Methods and Principles of Biological Systematics

Biological systematics is literally “the study of biological systems.” While this originally focused on classification, the field has broadened into discovery of phylogeny and use of phylogenies to understand the processes underlying diversification. In this broad definition of systematics, producing a classification and assigning names to groups (a field also known as taxonomy) is only one component of systematics. This chapter is divided into two sections. The first outlines how one goes about determining the history of a group, and the second then discusses how the history can be used to uncover patterns of diversification and to construct a classification. This field has become so large that it can stand as a course by itself, so the presentation here is necessarily brief and introductory. Many of the topics in this chapter are discussed at greater length in books by Baum and Smith (2012) and Stuessy et al. (2014), which should be consulted for additional details, references, and examples.

Discovering Phylogeny: How Phylogenetic Trees Are Constructed

As described in Chapter 1, evolution is not simply descent with modification, but also involves the separation of lineages. This process can be visualized with diagrams such as those in Figures 1.3 and 1.4. We can also think of phylogeny as the history of DNA molecules. DNA replicates in a semi-conservative fashion to produce two daughter molecules, each of which replicates again to produce its own two daughter molecules. This process has been going on in an unbroken chain since the beginning of life on Earth. Thus each DNA molecule present today must have come from an ancestral DNA molecule in an earlier generation. By tracing the history of DNA molecules, we can uncover the history of life.

Evolutionary Trees and What They Depict: Reading the Tree

The DNA replication process can be diagrammed as a tree (Figure 2.1). The tree shows ancestor-descendant relationships. Descendants 1 and 2 share with each other a more recent common ancestor (labeled “1-2” on the figure) than they do with descendants 3 and 4, and so 1 and 2 are interpreted as being more closely related to each other than either of them is to 3 and 4. Likewise, descendants 1...
through 4 are more closely related to each other (i.e., share a more recent common ancestor) than they are to 5 through 8.

DNA replication can thus be represented easily as a branching diagram. It is common to trace the history of one gene (a piece of DNA) and from that to infer the history of closely linked genes. From the history of genomes we can then infer the history of organisms that bear them, and from the history of individual organisms we can infer the history of species, genera, and families. It is worth keeping in mind that a diagram of a phylogenetic tree may be labeled to make it look like the history of species, but in fact it is an inference about species extrapolated from data on DNA sequences of individual genes, genomes, and cells. For example, because the large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco) is encoded in the chloroplast and we can sequence the large subunit of rubisco, we can infer the history of the entire chloroplast genome. Before DNA data were easily available, such inferences were made from morphological characters (see Box 2A).

The history of DNA molecules and thus the phylogeny is not changed by the orientation of the tree. For example, the tree in Figure 2.1 could be drawn with the most recent molecules at the bottom, top, right, or left, but the pattern of ancestor-descendant relationships would remain the same. In addition, the shape, or topology, of the tree is determined only by the connections between the branches. As long as the connections between the branches are unchanged, any branches can be rotated around the nodes; the inset in Figure 2.1 depicts exactly the same history as is shown in the main tree, but some of the branches have been rotated.

The evolutionary “story” thus can be described by starting at any branch tip in the tree and working toward the other branch tips. This means that the terms higher and lower have no meaning but simply reflect how we have chosen to draw the evolutionary tree. For example, some textbooks consider the Asteraceae an “advanced” family and Magnoliaceae and Nymphaeaceae “primitive” families. However, because the families have had exactly the same amount of time to evolve (i.e., the DNA has had exactly the same number of years over which to accumulate mutations), each is a complex mix of ancestral and derived characteristics, meaning that no extant plant family is more primitive or more advanced than the other.

Drawing a phylogenetic tree using DNA molecules as in Figure 2.1 is cumbersome, so we could replace the DNA molecules with ovals, as in Figure 2.2A, or we could leave out the ovals but still keep the labels for the ancestors,
as in Figure 2.2B. However, it is most common to draw phylogenetic trees without DNA molecules, ovals, or even labeled ancestors. Instead, the ancestors are simply represented by the nodes, or points where the branches join (Figure 2.2C). This is sometimes confusing for students because the lines are now more prominent visually than the nodes, but the pattern of inference is still the same. The history is traced backward from the terminals (current DNA molecules) to successively older nodes, which represent the common ancestors.

In the examples shown in Figures 1.3, 1.4, 2.1, and 2.2, we described evolution as though we were watching it happen. This is rarely possible, of course, so part of the challenge of systematics is that we must infer what went on in the past. The first step in making such inferences is to identify characteristics of organisms that are heritable. A heritable character is any aspect of the plant that can be passed down genetically through evolutionary time and still be recognizable. The most commonly used heritable characters are those of the DNA itself—the four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). However, morphological characters can also be used. A flowering plant’s petal color, inflorescence structure, and habit (general growth pattern) are all known to be under genetic control, and these characters are generally stably inherited from one generation to the next. Each character is then divided into character states, which are used in a way similar to DNA mutations to track evolutionary history. Many examples of such heritable characters are described in Chapter 4.

Sources of DNA Sequence Data

Sequencing determines the precise order of the nucleotides A, C, G, and T in a stretch of DNA. The DNA sequences from multiple organisms can then be aligned, and mutations can be detected by noting points at which sequences differ between two organisms. DNA sequence data are currently being generated in two ways: (1) a gene-by-gene approach, in which a gene of interest is chosen, isolated from a large number of organisms, and sequenced; and (2) a genomic approach, in which an entire chloroplast or nuclear genome is sequenced and the sequences of many genes from the genome are analyzed.

The plant cell contains three different genomes: chloroplast, mitochondrial, and nuclear (Table 2.1). Not only are these found in three locations in the cell, but they also differ in structure and gene content. Systematists use data from all three. The chloroplast and mitochondrion, and so also their genomes, are generally inherited uniparentally (usually maternally in angiosperms); the nuclear genome is biparental. The three genomes differ dramatically in size, with the nuclear being by far the largest—measured in megabases of DNA. The mitochondrial genome includes several hundred kilobase pairs (kbp) of DNA (200–2500 kbp), which makes it small relative to the nuclear genome but large relative to the mitochondrial genomes of animals (which tend to be about 16 kbp). The chloroplast genome is the smallest of the three plant genomes, in most plants ranging from 135 to 160 kbp.

Like the bacteria from which they are derived, mitochondria and chloroplasts have circular genomes. Large regions of noncoding DNA separate the genes in the mitochondrial genome, and their order in the genome is variable; in fact, their order changes so easily and frequently that many rearranged forms can occur even within the same cell. Rearrangements of the mitochondrial genome occur so often within individual
plants that they do not characterize or differentiate species or groups of species and are thus not especially useful for inferring relationships. The chloroplast genome, in contrast, is largely stable, both within cells and within species. The most obvious feature of the chloroplast genome is the presence of two regions that encode the same genes, but in opposite directions; these are known as inverted repeats. Between them are a small single-copy region and a large single-copy region (Figure 2.3).

Gene-by-Gene Approaches

The gene-by-gene approach is rapidly becoming of only historical interest, but we discuss it briefly here because it is still in use. For several decades, systematists have used the polymerase chain reaction (PCR) to amplify one gene at a time. They have then used this single gene as the basis for inferring the history of the cell and organism that bear it. The frequency of mutation of a gene determines how useful it is for addressing particular phylogenetic problems. In general, a rapidly mutating gene is needed to assess relationships among closely related populations or species. Genes that mutate more slowly can be helpful in studies of older groups; however, if a gene is changing slowly, it will be difficult to find mutations from which a phylogeny can be constructed. At a very low mutation rate, the level of variation will approach the expected level of sequencing error, and inferences will become unreliable. Conversely, if a gene is changing too fast, parallelisms and reversals will accumulate to the point that all phylogenetic information is lost; the history of the sequence will be obliterated. The latter problem is particularly acute in work with noncoding sequences or remotely related taxa.

Genes accumulate mutations at different rates, in part because gene products (proteins and RNAs) differ in how many changes they can tolerate and still function. Histones, for example, are proteins that generally cease to work if many of their amino acids are replaced with different ones. The internal transcribed spacer (ITS) of ribosomal RNA, on the other hand, can still fold properly even if many of its nucleotides are changed. Thus genes for histones do not accumulate mutations rapidly, whereas genes for the ITS do, reflecting the different functional constraints on their gene products. In general, chloroplast genes tend to accumulate mutations more rapidly than do mitochondrial genes in plants. It is more difficult to generalize about nuclear genes, which is hardly surprising, because there are so many of them.

When systematists first began using DNA characters, attention focused on chloroplast genes because of the ease of amplification and alignment and because many of them exhibit an appropriate level of variation. Some of the earliest insights from DNA data came from sequences of the gene for the large subunit of rubisco, mentioned above. This is still used from time to time but has been augmented with data from many other chloroplast genes. To track the history of the nuclear genome, many people used, and still use, the ITS. This is also easy to amplify because of its high copy number.
Both chloroplast genes and the ITS have their drawbacks. Because the chloroplast is maternally inherited, it will provide only part of the history of a group if there has been a history of hybridization and/or allopolyploidy. (Allopolyploidy refers to the doubling of the nuclear genome following hybridization. See Chapter 4 for discussion.) The high copy number of the ITS creates problems for data analysis because the thousands of copies in any cell are not all identical. In many studies, the ITS is treated as though it is a single marker, but in fact it has a complex pattern of evolution that is not easily interpreted. Thus it can provide an inaccurate picture of phylogenetic history.

Nuclear genes have been used less frequently for several reasons. First, they vary enough that new sets of PCR primers must be designed for each gene in each taxonomic group studied. In addition, many nuclear genes are duplicated or exist as part of a small set of genes (a gene family). Because of this, some preliminary work is often required to be sure that all the sequences used are from genes that are related by descent (orthologous genes) and not simply recent duplicates (paralogous genes). Fortunately, population genetic theory suggests that allelic variation should not be misleading in studies of closely related species, because alleles in one species should be more closely related to one another than they are to alleles in other species. Data on many nuclear genes support this expectation.

Genomic Approaches

It is now possible to sequence entire chloroplast genomes rapidly and cheaply, so phylogenetic studies are moving rapidly beyond focusing on single genes and instead use hundreds of genes. Sequencing technology is improving and changing at a remarkable rate, such that genomic studies are rapidly superseding studies of individual genes, and the techniques of molecular systematics are becoming those of genomics. Advances in sequencing methods are such that we did not attempt to describe here the range of technologies in the category of next generation (“next-gen”) sequencing. These are developing and disappearing so rapidly that whatever we wrote would have been obsolete before this book was published. For an introduction to current sequencing technology, see van Dijk et al. (2014).

Genome architecture has been used for many years to track phylogenetic relationships. Rearrangements of the chloroplast genome are rare enough in evolution that they can be used to demarcate major groups. For example, genome mapping has shown that one of the inverted repeats has been lost independently in a clade of papilionoid legumes, in all conifers, and in Euglena (a flagellated photosynthetic eukaryote unrelated to green plants). In another example, in all Asteraceae except subfamily Barnadesioideae, a portion of the large single-copy region is inverted relative to other angiosperms (Jansen and Palmer 1987). This provides evidence that Barnadesioideae is sister to all other Asteraceae.

DNA data are widely available and can be obtained quickly and cheaply if necessary. Sequences of both individual genes and whole genomes are available from GenBank, at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/genbank/), which is the repository of a vast amount of publicly accessible data.

Nuclear genomes are also becoming far more accessible for analysis because use of next-gen approaches is much less expensive than amplifying and sequencing one gene at a time. Entire nuclear genomes are being produced at a rapid rate and are available for search and download from a variety of websites, including GenBank and Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html), which is supported by the U.S. Department of Energy. While the taxonomic representation of entire nuclear genomes is still too sparse for addressing many systematic questions, they often provide useful sources of sequences for preliminary analyses or for using as outgroup taxa.

Other methods of sampling nuclear genomes are becoming common. One of these is to sequence just the transcribed genes, or transcriptomes. A second approach is to cut the nuclear genome into small pieces using restriction enzymes and then to sequence just the ends of the fragments, an approach variously known as RAD-seq (restriction site–associated DNA sequencing) or GBS (genotyping by sequencing). A third possibility is to use methods that enrich the pool of sequences for particular nuclear genes of interest.

