APPENDIX TWO

Generating a Morphological Data Matrix

Although the field of phylogenetics was founded on morphological data, most modern research uses molecular data, especially DNA sequences, which are abundant and relatively inexpensive to obtain. However, fossils can only be analyzed through the use of morphological data. Additionally, even when trees are inferred from molecular data, it often becomes necessary to build a morphological data matrix in order to study the evolution of morphological traits (Chapters 4 and 10). Therefore, despite the preeminence of molecular phylogenetics, it is important to know how morphological data are scored and assembled into data matrices.

Two steps can be recognized in the building of a morphological matrix, which we will call ***character encoding*** and ***character scoring.*** Character encoding involves deciding on the limits of characters and on the alternative states that are recognized for each character. For example, when you decide to score fur color using two states, brown/black and white, you have encoded one character (fur color). Character scoring involves looking at each taxon and assigning it a state for each encoded character. For example, you could score an otter as having brown/black fur. In practice observations made while scoring taxa often result in changes to character encoding, but it is still useful to distinguish these two steps.

Once a set of taxa has been selected for study, a systematist generally starts by looking for characteristics that appear to vary among them. Notice that, whereas many characters in a DNA sequence data matrix may be invariant, the way that morphological characters are selected means that constant characters are usually excluded from the outset.

Once some variation has been noted, the next challenge is to appropriately encode the characters. This is not straightforward. For example, imagine that

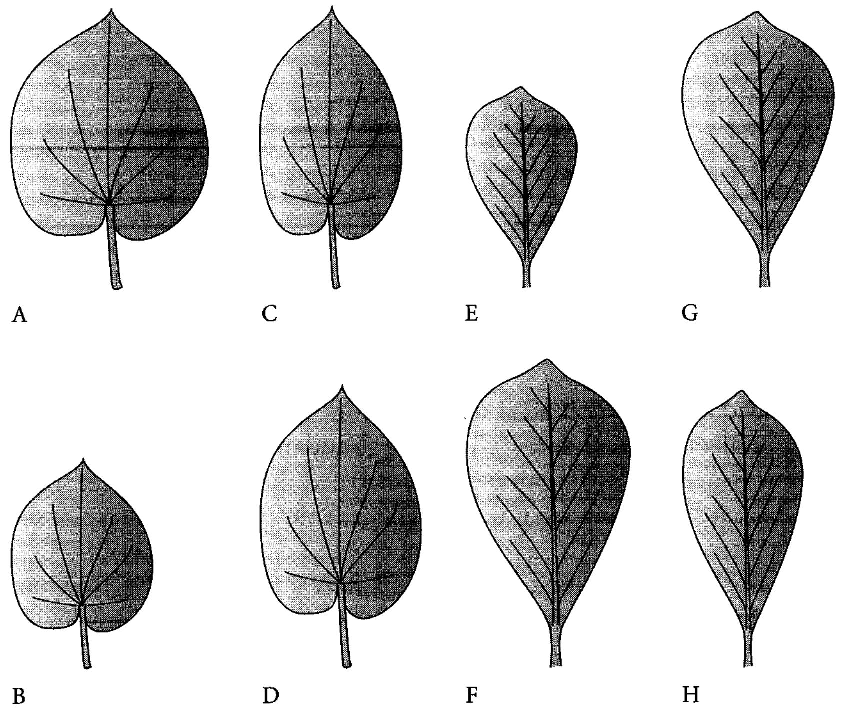

FIGURE A2.1 A hypothetical example of variation in leaf shape.

you observed that leaf shape and size differed among a set of eight plant species, as shown in Figure A2.1. How would you capture this variation?

Consider two of the numerous possible ways to encode this variation. (1) You recognize two basic leaf shapes, cordate (with the widest point in the lower half) and obcordate (with the widest point in the upper half), and two size classes. (2) You encode leaf length, leaf width, and the height of the widest point. The data matrix that might result from these two encoding schemes is shown in Tables A2.1 and A2.2. It is possible that the differences between these encoding schemes could alter the phylogenetic conclusions reached.

The decision among alternative encoding schemes is guided by a few, potentially conflicting, considerations. You want to capture as much of the variation as possible without "double counting." Scoring the same basic variation multiple times results in overweighting that variation to the point where it will dominate the phylogenetic results. For example, you might be concerned that by measuring both length and width of these leaves you might score one basic trait, leaf size, twice.

