CHAPTER THREE

What a Phylogenetic Tree Represents

Before you can use trees to organize knowledge of biodiversity and refine your understanding of evolution, you need to know what a tree diagram represents and to become comfortable with some of the conventions used to communicate phylogenetic information. As noted by Robert O'Hara in his seminal exploration of the concept of tree thinking, "just as beginning students in geography need to be taught how to read maps, so beginning students in biology should be taught how to read trees and to understand what trees communicate" (O'Hara 1997). In this chapter, we begin by clarifying how reproduction of individual organisms within populations is connected to ancestry and descent at the level of the tree of all life. We then describe how to read tree diagrams in order to extract their essential phylogenetic content.

THE CONTINUITY OF REPRODUCTION FROM THE POPULATION TO THE TREE OF LIFE

A phylogenetic tree depicts the evolutionary ancestry of a set of tips. The tips are typically living species or groups of species, but can also be fossil organisms, individual organisms, genes, or populations. There is only one basic kind of ancestry in biology: that which links parents, or more generally ancestors, and their children, or descendants. A pedigree or family tree is a depiction of the ancestor-descendant relationships within a population. A phylogenetic tree shows the same thing, but at a larger scale and in less microscopic detail. To understand what a phylogenetic tree depicts, therefore, we need to conceptually bridge the reproduction of organisms within populations and the branching of evolutionary lineages to create a phylogenetic tree.

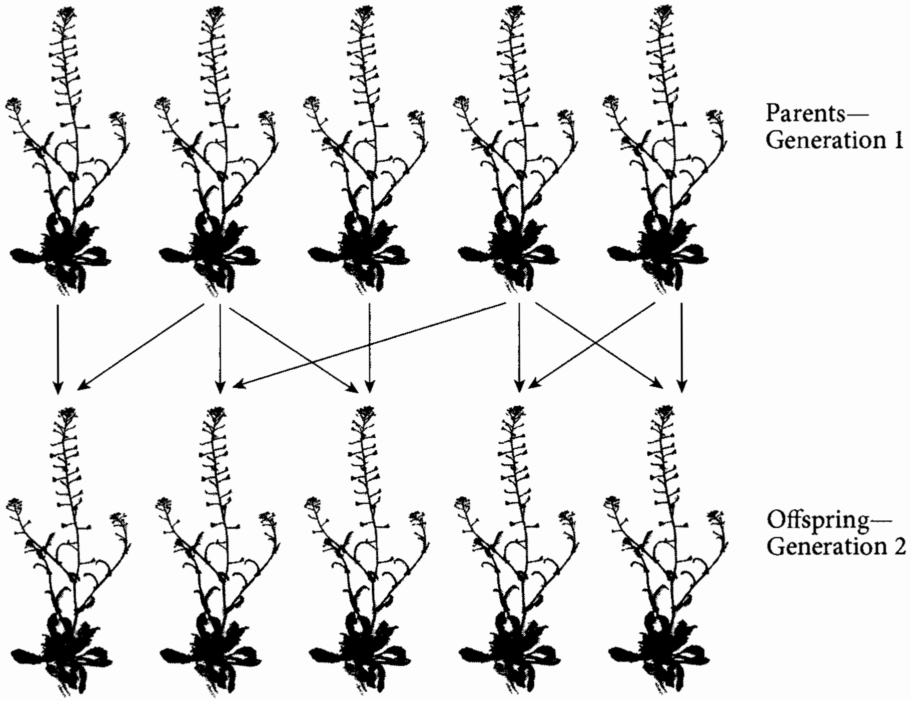

FIGURE 3.1 Two generations of shepherd's purse (*Capsella bursa-pastoris*) plants.

Start by imagining one generation of plants of a particular species; for example, the flowering plant shepherd's purse, growing side by side in a meadow and producing offspring by exchanging pollen. If we focus on five individual plants in the parental generation and the offspring generation, the pedigree could look like that shown in Figure 3.1. Here we have assumed that each individual has two different parents although self-pollination can occur.

If we now expand our image to encompass all plants in a population and several generations, it might look something like Figure 3.2. Notice that each individual has two parents, but gives rise to a variable number of offspring in the next generation.

Imagine taking such a pedigree and getting rid of the organisms so that only the descent relationships were retained, as shown in Figure 3.3. These parent-offspring connections can be thought of as the glue that holds the population together. When we get to evolutionary timescales, which typically entail hundreds of thousands or millions of generations, individual organisms are too transient to be of concern, except as the vehicles through which the lines of descent pass. The lines of descent are what we most care about.

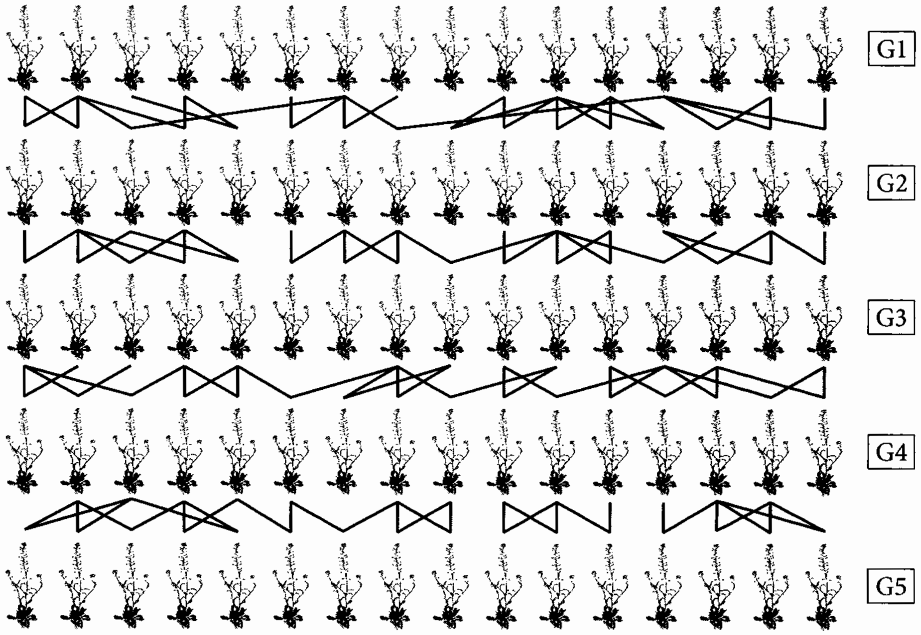

FIGURE 3.2 Multiple generations in the shepherd's purse population.

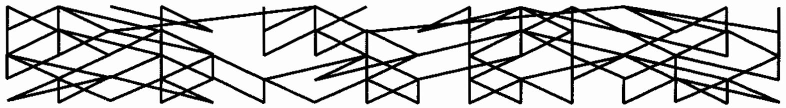

FIGURE 3.3 A representation of the generations in Figure 3.2 showing only lines of descent.

Instead of visualizing one small part of a single field of shepherd's purse over five years, expand your field of view to include many more individuals and generations. For example, Figure 3.4 is derived from a similar diagram as the preceding but now includes about 250 individuals and 80 generations. If

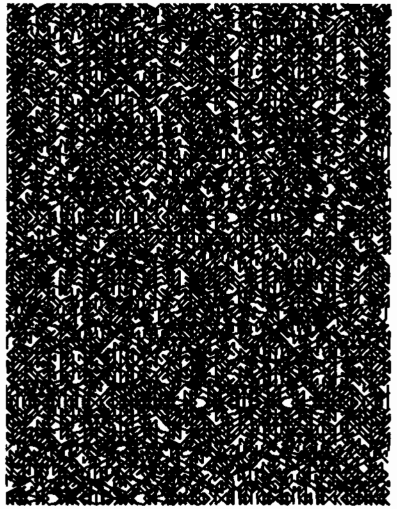

FIGURE 3.4 Many individuals and many generations of the shepherd's purse population.

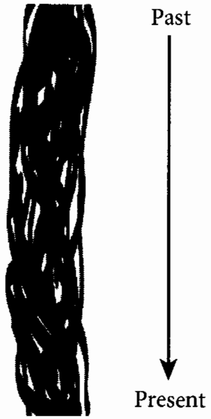

FIGURE 3.5 A lineage containing multiple populations connected by gene flow.

you zoomed out further and tried to represent a typical population of several thousand individuals that persists for hundreds or thousands of generations, all you would be able to see would be a fuzzy line. This line represents the genetic continuity through parent-offspring descent of a single well-demarcated population.

Individual populations may be fairly isolated for some period of time. However, on evolutionary timescales, seeds and pollen will occasionally carry genes between the distinct populations that make up a typical species. This gene flow between populations has the effect of "braiding" population lineages together. The graphic in Figure 3.5 might help you to visualize this braiding. Zooming out still further, this would probably look, again, like a fuzzy line. This is what is usually understood by an "evolutionary lineage."

During evolution, evolutionary lineages may diverge or "split." A technical term for such events is ***cladogenesis***, which refers to the origin (genesis) of new branches (*clados* is Greek for branch). Because lineages are sometimes equated with species, splits of this kind are sometimes called speciation events. However, considering the great controversy as to the exact meaning of "species" (see

Chapter 6), we avoid the term speciation in the following, instead using the less loaded term ***lineage splitting***.

Lineage splitting occurs when populations or groups of populations become isolated from one another so that they are no longer able to exhange genes via sexual reproduction. This might happen when a few individuals disperse to an isolated region (e.g., an island) or when a formerly contiguous range is divided by geological or climatic events (e.g., mountains, rivers, patches of inhospitable environments) that prevent gene flow. If the barrier to gene flow between the two populations remains intact for a long time, the isolated populations will begin evolving independently—a mutation arising in one population lineage will not spread to the other. As a result, the two populations will begin to acquire biological differences.