Sequence Alignment

Sequence alignment is a critical step in phylogeny reconstruction. Alignment is the stage at which the scientist makes the initial assessment of similarity of nucleotide sites. It is by far the most difficult part of using sequence data, and it is hard to automate. A poor alignment will lead to a meaningless phylogenetic tree.

In a sequence alignment, the columns correspond to nucleotides at homologous positions, and the rows correspond to homologous sequences of DNA (see Figure 2.4). An alignment allows us to track mutations in the DNA over evolutionary time. Although DNA is usually copied faithfully through many rounds of cell division, the cellular machinery makes occasional mistakes—mutations—that get transmitted to subsequent generations. One common mutation is substitution of one base pair for another, also known as a point mutation. Two other common mutations are insertion and deletion, both of which result in sequences of two different lengths. Because we often cannot tell whether the length change is an insertion in one sequence or a deletion in the other one, such length changes are usually called indels (insertion or deletion). Other sorts of mutations include changes in the order of the bases, such as inversion.

Figure 2.4A shows two sequences that are similar at all sites except the position near the middle, a G in sequence 1 and an A in sequence 2. Figure 2.4B shows an-
been many changes to the original algorithm of Needle - has remained the same.

man and Wunsch, the basic idea of alignment scoring entails.

ment score in addition to the three gaps identified in (C).

Figure 2.4

(A) Two DNA sequences, each 27 base pairs long, aligned so that, for the most part, the bases are the same in any one column. This alignment shows a single difference, in red—a G-A substitution. (B) An alignment of the two DNA sequences from (A) but with the bottom sequence shifted one base to the right. This alignment postulates 20 differences between the sequences, in addition to the gaps at either end. (C) Two DNA sequences that differ by two substitutions. One sequence is also three base pairs longer than the other. (D) Another alignment of the two DNA sequences in (C) but one that postulates four gaps in addition to the three gaps identified in (C).

Figure 2.4C

In fact, if gaps are “free,” then you can align anything with anything just by inserting enough gaps. To prevent this nonsensical result, a number called a gap penalty is subtracted from the total score. The alignment in Figure 2.4C would have 22 matches, 2 mismatches, and three gaps. If we decide that we should subtract 3 points from the score for every position with a gap (i.e., a gap penalty of 3), then the total score for the alignment in Figure 2.4C would be 11 (i.e., 22 – 2 – 9), whereas the one in Figure 2.4D would have 22 matches, no mismatches, and 7 gaps, for a total score of 1.

The alignment scoring process builds in assumptions about the underlying mutational process. For most regions of most genomes, we assume that the probability of a base pair substitution is greater than the probability of an insertion or deletion. Therefore, the gap penalty is usually assumed to have a larger absolute value (i.e., a larger number is subtracted from the alignment score) than the cost for a mismatch. Likewise, it is common to assume that a transition (substitution of a purine for a purine, or a pyrimidine for a pyrimidine) is easier and therefore more likely than a transversion (substitution of a purine for a pyrimidine or vice versa).

Many computer programs will produce alignments. For sequences from closely related plants, alignment may not be a serious problem. However, as more distantly related plants are compared, the sequences accumulate length mutations as well as nucleotide substitutions, and alignment becomes more difficult. For genes encoding RNAs, alignment sometimes can be guided by models of the secondary structure of the gene product (the way the molecule folds). In this case the secondary structure is used as a template and the sequence is mapped on it. This method ensures that the proposed alignments maintain the structure of the molecule. (Methods for inferring secondary structure, however, have their own limitations.)

In protein-coding genes, alignments must consider the structure of the protein. The DNA sequence of such genes is read in groups of three bases, or codons, with each codon specifying a particular amino acid. Most commonly, insertions or deletions occur in sets of three bases as well, corresponding to the gain or loss of an amino acid. Alignments need to incorporate this fact. Addition or subtraction of a single base (rather than a set of three) will change the entire structure of the protein encoded by the sequence by changing the start point of the subsequent codons (the reading frame). For example, the sequence AAATGACTTAC codes for the four amino acids lysine-leucine-threonine-tyrosine (K-L-T-Y, using their one-letter abbreviations, as in Figure 2.5). If a single base were lost from the first lysine codon—leaving, for example, AATGACTTAC—the protein would change to an asparagine followed by a stop codon. The stop codon

other alignment of the same two sequences. Intuitively, you can see that the second alignment is not as “good” as the first one, because fewer base pairs match up. How do you tell a computer program to find the alignment with the highest possible number of matches? In a highly influential paper in 1970, Needleman and Wunsch came up with a scoring scheme and an algorithm that would allow a computer to find the best scoring alignment. They gave a certain number of points to a match and subtracted points for a mismatch. For example, if we give 1 point for every match in Figure 2.4A (26 matches) and subtract 1 point for every mismatch (1 in this case), this alignment will have a score of 25, whereas the alignment in Figure 2.4B will have a score of −14 (6 matches minus 20 mismatches), not counting the gaps at either end. Application of the algorithm would indicate that the alignment in Figure 2.4A is preferred because it has a higher alignment score than the one in Figure 2.4B. While there have been many changes to the original algorithm of Needleman and Wunsch, the basic idea of alignment scoring has remained the same.

One problem comes up when sequences are not of identical lengths. For example, in Figure 2.4C, the sequences have 2 mismatches and a 3-base-pair gap. How do you score the gap? One possibility would be for the gap not to count at all. However, if that were the case, then there would be no way to distinguish between the alignment in Figure 2.4C and that in Figure 2.4D. In fact, if gaps are “free,” then you can align anything with anything just by inserting enough gaps. To prevent this nonsensical result, a number called a gap penalty is subtracted from the total score. The alignment in Figure 2.4C would have 22 matches, 2 mismatches, and three gaps. If we decide that we should subtract 3 points from the score for every position with a gap (i.e., a gap penalty of 3), then the total score for the alignment in Figure 2.4C would be 11 (i.e., 22 – 2 – 9), whereas the one in Figure 2.4D would have 22 matches, no mismatches, and 7 gaps, for a total score of 1.

The alignment scoring process builds in assumptions about the underlying mutational process. For most regions of most genomes, we assume that the probability of a base pair substitution is greater than the probability of an insertion or deletion. Therefore, the gap penalty is usually assumed to have a larger absolute value (i.e., a larger number is subtracted from the alignment score) than the cost for a mismatch. Likewise, it is common to assume that a transition (substitution of a purine for a purine, or a pyrimidine for a pyrimidine) is easier and therefore more likely than a transversion (substitution of a purine for a pyrimidine or vice versa).

Many computer programs will produce alignments. For sequences from closely related plants, alignment may not be a serious problem. However, as more distantly related plants are compared, the sequences accumulate length mutations as well as nucleotide substitutions, and alignment becomes more difficult. For genes encoding RNAs, alignment sometimes can be guided by models of the secondary structure of the gene product (the way the molecule folds). In this case the secondary structure is used as a template and the sequence is mapped on it. This method ensures that the proposed alignments maintain the structure of the molecule. (Methods for inferring secondary structure, however, have their own limitations.)

In protein-coding genes, alignments must consider the structure of the protein. The DNA sequence of such genes is read in groups of three bases, or codons, with each codon specifying a particular amino acid. Most commonly, insertions or deletions occur in sets of three bases as well, corresponding to the gain or loss of an amino acid. Alignments need to incorporate this fact. Addition or subtraction of a single base (rather than a set of three) will change the entire structure of the protein encoded by the sequence by changing the start point of the subsequent codons (the reading frame). For example, the sequence AAATGACTTAC codes for the four amino acids lysine-leucine-threonine-tyrosine (K-L-T-Y, using their one-letter abbreviations, as in Figure 2.5). If a single base were lost from the first lysine codon—leaving, for example, AATGACTTAC—the protein would change to an asparagine followed by a stop codon. The stop codon
would prevent the remainder of the protein from being synthesized.

Nucleotide substitutions may or may not affect the protein produced. For example, at position 10 in the alignment shown in Figure 2.5, some taxa have an A and some have a C. This variation in the character state has an effect on the amino acid produced at that position; *Aristida, Stipagrostis, Ertragrostis*, and *Enneapogon* all have glutamine (Q) at that position in rubisco, whereas the other taxa have lysine (K) there. In contrast, the variation at position 24 (G or A) does not change the protein, because both AAG and AAA encode lysine. Note that both position 10 and position 24 provide potentially useful phylogenetic information. In this case, for example an A and some have a C. This variation in the character state has an effect on the amino acid produced at that position; *Aristida, Stipagrostis, Ertragrostis*, and *Enneapogon* all have glutamine (Q) at that position in rubisco, whereas the other taxa have lysine (K) there. In contrast, the variation at position 24 (G or A) does not change the protein, because both AAG and AAA encode lysine. Note that both position 10 and position 24 provide potentially useful phylogenetic characters, even though only one has a biological effect on the protein.

For this discussion, assume that we have produced DNA sequences for a particular gene from four groups of imaginary plants and have aligned them (Figure 2.6A). This particular alignment shows that most of the plants have the same nucleotide at positions 1–7, 9, 10, 12–15, 17, and 18, but they differ at positions 8, 11, and 16 (highlighted in green). These three variable positions then provide the information that allows us to begin to define groups and infer the evolutionary history.

The matrix in Figure 2.6A allows us to infer that at some time in the past there were mutations at three positions. However, the matrix does not provide the direction of the mutation. For example, the ancestral nucleotide at position 8 could have been a G that later mutated to a T, or the T could have been ancestral and later mutated to a G. The matrix can also be represented as a network (Figure 2.6B) showing the relative positions of the three mutations. In Figure 2.6B, all plants designated by the same shape are drawn as though they arose at the same time. This arrangement generally indicates ambiguity; for the purposes of this simplified example, we have not provided any information on their order of evolutionary origin. We define the length of the network by the number of mutations it represents, in this case three. Even though the network looks somewhat like a time line, it is not: it could be read from left to right, from right to left, or perhaps from the middle out-
ward. To turn the network into an evolutionary tree, we must determine which changes are relatively more recent and which occurred further in the past. In other words, the tree must be rooted. Rooting polarizes the character changes, giving them a specific direction.

**Rooting**

Rooting a phylogenetic tree is critical for interpreting how plants evolved, and different rootings suggest different patterns of change (different character polarizations). If you imagine that the network is a piece of string, you can keep the connections exactly the same, even when you pull down a root in different places. The network from Figure 2.6B is redrawn in Figure 2.7 and rooted in three different places. Notice that the length of each tree (or cladogram) is the same as the length of the original network—three mutational steps—and that all the connections are the same but that the order in which the character-change events occur differs considerably.

For example, in the rooting shown in Figure 2.7A, the ancestral plants had G at position 16, position 8, and position 11. In contrast, Figure 2.7B suggests that the ancestral plants had A at position 11, T at position 8, and C at position 16. In Figure 2.7C, the tree is rooted in such a way that the ancestor must have had C at position 16 and G at position 11. However, notice that we can’t tell what the base was at position 8. It could have been a G and then mutated to a T in the ancestor of the squares + diamonds, or it could have been a T that mutated to a G in the ancestor of the circles + stars.

How is the position of the root determined? One common suggestion is to use fossils, but this approach is problematic. The fact that an extinct plant has been fossilized does not mean that its lineage originated earlier than those of plants now living; we know only that it died out earlier. In determining evolutionary history, we are interested in determining when lineages diverged from one another (i.e., when taxa originated). When taxa die out is interesting to know, but that fact by itself does not help in determining their origins. Fossils are, of course, extremely useful when included as additional taxa in a phylogeny. They often have combinations of character states that no longer occur in extant taxa and can affect the overall structure of the tree, sometimes in surprising and informative ways.

In general, evolutionary trees are rooted using a relative of the group under study: an outgroup. When selecting an outgroup, one must assume only that all ingroup members (members of the group under study) are more closely related to one another than to the outgroup; in other words, the outgroup must have separated from the ingroup lineage before the ingroup diversified. Often several outgroups are used. If an outgroup is added to a network, the point at which it attaches is determined as the root of the tree.

