1 Inferring trees from a data matrix

Last week, we considered a character matrix shown in table 1. Taxon A is the outgroup, and the table uses colors to indicate the information content of the characters. Blue for characters with only autapomorphies, green characters for with a synapomorphy that support the assumed separation between the ingroup and outgroup, and red for the character with a synapomorphy that can inform the other parts of the tree.

Table 1: Table 1 with color-coding of character types

	Character #									
Taxon	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1
D	0	0	0	0	1	1	1	1	1	0

1.1 Outgroup polarization – recap

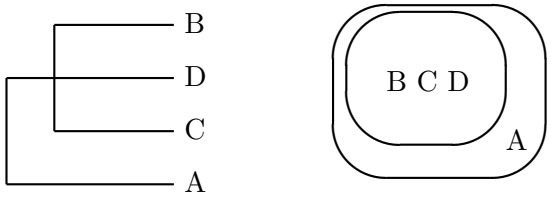
The most common way of determining the orientation of phylogenies (and, thus, characters) is the outgroup method. Usually we have a group of taxa that we are interested in, and we can identify organisms that are "more distant phylogenetically" than members of the group of interest. For example if we are interested in the African great ape phylogeny we might feel comfortable in assuming that human, chimp, and gorilla (the 3 African great apes) are more closely related to each other than any of them are to orangutan. Thus we can use orangutan as an **outgroup** in the analysis. In effect we treat as outgroup's character states as if they were the plesiomorphic states. We know that orangutan is not identical MRCA of orangutan and the African apes, but it turns out that this does not cause problems for phylogeny reconstruction (though it does make the interpretation of character evolution more complex – we'll discuss this later in the course). In the previous example the African apes are the **ingroup**.

In short, we assume (or use outside, independent evidence) to find an outgroup. When we do this, we **must** be confident that the *ingroup is monophyletic with respect to the outgroup.*

Going back to the ape example, not long ago many scientists felt that orangutan might be the closest living relative to humans. If we still view this as an open question, then it would be inappropriate to use orangutan as an outgroup with human, gorilla, chimp as the ingroup – we would move to another taxon such as an old world monkey (e.g. a baboon). Using very distantly related outgroups can lead to problems because it may be difficult to make homology statement over long stretches of time.

Consider the data in table 1 again. If this data were being analyzed by Hennig, then he

Figure 1: The assumption of A as the outgroup



would identify an outgroup. For the sake of argument, we will say that we have identified taxon A as the outgroup. So we start the analysis "knowing" that the B, C, and D form a monophyletic group, thus the tree we start with is shown in figure 1; The polytomy (before we look at the data) is interperted as a soft polytomy (representing uncertainty about the phylogeny rather than a statement that B, C, and D were created by a single speciation event).

2 Phylogenetic inference by Hennigian character analysis

Recall that Hennig pointed out that synapomorphies alone are needed to recognize monophyletic groups – in particular symplesiomorphies are not helpful. Consulting table 1, we see that only one character (#10) tells us something new about the relationships between taxa.

The apomorphies in characters 1-5 identify taxa as having acquired a new state, but in all of these case the states are unshared – autapomorphies. These characters reinforce our belief that taxa A, B, C, and D are distinct (if that were an open question), but the only groups that the apomorphies support are single taxa "groups." These are trivial groupings – ones that will be found an every possible tree for these 4 taxa.

The green characters (6-9) point to the existence of a monophyletic group B + C + D, but this was the entire ingroup. So these characters fall on the internal branch of the tree shown in figure 1. This is a branch that we already knew was in the tree.

Finally we come to character # 10. The apomorphic state for this characters points to grouping of B + C. This grouping is not present in the tree shown in figure 1, so we have learned something about the phylogeny from this character matrix (or we made a mistake making the homology statements during the construction of the matrix, and have inferred something incorrect about the phylogeny).