It should be noted that in this scenario, technically called ***allopatric*** divergence ("allo-" = different; "patria" = homeland), the distinct lineages are initially formed by extrinsic geological events and only later evolve differences as a consequence of being genetically isolated. This is thought to be the predominant mode of lineage splitting. However, in some circumstances lineage splitting can be facilitated by selection for ecological specialization within a population. In such cases, lineage splitting can occur without complete geographic isolation, a phenomenon called ***sympatric*** ("sym-" = same) divergence. For example, sympatric lineage splitting has been proposed for several fish groups, including the cichlids of Lake Victoria. However, because sympatric divergence is less common, we will focus on the allopatric case and how it leads to the production of new evolutionary lineages.

If population isolation is transient, then after the geographic barrier disappears, genes will flow again between the daughter populations, "braiding" them back into a single lineage. However, if the lineages remain isolated, the organisms in the isolated populations will tend to accumulate differences from each other in morphology (physical makeup), physiology, and behavior (Chapter 4). Eventually, these differences may make it impossible for individuals from the two allopatric lineages to mate successfully and/or to create viable offspring with one another. At this point, the separation of the lineages ceases to be dependent on the persistence of an original geographic barrier: reproductive isolation has shifted from being extrinsic, due to geography, to being intrinsic, due to the biological traits of the organisms.

It is a useful simplification at this point to assume that once lineage splitting has been completed, the two descendant lineages will remain isolated (exceptions are discussed in Chapter 6). This means that, once they have diverged, lineages do not exchange genes by hybridization. This is the underlying reason why evolution can be modeled as a tree rather than as a net. While some groups of organisms, for example, some microbes, do transfer genes between distant relatives, it can still be useful to think in terms of trees. For example, even when the organismic relationships are netlike, the genes themselves will typically have treelike histories (Chapter 6).

Figure 3.6 shows what we might see if we followed the fate of one initial lineage long enough to see it give rise to four living descendant lineages. This example also includes three lineages that were established but then went extinct before the end of the observation period.

In the left panel we have maintained the direction of time from the previous figures, with descendants below their ancestors. When we start looking at longer time frames, it is common to invert the arrow of time, placing the past at the bottom and present at the top. This convention probably arose because older (ancestral) fossils tend to lie in lower strata than fossils of lesser age. Also, the resulting figure, such as that in the right half of Figure 3.6, looks more like a living tree.

These diagrams show simple phylogenetic trees. Because the lines depict actual parent-offspring descent within lineages, they show that at least some of the organisms in the ancestral population are direct lineal ancestors of organisms in the (living) descendant populations. If there are $N + 2$ intervening generations then you can think of the individuals in the ancestral population as being (great × N) grandparents of the living individuals.

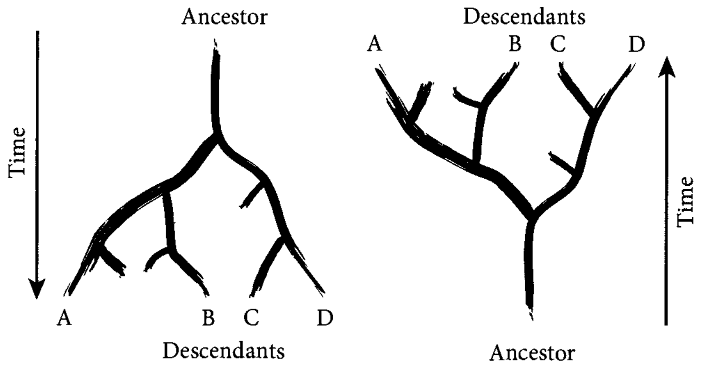

FIGURE 3.6 The branching of an ancestral lineage into four descendants.

If we were to zoom out as far as we could go from these four living species, we would ultimately find the entire tree of life, whose last common ancestor lived in the truly ancient past, over 3 billion years ago. As with the small tree in Figure 3.6, some long-extinct organisms are direct lineal ancestors of all organisms alive today. While it can be hard to imagine that there are organisms that are the (great × N) grandparents of such different organisms as humans, oak trees, mushrooms, and bacteria, this is what the tree-of-life model implies. As summarized in Chapter 2, there is abundant evidence supporting the tree model and descent from common ancestry.

Except in rare cases, such as laboratory studies of viruses or bacteria, we are not able to watch lineages evolve. Instead of starting from one ancestor and observing evolution occurring in a forward direction, phylogenies are generally approached in the reverse direction. We start from a sample of living tips of the tree and ask, how are they connected back through time? We are effectively taking a cross section of the evolving tree at the present and using information in this time slice to learn about earlier periods of time.

In thinking about evolutionary connections among the tips, it goes without saying that the future of the tree can be ignored. While we should always remember that evolution is ongoing (albeit too slowly to see except in some rapidly evolving groups), trees depict history only. Perhaps less obviously, we do not need to have direct knowledge of any ancestors. Although a tree implies the existence of certain ancestors, and even implies that those ancestors had certain combinations of traits (see Chapter 4), tree thinking is primarily concerned with understanding the evolutionary connections among tips. While ancestors must have existed, we never need to directly interact with ancestors to reconstruct or utilize trees.

The preceding might lead you to wonder about fossils. Are they on the tree and, if so, where are they? While some fossils might be actual ancestors of living species, the best way to approach fossils is to think of them as organisms that were collected in the present (as indeed they were), but stopped evolving a long time ago. That is to say fossils are best viewed as tips of the tree that have a shorter branch (in units of time) connecting them to the (inferred) ancestors. They are treated as living forms that have undergone no evolution in the millions of years since they were entombed in rock.

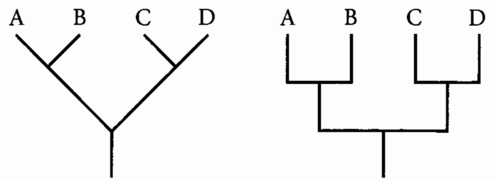

FIGURE 3.7 Two alternative representations of the same four-species tree.

In the case illustrated in Figure 3.6, we will assume that no fossils have been found. Thus, only the four living species are available for study. In this case, all the relevant information in Figure 3.6 can be summarized as either of the simplified tree diagrams shown in Figure 3.7. These show that the initial lineage-splitting event gave rise to two lineages, one of which later split to give rise to descendant species A and B, whereas the other gave rise to C and D. As discussed more fully in Chapter 5, this means that A and B are more closely related to each other than to C and D (and conversely C and D are more closely related to each other than to A and B).

ASEXUAL ORGANISMS

The preceding characterization of phylogenetic trees applies to organisms that reproduce sexually. It is sex, and the resulting potential for gene flow, that glues local populations into cohesive evolutionary lineages. But many different types of organisms reproduce asexually or ***clonally***. So how should we visualize descent relationships in strictly asexual organisms or in a clone of cells?

In asexual organisms each organism has only one parent, which contrasts with sexual organisms in which two parents are required for reproduction. Asexual reproduction, unlike sexual reproduction, is treelike down to the level of individual organisms (or cells within a developing organism). Consider the growth of an asexual aphid population from a single founder arriving on a host plant, as shown in Figure 3.8. If you look closely you will see that it has a perfectly treelike form—lineages split but never merge. We can therefore show the evolutionary history of these aphids in a tree form. Families derived from a single ancestral organism are equivalent to sets of species descended from a single ancestral species. Thus, all the conventions and manipulations described later

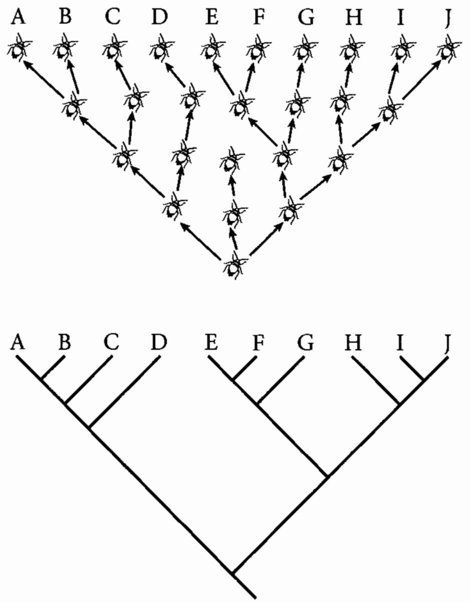

FIGURE 3.8 Treelike history of organisms in asexual lineages, as illustrated with a hypothetical aphid population.

in this chapter can be applied easily and naturally to asexual organisms, right down to the level of individual clones.

TREE TERMINOLOGY AND CONVENTIONS

A tree diagram is made up of lines, called ***branches*** (or ***edges***), connected at nodes. To be considered a tree in the formal sense, the diagram needs to be ***directed***, meaning that time runs in one direction along each branch, and ***acyclic***, meaning that lineages that diverge never subsequently fuse. Figure 3.9 shows a simple rooted tree with some of its parts labeled. The version on the left is in a rectangular format, whereas the tree on the right is in a diagonal format.