In Figure 2.8A, an imaginary outgroup, a yellow polygon, is added to the matrix from Figure 2.6A. For the particular gene sequence in the matrix, the yellow polygon has the same set of nucleotides as the star plants. With this information, the polygon can be added to the network as an outgroup, as in Figure 2.8B. Because the polygon attaches among the star species, the tree can be rooted and redrawn as in Figure 2.8C. This tree corresponds to the
Methods and Principles of Biological Systematics

rooted tree in Figure 2.7A and strengthens the hypothesis that Figure 2.7A accurately reflects evolutionary history. With this tree we can illustrate the same point made in Figure 2.1—the tree can be drawn in different ways and still reflect the same evolutionary history. Comparing Figures 2.9A and B with Figure 2.8C shows that we can rotate the branches of the tree around any one of the nodes without affecting the inferred order of events.

With a rooted tree (and only with a rooted tree), we can determine which groups are monophyletic (made up of an ancestor and all of its descendants) and form a clade. In the example laid out in Figure 2.8C, the diamond plants are monophyletic. The evidence for their monophyly is provided by the shared derived (synapomorphic) A at position 11. But notice that the interpretation of monophyly depends on rooting. If Figure 2.7B were the correct rooting, then A at position 11, T at 8, and C at 16 would be the ancestral character states (usually called symplesiomorphies, or shared ancestral states) rather than derived synapomorphies. In Figure 2.7B, the species indicated by diamonds and squares do not share any derived character and do not include all the descendants of their common ancestor; some of those descendants went on to become the circle and the star plants. Therefore, if Figure 2.7B were correct, the diamond plus square species would not be a monophyletic group (as they are with the rooting in Figure 2.7A). Instead, they would constitute a paraphyletic group, one that includes a common ancestor and some, but not all, of its descendants.

A character state that is derived at one time may become ancestral later. In Figure 2.8C, C at position 16 is a shared derived character of a large group. It is a synapomorphy and indicates monophyly of the circles + squares + diamonds. For the clade of squares + diamonds, though, C at position 16 is an ancestral, or plesiomorphic, character: a character that all of the species in the group inherited from their common ancestor and thus indicates nothing about their relationships with one another. Plesiomorphic character states cannot show evolutionary relationships in the group being studied because they evolved earlier than any of the taxa being compared and have merely been retained in the group’s various lineages.

Homoplasy

As the preceding discussions have shown, determining the evolutionary history of a group of organisms follows a series of standard steps. First, DNA from multiple plants is sequenced and recorded as an alignment (see Figure 2.6A). Second, a branching network is constructed (see Figure 2.6B). Third, by inclusion of an outgroup, the network can be rooted to produce an evolutionary tree, cladogram, or phylogeny (see Figure 2.8C).
Two phenomena, however, make it much harder to determine evolutionary history in practice: parallelism and reversal, which are sometimes referred to together as homoplasy. Parallelism is the appearance of similar character states in unrelated organisms. (Many authors make a distinction between parallelism and convergence, but for this discussion we will treat them as though they were the same.) A reversal occurs when a derived character state changes back to the ancestral state.

To provide a clear example, we will divide the group that we have called star plants into white star plants, red star plants, and gold star plants. Let us assume that the red star and white star plants have an A at position 3, whereas all the rest of the plants have a C (including the polygon plants outgroup). Let us further assume that the white star plants have a T at position 8. We can revise the alignment in Figure 2.8A to create Figure 2.10A, which gives the same information as the network in Figure 2.10B.

**Figure 2.9** Two different ways to draw the tree in Figure 2.8C. Note that the length does not change, nor does the hypothesized order of events.

2.10B. Counting the number of changes on this network (its length), we find five: one each at positions 3, 11, and 16 and two at position 8. According to this network, there have been two parallel mutations at position 8.

If we were to look solely at position 8 and create groups based on the bases at that position, the plants with a T would be considered polyphyletic. Polyphyletic groups have two or more ancestral lineages in which the parallel character states evolved. (Although we distinguish here between paraphyletic and polyphyletic groups, many systematists have observed that the difference is slight and simply call any para- or polyphyletic group nonmonophyletic.) Why not draw the network in such a way that there is only one change from G to T? Such a network is shown in Figure 2.10C. Now we have one change at position 8, but that change then requires two changes at positions 3 and 16, making the network six steps long. There is no way to draw the network without at least one character changing more than once.

Each of the networks can be converted into a phylogeny by rooting at the polygon, but each makes different suggestions about how the plants have evolved. In Figure 2.10B, positions 3 and 16 have been stable over evolutionary time, whereas there have been two independent mutations at position 8. In Figure 2.10C, we postulate that positions 3 and 16 have mutated twice over evolutionary time, while position 8 has changed only once. By drawing either of these networks, we are proposing a hypothesis about how evolution has happened—about which genetic changes have occurred, at what frequency, and in which order. You can see that in fact you could construct quite a few different trees from these data, all of them producing different ideas about evolution. How do we determine which one is correct? There is no way to be certain. No one was there to watch the evolution of these plants. We can, however, make educated guesses, or inferences, and some such inferences seem more likely than others to be correct.

There are two general ways to infer a phylogenetic tree. One way is to use a clustering algorithm such as neighbor joining. An algorithm is a series of repetitive steps that allow a large problem to be broken into a set of smaller problems. A second general way to produce a tree is to use an optimality criterion (plural, criteria), a specific rule that helps you determine which tree is best. Three such criteria are in common use. First is the criterion of tree length, in which shorter, or more parsimonious, trees are preferred to longer ones. Second is the maximum likelihood criterion, in which the preferred tree is the one that maximizes the likelihood of the data, given the tree. Third is the Bayesian posterior probability, in which the preferred tree is the one with highest posterior probability. We will discuss algorithms first and will discuss the optimality options in Methods That Use an Optimality Criterion (see page 27).
Tree Building:
Clustering Algorithms and Neighbor Joining

There are many different clustering algorithms, but most share a number of features. In general, they start with a dissimilarity, or distance, matrix. A dissimilarity matrix for the sequences in Figure 2.10A is shown in Figure 2.11A. While it is possible to record the dissimilarity of two sequences as the numbers of mutations separating them, as shown above the diagonal of blank boxes, dissimilarity is more commonly normalized by length of sequence. In this case, the numbers of mutations distinguishing two sequences are divided by 18, which is the total sequence length, as step 1. Step 2 is to join the two least dissimilar sequences to form one node with two branches. Step 3 is to remove those two from the matrix and replace them with a single value that represents them. Step 4 is to recompute the matrix. These steps are repeated until all sequences have been joined.

The most commonly used algorithm for constructing phylogenetic trees is the neighbor-joining algorithm. While a full description of this method is beyond the scope of this book, a hallmark of the algorithm is that it accounts for the possibility that the rate of molecular evolution may be different in different sequences. While it starts with a dissimilarity matrix, the matrix is first corrected to account for differences in the mutation rate. The steps are then repeated until a full tree is produced.

The result is a tree in which the lengths of the branches are distances, rather than particular mutations (Figure 2.11B). Another valuable property of the neighbor-joining algorithm is that the sequences that are most similar may not necessarily be sisters in the tree. There are two possible reasons why two sequences might be similar to each other. First, and most intuitive, is the possibility that
the two are indeed very closely related and share a recent common ancestor. The second possibility is that the two are not particularly closely related but that they just have not changed very much over time. An effective tree-building method must distinguish between these two possibilities, and neighbor joining is able to make such a distinction. For example, in Figure 2.11A the red star and green circle plants are not very dissimilar (0.111, or 2 mutations apart), which might make you think that they could be closely related. However, looking at Figure 2.11B reveals that they are on different branches entirely but simply have not changed much since the root of the tree. An older method that was once widely used, the unweighted pair group method with averaging, or UPGMA, is unable to distinguish a true relationship and a slow rate of change and is thus not appropriate for phylogenetic reconstruction.

In summary, neighbor joining is a very fast algorithmic method that produces large trees in a fraction of a second. In addition, it considers the possibility that different sequences may mutate at different rates. Another useful characteristic of neighbor joining is that it can accommodate different models of evolution, described in the next section.

**Models of Evolution**

In the description of the neighbor-joining method and in the examples earlier in this chapter, we have been proceeding as though we could see every mutation that ever happened. This, of course, is nonsense: DNA has existed for several billion years and has been mutating the entire time. Because of this long invisible history, we can be certain that many mutations have occurred that are no longer detectable. This is particularly true if the time since divergence of sequences is long, or if the mutation rate is high.

How can we address the problem of unseen mutations? Start with the assumption that mutation is a random process. To envision this, begin by thinking of raindrops falling on dry pavement. When there are not many drops, you will be able to see where each one has fallen, and they will make a random pattern. You can count them all, probably fairly accurately. However, as raindrops continue to fall, sooner or later a drop will land right on top of a previous drop, and if this happens very often, you will quickly lose count of exactly how many drops have fallen at a given spot. If you now try to use the number of spots to estimate the number of raindrops, your estimate will be too low because you will miss all the ones that have fallen directly on top of each other.

Like the pattern of random raindrops, it is easy to locate the position of mutations in a DNA sequence if there are not very many mutations (Figure 2.12A). However, given enough mutations, sooner or later a base pair that has mutated once will mutate again. At this point you lose any evidence that the original mutation occurred. For example, a base that was originally A could mutate first to G and then to T. Unless you have an individual plant with the intermediate state (G), you will have no way of knowing that it ever occurred; you will only see the A and T. Your estimate of the number of mutations, and hence the distance, will be too low. This pattern of multiple mutations at the same site is sometimes called multiple hits. This can be diagrammed by graphing the dissimilarity between the two sequences against time (Figure 2.12B).
The actual number of mutations increases without limit over time, but the observed number accumulates more slowly and eventually approaches an asymptote at 75% distance. Because there are only four base pairs, there is a 1/4 (25%) chance that two base pairs will match purely at random. Thus two random sequences will always be approximately 25% similar and 75% dissimilar. At 75% dissimilarity, the actual divergence cannot be distinguished from random noise.

Basic probability theory can be used to calculate the probability of multiple hits for a given pair of sequences, and hence can come up with a corrected measure of dissimilarity that approximates the actual value. Such an estimate of the probability of multiple hits is known as a model of evolution. In other words, if one observes a pairwise dissimilarity between two sequences of, say, 1%, what is the probability that there have been additional invisible mutations? Inspection of the graph in Figure 2.12B suggests that at 1% dissimilarity, the actual value and the observed value are very similar. Conversely, if one observes a value of 30% dissimilarity, one can expect to have missed quite a few mutations, and the actual value is probably a good deal higher than 30%. This can also be seen in Figure 2.12C, which shows a sequence sampled at three different times. If we look at time 2, we can see that between time 1 and time 2, the sequence underwent four mutations—C to T, G to A, T to C, and another G to A. Between times 2 and 3, the first mutated position changed from T to G, the second and fourth positions did not change, and the third position changed from C back to T. If the sequence at time 2 were not observed, three mutations would be “invisible” and we would think that there had only been three rather than six changes.

A model of evolution is thus a correction factor to estimate the real dissimilarity between two sequences. In the context of a clustering algorithm such as neighbor joining, the pairwise distance from the original sequences can be adjusted according to a model of evolution and then the algorithm can be applied. For example, suppose one has two sequences that differ at 20% of their sites. One assumes that mutation is random, that nucleotide frequencies are equal, that all substitutions are equally probable, and that all sites are free to vary. (These are the assumptions of the Jukes-Cantor model, named after the authors who first proposed it.) One can then calculate what the actual percentage of changes must have been to have an observed dissimilarity of 20%. The actual percentage of changes is the corrected dissimilarity, or branch length. The assumptions of the Jukes-Cantor model are oversimplified, and more realistic models have been developed, but all serve the same general purpose.

What happens if you don’t correct for multiple mutations at the same site in a molecule? If all sequences are fairly similar, i.e., if you are at a point near the origin in the graph in Figure 2.12B, then the observed dissimilarity is not far from the actual dissimilarity and correcting for multiple mutations does not make much difference. However, if the dissimilarities between sequences are large and, in particular, very unequal, uncorrected distances will produce a particular tree topology, whether it is correct or not. This can be seen Figure 2.13. In this figure, we consider only four sequences, and we see in Figure 2.13A that sequences 1 and 4 have accumulated more mutations than sequences 2 and 3 and, furthermore, that there are not many mutations along the internal branch. Such a pattern can be caused by a large difference in the rate of mutation. Because sequences 1 and 4 have accumulated
more mutations, the chances of their having multiple hits goes up. Because there are only four nucleotides, some of those mutations will lead to identical bases in both sequences, purely by chance. In this example, note that positions 2, 9, and 11 are identical in sequences 1 and 4 but that these similarities come from independent mutations; that is, sequences 1 and 4 are different from their ancestral sequences at those positions. The rapidly changing sequences can thus appear to be closely related even if they are not.