Figure 2 shows the tree that Hennigian inference would prefer. Tick-marks on branches are paired with numbers. These numbers indicate the number of the character(s) in the matrix that support that branch. Each character changes from the plesiomorphic state to the apomorphic state on the branch with the corresponding tick mark. Figure 3 shows a

Figure 2: The tree preferred by Hennigian analysis of the data in table 1

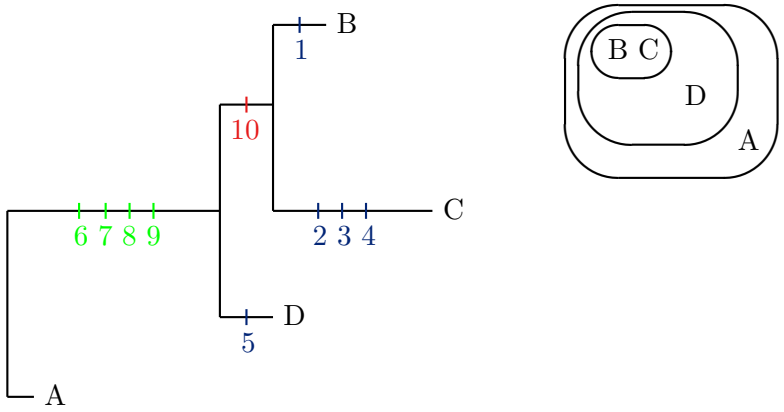
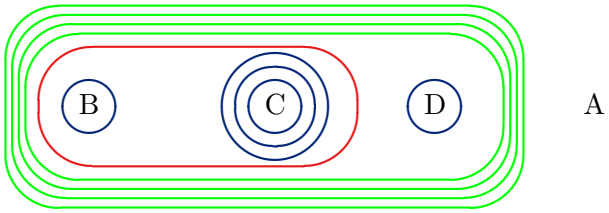


Figure 3: The tree preferred by Hennigian analysis of the data in table 1



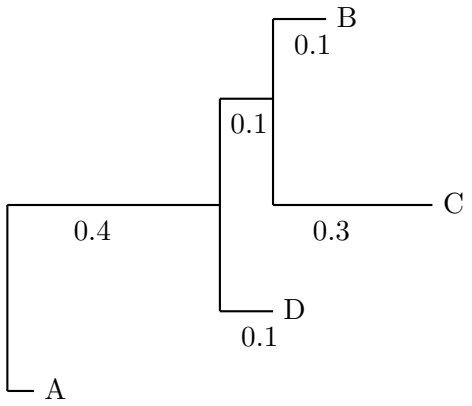
hierarchical characterization in which each character results in a line surrounding the taxa with apomorphic states. The lines are colored according to the character colors (but not labelled to avoid cluttering the figure).

2.1 Lack of homoplasy

Note that there does not appear to be any homoplasy when we map the data onto the tree shown in figure 2. Each character can be mapped on to the tree with a single transition from the plesiomorphic state to the apomorphic state. This is the example of a best case scenario. If we view the initial construction of the matrix as a primary hypothesis of homology, and the construction of the tree as the secondary test of homology, then we would say that the tree gives us no reason to question our primary homology statements.

This lack of homoplasy can be seen in the figure 3 by the fact that none of the lines have to cross each other. Each character corresponds to a grouping that is **compatible** with the grouping implied by every other character. There is a one-to-one correspondence between a (variable) character in the matrix and the branch on the tree that it "supports" – we say that a character supports a branch when the character changes state across the branch.

Figure 4: The tree figure 2 with branches expressed as the proportion of characters that change across the branch.



This means implies that an evolutionary event that occurred in the lineage that the branch represents

2.2 Path lengths = character divergence

Also note that if we label the branches with the proportion of the characters in the matrix that change across each branch then we get the tree show in figure 4.

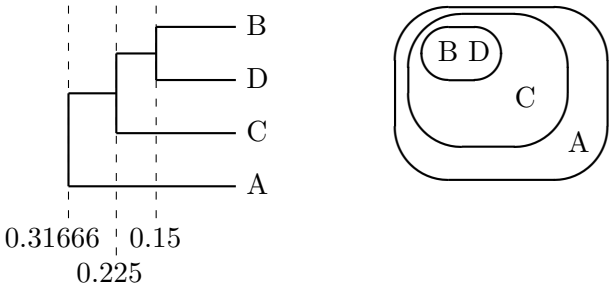
A *path* on a tree is a set of branches that you have to follow when you move from one taxon to another. A path length can be defined as the sum of lengths of all of the branches along the path – in the context of a leaf-to-leaf path, this is referred to as a patristic distance. Interestingly, if we use the tree in figure 4 to construct a path length matrix for all pairs of taxa, then we will obtain a matrix that is identical to the pairwise distance matrix (table 2) that we calculated directly from the character data.