In an evolutionary context, the labels at the top of a tree could be individual species, individual organisms that represent particular species, or sets of related

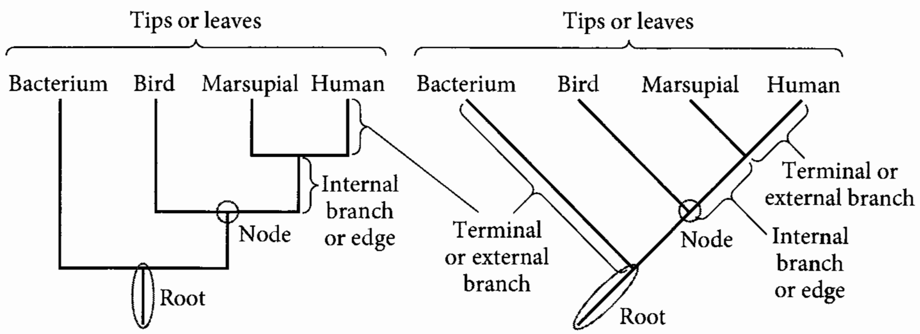

FIGURE 3.9 **Terms associated with phylogenetic trees.**

species that constitute one branch of the tree of life. In some situations they could be individual genes. The most common terms for the items represented by these labels are ***tips*** or ***leaves***, but you might also see them called ***terminals***. If the tips have scientific names, they may be called taxa (singular = taxon). The branches represent evolving lineages, whereas ***nodes*** correspond to lineage-splitting events. A node marks the last common ancestor of organisms in the daughter lineages. Whereas the ***internal branches*** (or ***internodes***) connect two nodes, ***external branches*** connect a tip and a node. The ***root*** of the tree is a special node that marks the point where time enters the diagram. The root is usually indicated by an external branch whose tip is unlabeled—generally drawn on the opposite side of the diagram to the labeled tips.

When describing trees, it can be useful to have a way to refer to a piece of a tree that is descended from one particular ancestral lineage. A ***clade*** is a piece of a phylogeny that includes an ancestral lineage and all the descendants of that ancestral lineage. Alternatively, just focusing on living taxa, a clade can be defined as a set of tips that comprise all the living descendants of one particular ancestral node. Clades have the property of ***monophyly*** (from the Greek for "single clan") and, thus, may also be called monophyletic groups. As shown in Figure 3.10, a clade or monophyletic group is easy to identify visually: it is simply a piece of a larger tree that can be cut away from the rooted part of the tree with a single cut. If one needs to cut the tree in two places to extract a set of tips, then that group is non-monophyletic and is not a clade. See Chapter 5 for further discussion of non-monophyletic groups.

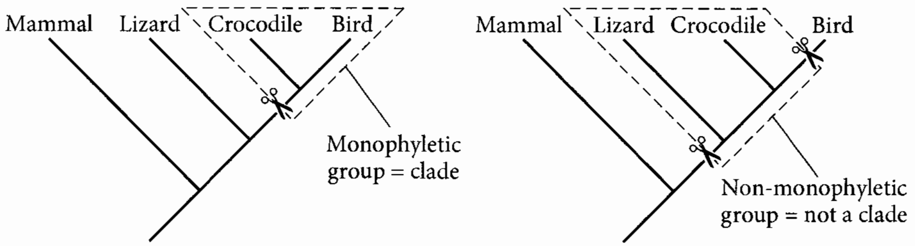
Mammal
Lizard
Crocodile
Bird
Monophyletic group = clade
Mammal
Lizard
Crocodile
Bird
Non-monophyletic group = not a clade

FIGURE 3.10 **Distinguishing a monophyletic group (or clade) from a non-monophyletic group.** Monophyletic groups can be separated from the root by a single cut, whereas separating non-monophyletic groups requires at least two cuts.

By analogy to family trees, we may refer to the two descendants of a single node as ***sister groups*** or ***sister taxa***. This convention provides a useful way to verbally describe a tree topology. For example, in Fig. 3.10, bird and crocodile are sister taxa and the lizard lineage is the sister taxon to the bird+crocodile clade. Note that the sister taxon relationship is unique; a taxon can have one and only one sister taxon. If an ancestral lineage branched simultaneously into three or more daughter lineages (discussed further below), then the daughter lineages would not have sisters.

TREE TOPOLOGY

The most basic information in a tree is the relative branching order, or ***topology***. Tree topology, that is, which lineages lead to which tips, is an important predictor of the distribution of traits among tips (Chapter 4). Tree topology also tells you which organisms are more or less closely related to each other (Chapter 5). Here, we will focus on how to correctly read tree topology because many students find this challenging. Later in this chapter we will discuss tree diagrams that include information on the amount of evolution occurring on a branch and/or the duration of branches.

If you recall the way that a phylogeny "grows" by ancestral lineages splitting, it is arbitrary which descendant lineage is shown on each side of the figure. Trees that are topologically equivalent can look quite different when different branches are positioned to the right or left. We can "spin" parts of a tree around

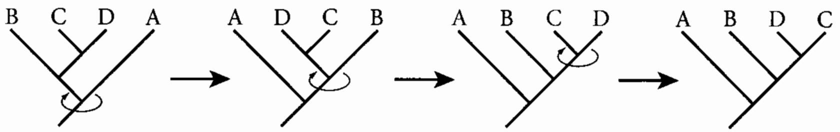

FIGURE 3.11 Rotating branches at nodes does not alter relationships. The trees can be interconverted by rotating the indicated branch.

any internal branch without changing the topology. So long as you can get from one tree to another by rotating nodes or reshaping branches, but not cutting and reattaching branches, those trees have the same topology. For example, Figure 3.11 shows four equivalent trees with an indication of which nodes need to be rotated to get from one to the other. In each tree, (C, D) is a clade and so is (B, C, D).

Notice in Figure 3.11 that in each tree the tips are ordered differently despite having the same topology. This shows us that the left-to-right ordering of tip labels is arbitrary and should not be used to extract information from a tree diagram. In particular, the ordering of tips should never be taken to convey information about evolutionary "advancement" (Chapter 2).

For many people the kinds of mental gymnastics needed to see the equivalence of different tree topologies is challenging. These skills can be developed by playing with computer programs that allow branches to be moved around graphically (e.g., Mesquite) or by manipulating physical models of trees (e.g., made from pipe cleaners). Nonetheless, it is helpful to also know about some formal rules that can be applied to determine if two trees have the same topology.

One method is to imagine that the lineages of a tree are a set of paths with signposts at each junction, indicating the tips (villages, if you will) that are served by each alternative route. If you walked up from the root on the first tree in Figure 3.11, the first junction you would come to would have one sign pointing to village A and one pointing to B, C, and D (see Figure 3.12). We can write this out in the *splits* format: A|BCD. All that matters in this convention is which villages are clustered on each side of the vertical line. Thus, A|BCD is the same as A|DCB, BCD|A, DBC|A, and so on.

If you then walked up the BCD path, you would come to a B|CD signpost (Figure 3.12). The tree also has a C|D split, but this does not add any information that was not given by the B|CD split. You already knew that there was a road leading to C and D, so there must be a C|D signpost. Therefore, the pertinent information in the tree is summarized by the two splits: A|BCD and B|CD. If you do the same exercise on the other trees in Figure 3.11, you will see that each has the same two splits, showing that they all have the same topology.

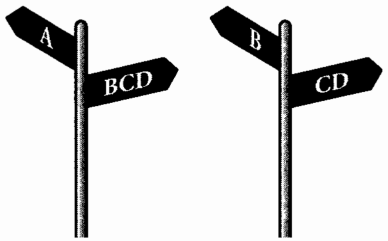

FIGURE 3.12 Trees as road signs. Branching points of trees are like forks in the path of evolution, with either side leading to a different clade or tip. The figure shows the two road signs or "splits" that are shared by all the trees in Figure 3.11.

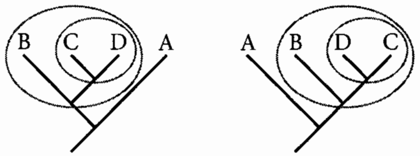

FIGURE 3.13 Comparing trees by clade composition. The two clades (CD) and (BCD) are the same, showing that these two trees are equivalent.

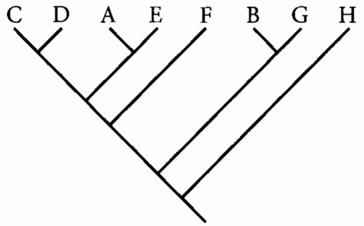

FIGURE 3.14 An eight-taxon tree.

A slightly different method is to list clades: sets of tips that are descended from a particular internal node. For example, because the two trees in Figure 3.13 contain the same two clades, (BCD) and (CD), we can see that they have the same topology.

The clade convention is not only useful for seeing if two trees are equivalent, it also provides a simple way of writing a tree topology in text format, by listing all the clades within parentheses. For example, the trees in Figure 3.13 can be rewritten in so-called Newick format as (A(B(CD))). This convention can be scaled up to an indefinite number of tips. For a slightly more complex example, the tree in Figure 3.14 can be written as (H((GB)(F((EA)(DC))))). This tree description is not easy to read by humans, but is a standard way to input trees into computer programs.

DIFFERENT TREE STYLES

There are several alternative styles in which trees can be drawn (Figure 3.15). Most trees drawn so far in this chapter have been in the ***diagonal*** or ***rectangular*** format and have had the root at the bottom so that time points up. These same formats are sometimes used in a different orientation. For example, Figure 3.15 shows the same topology drawn in a diagonal-up, diagonal-down, and rectangular-right format. This figure also shows one additional tree style, the circle tree. A ***circle tree*** has only one orientation: time always runs outward from the middle.