An uncorrected neighbor-joining analysis will infer the tree in Figure 2.13B instead of the real tree (in Figure 2.13A). Note that sequences 1 and 4 are now placed together and sequences 2 and 3 are together. In other words, the algorithm will generate the wrong tree. In this case, the long branches are said to “attract” each other, and the phenomenon is known as long-branch attraction. Even worse, the more data you collect, the more correct the wrong tree will appear to be; that is, you will become increasingly confident that the wrong answer is correct. While this example is artificial, such trees are surprisingly common in the study of plant evolution. All the problems with long branches are worse if the internal branches of the tree are short. This issue appears often with rapid radiations of species in which there was a burst of speciation and little time for mutations to accumulate, followed by a longer time in which lineages diverged.

A version of the long-branch problem can also appear with individual long branches, which have accumulated many mutations relative to other sequences in the tree. Random similarities to various other sequences can lead to many placements that are almost equally good. Such branches appear to “wander” in trees. Because the branch leading to the outgroup is often the longest in the tree, the position of the root itself may be uncertain because of the long branch.

Sometimes the long-branch problem can be solved and sometimes it can’t. While model-based methods mitigate this problem, they do not eliminate it entirely. If there are close relatives to the species on the long branch, these can be included and will have the effect of breaking up the branch. However, in some cases (e.g., the root of the angiosperm tree) none of the extant relatives are particularly recent. Thus some problems are inherently difficult and possibly not solvable by current methods or data. We will return to this unfortunate property of trees when we talk about parsimony algorithms on pages 27–28.

Should You Believe the Tree: Bootstrapping and Comparison of Trees

An evolutionary tree is simply a model or hypothesis, a best guess about the history of a group of plants. It follows that some guesses might be better, or at least more convincing, than others. A computer will produce a tree for any data you give it. How can you tell whether the tree is a good reflection of the data? (Note that you have no way of knowing whether the tree is accurate or not—in most cases the phylogeny is unknown and must simply be estimated.)

One way to assess the strength of the tree would be to randomize the entire data set, construct a new tree, repeat this many times, and determine whether the tree you have produced is significantly different from one of the random trees. Would this be an interesting result? Probably not. Generally we are interested in saying something more specific than simply that our data are nonrandom. Usually instead we are interested in assessing whether the tree is a good reflection of the data, or if it would be
produced by a similar sample of data. We are also generally interested in knowing how well the data support particular parts of the tree: we want a way to assess support for particular groups, or clades.

One very common statistical method of resampling is bootstrap analysis, which randomizes columns in the matrix with respect to taxa. To create a bootstrapped data set, start with the original alignment. As an example, begin with the matrix in Figure 2.10A, and choose one position (column) at random. Make that column the first position of a new alignment. Go back to the original alignment and choose another column at random. Make that the second position of the new alignment. Continue until the new alignment is the same length as the original one. Because you are returning to the original matrix each time to choose a new column, some characters may be represented several times in each new matrix, while others may be omitted entirely. In other words, sample the original alignment with replacement. Figure 2.14 shows the new matrix constructed from sampling randomly with replacement from the matrix in Figure 2.10A; note that character 17 from the original matrix has been selected four times, whereas several characters (e.g., 5 and 6) have been missed by the random selection process.

Once a bootstrapped data matrix is created, it can be analyzed by neighbor joining or by the parsimony or maximum likelihood methods described below. A hundred or more bootstrapped matrices are created and analyzed, producing a set of at least 100 phylogenetic trees. The large set of trees is then summarized by a majority-rule consensus tree. If a particular clade is present in the majority (more than 50%) of the bootstrap trees, then this clade is included in the majority-rule consensus tree, along with an indication of the bootstrap value, that is, the percentage of bootstrap trees showing that clade. For example, Figure 2.15 shows a recent phylogeny of the Lamiales (Refulio-Rodriguez and Olmstead 2014). Above each branch are three numbers, the second and third of which are bootstrap values. Higher values are better, so nodes that appear in 100% of the bootstrap trees, such as those at the deeper nodes of this particular tree, are more strongly supported by the data than nodes with lower values. It is important to keep in mind that bootstrap values are not conventional probability values, so a bootstrap value of 85% does not indicate that the clade has an 85% chance of being correct. It merely provides a rough gauge of how well the data support a given result.

**Methods That Use an Optimality Criterion**

We have talked a great deal about one common method of tree construction, the neighbor-joining method. This is an algorithmic method, in which the data are simply fed through a series of repetitive computational steps and a single tree appears at the end. This approach is limited, however, because it provides no way to consider trees that may be nearly as good as the one you arrived at. As in all aspects of science, it is important to consider alternatives to your results. In this case, it is often helpful to have a number by which you can rank trees as the best, second best, third best, etc. Also, many systematists wish to develop a tree that is consistent with a particular statistical or philosophical criterion. Such trees are said to use optimality criteria. The most common of these are parsimony, maximum likelihood, and Bayesian approaches.

**Parsimony** In parsimony analyses, the optimality criterion is tree length, or number of mutational steps. Specifically, the shortest tree for the data is preferred. The parsimony method asks, What is the simplest explanation of the observations? By asking this question, we apply a rule that is used throughout science, known as Occam’s razor: Do not generate a hypothesis any more complex than is demanded by the data. Applying this principle of simplicity, or parsimony, leads us to prefer the shorter network. The fact that it is shorter does not make it correct, but it is the simplest explanation of the data. The network in Figure 2.6 is in fact a parsimony network—the shortest network that can be constructed for those data. The fact that we prefer the network in Figure 2.10B over the one in Figure 2.10C is also based on the principle of parsimony.
In the example in Figure 2.10, in which there are few characters and little homoplasy, it is easy to construct the shortest network that can link the organisms. In most real cases, however, many networks are possible, and it is not immediately obvious which one is the shortest. Computer algorithms compare thousands of trees, calculate their lengths, and determine which is shortest. If the taxa are numerous, only heuristic algorithms can be used. These algorithms may not succeed in finding the shortest tree or trees, because of the large number of possible trees.

Branch lengths in parsimony are recorded as the total number of substitutions, and they are not normalized according to the length of the matrix (generally a DNA sequence). Because of this lack of normalization it is difficult to assess whether branch lengths are particularly short or particularly long. (In other words, if I see a clade supported by a branch length of three mutations, should I be impressed or concerned?) Sites that appear to be the same in all sequences (invariant sites) are considered to be *parsimony uninformative* and are simply ignored. Note that this is quite different from the way model-based methods such as maximum likelihood and Bayesian analyses (see below) treat such sites.

Tree length does not provide any method for considering multiple mutations at the same site. For some proponents of this method, this feature is a virtue because it uses only data that can be observed directly and does not rely on inferences. For detractors, however, the inability to consider multiple mutations (a model of evolution) means that the method ignores much of what we know about the mutational process. All users of this method know that it is particularly susceptible to the problem of long-branch attraction. Because long-branch attraction is especially problematic with DNA sequence data, it is becoming less and less common to use parsimony approaches. Conversely, parsimony is particularly suited to morphological data for which models of evolution are hard to apply. Thus parsimony methods are still valuable when incorporating fossils into phylogenetic analyses.

The parsimony method has been widely used, is easily applicable to morphological changes, and is also possibly the most intuitive of tree reconstruction methods. Parsimony methods work well when evolutionary rates are slow enough that chance similarities (due to the evolution of identical derived character states independently in two or more lineages) do not overwhelm character states shared by the common ancestor. At higher rates of change, however, parsimony methods are particularly susceptible to long-branch problems, described above for uncorrected neighbor-joining analyses. Support for parsi-
It is possible to build a phylogeny from morphological characters, although it is rarely done now that DNA data are so common and easy to obtain. However, DNA data are not always available, either because the plant is represented by a single herbarium specimen with limited material, or because the plant is fossilized. In such cases, a phylogeny can still be constructed. This is usually done using the parsimony method (see page 27) because of the difficulty of constructing a model of evolution for most morphological characters.

The basic principles are the same as those outlined in the main text, but instead of starting with an alignment of DNA sequences, the analysis starts with a matrix of morphological characters, in which the rows correspond to individual plants or groups of plants and the columns correspond to characters. Individual cells in the matrix then carry information on character states. When we describe the variation among similar morphological structures by dividing the character into character states, we are in effect putting forward a hypothesis of underlying genetic control, even though the assumption is rarely stated in these terms. For example, if two species differ in the color of their flowers, we may score the character petal color as having two states, red and blue. By scoring it this way, we are hypothesizing that the genes underlying petal color have switched, over evolutionary time, to produce either red flowers from a blue-flowered ancestor or blue flowers from a red-flowered ancestor. In this instance, we know that there are in fact genes (e.g., components of the anthocyanin pathway) that control petal color, and thus the inference of two states controlled by a “genetic switch” is probably a reasonable one.

In many cases we have no idea of the genetic mechanisms that control the character states observed. In proposing hypotheses about the nature of the underlying switches, often all we can do is be sure that the character states really are distinct. For quantitative characters such as leaf length or corolla tube width, this means graphing the quantitative data (i.e., the measurements) to be sure that the measurements of the species we are studying do not overlap. If the measurements do overlap, then the assumption of underlying genetic switches—and therefore division into character states—is unsupported by any evidence. In these cases, the characters in question should be omitted from phylogenetic analysis unless the overlap is caused by only a few individuals, in which case the character could be scored as polymorphic for that species and retained in the analysis. Even though such overlapping characters might reflect genetic changes over evolutionary time, overlap makes it difficult, given our current state of knowledge, to extract any reliable information on the underlying gene changes (although methods of dealing with plants with variable characters have been developed).

It may be harder than you think to determine which structures of one plant can usefully be compared with structures of a different plant. Two structures may be deemed to be similar if (1) they are found in a similar position in both organisms, (2) they are similar in their cellular and histological structure, and/or (3) they are linked by intermediate forms of the structure (either by intermediates at different developmental stages of the same organism or by intermediates in different organisms). These three statements constitute Remane’s criteria of similarity.

Remane (1952) actually called this list the “criteria of homology.” In this book, however, we use the term homology in a more restricted sense, to mean identity by descent. In other words, if we say that a character is homologous among a group of species, we mean that all those species inherited that character from a common ancestor. Under this definition, observing similarity is only the first step in determining homology, since not every observed similarity is the result of homology.* (For example, structural similarities can evolve independently in unrelated plants that live in similar environments.) This text follows the viewpoint held by the many phylogenetic systematists who argue that homology can be determined only by constructing an evolutionary tree.

Morphological characters may be divided into two states: binary (having two states) or multistate (having three or more states). Binary characters are interpreted as representing a single genetic switch—“on” producing one state (e.g., tricolpate pollen), and “off” resulting in the other state (e.g., one-grooved, or monosulcate, pollen). Over evolutionary time, of course, such characters can continue to change. For example, tricolpate pollen is modified in some Caryophyllales, where it is spherical with many pores evenly spaced around it (e.g., pantoporate, looking rather like a golf ball). If we were to include the character pollen colpi in a matrix containing some taxa with pantoporate pollen, then pollen colpi would have three states (monosulcate, tricolpate, and pantoporate), making it a multistate character. Multistate characters raise a difficult question: How many genetic switches are involved?

It is possible that monosulcate pollen changed to tricolpate pollen, which then changed to pantoporate pollen (Figure 2.16A). This scenario implies two genetic switches that must have occurred in order; that is, pantoporate pollen could arise only after tricolpate pollen did. If we accept this series of events, the multistate character is considered ordered. If we decide to allow for reversals of character states—that is, if we consider

---

*You should be aware that homology has several different meanings, and when you read the literature, it is worth checking what particular authors mean when they use the term.
Chapter 2

Parsimony trees are generally assessed by bootstrapping. In the tree in Figure 2.15, the rightmost number at each branch is the parsimony bootstrap value.

