxa with long branches share the same kind of codon usage/nucleotide composition bias.

To address this problem, get more sequence data: use more genes from more taxa, and work hard to find sequences that have the right level of variability -- good signal, not much noise. More genes will provide more characters to help resolve relationships, and if you choose well they will provide more signal without much more noise. This strategy will address both the lack of strong support and the possibility of long branch attraction. Adding sequences from more taxa will will further to address long branch attraction, if you get sequences from "intervening" taxa, because these taxa will break up the long branches into shorter branches, making saturation less likely to occur because you can infer intermediate states, as well.

d. Suppose now that the gene you chose gives the same phylogeny as above, but the node supporting a primate-mouse clade has a bootstrap value of 95 and a Bremer support of 9. Do you have more or less confidence in your phylogeny? Why? Is it possible still that you do not have the correct taxonomic phylogeny? If so, why might you have gotten this tree?

More confidence. Higher bootstrap and Bremer values mean that there is stronger character support and less homoplasy in the sequences, so this tree is not likely to be the most parsimonious just due to the chance effects of noise/homoplasy.

It is possible, however, that your tree is still wrong. First, if there is extreme saturation of two of the sequences, the wrong tree can arise due to long branch attraction and be strongly supported by saturated characters. If so, get more appropriate genes and more taxa into your analysis.

Second, your tree could be the correct phylogeny for your gene but not for the taxa that carry them. Incongruence of gene trees with species trees could occur for a number of reasons. It could be due to horizontal transfer of genes across species boundaries, causing more distantly related taxa to carry closely related genes. It could be due to the use of paralogous duplicated genes rather than orthologous sequences for some taxa; use of a paralogous gene can make a closely related taxon appear to have diverged more anciently from another taxon to which it is closely related. Finally, it could be due to ancestral polymorphism/lineage sorting: if the ancestral population of these three orders had two alleles of the gene you are using at the time the first cladogenesis event occurred, one of the alleles will be lost in each lineage, entirely due to the chance effects of drift. Quite frequently, the same allele will be lost in one taxon and a more distantly related one, while the other allele is lost in a more closely related taxon. The result will be that the distantly related species will carry more closely related alleles.

The gene tree is accurate but the species tree is wrong.

In all three of these cases, the solution is to get more genes and add them to your analysis. It is unlikely that a large group of genes will all have been transferred by horizontal transfer, especially if they are from different chromosomes. Because lineage sorting is a form of drift, trees for different genes will be produced with equal probability. If you use lots of genes, the genes for which lineage sorting has occurred will give equal support to competing trees, but the (presumably greater) number for which lineage sorting did not occur will all favor the correct taxonomic tree. To prevent false orthology, try to use genes that are thought to be single-copy genes rather than members of large gene families.