The four trees in Figure 3.15 are all equivalent to one another. This can be established using the clade or signpost methods or using mental gymnastics to convert one to the other by twisting and bending branches. In all cases, the tree shown has the topology (A(B(C,D))).

The choice among different tree formats is guided by practical and stylistic issues rather than biological factors. Diagonal trees are efficient because few lines need to be drawn. Psychological research has shown that diagonal trees may confuse students, who sometimes misinterpret the long diagonal line from the root to taxon D as indicating that taxa A, B, and C descended "from D." This confusion is easily overcome by recalling the meaning of a tree diagram and by becoming fluent in converting topologies from a diagonal to a rectangular format.

Rectangular trees are tidy and provide horizontal lines on which to insert text. Circle trees are useful when one wishes to include many tips in a compact space. They also have the more subtle advantage of not being easily misinterpreted as implying that evolution was headed toward one privileged tip. However, because they are tricky to draw and take practice to read, they are used only when the other formats are incompatible with the available space.

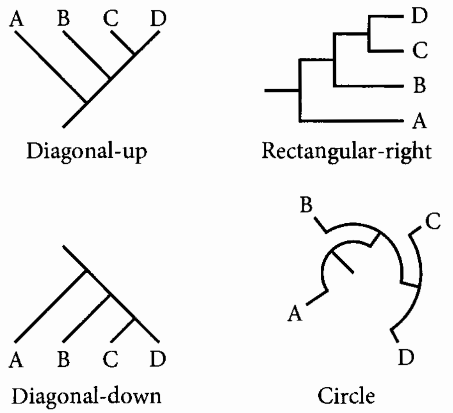

FIGURE 3.15 Four alternative representations of the same topology.

Because you may encounter a diversity of tree formats, it is important to be able to tell whether trees in different formats contain the same information. In this book we intentionally use a mixture of tree formats to give you practice in working with these different forms.

MERGING AND PRUNING

The entire tree of life is very, very big, including several million known living species. A tree depicting the relationships among a single representative of every species would be bewilderingly large. Furthermore, within a single named species, multiple tips can be recognized: subspecies, populations, or individual organisms. Also, there is no obvious limit to the number of fossils we might eventually discover, and each fossil form is best considered a tip, as discussed earlier in this chapter. In light of the immensity of the grand tree of life, the tree metaphor has utility only because it is resilient to certain simplifications—resilient in the sense that statements based on a small piece of the tree will be true for the tree as a whole. Two ways of simplifying trees are especially important: ***pruning*** and ***merging***.

Phylogenetic trees only depict relationships among the terminals that are included in the diagram. Nonetheless, the treelike form has the desirable property that pruning tips off a tree does not change the relationships of the remaining tips. For example, given the tree on the left in Figure 3.16, the pruned tree on the right correctly represents the phylogeny for the remaining species.

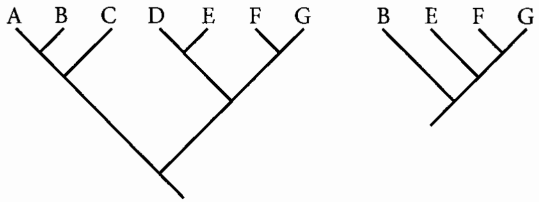

FIGURE 3.16 Pruning does not alter relationships of the included tips. Removing three tips (A, C, D) from the tree does not change the relationships among the remaining taxa (B, E, F, G).

All we have done is cut off three tips (A, C, and D) and then straightened the remaining branches. Two trees are said to be ***compatible*** if there is one larger tree topology that can be converted to either of the two trees by selective pruning of branches (one of the trees can be identical to the larger tree).

The resilience of trees to pruning is an important feature that explains why they are such good devices for communicating information. Because of this property it is possible to make accurate statements of evolutionary kinship without having to list every species that ever lived. Or, conversely, adding a newly discovered species to a well-established tree has no effect on the relationships among the species that were already included.

Stability in the face of pruning takes advantage of the fact that each tip is connected to the rest of the tree (and hence to all other tips) by only one connection. This can be illustrated by imagining a literal tree (Figure 3.17). A squirrel seeking to climb from tip A to tip Z along branches (without jumping) follows exactly the same path regardless of how many other tips and branches have been pruned off. This is a manifestation of the acyclic nature of a tree.

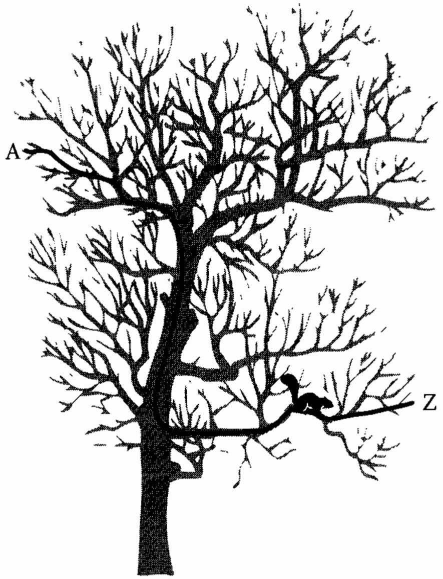

FIGURE 3.17 The path of a squirrel from leaf A to leaf Z is the same regardless of how many other branches are on the tree.

Figure 3.18 provides an illustration of tree pruning. In the upper panel we show the tree that reflects the currently accepted relationships among these mammals (all except the tenrec are in a clade called Afrotheria), and we have pruned off several tips to yield the smaller tree to the right. In the lower panel, we have taken the same "full" tree and have removed a different set of branches. While the two pruned trees may look different, they have the same topology for the species that are included. The three trees are compatible with one another.

Determining by eye whether a small tree is a validly pruned version of a larger tree takes practice. A modification of the clade method for testing

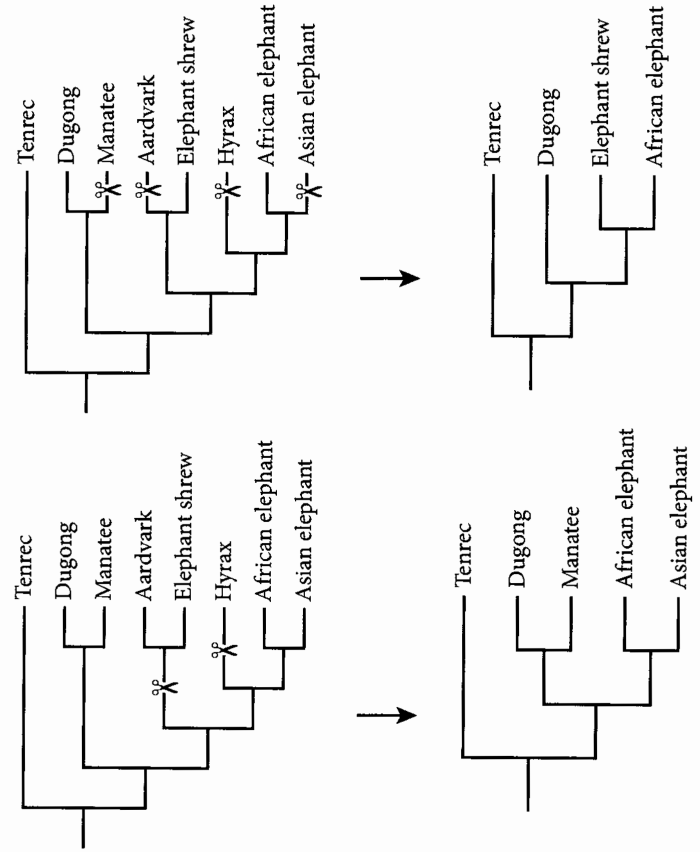

FIGURE 3.18 Two pruned versions of the same larger tree.

topological identity can be used to see if trees with different numbers of tips are compatible. The rule can be broken down into the following steps: (1) Find a clade on the smaller tree and note the tips that are included. (2) Find the smallest clade on the larger tree that includes all those tips. (3) Note any tips that are in the clade on the larger tree that were not in the clade on the smaller tree. (4) If these extra tips occur anywhere on the smaller tree, then the two trees are incompatible: the smaller tree is not a pruned version of the larger tree. (5) If, after considering all clades in the smaller tree, you do not find any cases of incompatibility, then the smaller tree is a validly pruned version of the larger tree.

For example, the upper pruned tree in Figure 3.18 has a clade that includes just African elephants and elephant shrews. The smallest clade on the unpruned tree that includes these two tips also includes the aardvark, hyrax, and Asian elephant tips. However, since none of these three taxa are anywhere on the pruned tree, we can conclude that the African elephant + elephant shrew clade is compatible with the full tree. Repeating this procedure for all the clades can show that the pruned trees in Figure 3.18 are compatible with the full trees to the left.

In addition to being resilient to pruning, trees can also be simplified by merging a clade into a single tip. Regardless of how large a clade is merged, the basic topology of the tree remains the same. For example, instead of displaying every species in clade H in the tree on the left in Figure 3.19, we can redraw the tree with 'H' merged into a single terminal.

It is important to understand that merging is a valid maneuver for clades but not for non-monophyletic groups. To see why, consider a couple of examples using the tree in Figure 3.19. First, imagine that you renamed the non-monophyletic group C + D + E as "I." Wherever you placed I on the merged tree would imply an incorrect placement for at least one tip. For example, if "I" were placed as the sister taxon to F + G, it would incorrectly suggest that C is more closely related to F + G than it is to A + B. This is a major reason why the current convention is to give formal taxonomic names only to monophyletic groups of organisms (Chapter 5).