In summary, parsimony is a mutation-counting method that uses the optimality criterion of tree length (numbers of mutations, or steps). Branch lengths are numbers of mutations rather than mutations per length of sequence. Branch lengths cannot be corrected for multiple hits, making parsimony particularly prone to artifacts created by long branches. Parsimony is used less frequently for nucleotide data but remains the most commonly used method for morphological data (Box 2A). It also underlies

BUILDING A PHYLOGENY FROM MORPHOLOGICAL CHARACTERS (continued)

Consider the possibility that pantoporate pollen might switch back to tricolpate and tricolpate back to monosulcate—the character is still ordered. It requires two evolutionary (genetic) steps to go from monosulcate to pantoporate pollen, and two steps to go from pantoporate to monosulcate pollen.

If we didn’t know anything about the plants involved, we might want to consider the possibility that monosulcate pollen changed to tricolpate pollen and, in an independent event, monosulcate pollen also changed to pantoporate pollen (Figure 2.16B). This sequence would suggest that there is a genetic switch that allows change from monosulcate to tricolpate pollen, as well as a switch that allows change from monosulcate to pantoporate pollen, but that a change from tricolpate to pantoporate pollen is impossible. The character in this case is still ordered, but in a different way than in Figure 2.16A. If reversals are possible, then two steps are required to get from tricolpate to pantoporate pollen, and two from pantoporate to tricolpate pollen.

With morphological characters and character states, we are usually unsure of which switches are possible, so it is common to treat multistate characters as unordered (Figure 2.16C). This method is sometimes called Fitch parsimony. In the case of an unordered character, we postulate only one switch between any two states. Note, for comparison, that DNA sequence characters are multistate characters with four states (A, T, G, C). To treat these as ordered would be nonsensical; adenine does not need to change to cytosine before changing to guanine. DNA characters are always treated as unordered and fully reversible.

In parsimony analyses of morphological characters, it is common to assume that each character state change represents a single genetic switch and that all state changes are equally likely. Although this approach sometimes is described as “unweighted parsimony,” in fact it assumes that all characters are equally likely to change and weights them accordingly. Mechanisms exist to give added weight to particular changes or particular directions of change, based on prior knowledge of the character. For example, it is possible to build into a morphological analysis the assumption that presence or absence of trichomes may be more likely to change over evolutionary time than numbers of anthers.

Also implicit in parsimony analyses of morphology is the assumption that all characters of organisms evolve independently, that is, that change in one character does not increase the probability of change in another character. This assumption may be violated frequently; for example, a change in flower color might well lead to a shift in pollinators, which would then increase the probability that corolla shape would change. Violations of this assumption obviously affect character weighting, in that the likelihood of change for any two characters is not the same.
In current methods of phylogenetic analysis, a network is constructed and is then polarized by determining where the outgroup attaches. However, in the original concept of phylogenetic analysis proposed by Willi Hennig (see Chapter 3), characters were polarized first using one or several outgroups, and the phylogeny was constructed afterward.

Consider, for example, the imaginary plants presented in Figure 2.17 and their imaginary morphological character states listed in Table 2.2. In Table 2.2, the character states of the outgroup are assumed to be ancestral (plesiomorphic) and are represented by a 0, while the derived (apomorphic) states are represented by a 1. These character states are then used to produce a character × taxon matrix (Table 2.3). Next, a phylogenetic tree is constructed in which the taxa are grouped according to evidence provided by shared derived character states (synapomorphies). The presence of a derived character state in two taxa suggests that they share a unique common ancestor in which the apomorphy first evolved; the two taxa are assumed to have inherited the apomorphy from this ancestor.

The phylogenetic tree, then, represents the simplest hypothesis that explains the pattern of derived character states, following the principle of parsimony.

A hypothesis of evolutionary relationships for species I, II, and III of Figure 2.17 is presented in Figure 2.18. Species II and III are hypothesized to share a unique common ancestor because they share the derived states of characters 3 and 5, that is, they both have opposite, petiolate leaves (see Table 2.2). These are hypothesized to have evolved in their common ancestor. Similarly, the shared possession of solitary flowers (character 9) supports the recognition of a more inclusive monophyletic group containing species I, II, and III.

The presence of pubescent (hairy) stems in species I and III is homoplastic; that is, hairy stems are hypothesized to have evolved in parallel in these two species, so their similarity is not based on common ancestry. Note, however, that hairy stems could have evolved in the most recent common ancestor of all three species.
an older method of phylogeny reconstruction known as Hennigian argumentation (Box 2B).

**Maximum likelihood**  Maximum likelihood is a powerful approach used in many aspects of statistics, including construction of trees, and is based on a model of the probability of change from one character state (in this case) to another. This model is used to calculate the likelihood that a given branching diagram will lead to the particular set of data observed. The best tree is the one that maximizes the likelihood of the data, given the tree topology, a set of branch lengths, and a model of evolution, so this model is known as the maximum likelihood method. In other words, the method starts with a tree and then asks, If this were the real tree, what is the probability that it would have produced my set of data?

For example, consider the tree in Figure 2.19. This is a tree topology and a set of arbitrarily chosen branch lengths. The terminals represent position 8 in the matrix from Figure 2.6. The likelihood calculation then proceeds as follows: Suppose the base at node X is A and the one at node Y is also A. Given the particular set of branch lengths and assuming that mutation is proceeding under a Jukes-Cantor model, compute the probability that that A will change to T in the diamond plants, T in the square plants, G in the circle plants, and G in the star plants. Then compute the probability for the case where the base at node X is A but the one at Y is G; record that

![Figure 2.19](image)

**TABLE 2.3** Character × taxon matrix for the three hypothetical species of Figure 2.17, based on characters in Table 2.2

<table>
<thead>
<tr>
<th>Characters</th>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species I</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Species II</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Species III</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Outgroup(s)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 2.18** A phylogenetic tree illustrating hypothesized evolutionary relationships of the three imaginary species of Figure 2.17. Synapomorphies are indicated on the tree by the numbered green lines.

**FIGURE 2.19** Maximum likelihood computation for position 8 from the matrix in Figure 2.6. The goal is to compute the likelihood of observing the bases shown, given the branch lengths shown, testing for all possible combinations of bases at positions X and Y. This computation is then repeated for all sites in the alignment and for various sets of branch lengths.
probability. Continue until you have done the probability calculations for all possible combinations of nucleotides at nodes X and Y, and add up the probabilities of all possible combinations. Then do a similar calculation for every site in the alignment. Then change the branch length and repeat the computations. The product of all these very small probabilities is a very tiny number, beginning with a decimal point followed by several thousand zeros. Likelihoods are thus generally expressed as logarithms or negative logarithms, and the highest (maximum) likelihood is the one with the smallest absolute value.

Note that you can estimate the probability of change even for sites that appear to be the same for all sequences (e.g., position 1 in the matrix in Figure 2.6). There is a probability that these sites have not changed at all from their common ancestor, but there is also some probability that “invariant sites” have simply changed to the same state but from different ancestral nucleotides. For example, the nucleotide at ancestor X could have been C and changed to A independently in both the diamond and square plants. Or the nucleotide at X could have been A, changed to something else, and then changed back to their original state (e.g., A to C to A). The probability of such invisible changes gets higher as the branch lengths get longer. This is different from the way that parsimony methods treat invariant sites; in parsimony methods such sites are assumed to be the same in the ancestors and descendants, and are assumed not to have changed over evolutionary time.

Assessment of support for a maximum likelihood tree is done via bootstrapping as in neighbor joining and parsimony. For many years maximum likelihood algorithms were so slow that this was very hard to do, but rapid algorithms now make bootstrapping a maximum likelihood tree possible and indeed routine. In the tree in Figure 2.15, the middle number on each branch is the maximum likelihood bootstrap value. Because maximum likelihood uses a model of evolution, it is much less susceptible than parsimony or uncorrected neighbor joining to long-branch attraction. However, it has a weak tendency to fall prey to the opposite problem, long-branch repulsion. In this case, if the long branches really are sisters (e.g., if there is a speedup of the evolutionary rate of a clade), maximum likelihood can tend to force them apart.

In summary, the maximum likelihood method uses likelihood as its optimality criterion. The preferred tree is the one that maximizes the likelihood of producing the observed data, given a particular model of evolution. Branch support is generally evaluated by bootstrapping. While it is less susceptible than uncorrected neighbor-joining or parsimony methods to artifacts caused by long branches, it is not wholly immune. It requires more computational power than neighbor joining but is still reasonably fast for moderate-sized data sets (several hundred sequences).

**Bayesian methods** Bayesian statistics is a fundamentally different branch of statistics from the more familiar frequentist statistics that are taught in common university courses. The approach dates to the eighteenth century and is distinguished from frequentist approaches because it explicitly takes into account prior information. In other words, rather than assuming that nothing is known about a system, it explicitly incorporates knowledge accrued before the data analysis in question.

In the context of tree building, Bayesian analyses are model based, like maximum likelihood methods. The optimality criterion is the **posterior probability**; the preferred tree is the one with the greatest posterior probability. In contrast to the likelihood, which is the probability of the data given the tree, the posterior probability is the probability of the tree given the data, so it represents a more intuitive measure of probability. According to Bayes’ theorem, the posterior probability is calculated as the likelihood multiplied by the prior probability (or just “prior,” for short) that the hypothesis (in this case the tree) is true; that product is then divided by a constant. The center of disagreements between frequentist and Bayesian statisticians is, first, whether a prior should be incorporated at all and, second, how that should be done.

We have already discussed how likelihoods are computed. How do we assign prior probabilities to a tree or set of trees? Two approaches are in common use. The first (sometimes called “flat priors”) is to assume that the prior probabilities of all possible trees are the same. In this case, the posterior probability is simply the likelihood multiplied by a constant, so the Bayesian tree is the same as the likelihood tree, although the interpretation of the two is different. A second approach is to permit the priors of all trees to be almost but not quite the same; this is the approach implemented in the program MrBayes (http://mrbayes.sourceforge.net/index.php; Ronquist et al. 2012), which is the most common program used to compute Bayesian trees.

Rather than try to identify a single best tree, as maximum likelihood methods do, Bayesian analyses examine a set of “credible trees.” Within that set, the frequency of each clade is then determined. This means that the estimate of support is built into the analysis, rather than determined afterward by bootstrapping.

The steps in a Bayesian analysis are as follows: First, assign priors. Sample tree space for each tree, computing the likelihood and multiplying by the prior. Then, when you have sampled enough trees (see the next paragraph for how to know how many is enough), remove the suboptimal trees that were computed at the start. Finally, calculate a consensus tree. The result is generally a topology quite similar to that from a maximum likelihood analysis with an assessment of support.

Although the steps in a Bayesian phylogenetic analysis are conceptually straightforward, the tree search methods are computationally intractable. Thus a Bayesian search is carried out using a Markov chain Monte Carlo (MCMC) method. MCMC is a “robot” with a series of rules for
searching through many combinations of branch lengths, tree topologies, and characters at ancestral nodes. Effectively, the computer algorithm wanders around in tree space computing posterior probabilities as it goes. The algorithm (1) starts with a tree; (2) modifies all parameters a little bit and evaluates the tree again; (3) following some preset rules, accepts or rejects the new tree. Unlike tree evaluation in maximum likelihood and parsimony methods, the new tree may or may not be better than the previous ones. Periodically all the parameters are changed substantially in an effort to prevent the algorithm from getting stuck in one small set of trees.

Each new tree is called a generation. The first trees computed are not yet optimized so are usually discarded. After these initial trees, posterior values generally settle down and vary around a mean. Once the values have settled down, it is common to sample every 1000 trees for several million generations. Finally all the sampled trees are combined into a majority-rule consensus tree, as would be done for a bootstrap analysis. The value on any given branch is the posterior probability.

How are Bayesian posterior probability values compared with bootstrap proportions? Because the credible set of Bayesian trees is not constructed or assessed in the same way as a set of bootstrap trees, it is not surprising that the two sorts of values cannot be directly compared; they are simply measuring different things. A general observation is that Bayesian posterior probabilities are higher than maximum likelihood bootstrap values for the same branches on trees. Recall that bootstrap analysis resamples the data matrix, whereas Bayesian analysis does not. Therefore a node supported by only one or two mutations will necessarily have a low bootstrap value, because those mutations will be missed by the resampling procedure some of the time. The Bayesian analysis will not be misled in the same way. A comparison between Bayesian posterior probabilities and bootstrap values can be seen in the Lamiales tree in Figure 2.15, where the posterior probability, the maximum likelihood bootstrap value, and the parsimony bootstrap value are shown from left to right along the branch for each node. For many nodes, the posterior probability (× 100) is considerably higher than either of the two bootstrap values.