Recall that UPGMA (Unweighted Pair Group Method with Arithmetic mean) inferred a tree by summarizing the character matrix into a "dissimilarity" (or "difference" or "distance") matrix shown in table 2. By repeatedly clustering the closest pair of taxa (or "operational taxonomic units" – OTU's), we produced the tree shown in figure 5.

Table 2: The pairwise distance matrix for the characters shown in Table 1

	Taxon			
Taxon	A	B	C	D
A	-	0.6	0.8	0.5
B	0.6	-	0.4	0.3
C	0.8	0.4	-	0.5
D	0.5	0.3	0.5	-

Figure 5: the UPGMA phenogram and hierarchy from distance matrix in Table 2



Even though the Hennigian-based tree did not use the distance matrix directly it was able to explain the distance matrix perfectly (while the distance-based approach fails). The patristic distances from the Hennigian tree in figure 4 perfectly return the character-based distances shown in Table 2. However if we use calculate a taxon-to-taxon distance matrix from the patristic distances from the UPGMA phenogram shown in figure 5, we get the matrix shown in table 3

Table 3: Patristic distance matrix from the phenogram shown in figure 5

	Taxon			
Taxon	A	B	C	D
A	-	$0.633\overline{3}$	$0.633\overline{3}$	$0.633\overline{3}$
B	$0.633\overline{3}$	-	0.45	0.3
C	$0.633\overline{3}$	0.45	-	0.45
D	$0.633\overline{3}$	0.3	0.45	-

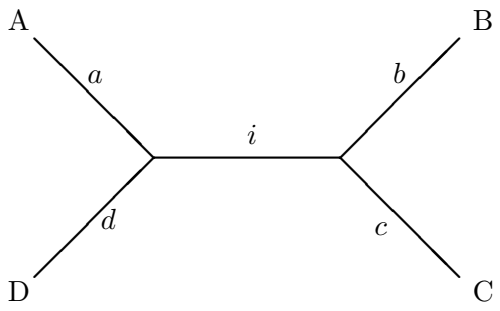
We noted that UPGMA has problems as a phylogenetic inference method when the rates of evolution are not equal on all parts of the phylogeny. Note that this tree appears to be such a case. The branch leading to C is 3 times as long (3 times as many characters changed) as the terminal branch leading to taxon B or to D. In fact the path from B to D is short – these two taxa are very similar, but mainly because of sharing states which are symplesiomorphies. UPGMA does not attempt to discriminate between primitive and derived similarities, and the result is that it groups B and D because their raw distance was lower than the B to C distance.

2.3 Accurate estimation from distances

Can we accurately predict the tree if we have good estimates of the evolutionary distance between taxa?

Yes. If our distances were directly proportional to the time to the MRCA between each pair of taxa, then UPGMA would always yield the correct phylogeny. But that is not too helpful,

Figure 6: Tree with edge lengths



though because we don't observe these times – we can only observe characters data that gives us some clues about the history.

What if we knew the true number transitions that had occurred between species – would that be enough information to correctly infer the phylogeny? The distance matrix shown in table 2 represents an example of such knowledge – As demonstrated by the fact that we could recover it as the patristic matrix from the Hennigian tree in figure 4 (every change on the phylogeny in that figure contributes to an increase in the evolutionary distances between leaves of the tree).

A computer scientist named Peter Buneman (1971) showed that a dissimilarity matrix that accurately measures the evolutionary distance between taxa *is* sufficient knowledge to infer a phylogeny. Rather than just group the smallest distance in a cluster (as UPGMA did), Buneman's method works on quartets (groups of four taxa) and chooses the tree based on the sum of two distances:

Table 4: Correspondence between tree and pairs of distances in a method based on Buneman's four point condition

Tree	distance sum	sum on tree in figure 6	from table 2
$(A+B)\|(C+D)$	$d_{AB}+d_{CD}$	$a+b+c+d+2i$	1.1
$(A+C)\|(B+D)$	$d_{AC}+d_{BD}$	$a+b+c+d+2i$	1.1
$(A+D)\|(B+C)$	$d_{AD}+d_{BC}$	$a+b+c+d$	0.9

Note that when the sum associated with the true is the smallest (and the internal edge length can be estimated by half the distance between that sum and the other two).