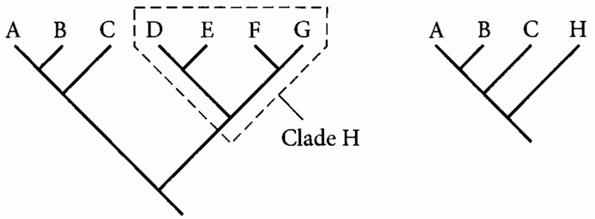

FIGURE 3.19 Merging a clade into a single tip. The two trees are identical given that taxa D–G are merged into clade H.

THE TIME AXIS

A rooted tree follows the fate of one ancestral lineage through a series of lineage branching events, usually leading to a set of living species. The tree is a historical chronicle: the nodes and branches represent ancestral populations that lived at some particular time in the past. A tree diagram must, therefore, contain some implicit information on the relative timing of different lineage-splitting events. However, you should be careful not to read too much temporal information into a tree diagram.

Two nodes that are on the same path from the root have a fixed relationship to one another. The node closer to the root represents a population of organisms that is ancestral to the other node, and therefore lived earlier. For example, in Figure 3.20, node *b* is a descendant of node *a*. The latter must represent a population of organisms that lived after the former. Node *c* is also a descendant of node *a*, and must, likewise, have lived after node *a*. To make this easier to see, the figure on the right adds arrowheads pointing along lineages from ancestors to descendants.

While the tree contains information about the relative ages of ancestral and descendant nodes, this diagram does not contain information about the relative ages of nodes that are not on the same path from the root. For example, Figure

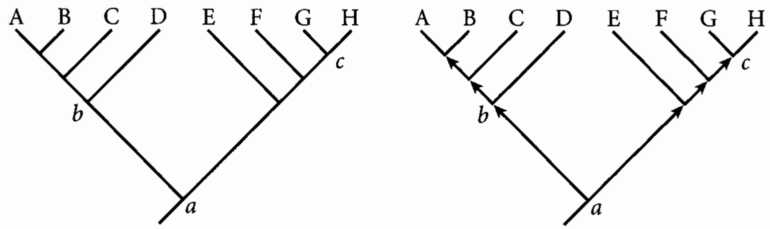

FIGURE 3.20 Trees contain information about the relative age of nodes that are on the same path from the root. We can infer that nodes *b* and *c* are both younger than node *a*, because they are both descendants of node *a* (as indicated by the arrows in the right panel). However, without explicit temporal information, we cannot determine the relative ages of nodes *b* and *c*.

3.20 does not contain information on the relative ages of nodes *b* and *c*. It might be tempting to infer that node *c* lived after node *b* because there are two nodes between *a* and *c* but no intervening nodes between *a* and *b*. This reasoning is flawed because this diagram just shows tree topology. There is no information about branch duration in this diagram. For example, the three branches between nodes *a* and *c* could each represent a short period of time, summing to less total time than the single internal branch between nodes *a* and *b*.

A convenient way to summarize the ages of nodes (when known) is to draw branch lengths proportional to time, usually with an associated scale to allow one to read off the estimated age of an internal node. Such diagrams are called ***chronograms***, because they contain information on time, as contrasted with ***cladograms***, which only depict topology and clade membership. Figure 3.21 shows a chronogram that matches the cladogram in Figure 3.20. It is assumed that all the tips shown are extant (i.e., still living), meaning that they lived zero millions of years ago. By dropping a line from internal nodes to the timescale, we can see that node *c* is older than node *b*. Chapter 11 introduces molecular dating methods that may be used for constructing chronograms.

An intermediate situation between a cladogram and a full chronogram is encountered when certain nodes or tips within a tree are assigned ages, but branch lengths are not drawn proportional to time. Let us start with a case where certain tips are fossils of known age, as shown in Figure 3.22. Given that tip F is a fossil that is dated at 55 Ma, what else can we infer?

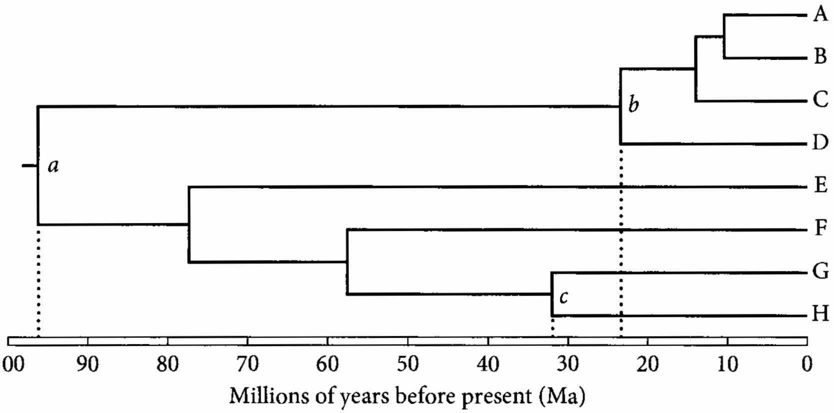

FIGURE 3.21 A chronogram showing the timing of the branching events. This figure shows that node *c* predates node *b*, something that could not be inferred from the cladogram depicted in Figure 3.20.

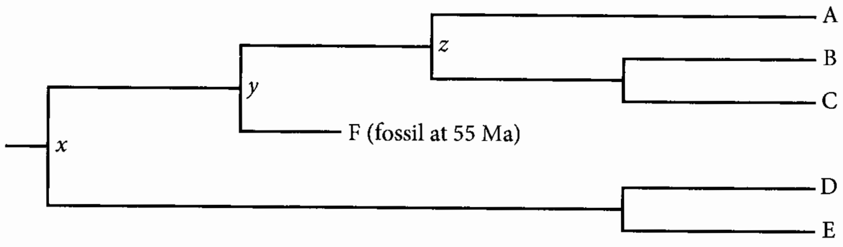

FIGURE 3.22 Using a dated fossil to place limits on nodal ages. Given that F is dated at 55 Ma, we can infer that nodes *x* and *y* are both at least 55 Ma. We cannot, however, constrain the age of node *z* or any of the unlabeled nodes without making additional assumptions.

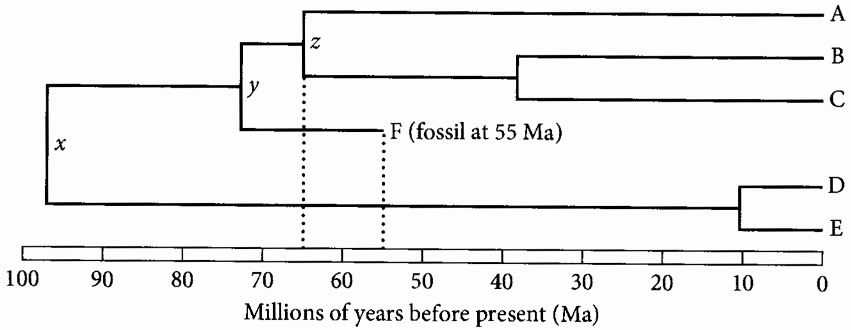

FIGURE 3.23 A chronogram showing that fossil F lived after node *z*.

Because nodes *x* and *y* are ancestral to F, it is valid to assume that they existed at least 55 Ma. It might be tempting to infer additionally that node *z* is younger than 55 Ma, based on the reasoning that it occurred after the origin of the F lineage. However, such reasoning would be invalid. The branch between node *y* and F could be of long duration, while the one leading from node *y* to node *z* could be of short duration. In that case node *z* could predate 55 Ma, as shown in Figure 3.23. Thus, based only on Figure 3.22, we have no direct information on the age of node *z* or either of the other two unmarked nodes.

BRANCH LENGTH AND EVOLUTIONARY RATE

While topology suffices for some purposes, such as delimiting taxonomic groups or inferring evolutionary relationships (Chapter 5), some downstream uses of trees require information on the length of branches, where length usually represents the relative probability that a character would change state on a particular branch. For example, if you are quantitatively analyzing patterns of trait evolution (Chapter 10), knowing that more evolutionary changes tend to occur on some branches than others can have a substantial effect on your conclusions. In many cases, expressing branch length in units of time (i.e., in the form of a chronogram) is all that is needed. In other cases, it can be useful to draw branches such that their length is proportional to the amount of evolution that is inferred to have occurred on the branch, most often expressed as the average number of changes occurring to each character used in a particular analysis (Chapter 8). Trees with branch lengths drawn proportional to the amount of evolution are called ***phylograms***.

Figure 3.24 shows a sample phylogram from a study of the evolution of cotton (*Gossypium*) and its wild relatives. The scale bar at the bottom of the diagram indicates that the branch lengths are proportional to the amount of evolutionary change. Phylograms are among the most common tree diagrams in research literature, but they are less common in secondary literature and in textbooks, where cladograms and chronograms predominate.

The branch lengths on a phylogram relate to a specific set of traits, most commonly the gene sequences that were used to infer the tree (Chapters 7–8). For example, if we were considering the evolution of the hemoglobin protein, then length might be drawn proportional to an estimate of the proportion of amino acid sites that changed on each branch. In the case of Figure 3.24, the branches are drawn proportional to the number of substitutions estimated to have occurred at each site in a portion of the plastid (i.e., the chloroplast) genome of these plants. The length of a branch is its duration multiplied by its average rate of evolution. Because the rate of evolution can vary across branches, phylograms are not the same as chronograms.