In summary, a Bayesian analysis is model based and is similar in many respects to a maximum likelihood analysis because the likelihood is one component of the posterior probability. Branch lengths are in substitutions per site. The analysis itself tends to be slower than the other methods. Bayesian support values are generally interpreted as conventional probability values, so any value below 90% or 95% indicates poor support.

Comparing Trees from Different Methods and Sources of Data

In theory, every method of analysis and every source of data should give the same history for a group of organisms. In fact, this has proved true for many groups of plants. The monophyly of such families as the Poaceae, Onagraceae, Ericaceae, Asteraceae, Fabaceae, and Orchidaceae, and of other large groups of plants such as angiosperms, monocots, and asterids, has been supported by virtually all phylogenetic analyses of many kinds of data, including morphology, chloroplast gene sequences, and nuclear gene sequences. This list of robust groups is long. While many papers in the systematics literature wrestle with situations in which phylogenies disagree, it is important to recognize that in many cases phylogenetic data are quite consistent.

Even though many aspects of plant phylogeny are well supported, not all phylogenetic trees are identical. We have already mentioned a number of reasons for this. First, the different analytical methods have different underlying assumptions and thus are subject to different analytical artifacts, particularly when dealing with nucleotide data. Because of this, it is common to use several analytical methods to determine which aspects of the result are sensitive to the method of analysis and which are robust. Often analyses will begin with a quick neighboring analysis just to guide subsequent thinking and data exploration and perhaps additional sampling. These fast preliminary analyses can then be followed with more extensive analyses using the various optimality criteria.

Second, trees may differ if they use different sorts of characters. For example, phylogenies based on morphology or chloroplast, mitochondrial, or nuclear DNA nucleotide sequences can be (and often are) compared. If these phylogenies show similar groups, we can be more confident that they reflect the true order of events. For example, chloroplast, mitochondrial, and nuclear DNA each support three major clades in the Rosidae (Sun et al. 2015). All data support the monophyly of Fabidae (Fabales plus some rosids), Malvidae (Malvales plus other rosids), and the clade made up of Celastrales, Oxalidales, and Malpighiales (COM), but the relationships among the three clades are different when chloroplast gene sequences are used than when mitochondrial or nuclear gene sequences are used (Figure 2.20A). Based just on these three trees, a majority-rule consensus tree would appear as the one on the left in Figure 2.20B, reflecting that two-thirds of the trees place the COM clade as sister to Malvidae.

One way to show groups that are found by all methods of analysis, or among different sorts of data, is a strict consensus tree, which displays only monophyletic groups that are common to all trees. Lineages whose positions are uncertain are then drawn as though they arose at the same node in the tree diagram. In Figure 2.20, the different sets of data find that COM could be sister to the Fabidae or sister to Malvidae, so the strict consensus tree shows the branching order of the three clades as being ambiguous (Figure 2.20B, right). Although it appears that the lineages arose at the same time, in fact the diagram indicates that we cannot tell when they originated.
As the example in Figure 2.20 shows, the history of any particular gene may not reflect the history of the organisms bearing that gene. Nuclear genes may or may not track the history of the nucleus, and chloroplasts and mitochondria may or may not have a history different from that of the nucleus. There are three main reasons for these differences:

1. Mutation is a random process; therefore the phylogeny reconstructed for a particular gene may differ from those for other genes by chance alone, as discussed above in Models of Evolution.

2. Polymorphisms in an ancestral species can be lost in descendant species. By chance, this can result in a history of the genes that is actually different from that of the nucleus. There are three main reasons for these differences:

3. Hybridization or introgression (described in Chapter 5) may transfer some DNA into a different lineage. This is particularly true in the case of chloroplasts and mitochondria, which are not linked to particular nuclear genomes. In the example of the Rosidae in Figure 2.20, Sun et al. (2015) suggest that hybridization might explain the differences in the gene phylogenies. Allopolyploidy (duplication of an entire nuclear genome) also often accompanies hybridization.

We now have multiple gene trees for many groups of organisms, and it is possible that none of those gene trees will be exactly the same as the species tree. For example, many plant taxa have been shown to have the “wrong” chloroplast, presumably because of introgression. In a recent striking example, many species of willows (Salix) have similar chloroplast sequences, even though the species are morphologically distinct, a pattern that is explained by widespread hybridization (Percy et al. 2014).

Individual nuclear genes may also have quite different histories. Figure 2.22 shows two nuclear gene trees from the genus Linaria (toadflax, family Plantaginaceae) (Blanco–Pastor et al. 2012). One tree is for the internal transcribed spacer of the rRNA (ITS), and the other is for...
FIGURE 2.22 Two distinct phylogenies for members of the genus Linaria (Plantaginaceae). (A) Tree for the nuclear sequences of the internal transcribed spacer of the rRNA, or ITS. (B) Tree for the nuclear gene alanine:glyoxylate aminotransferase1, or AGT1. Branches highlighted in blue or green indicate species that are morphologically placed in the same subsection. Note that the positions of these taxa are different in the two phylogenies, as would be expected if the history of the genus involved hybridization and incomplete lineage sorting. (After Blanco-Pastor et al. 2012.)

alanine:glyoxylate aminotransferase1 (AGT1). Although there are clearly similarities between the trees, there are notable differences, as can easily be seen by comparing the positions of the green and blue branches of the trees. In this case, the authors attributed the differences to a combination of hybridization and incomplete lineage sorting (see Chapter 5 for more information on these important biological processes).

Hybridization followed by polyploidy can also lead to complex patterns of gene trees, as has been documented in the case of cotton, wheat, and many other crop plants (Renny-Byfield and Wendel 2014). Hybridization can be, and often is, accompanied by spontaneous doubling of the genome, also known as polyploidy. Organisms, such as humans, with two sets of chromosomes are described as being diploid. Organisms with more than two sets are known as polyploid, with the appropriate Greek prefix used to indicate the numbers of chromosome sets. For example, plants with four sets of chromosomes are tetraploid, and those with six sets are hexaploid. Polyploidy that is the result of hybridization is sometimes referred to as allopolyploidy. An allotetraploid contains four sets of chromosomes, one pair from each diploid parent.

Figure 2.23A illustrates a phylogeny of diploid species A through F and a hybridization event between B and D that leads to an allotetraploid containing genomes from both the B and D parents. When we sequence genes from the diploids and tetraploids, the gene tree will look like the one in Figure 2.23B. DNA sequences from the tetraploid plant will appear in two places in the gene tree.

Gene copies from parent B will group with that parent, and the copies from parent D will group with that parent. The phylogeny thus provides two pieces of information: it shows that hybridization and polyploidy occurred, and it identifies the parents of the cross.

Using Phylogenetic Trees

Phylogenetic trees have proved to be a remarkably powerful tool for comparative biologists in general, not just systematists. While phylogenies form the basis of most current classifications, they are also useful for studying how characters of plants evolve over time, which is accomplished by mapping characters on trees. In addition, phylogenies allow us to estimate the dates of major events in evolutionary history so those dates can be correlated with geological changes. We will discuss each of these topics in turn.

Constructing a Classification

A classification can be constructed in many ways. For example, plants can be classified on the basis of their medicinal properties (as they are in some systems of herbal medicine) or on the basis of their preferred habitats (as they are in some ecological classifications). Phylogenetic classification, as we use in this book, attempts to arrange organisms into groups on the basis of their evolutionary relationships.

The theory of classification is a topic with which systematists have been wrestling for centuries, and their struggles have led to a broad and frequently contentious literature (see Chapter 3). The principles of phylogenetic classification outlined here are commonly but not universally held. In general, however, classification has several goals. A classification is a common vocabulary designed to aid communication. Therefore a classification should be stable; names that are frequently changed become useless for communication. In addition, a classification should be predictive; that is, the name of a plant should help you to learn more about that plant and guide you to its literature.

Systematists generally agree about the goals of classification, but they may disagree profoundly about how to reach those goals. In this text, we take a particular point of view, using phylogenetic classifications throughout, as opposed to phenetic or evolutionary school classifications.
(see below and Chapter 3). As far as possible, we recognize monophyletic and avoid paraphyletic or polyphyletic groups. In the few cases in which a nonmonophyletic family or subfamily has not yet been divided into monophyletic units, we have placed the taxon name in quotation marks. The monophyly of many genera of angiosperms is questionable, but fewer phylogenetic analyses are available at this level, so generally we have not tried to indicate possible or probable paraphyly or polyphyly of genera.

A classification can be derived directly from a phylogeny. For example, the shared derived character states in Figure 1.4 can be arranged in a hierarchy from more inclusive (e.g., stems woody or petals red) to less inclusive (e.g., leaves hairy, seed coat spiny). This arrangement then allows the plants themselves to be arranged in a hierarchical classification that reflects their evolutionary history. The plants could be divided into two groups: one with the shared derived character state red petals and the ancestral character state of herbaceous stems, the other with the shared derived character state woody stems and the ancestral character state white petals. Each of these groups could also be divided into two groups.

The biological diversity on Earth is the result of geographical descent with modification, and monophyletic groups owe their existence to this process. It is appropriate, therefore, to use monophyletic groups in biological classifications so that we may most accurately reflect this genealogical history. Classifications based on monophyletic groups are more predictive and of greater heuristic value than those based on overall similarity or on idiosyncratic weighting of particular characters (Farris 1979; Donoghue and Cantino 1988).

Phylogenetic classifications, because they reflect genealogy, will be the most useful in biological fields such as the study of plant distributions (phytogeography), host-parasite or plant-herbivore interactions, pollination biology, and fruit dispersal and in answering questions related to the origin of adaptive characters. Because of its predictive framework, a phylogenetic classification can direct the search for genes, biological products, biocontrol agents, and potential crop species. Phylogenetic information is also useful for making conservation decisions. Finally, phylogenetic classifications provide a framework for biological knowledge and a basis for comparative studies linking all fields of biology (Cracraft and Donoghue 2004).

Constructing a classification involves two steps. The first step is delimitation and naming of groups. In a phylogenetic classification this step is uncontroversial: named groups must be monophyletic. The second step involves ranking the groups and placing them in a hierarchy. This step remains problematic.

**Grouping: Named groups are monophyletic** A phylogenetic classification reflects evolutionary history and attempts to give names only to groups that are monophyletic, that is, composed of an ancestor and all its descendants. For example, we infer that the Asteraceae are monophyletic because they have flowers in heads. The Asteraceae plus other plants that share the character state fused petals are also monophyletic; this group also has a name, the Asteridae (or the asterid clade). Similarly, the entire group of plants with tricolpate pollen is monophyletic and is known as the eudicots (or the tricolpate clade). This group has been given a formal Latin name, Eudicotyledoneae, under the PhyloCode (see Cantino et al. 2007). In phylogenetic classification, paraphyletic groups are not named. Naming a group that includes species from two clades would imply that those species are closely related even though they are not.

This book contains examples of named groups of plants that we now believe to be paraphyletic. One is “Caesalpinioideae,” a group of legumes that traditionally includes the North American redbuds (Cercis) and the sensitive pea (Chamaecrista) among many others. But the redbuds and sensitive peas are more distantly related to one another than either of them is to other legumes. Without quotation marks, the name Caesalpinioideae implies a closer relationship than actually exists.

As phylogenetic data accumulate, some families previously thought to be monophyletic are being found to be actually paraphyletic; examples include Apocynaceae and Capparaceae, as traditionally delimited. Over the last decades, these families have been recircumscribed to recognize monophyletic groups: Apocynaceae has been combined with Asclepiadaceae, and Capparaceae has been divided into Capparaceae sensu stricto (s.s., meaning, in the strict, or narrow, sense) and Cleomaceae.

**Naming: Not all groups are named** A phylogenetic classification attempts to name only monophyletic groups, but the fact that a group is monophyletic does not mean it needs to have a name. The reasons for this are practical. We could put every pair of species into its own genus, every pair of genera into its own family, every pair of families into its own superfamilly, and so on. But such a classification would be cumbersome; in addition, it would not be stable, because our view of sister species would change each time a new species was described, and our view of the entire classification would have to shift accordingly.