TABLE A2.1 One possible coding scheme for leaf shapes in Figure A2.1

Taxon	Leaf shape (0 = cordate; 1 = obcordate)	Leaf size (0 = small; 1 = large)
A	0	1
B	0	0
C	0	1
D	0	1
E	1	0
F	1	1
G	1	1
H	1	0

TABLE A2.2 An alternate possible coding scheme for leaf shapes in Figure A2.1.

Taxon	Leaf length (0 = short; 1 = long)	Leaf width (0 = wide; 1 = narrow)	Height of widest point (0 = below middle; 1 = above middle)
A	1	0	0
B	0	0	0
C	1	0	0
D	1	0	0
E	0	1	1
F	1	0	1
G	1	0	1
H	1	1	1

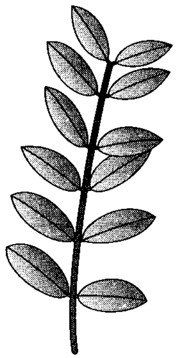

FIGURE A2.2 A compound leaf.

Another important consideration is that the character states recognized should really be versions of the same character. This is not always easy to decide. Suppose that close relatives of these plants have compound leaves (Figure A2.2). Should their "leaf shape" be encoded based on the individual leaflets or the outline of the whole compound leaf? The answer to this question will depend on your understanding of the developmental changes that resulted in a transition from compound to simple leaves, or vice versa.

Once characters are defined, the next question is how to delimit character states. If the variation is rather discrete between taxa, with little variation within taxa, as illustrated by leaf shape in the preceding example, then it may be easy to delimit character states. However, most morphological traits are inherently continuous and variation within taxa is common. Thus, it can be difficult to divide continuous variation into the discrete states needed for phylogenetic analysis. To illustrate the challenge, Figure A2.3 shows hypothetical data on leaf length in 10 species.

You might see the data in Figure A2.3 as being composed of three "clusters" of taxa corresponding to three states: small (A and E), medium (D, F, and H), and large (B, G, and I). In that case, you might score C as polymorphic for small-plus-medium and J as polymorphic for medium-plus-large. Alternatively, you could recognize two classes (small and large), or five size classes (A and E; C; D, F, and H; J; B, G, and I). Unfortunately, there is no well-grounded theory to tell you which of these encoding schemes will yield the best estimates of the phylogeny of these species.

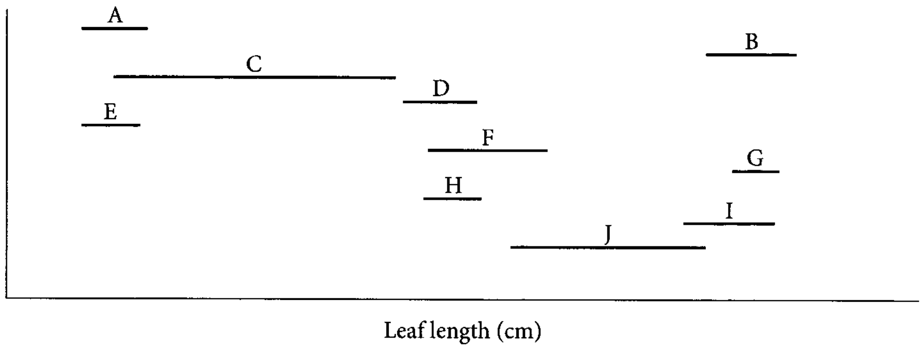

FIGURE A2.3 Variation in leaf length within and between 10 hypothetical taxa (A–J).

Taken together, you can probably see that there are many somewhat subjective decisions that must be made in encoding morphological data and that these are likely to be adjusted by observations made while scoring individual taxa. While subjectivity is something that makes scientists uncomfortable, this fact does not invalidate morphology as a source of phylogenetic data. So long as different encoding schemes capture the actual variation among taxa, then they should yield similar, if not identical, estimates of the tree. It is considered good practice to try a few different schemes and see if the phylogenetic conclusions remain the same. If they do, and if statistical approaches such as the nonparametric bootstrap (Chapter 9) indicate high support for clades, it is possible to achieve quite high confidence in the conclusions of a morphological phylogenetic analysis. For example, the relationships of many extinct groups that are known only from fossils are now known with a high degree of certainty.