If two sister lineages are of different lengths, and if both have living representatives, then the rate of evolution must have differed between them. The longer branch has accumulated more changes and, thus, has evolved at a faster rate than the short branch. On Figure 3.24, *Gossypium* and *Hibiscus* are part of a clade called the Eumalvoideae with long branches, whereas the baobab, *Adansonia*, is in a clade called Bombacoideae with relatively short branches. How should this be interpreted? All one can safely assume is that for *this gene* the rate of molecular evolution was higher in Eumalvoideae than in Bombacoideae. It could be the case that all genes evolved more rapidly in Eumalvoideae

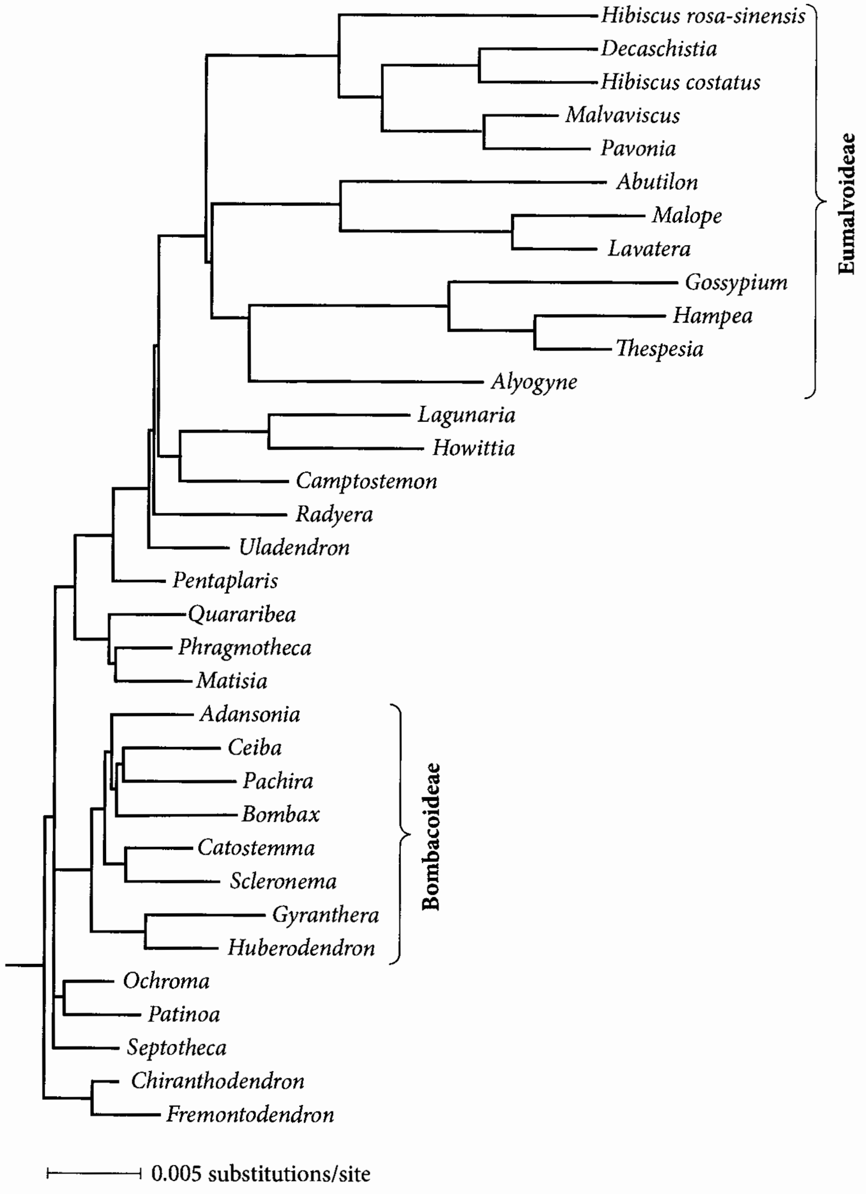

FIGURE 3.24 **Example of a phylogram.** The scale bar indicates the relationship between the branch lengths shown and the average number of substitutions that occurred at each site in the sequence. Adapted from Baum et al. (2004).

than in Bombacoideae, but this conclusion should not be drawn from a single phylogram.

COMMUNICATING PHYLOGENETIC UNCERTAINTY

Up until this point, all the trees we have shown were ***binary***: ancestral lineages split into just two descendant lineages. The splits are ***dichotomous*** (Greek for "cutting into two"). A fully binary tree is also called ***fully resolved***. In parts of the tree of life where lineage splitting is a rare event, it is probably reasonable to assume that all lineage-splitting events are dichotomous. But it is easy enough to imagine cases in which an ancestral lineage splits more or less simultaneously into multiple descendants. For example, a widespread species might become fragmented into multiple isolated populations as a result of a change in the climate. If several of these populations persisted to establish new lineages, the result would be a node with more than two descendant lineages, a ***polytomy***. Figure 3.25 compares binary and polytomous trees drawn in either diagonal or rectangular format.

It is certainly possible that true phylogenies have polytomous nodes, so-called ***hard polytomies***. More commonly, polytomies in tree diagrams indicate uncertainty as to the correct branching pattern. Recall that the trees that appear in research publications, textbooks, or websites are inferred from data: they are hypotheses of actual evolutionary relationships. Thus, even if the true tree were fully binary, the data might be insufficient to resolve all the relationships. Polytomies that are used to communicate uncertainty in the tree topology (rather than ancestral lineages actually splitting into multiple descendants) are called ***soft polytomies***.

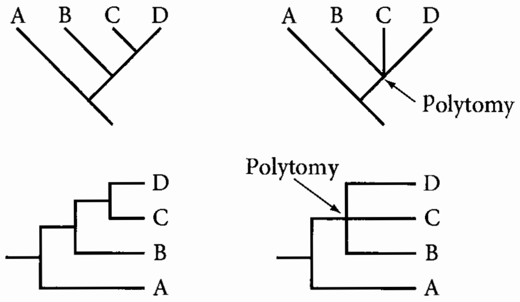

FIGURE 3.25 Polytomies express uncertainty in phylogenetic relationships. The clade containing B, C, and D is collapsed into a polytomy in the trees on the right.

Consider a study that has ruled out most of the possible trees, but cannot rule out the two trees in Figure 3.26. The remaining uncertainty can be captured with a ***consensus tree*** using the two possible trees as ***input trees***. The consensus tree shown in Figure 3.26 is a ***strict consensus tree***: a tree composed only of clades that occur on all input trees. Both input trees include the clades (ABC), (DEFG), and (FG), and therefore the consensus tree includes just these three resolved clades. Internal branches that are not present on all the input trees are collapsed into a polytomy. The clade (DE), for example, occurs only in one of the input trees and is therefore not shown on the consensus tree. There are other kinds of consensus trees, but we will not describe them here.

You may be wondering how to interpret polytomies if they can represent either an ancestral lineage splitting simultaneously into multiple descendant lineages or phylogenetic uncertainty. The safest interpretation is to view the polytomy as an indication of uncertainty, where that uncertainty includes the possibility of a hard polytomy.

Polytomies in tree diagrams represent complete uncertainty. However, it is common for an analysis of real data to yield a tree whose branches receive different levels of support by those data. Phylogeneticists have developed a

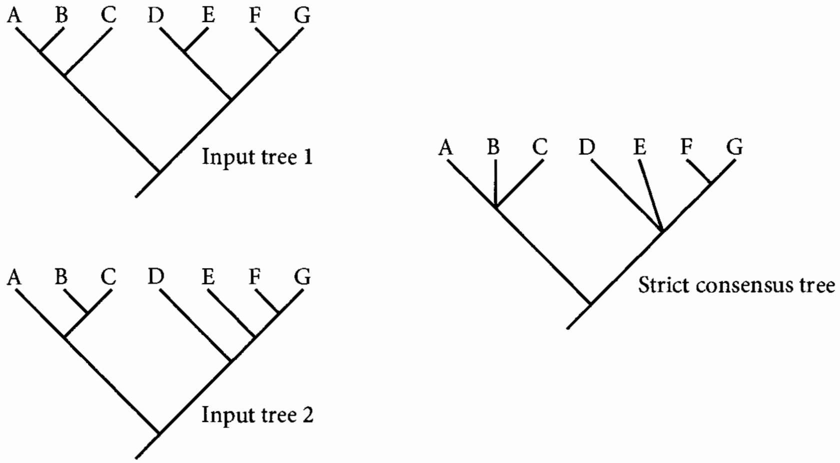

FIGURE 3.26 Combining two resolved input trees into a strict consensus tree with polytomies. The consensus tree contains only clades that are present in both input trees.

number of different ways to annotate a tree to indicate the degree of confidence that should be associated with each clade. The strength of support for a particular clade is typically indicated by placing a number on the branch subtending that clade. The most commonly used measures are ***bootstrap percentages*** (also called ***bootstrap scores***), which range from 0 to 100% and ***posterior probabilities*** (also called ***clade credibilities***), which range from 0 to 1.0. The meaning of these numbers is explained in Chapters 8 and 9. For now, it will suffice to know that the higher the number, the more strongly the data support the clade descended from the annotated branch (or node). While the thresholds are subjective, a rough rule of thumb is that clades with bootstrap scores greater than 80% or posterior probabilities greater than 0.95 are considered well supported.

