1 Inferring trees from a data matrix

Consider the character matrix shown in table 1. We assume that the investigator has conducted primary homology analysis such that:

1. the characters (columns) contain codes for aspects of the organisms that are thought to be comparable (we assume that the character homology statements are correct).
2. the character states are described with sufficient detail that we expect organisms with the same state to both being displaying the same evolutionary innovation (we assume that the character state homology statements are correct – satisfying Remane's "special similarity" and continuation criteria).

Table 1: A simple character matrix

	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

2 Clustering by distance

The most obvious way to infer a tree of taxa that describes this data is to cluster taxa based on similarity. We can produce a pairwise distance matrix for this set of taxa that reveals the proportion of characters for which any two taxa differ. This is shown in Table 2.

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	-

2.1 Side-note about distance matrices

Note that the distance matrix is symmetric because the distance from taxon A to taxa B is the same as the distance from B to A. The distance matrix summarizes the amount of

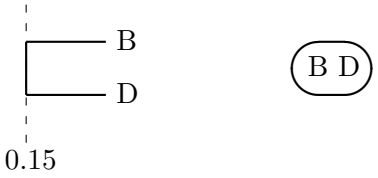
divergence in character states between two taxa, but is not a complete description of the information in the character matrix. It is possible to convert a character matrix into a distance matrix, but we can't invert this mapping. Another way to say this is to note that many different character matrices can map to the same distance matrix. The implication of this line of thought is that we may have more power to infer trees if we use the character matrix directly. This turns out to be the case, but for now we will pursue a simple method based on the matrix of pairwise distances.

2.2 UPGMA

Unweighted Pair Group Method with Arithmetic mean – a technique from numerical taxonomy (phenetics).

We expect close relatives to be similar to each other, so we could construct a tree by progressively grouping the pair of taxa that are closest – those 2 with the smallest distance. If we do this using the distance matrix in Table 2 we see that the smallest distance is 0.3 and is this is the distance between taxa B and D. This gives us the small tree shown in figure 1
Note that we set the node that connects B and D back in time a distance (0.15) that is half

Figure 1: round 1 of UPGMA from distance matrix in Table 2



the distance between B and D. This means yields a path from B to the MRCA of B and D then to D with a total length of 0.3 – which is the observed distance between the taxa.

We have a start on constructing a tree, but note that to take the next step we have to have distances between the other taxa (A and C) to the new group (B+D). In UPGMA, we create these distances by taking arithmetic averages. So

$$\begin{aligned} dist(A \leftrightarrow (B+D)) &= \frac{dist(A \leftrightarrow B) + dist(A \leftrightarrow D)}{2} \\ dist(C \leftrightarrow (B+D)) &= \frac{dist(C \leftrightarrow B) + dist(C \leftrightarrow D)}{2} \end{aligned}$$

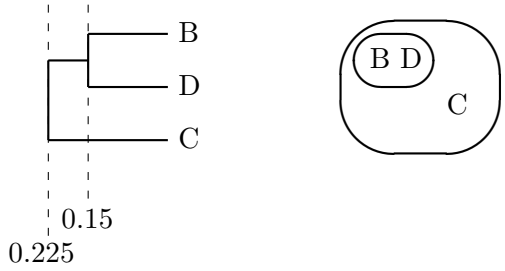
This allows us to construct a distance matrix for the second "round" of UPGMA. See table 3

At this stage the smallest distance is between C and the group (B+D). So this next group on the tree will lead to B, C, and D, and the internal node will have a depth of $\frac{.45}{2}$ (or .225). This gives us the small tree shown in figure 2
If we consider the distance matrix again, we find that we have only one distance:

Table 3: The second round pairwise distance matrix derived from perfect4dist after grouping B and D.

Taxon	Taxon		
	A	(B+D)	C
A	-	0.55	0.8
(B+D)	0.55	-	0.45
C	0.8	0.45	-

Figure 2: round 2 of UPGMA from distance matrix in Table 2



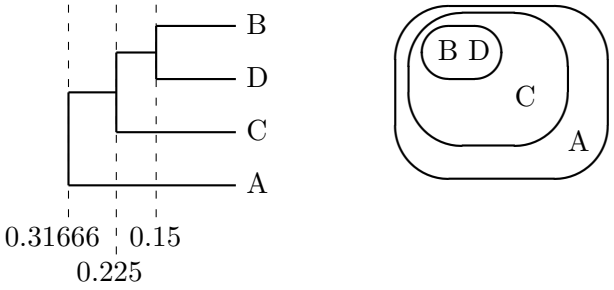
$$dist(A \leftrightarrow (B + C + D)) = \frac{dist(A \leftrightarrow B) + dist(A \leftrightarrow C) + dist(A \leftrightarrow D)}{3}$$
$$\approx 0.6333333$$