In practice, many monophyletic groups are not named. For example, the genus Stenanthium (Melanthiaceae) is monophyletic and contains four species (Zomlefer et al. 2001; Zomlefer and Jud 2002; Wofford 2006). Although the data show that these four species fall into two monophyletic pairs, the two pairs of species are not named, and few systematists would consider doing so. In another example, over half of the genera of the grass family fall into a single large clade that contains four large, traditionally recognized subfamilies plus two smaller ones. Although agrostologists refer to this clade as the PACMAD clade (an acronym for Panicoideae-Arundinoideae-Chloridoi-
deae–Micrairoideae–Aristidoideae–Danthonioideae), it has no formal Latin name.

No formal rules exist for deciding which monophyletic groups to name, but several criteria have been suggested, and some criteria are used despite not being formally articulated. A major criterion—perhaps the major criterion—is the strength of the evidence supporting a group. Ideally, only clades with strong support should be formally recognized and named in classifications. This makes sense if a classification is to function as a common vocabulary.

Names are most useful if they can be defined, and the more precise the definition the better. In other words, if a clade is to be named, it should have a set of characters by which it can be distinguished from other clades, or diagnosed. This criterion is also important for nomenclatural stability: if the meaning of a name shifts every time a new phylogeny is produced or a new character is examined, the name becomes effectively meaningless. Similarly, if, for example, the only way a field biologist can identify an organism is by knowing whether it has an alanine or a serine at position 281 in its rubisco large subunit molecule, he may not find the classification much help. If, on the other hand, he knows that the organism is a grass with a particular spikelet structure, he can easily and reliably infer many aspects of its biology. (Lack of an obvious morphological synapomorphy is one of several reasons that the PACMAD clade of the grasses is not given a name.) The characters used for classification do not have to be those used for identification, but many systematists prefer to name clades that are easily recognized morphologically.

Another criterion for naming is size of the group. Human memory is easily able to keep track of small numbers of items (in the range of three to seven) (Stevens 1998), but to organize and remember larger numbers of items requires additional mnemonic devices. (As an example, consider how many nine-digit zip codes you can remember compared with the five-digit variety, or with seven-digit telephone numbers.) Dividing a large group into smaller groups is a way to organize one’s thinking about large numbers of taxa. For example, the genus Ste- nanthium could be redefined to include only Stenanthium graminicum and S. diffusum, and a new genus could be described to include S. densum and S. leimanthoides. There seems little reason to do this, however, because four species is not a difficult number to keep track of. That said, there seems little reason to divide a large group if well-supported, morphologically distinct clades cannot be identified within it, and even then there is no necessity to divide it.

A third naming criterion is nomenclatural stability. A classification is ultimately a vocabulary, a means of communication. It cannot function this way if the meanings of the names continually change. Thus, given a set of well-supported, diagnosable, monophyletic groups, groups that have been named in the past can—and we would argue should—retain their names. This is yet another argument against formally naming the PACMAD clade of the grasses, because it would entail an unnecessary set of changes affecting long-standing taxonomic usage (Backlund and Bremer 1998; Stevens 1998).

**Ranking: Ranks are arbitrary** Having decided which monophyletic groups to name, we still have the question of exactly how to name them. The groups could, for example, be numbered, and a central index could list what is included in each numbered group. This approach is similar to the system used by telephone companies to organize telephone numbers. The difficulty, of course, is that without a telephone book (a central index) and/or an excellent memory, the system is inaccessible.

Biological classification attempts to provide a working vocabulary that conveys phylogenetic information, yet can be learned by biologists who are not themselves primarily systematists. Because a phylogeny is similar in structure to a hierarchy, in which small groups are included in larger groups, which themselves are included in still larger groups, it makes sense for the classification to reflect the hierarchy.

Botanical classification uses a system developed in the eighteenth century in which taxa are assigned particular ranks, such as kingdom, phylum, class, order, family, genus, and species (i.e., Linnaean ranks; see Chapter 3 and Appendix 1). A classification of named monophyletic groups should be logically consistent with the phylogenetic relationships hypothesized for the organisms being classified. So, the categorical ranks of a Linnaean classification are used to express sister-group relationships.

Even though monophyletic taxa are considered to represent real groups that exist in nature as a result of the historical process of evolution, the categorical ranks themselves are only mental constructs. They have only relative (not absolute) meaning (Stevens 1998). In other words, the familial level is less inclusive than the ordinal level and more inclusive than the generic level, but no criteria are available to indicate whether a particular taxon, such as the angiosperms, should be recognized at the level of phylum, class, or order.

In **Figure 2.24**, a cladogram of imaginary taxa A through E is converted into a hierarchical classification according to Linnaean categorical ranks. Note that subgenus DE is nested within genus CDE, which is in turn nested within family ABCDE. (But, we could have treated clade ABCDE as an order, clade CDE as a family, and clade DE as a genus.) Often, however, to fully express the sister-group relationships (in the cladogram), one needs more ranks than are available (in the taxonomic hierarchy), even after creating additional ranks by use of the prefixes *super-* and *sub-*.

Even though ranking is arbitrary, the criteria described here for deciding which groups to name can also be ap-
Some systematists have proposed abandoning the Linnaean system altogether and replacing it with a “phylogenetic taxonomy.” Full exploration of this possibility is beyond the scope of this text, but we will briefly address the arguments against the use of Linnaean ranks here. Because rank is arbitrary, a genus in one family may not be the same age as, encompass the same amount of variation as, or indeed have anything in common at all with a genus in another family—other than the fact that they are both monophyletic groups. Trained systematists are generally aware of this (Darwin was, for example) and realize that genera, families, and so on are not comparable units (Stevens 1997). Some scientists, however, frequently use such categories as though they were real. For example, it is common to measure plant diversity by listing the number of families represented by a local flora, even though family does not say anything in particular about a unit.

If rank is arbitrary, then one logical step would be to eliminate ranks altogether. Taxa would be placed in named groups, but the groups would not be designated as genus, family, order, or any other rank. Such categorization already exists informally, particularly among groups above the level of orders. The eudicots, for example, are widely recognized as monophyletic but are not given a particular Linnaean rank. Similarly, few systematists worry about whether the angiosperms should be recognized as a division, class, subclass, superorder, or other rank; they are clearly monophyletic and designated by the non-Linnaean name angiosperm. Alternatively, rank could be based on absolute age of a clade, but this would cause major disruptions, even if age could be assigned unambiguously.

Eliminating ranks becomes more problematic among orders, families, and genera. Groups assigned to those ranks are familiar, their names are in common use, and they tap into the implicit hierarchy we use in language, so an entirely new sort of nomenclature is unlikely to be accepted rapidly or without protest. Nonetheless, an alternative system of phylogenetic nomenclature, known as the PhyloCode, is being developed. The PhyloCode is designed entirely outside the rules of the International Code of Nomenclature for Algae, Fungi and Plants (ICN), which governs the use of Linnaean ranks and has long been used by all plant taxonomists (see Appendix 1). It is an alternative nomenclatural system rather than a revision of the existing system (see the PhyloCode website at www.ohiou.edu/phylocode).

Many phylogenies are only partially resolved, so precise placement of taxa is impossible given the available data. This means that some species cannot be placed for certain in a genus, and some genera cannot be reliably assigned to a family. The current system allows uncertain placements above the rank of species to be reflected by the category incertae sedis—literally “of uncertain position.” An alternative would be a rank-free system, in which neither placement in a larger group nor naming all branches of a dichotomy or polytomy would be necessary.
The authors of this text have been involved in reclas\-\-sifications of genera, families, and orders on the basis of phylogenetic data and have found that—as long as the phylogeny is clear—use of the standard Linnaean hier\-\-archy is quite easy (especially when it is supplemented by unranked informal names or PhyloCode names). When the phylogeny is unclear, one should generally wait for more data before modifying the classification.

**Comparing phylogenetic classifications with those derived using other taxonomic methods** Not all tax-\-\-onomists use phylogenetic methods, although this is the majority approach. A few systematists have held the view that although evolution has occurred, parallelism and reversal are so common that the details of evolutionary history can never be deciphered. This point of view led to a school of systematics known as **phenetics**. Pheneticists argued that because evolutionary history can never be unequivocally detected, organisms might best be classified according to overall similarity. Thus they placed similar organisms together in a group and very different organisms in different groups (Sneath and Sokal 1973). This approach is discussed in more detail in Chapter 3.

The school of **evolutionary taxonomy** also differed from phylogenetic taxonomy in its approach to classification. In this school, the morphological similarity within a group was of utmost importance, and monophyly and par\-\-aphyly (in the strict cladistic senses of those words) were secondary. Thus a group could be recognized on the basis of some combination of derived and ancestral, unique and shared characters (Figure 2.26). Importance was given to the recognition of “gaps” in the pattern of variation among phylogenetically adjacent groups (Simp-\-\-son 1961; Ashlock 1979; Cronquist 1987). Characters considered to be evolutionarily (or ecologically) significant were stressed, and the expertise, authority, and intuition of individual systematists were central. This sort of classification led to recognition of the classical group “dicots” as distinct from monocots. The morphological similarity of the “dicots” was afforded more importance than the fact that “dicots” are not a monophyletic group.

It has been said that systematics is as much an art as a science (although this statement begs the question of how one might define art and science), in part because so many aspects of the discipline have seemed to have no objective basis. One fortunate result of phylogenetic systematics is that at least one major aspect of systematics—the delimitation of groups—has become formalized such that there is general agreement on how it should be done. Whereas phenetic and evolutionary classifications were ambiguous about grouping criteria, phylogenetic classifications are precise. A named group can be taken as monophyletic, including all descendants of a single common ancestor.

**Describing Evolution: Mapping Characters onto Trees**

A major aspect of plant systematics is understanding the evolution of particular characteristics—morphological, physiological, environmental, or other—of plants. Phylogenies can be used to describe the evolutionary process and to develop hypotheses about adaptation, morphologi-\-\-cal and physiological change, or biogeography, among many other uses. This is the kind of study that system-\-\-atists frequently engage in because the details of character evolution lead to hypotheses about how natural selection has worked. In addition, when constructing classifications, one frequently wants to know what morphological characters can be attributed to and distinguish a partic-\-\-ular monophyletic group.

Making use of a phylogeny in this way requires that changes in particular characters be assigned to particular
nodes in the phylogenetic tree. For example, we may wish to determine whether a group of cold-tolerant plants in the Arctic is derived from a different cold-tolerant Arctic group or whether it is derived from cold-intolerant plants from lower latitudes. This requires that we create a hypothesis about the nature of the plants represented by each ancestral node on the tree and estimate whether they are cold-tolerant or cold-intolerant. Generating this sort of hypothesis is known as mapping the character onto the tree.

The characters mapped onto the tree may be morphological characters, physiological characters such as drought tolerance or high-efficiency photosynthesis, environmental characters such as mean precipitation in July or density of shade, or biogeographic characters such as location of plants on particular continents or in particular biomes. The characters may be quantitative or qualitative (although for methodological reasons, most are qualitative). The process of identifying homologous structures, defining characters, and delimiting character states is the same as that described for building a morphological classification (see Box 2A).

The process of mapping characters is generally separate from that of generating a tree. Mapping can be done using only the topology of the tree or can also incorporate information on branch lengths. It can be done in a parsimony, likelihood, or Bayesian framework. Note that in this sort of study the tree is taken as given. This means that the method used to produce the tree does not need to be the same as the method used to map characters onto it. It does not matter how the tree was produced, and indeed the data used to produce the tree are not important.

**Mapping in a parsimony framework** It is perhaps easiest to start by thinking about simple parsimony mapping using only the topology of the tree, that is, ignoring branch lengths for the time being. Consider a group of plants for which the phylogenetic tree is known; a good example is the Ericaceae, for which much information on branch lengths for the time being. Consider a group of plants for which the phylogenetic tree is known; a good example is the Ericaceae, for which much information is available (Figure 2.27). Assume for the purposes of this discussion that each of the terminal genera really is monophyletic, as demonstrated by studies of multiple species of each. Then consider a study that is concerned with the gain or loss of fused petals, which are intimately connected with the evolution of pollination systems. Figure 2.27 shows the observed character states (fused or free petals) for the genera.