As an example, consider Figure 3.27, which shows part of a phylogram from a scientific paper (Medina et al. 2001). While most of the sampled tips may be unfamiliar, you may know a number of the major clades represented. Each branch of the tree has an associated bootstrap score. Values less than 50% are typically not provided. Knowing that all internal branches have bootstrap scores, and that these scores refer to the clade descended from that branch, you can see which number refers to which clade. This allows us to determine which results were strongly supported by this analysis. For example, Figure 3.27 provides quite strong support (93%) for a clade comprising both Bilateria (the clade that includes the vast majority of living animal species, including us) and Cnidaria (e.g., jellyfish and corals). In contrast, this study provides relatively weak support (77%) for the monophyly of Cnidaria. Likewise, while this tree contradicts the monophyly of sponges, it does so only weakly; this is because the support for Calcarea forming a clade with other animals rather than the remaining sponges is only 66%. While it can take practice to read trees in this manner, an ability to do so opens a wonderfully rich array of scientific literature that summarizes phylogenetic data using these conventions.

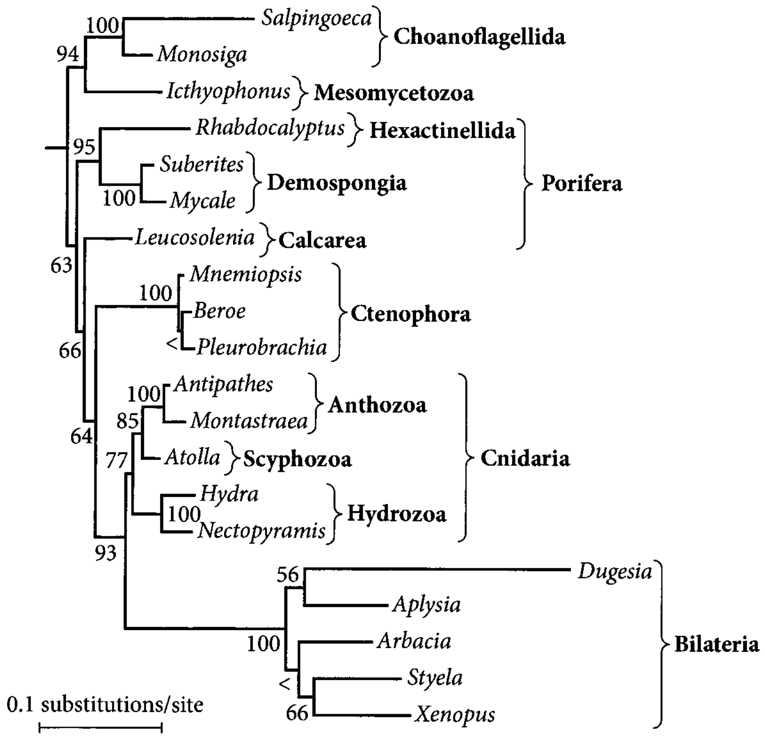

FIGURE 3.27 **A phylogeny of animals and their closest relatives.** Numbers are bootstrap scores (scores <50% are marked "<"). Scale bar as in Figure 3.24. Adapted from Medina et al. (2001).

UNROOTED TREES

Rooted trees contain information about the flow of time, which allows us to discern the pattern of descent from common ancestry and the direction of trait evolution (Chapter 4). Rooted trees are therefore essential for most downstream uses of phylogenies. While it is only necessary to understand rooted trees in order to read much of the secondary phylogenetic literature, you will need some understanding of ***unrooted trees*** and how trees are rooted if you plan to read the primary phylogenetic literature. Additionally, developing a clear sense of how rooted and unrooted trees differ can help you to achieve a deeper understanding of trees in general.

An unrooted tree is a tree without a defined root. In an unrooted tree the branches represent evolutionary lineages, but unlike a rooted tree, we do not know which way evolution preceded along the lineage. Because a clade comprises an ancestor and all its descendants, we need temporal information to identify clades. Thus, the internal branches of an unrooted tree do not denote clades but rather split the taxa into two sets of lineages that are attached (directly or indirectly) to the two ends of the branch.

As an example, consider a rooted tree for selected archosaurs (Figure 3.28a). Figure 3.28b shows an unrooted version, obtained by removing the root and collapsing the lowermost internal branch into a polytomy. The figure may seem to imply that the true root is on the crocodile lineage or between a crocodile + pterosaur clade and the dinosaurs, but this cannot be assumed. Because this is an unrooted tree, in the absence of extra information, we should be open to the possibility that the true root lies along any branch of the tree.

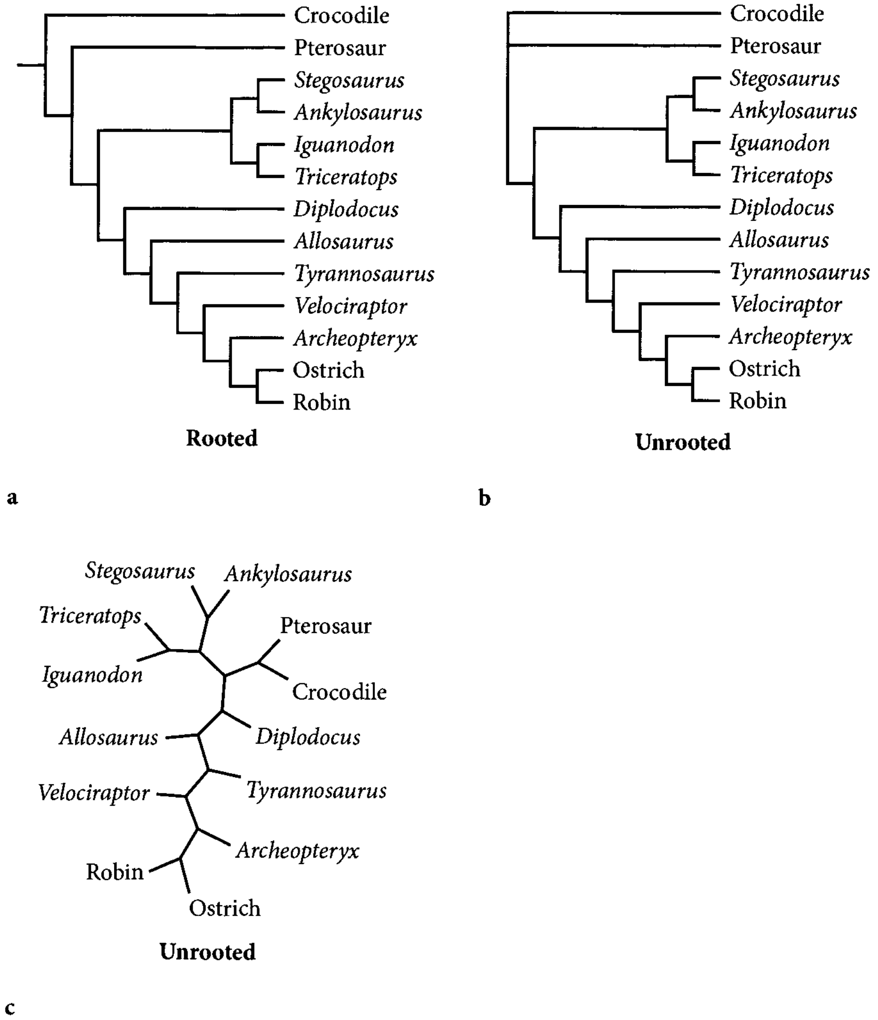

FIGURE 3.28 The same tree topology in rooted (a) and two different unrooted (b, c) tree formats.

To avoid implying a root, the tree has been redrawn in a spread-out style (Figure 3.28c). Like the other trees in Figure 3.28, this tree is binary in that each node has three branches, one corresponding to the ancestral lineage and two to descendant lineages. However, because the trees in Figure 3.28b and c are unrooted, we cannot tell which lineages are ancestral and which are descendant.

To see that the three trees in Figure 3.28 are topologically identical, confirm for yourself that to get from one to the other you need only remove the short branch leading to the root and then unbend, resize, and rotate branches. No additional branches beyond the root branch have to be cut. Once again, imagine that the trees are made of pipe cleaners (but whose length can change) and you can rearrange the first tree to yield the second or the third.

If this physical modeling is difficult, try the approach of listing clades (see earlier section on Tree Topology) to establish the topological identity of these three trees. First list the clades in the rooted tree. Then list the taxa on the unrooted tree that are separated from each other by an internal branch, using a vertical line to indicate which taxa are on which side of the internal branch. These are splits, also called ***bipartitions***. For example, as shown in Figure 3.29, there is one internal branch that divides ***Stegosaurus, Ankylosaurus, Triceratops, Iguanodon, Pterosaurus***, and crocodile on the one side from ***Diplodocus,***

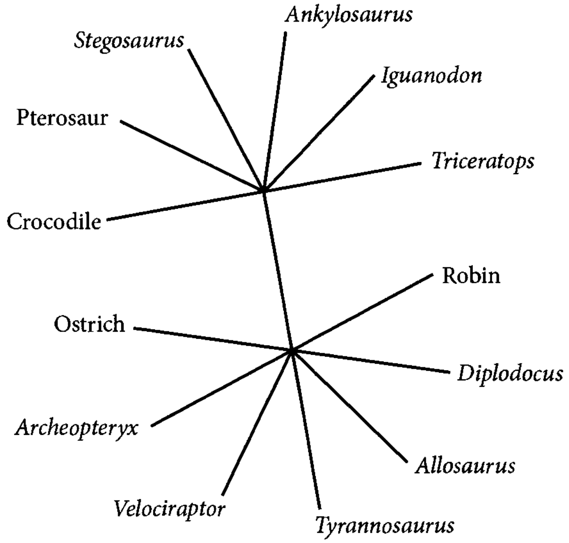

FIGURE 3.29 One of the splits (a bipartition) in the archosaur tree (Figure 3.28).