Thus we can place the root of the tree at $\frac{0.6333333}{2}$ (or 0.3166667), and attach taxon A to complete the tree, see figure 3.

problems with UPGMA

We have just constructed the upgma tree for the characters matrix shown in table 1 (via the distance matrix shown in table 2). UPGMA is a phenetic technique. Recall that the numerical taxonomists (or pheneticists) were more interested in producing clustering tools that summarize the similarity between groups. They tended to think the phylogenetic relationships were too difficult to estimate to serve as a basis for taxonomy. We can ask how well the tree (or "phenogram" in this case) represents the distances in the distance matrix.

Figure 3: the UPGMA tree from distance matrix in Table 2



The answer is that it does not do a perfect job. For instance, just looking at the tree you would expect that

$$dist(C \leftrightarrow B) = dist(C \leftrightarrow D),$$

and

$$dist(A \leftrightarrow B) = dist(A \leftrightarrow C) = dist(A \leftrightarrow D),$$

but neither of these sets of equations apply to the distance matrix in table 1. Thus has been some distortion of the distance matrix as we fit it onto a tree. This could be caused by the fact that we have a very small sample of characters – so random errors in the data matrix obscure the expected amount of character change between taxa. Another possibility is that UPGMA, while a straightforward way to construct a tree, is too simplistic and does not handle the complexities of real data. This second possibility has been demonstrated convincingly. In particular UPGMA is very sensitive to changes in the rate of character evolution – and these changes seem to be common in evolution. Thus UPGMA is not commonly used anymore.

3 Phylogenetic inference by Hennigian character analysis

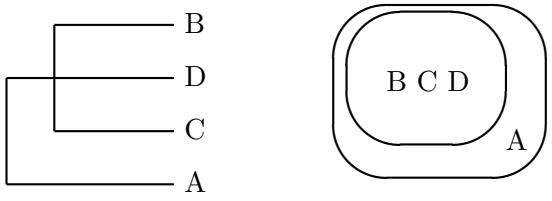
Note that in the previous section we did not use the terms "apomorphy" or "pleisiomorphy." That is because UPGMA does not come out of the phylogenetic systematics school that Hennig was crucial in developing. Recall that Hennig pointed out that synapomorphies alone are needed to recognize monophyletic groups – in particular symplesiomorphies are not helpful. But from table 1 alone we cannot determine which states are apomorphic and which are plesiomorphic.

3.1 Outgroup polarization

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.*

Figure 4: The assumption of A as 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 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 4; 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).

For convenience, I coded every character in the data matrix in table 1 such that the character stated displayed by taxon A is denoted by 0. Thus, using the outgroup (A) to polarize the characters, we assume that 0 is the plesiomorphic state and 1 is the apomorphic state. Recall that Hennig pointed out that we can only learn about the phylogeny through apomorphic states, specifically from shared, apomorphic states – synapomorphies.

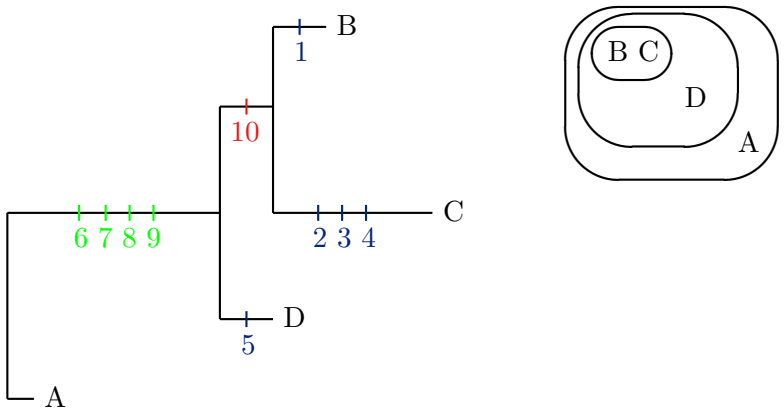
Table 4 shows the same data as table 1, but the characters have been color-coded. Blue for characters with only autapomorphies, green characters for with a synapomorphy that support the assumed separation between the ingroup and outgroup, and red for the character with a synapomorphy that can inform the other parts of the tree.