It seems trivially obvious from looking at the distribution of characters and character states that free petals must have evolved once in the lineage leading to *Rhododendron* subsection. *Ledum* (Labrador tea) and again in the lineage leading to *Vaccinium* sect. Oxycoccus (cranberries). Phrased another way, the ancestor of *Vaccinium* sect. Oxycoccus and all other vacciniums (blueberries) had fused petals, as did the ancestor of *Rhododendron* subsections *Ledum* and 3. However, examine this “obvious” conclusion more closely. If we were studying only species of *Vaccinium*, we would have no way of knowing whether fused petals were ancestral or derived (Figure 2.28A). There must have been one genetic change, but it could have happened as easily in the lineage leading to *Vaccinium* sect. Oxycoccus as in the lineage leading to the other *Vaccinium* species. Only by reference to the outgroup *Epacris* can we determine when petal fusion was lost. Because *Epacris* has fused petals, free petals must have originated within *Vaccinium* sect. Oxycoccus. This is the same as saying that the ancestor of other *Vaccinium* species plus *Vaccinium* sect. Oxycoccus had fused petals. If we were to postulate that the ancestor had free petals (Figure 2.28B), we would need two changes to fused petals: one in *Epacris* and one in the blueberries. The same argument applies in the case of *Rhododendron* subssect. 3 and *Rhododendron* subssect. *Ledum* (see Figure 2.27).

Now suppose we were studying only species of *Vaccinium*, but this time, instead of using *Epacris* or other Ericaceae as outgroups, we used only *Rhododendron* subssect. *Ledum*. This could easily be the case if material of the other genera were hard to obtain or if those genera were extinct and we did not know they had existed. Now the most parsimonious conclusion is that the ancestor of all vacciniums had free petals and that in response to some unknown selective pressure, there was a change to fused petals in *Vaccinium* sect. Oxycoccus. Figure 2.27 shows the observed character states (fused or free petals) for the genera. Assum
conclusion from the one reached in the previous example, and the only difference is the genera included in the analysis. Additional outgroups (Figure 2.29B). For example, consider petals (Figure 2.29A, bottom). This is exactly the opposite conclusion from the one reached in the previous example, and the only difference is the genera included in the analysis.

One might try to improve the situation by using additional outgroups (Figure 2.29B). For example, consider the same study of *Vaccinium*, but now use both *Rhododendron* subsect. *Ledum* and *Rhododendron* subsect. 3 as outgroups. In this case the direction of change is ambiguous. It is as simple to postulate that the ancestor of the group had fused petals and that there were two changes to free petals as it is to postulate that the ancestor had free petals and that there were two changes to fused. These two choices are known as equally parsimonious reconstructions. For many characters on many trees, there are multiple equally parsimonious reconstructions. In other words, there are multiple equally good hypotheses about the direction and timing of character state change.

Ambiguity can also come from including taxa for which the character state is not known. Suppose, for example, two new taxa are discovered such that, on the basis of other characters, new species 1 is clearly sister to blueberries (other *Vaccinium*) and new species 2 is clearly sister to *Vaccinium* sect. *Oxycoccus* (Figure 2.30). In addition, suppose that it is unclear whether the petals are fused or free.

This type of ambiguity is more common than you might think; it can occur when the original description is vague and/or illustrations are unclear, or when the original plant is known only from fruiting material. In this example we do not know what the ancestral state was for *Vaccinium*, so we cannot make any hypothesis about the direction of evolutionary change. Nor can we be sure that the character fused petals is a synapomorphy for the genus.

**Mapping in a likelihood or Bayesian framework** In the process of producing a likelihood or Bayesian tree, the analysis automatically assigns the probability of each ancestral state for each nucleotide at each ancestor (see Maximum Likelihood and Bayesian Methods on pages 32–34). In the example described above (see Figure 2.19), we estimated the likelihood of the observed data if ancestor X had an A at a particular position, and then we estimated the probability of the data if X had a C or a G or a T. Using exactly the same process, we could esti-
mate the likelihood that we would observe any other set of characteristics given a particular ancestral state. This would lead to a maximum likelihood estimation of a particular character for each ancestor in the tree. Because we would be working in a likelihood framework, however, the ancestral states would not be simply binary (e.g., petals either fused or free, or plant cold-tolerant or cold-intolerant), but rather the likelihood that a particular state of an ancestor would lead to the states of descendants that we see today. An example is shown in Figure 2.31, which shows corolla color mapped onto a portion of the phylogeny of the genus *Ruellia* (Acanthaceae) (Tripp and Manos 2008).

The common ancestor of the clade that includes *R. beyrichiana* through *R. geminiflora* is inferred to have had purple flowers, a result that likely would have appeared also if the mapping had used a parsimony algorithm. However, the corolla color of the common ancestor of the entire clade is uncertain in most trees (indicated by the large gray region of the pie chart on the far left), but there is a small probability, indicated by the very small pie slice, that the corolla might have been purple. The pie charts at the other nodes also indicate considerable uncertainty about the ancestral state.

Also unlike many parsimony methods, maximum likelihood methods take into account the lengths of the branches. Thus if a branch is short (i.e., there were few mutations between an ancestor and a descendant), then change of the character state is less likely. Conversely, estimation of ancestral states over a very long branch becomes highly uncertain.

Because the Bayesian posterior probability is computed from the likelihood multiplied by the prior probability, Bayesian estimates of ancestral character states are intimately connected to likelihood values. Similar to maximum likelihood reconstructions, Bayesian approaches to mapping ancestral states can incorporate uncertainty in the tree topology and in parameters of the underlying model of evolution. While this often means that the ancestral character state appears ambiguous, most system-
Assigning Dates to Nodes in Phylogenies

A phylogenetic tree is a diagram that depicts time. As such it can be used to assess both relative time (Did event X occur before event Y?) and absolute time (How many years ago did event X occur?).

Many questions in evolution can be answered with knowledge of only relative time. The actual date of an event is not especially relevant, only that one event occurred before or after another. In general, ancestors deeper in the tree must have existed before ones at more shallow nodes. Likewise, if characters are mapped onto the tree, then the phylogeny provides information on the relative timing of character state change. Thus the phylogeny of the Ericaceae in Figure 2.27 shows that free petals originated after fused petals in this group, even though we don’t know the exact dates for any of the nodes of the tree.

Although the tree alone can answer questions of relative time, many evolutionary questions require more precise data on how fast mutations are occurring and what geological date should be assigned to a particular node. For example, a dated phylogeny is needed to determine whether Gnetaceae underwent most speciation during the Cretaceous, or whether Malvaceae is old enough to have been affected by the breakup of Gondwana. If speciation in Gnetaceae or Malvaceae is more recent than about 80 million years ago, then it was probably not affected by the breakup.

When molecular data first became available, there was some hope that DNA might mutate in a clocklike fashion, so every mutation would equal one “tick of the clock.” If this were the case, then we would only have to produce a molecular phylogeny and count the number of mutations and we could estimate time. However, DNA does not provide a well-behaved clock. Different genes in different lineages have rates of mutation that are far more different than would be expected purely by random chance. This means that over the same period of time, some branches of a tree accumulate mutations more rapidly than others and thus appear as longer branches on a phylogenetic tree. For example, woody plants in general tend to have slower rates than herbaceous plants (Smith and Donoghue 2008).

Methods for assigning dates to phylogenies involve a series of steps. First, basic probability theory is used to test whether the mutation rate in any particular lineage is behaving in an approximately clocklike fashion. In some cases, the DNA sequences all mutate in a way consistent with the same underlying rate, but in most trees there is significant variation in mutation rate from one branch to the next. In the latter case, the rate of molecular evolution must be made predictable throughout the tree. There are many ways to do this, and all involve complex sets of assumptions. In one set of methods, known as rate-smoothing methods, the rate of molecular evolution is permitted to vary from one branch to the next, but the amount of change is specified by a number known as a smoothing parameter. Perhaps not surprisingly, changes in the smoothing parameter can give different estimates of the date of an internal node.

Once the rate of mutation is forced to be approximately constant throughout the tree, the next step is to determine how many mutations have occurred in a particular period of time. In other words, if a mutation is one tick of a clock, how fast is the clock ticking? Answering this question requires that the clock be calibrated. To do this, we must assign one or more fossils a position in the phylogeny based on whatever characters are available. The date of the fossil then provides a minimal age of the node to which it attaches. For example, if we have a fossil that is dated to 50 million years ago (mya), and if that fossil is well enough preserved that we can be sure that it has the synapomorphy of a particular clade, then the clade must be at least 50 million years old. The fossil does not give us a maximum age, because the group could have been around since before 50 mya; we have no way of knowing.

The tree in Figure 2.32 shows the history of fern genus Osmunda, its sister genus Osmundastrum (which was formerly included in Osmunda), and two outgroups, To-dea and Leptopteris. Node 1 is known as the stem node, the point of origin of Osmunda, and node 2 is the crown node, the point of the earliest speciation events within the genus. Suppose you find a fossil that is dated to 153 mya and it has the synapomorphies of both Osmunda s.s. and the Osmundastrum + Osmunda clade. Should the fossil be attached to the stem or crown node? The generic synapomorphy may have arisen anywhere along the internode between the stem and crown node, but the earliest possible time that synapomorphy could have arisen was immediately after the stem node. Thus we would assign 153 million years as the minimal age of the stem node. Note that the stem node could be older—in fact, quite a lot older—and that 153 mya is merely the latest that speciation event could have occurred.

The methods just described separate the steps of making the tree clocklike and then calibrating it. However, in another set of methods, a Bayesian framework is used to incorporate over a broad range of possible rates and possible calibration points. These methods are becoming increasingly popular because they require less precise (and hence presumably more realistic) assumptions about calibration dates. Still other methods are being developed that are less heavily influenced by the date of the oldest fossil and that incorporate more extensive morphological data along with DNA data (see Grimm et al. 2015 for discussion).

Despite considerable improvement in methods for estimating dates of ancestral nodes, estimates of dates have huge errors associated with them. For example, in dating the phylogeny of Osmunda shown in Figure 2.32, Grimm et al. (2015) tested four different analytical methods,
The phylogeny. (After Grimm et al. 2015.)

The distinction is important for calibrating the earliest speciation event within Osmunda morphology has been interpreted in a way that is comparable to that of extant taxa. While some fossils are indeed astonishingly well preserved, others are less so. In addition, it is often hard to double-check the paleontological data, since the specimens usually have to be checked in the museums that hold them. In addition, the stratigraphy can be hard for a nonspecialist to assess.

In summary, current methods for dating phylogenies are constantly improving, but any assessment of the time of origin of a particular clade still needs to be viewed skeptically. Estimates of dates can vary by tens of millions of years, which makes correlating biological events with changes in geology or climate risky at best. Combining improved analytical methods with a better understanding of the fossil record and of morphological evolution provides the best hope for a temporal understanding of plant diversification.

FIGURE 2.32 Phylogeny of the genus Osmunda and relatives. The stem node of Osmunda is the node marking the split between Osmunda and Osmundastrum (node 1), whereas the crown node is the one marking the earliest speciation event within Osmunda (node 2). The distinction is important for calibrating the phylogeny. (After Grimm et al. 2015.)

which dated the split between Osmunda and Osmundastrum as 238, 182, 176, or 156 MYA. While the authors offer reasons to prefer the oldest date, the discrepancies are still greater than 25%. Likewise, estimates of the age of the crown node of the angiosperms mostly place the divergence between Amborella and all other angiosperm taxa in the range of 210–130 MYA, although some estimates are 270 MYA (Stevens 2001 onwards; Bell et al. 2010; Moore et al. 2010; Magallón et al. 2015). This is a very large period of Earth history.

All dates assume that the fossils used for calibration have been identified and dated correctly and that their morphology has been interpreted in a way that is comparable to that of extant taxa. While some fossils are indeed astonishingly well preserved, others are less so. In addition, it is often hard to double-check the paleontological data, since the specimens usually have to be checked in the museums that hold them. In addition, the stratigraphy can be hard for a nonspecialist to assess.

In summary, current methods for dating phylogenies are constantly improving, but any assessment of the time of origin of a particular clade still needs to be viewed skeptically. Estimates of dates can vary by tens of millions of years, which makes correlating biological events with changes in geology or climate risky at best. Combining improved analytical methods with a better understanding of the fossil record and of morphological evolution provides the best hope for a temporal understanding of plant diversification.

LITERATURE CITED & SUGGESTED READINGS

Items marked with an asterisk are especially recommended to those readers who are interested in further information on the topics discussed in this chapter.


© 2016 Sinauer Associates, Inc. This material cannot be copied, reproduced, manufactured or disseminated in any form without express written permission from the publisher.