Allosaurus, Tyrannosaurus, Velociraptor, Archaeopteryx, ostrich, and robin on the other. This could be written as: *Stegosaurus, Ankylosaurus, Iguanodon, Tricerotops, Pterosaur*, crocodile | other species. Table 3.1 lists all the splits seen in Figure 3.28b and c. For each split you need to look at the rooted tree (Figure 3.28a) and ask, Does the set of taxa before the line or the set of taxa after the line or both correspond to a clade on the rooted tree? If the answer is yes for each split on the unrooted tree, and if all clades on the rooted tree correspond to one of the splits, then the rooted and unrooted trees match.

Assuming that you know where the root should go on an unrooted tree, you can easily convert an unrooted tree back into a rooted tree. Rooting an unrooted tree just involves adding an additional node to one of the branches and reorienting the tree relative to that node. Figure 3.30 illustrates three ways to root the same unrooted tree.

You may wonder, How do I decide where to place the root on an unrooted tree? When reconstructing trees (Chapters 7 and 8), scientists generally use one of two methods to decide on how to root the trees. Most commonly, an analysis will include a group that is known to be outside of the group whose relationships are being studied: an ***outgroup***. When an unrooted tree is obtained from a study, the root is placed between the ingroup and the outgroup. Alternatively, in some situations, for example, when a ***molecular clock*** applies (Chapter 11), we can infer the position of the root based on the relative lengths of different branches.

TABLE 3.1 List of splits in Figure 3.28b and c

Stegosaurus, Ankylosaurus | other species
Iguanodon, Triceratops | other species
Stegosaurus, Ankylosaurus, Iguanodon, Triceratops | other species
Pterosaur, Crocodile | other species
Stegosaurus, Ankylosaurus, Iguanodon, Triceratops, Pterosaur, Crocodile | other species
Allosaurus, Tyrannosaurus, Velociraptor, Archaeopteryx, Ostrich, Robin | other species
Tyrannosaurus, Velociraptor, Archaeopteryx, Ostrich, Robin | other species
Velociraptor, Archaeopteryx, Ostrich, Robin | other species
Archaeopteryx, Ostrich, Robin | other species
Ostrich, Robin | other species

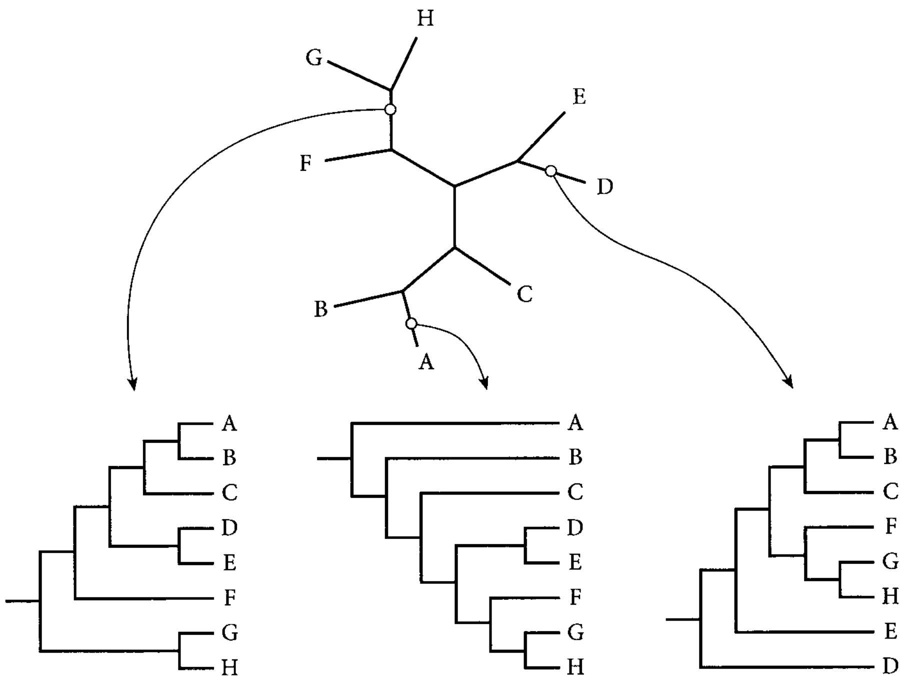

FIGURE 3.30 **Three alternative ways to root the same unrooted tree.**

TREE-TO-TREE DISTANCES

When working with two trees that are not identical, it may be useful to determine how similar they are to one another. For a number of purposes, it is valuable to quantify this as the degree of difference, that is, the ***tree-to-tree distance***. It is possible to quantify tree-to-tree distance while taking into account both topology and branch lengths. However, since branch lengths make the analysis significantly more complicated, we will only introduce methods that consider tree topology while ignoring branch lengths.

There are several methods to measure the distance between two tree topologies, of which we will mention two. For simplicity we will only consider trees that are fully resolved (see earlier section on Communicating Phylogenetic Uncertainty), although these basic approaches can be adapted to handle polytomous trees. The first way to quantify the distance between two tree topologies is to count the proportion of shared clades. For example, Figure 3.31 shows three trees, each with seven clades. Tree 1 has two clades in common with tree 2, (ABCDEFGH) and (AB), and five discordant clades. Tree 1 has three clades in common with tree 3, (AB), (GH), and (ABCDEFGH), and four discordant clades. The number of discordant clades is a measure of distance: tree 1 has a distance of 5 to tree 2 and a distance of 4 to tree 3. This suggests that tree 1 is more similar to tree 3 than to tree 2.

A problem with counting clades as a way to measure the distance between trees is that simple rearrangements can disrupt many clades simultaneously. If you compare trees 1 and 2 more closely, you will see that the entire difference between them is due to the movement of one clade, (AB), which is a sister group to taxon C in tree 1 and to taxon G in tree 2. This suggests that a more appropriate way to measure the distance between two topologies is to count the number of tree rearrangements needed to convert one topology into another.

There are several tree rearrangement methods, of which we will introduce one: ***subtree pruning and regrafting***, or SPR. As the name indicates, this maneuver entails cutting a piece off the tree and reattaching it in a new location. The word ***subtree*** rather than clade is used because single tips can be pruned and regrafted. Also, since SPR rearrangements are usually applied to unrooted trees (Figure 3.32), it is unclear which of the two subtrees is a clade.

While it may be difficult to calculate for large trees, it is usually possible to determine the minimum number of SPR rearrangements needed to convert one specific tree topology into another. This can be used as a measure of tree topology distance. In the case of Figure 3.31, converting tree 1 into tree 2 can be achieved by repositioning the (AB) clade. This shows that these two trees are separated by just one SPR, giving a distance of 1. Similarly, we can determine that it takes at least three SPRs to convert tree 1 into tree 3, giving those trees a distance of 3. Thus, in terms of SPR rearrangements, tree 1 is closer to tree 2 than to tree 3.

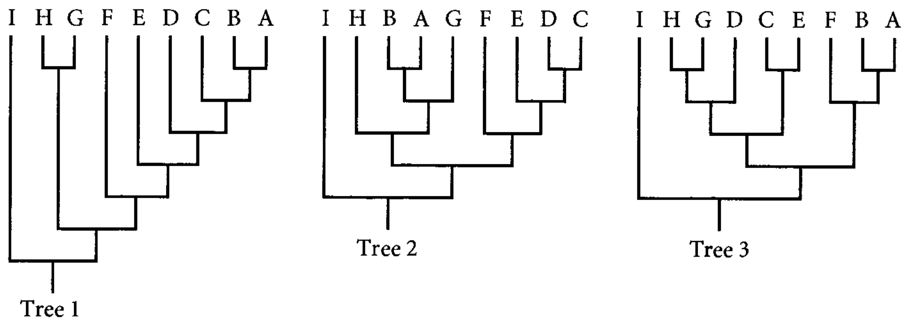

FIGURE 3.31 An example illustrating tree-to-tree distance. Tree 1 shares two clades with tree 2 and three clades with tree 3. It takes only one SPR rearrangement to convert tree 1 into tree 2 but three to convert it into tree 3.

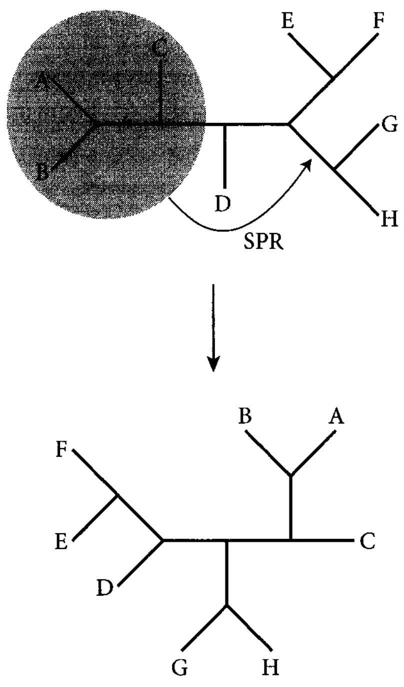

FIGURE 3.32 An example of subtree pruning and regrafting on an unrooted tree.

Before leaving tree-to-tree distances and tree rearrangement, it is worth noting that it is possible to convert one tree topology into *any* other tree topology by doing a series of SPR rearrangements. This tells us that there is a continuous, multidimensional "space" of tree topologies and that one can move through this space by rearranging trees SPR by SPR. As will be discussed more fully in Chapter 7, this ability to traverse tree space is what allows a computer program to search systematically for the optimal trees, even when the analyses include very many taxa.