Table 4: 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

We see that only one character (#10) tells us something new about the relationships between

Figure 5: The tree preferred by Hennigian analysis of the data in table 1



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 4. This is a branch that we already knew was in the tree.

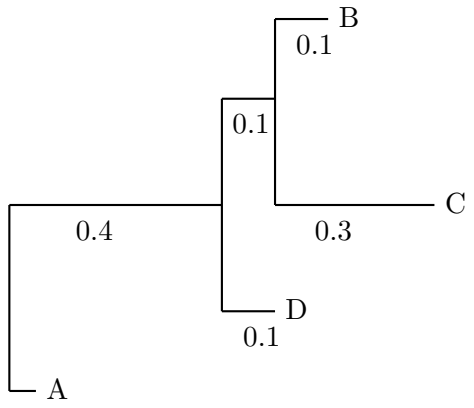
Finally we come to character # 10. The apomorphic state for this characters points to grouping of B + C. This grouping is not present in the tree shown in figure 4, so we have learned something about the phylogeny from this character matrix (or we made a mistake making the homology statements during the construction of the matrix, and have inferred something incorrect about the phylogeny).

Figure 5 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.

3.2 Lack of homoplasy

Note that there does not appear to be any homoplasy when we map the data onto the tree shown in figure 5. 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

Figure 6: The tree figure 5 with branches expressed as the proportion of characters that change across the branch.



tree gives us no reason to question our primary homology statements.

3.3 Path lengths = character divergence

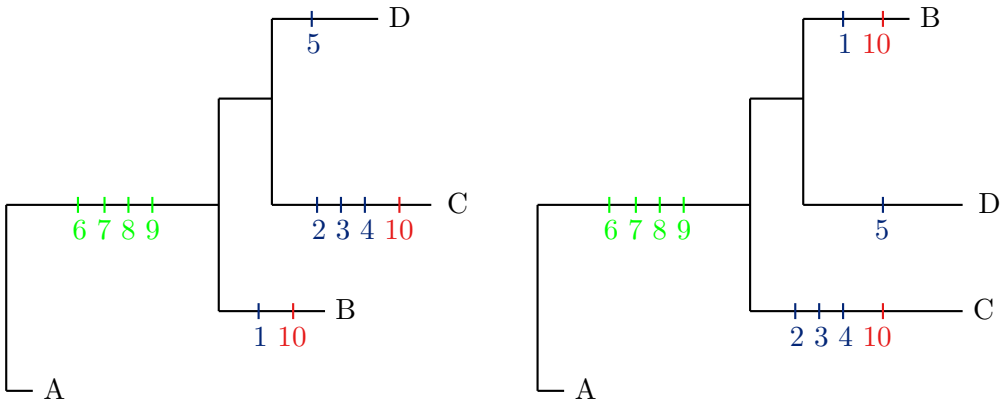
Also note that if we label the branches with the proportion of the characters in the matrix that change across each branch then we get the tree show in figure 6. A *path* on a tree is a set of branches that you have to follow when you move from one taxon to another. A path length can be defined as the sum of lengths of all of the branches along the path. Interestingly, if we use the tree in figure 6 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 from section 2.2 on the problems with the UPGMA tree, that there were some obvious problems with the UPGMA in terms of not being able to predict properties of the distance matrix. 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).

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.

3.4 Alternative trees require homoplasy

In this case of perfect, homoplasy-free data we were able to construct the tree by adding groupings every time we found a character that supported the grouping (in this case there

Figure 7: The bifurcating trees rejected by Hennigian analysis of the data in table 1



was only one helpful character, but that is because it was a tree of only 4 species). We might ask whether or not we can recognize incorrect trees. The other two bifurcating trees for these taxa are shown in figure 7. Note that in both of these trees there is an internal branch with no inferred character changes – these branches have no support in this character matrix. The other point to note is that in each of these trees, character #10 (the only character with a synapormorphy that gave us information about the relationships within B+C+D) can only be explained as a homoplastic character – with two independent acquisitions of the 1 character state. The conclusion is that the analyses suggested by Hennig can not only allow us to construct the tree, but given a tree we can:

- detect unsupported groupings (branches with no inferred changes), and
- detect disagreement between the tree and character matrix (mapping characters on rejected trees requires homoplasy)

References