CHAPTER SEVEN

Phylogenetic Inference with Parsimony

Given the importance of phylogenetic trees in modern biology, it is important to know enough about the reconstruction of phylogenetic trees, ***phylogenetic inference,*** to understand why we might (or might not) wish to accept trees generated by evolutionary biologists. Additionally, phylogenetic inference serves as an excellent example of the general principles that allow scientists to elucidate events that happened in the past. As a result, even if you are never likely to do phylogenetic research, it is worth knowing something about phylogenetic analysis so that you can appreciate the rigor of historical sciences such as evolutionary biology, geology, paleoclimatology, and cosmology.

Nowadays there are several alternative methods for phylogenetic inference, most of which proceed via the same basic steps: constructing a data matrix, identifying trees that are most compatible with the data matrix, and then conducting statistical analyses to evaluate how confident we should be in our phylogenetic conclusions. In this chapter we describe the first two steps in this process, focusing on the method of maximum parsimony and its historical predecessor, Hennigian inference. Parsimony is just one of a variety of methods for phylogenetic inference. It provides a useful starting point for understanding how phylogenetic trees are estimated and can serve as a foundation for the introduction of model-based methods (Chapter 8).

A BIOLOGICAL EXAMPLE: CARNIVORA

To provide a context for the discussion of methods of phylogenetic inference, we will use a simplified biological example, a study of the Carnivora. Carnivora

is a group of mostly meat-eating mammals, including dogs, cats, bears, weasels, mongooses, skunks, and seals. While these animals differ greatly in their external appearance and ecology, they share several skeletal features. For example, almost all species have enlarged side teeth, carnassials, which may be used for shearing meat (Figure 7.1), six incisors, and two well-developed canines in each jaw. Based on these and other traits, it has long been accepted that the Carnivora is a monophyletic group, a clade. But what are the relationships within Carnivora?

Before embarking on a study of Carnivoran phylogeny, we need to decide which species to select for our study and which traits to score. With around 250 carnivoran species, we cannot easily examine all living forms. How many and which species to include in a study is governed somewhat by the specific questions we wish to answer. Let us say that our pressing concern is to find out if the carnivores are divided into two monophyletic subgroups, the aquatic pinnipeds (seals, sea lions, and walruses) and the terrestrial fissipeds (all others). This is a long-standing hypothesis that predates the development of formal phylogenetic methods. To answer this question, we would have to sample representative species from most of the carnivoran families (dogs, cats, seals, sea lions, etc.). While a research scientist would likely include multiple representatives for each family (to test family monophyly and minimize the chances of artifactual results), we will use just one species from each of ten families (three pinnipeds and seven fissipeds) to simplify the example.

Now that we have chosen which taxa to include, we must decide which characters to use. Any trait that varies among tips and is thought to show some

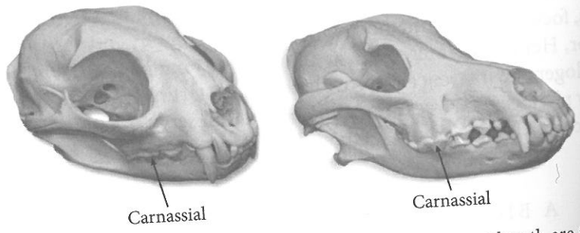


FIGURE 7.1 Bobcat and Mexican grey wolf skulls. Upper carnassial teeth are indicated. Images used with permission from www.SkullsUnlimited.com.

degree of heritability has the potential to provide phylogenetic information. Appendix 1 summarizes the main classes of data that may be used for phylogenetic analysis. Until the early 1990s, phylogenetic analyses were usually based on morphological traits. With the advent of modern molecular methods, almost all phylogenetic studies now employ DNA sequence data. However, it will be instructive to begin with a consideration of morphological characters. We will do this by extracting morphological data from a published study by Wyss and Flynn (1993).

There is one final consideration before we can begin to collect data: we must choose one or more ***outgroup*** taxa. Outgroups serve as a point of comparison with the ***ingroup*** (here, the carnivorans), allowing us to root our trees and determine the direction (polarity) of character change. Any taxon that is not a member of Carnivora could theoretically serve as a valid outgroup. However, the best outgroups are reasonably closely related to the ingroup so that traits are more easily compared between the ingroup and outgroup. For our analysis of morphological data we will use as an outgroup the creodonts, an extinct group of mammals. The reason that they can serve as an outgroup is not because they are extinct. Rather, it is because they fall outside of the Carnivora clade as indicated by the fact that they lack certain shared derived characteristics or synapomorphies of Carnivora, such as the bony auditory bulla that encloses the inner ear. A living group that is known to be outside the Carnivora clade would do just as well provided that it had characters that could readily be compared to those found in Carnivora.

Now we can proceed to score the ingroup and outgroup taxa for the morphological traits selected. Appendix 2 provides more information on the complexities of building a data matrix based on morphological data. The species are scored for each trait by observing an individual or a few individuals from that species and recording the form of that trait, its character state. The 12 characters and the states for each character are given in Table 7.1. We have assigned 0 to all of the character states present in the outgroup. However, no significance should be attached to this convention: choosing different labels for the states would not affect the results.

Moving through the tips, we record the character state for each character for each species to build a ***character state matrix***. For example, we observe that, for trait 6 (the tail), creodonts have the "elongated" state. We have chosen to represent this state with a 0, so creodonts are given a score of 0 for trait 6 in the

TABLE 7.1 Characters and character states for an analysis of carnivorans, with numerical representation (as used in Table 7.2) provided in parentheses

No.	Character	States
1	Complexity of the mucus-coated surfaces in the nose (maxilloturbinals)	Minimally branched (olfactory surfaces in nasal passage) (0); highly branching (olfactory surfaces excluded from the nasal passage) (1)
2	Bony spur by the auditory bulla (paroccipital process)	Straight and projecting (0); cupped around auditory bulla (1)
3	Number of lower incisors	2 (0); 3 (1)
4	Upper molar 1	Present (0); absent (1)
5	Baculum (bone within the penis)	Present (0); absent (1)
6	Tail	Elongated (0); short (1)
7	Hallux (5th digit, or dewclaw, on hind leg)	Prominent (0); reduced or absent (1)
8	Claws	Nonretractable (0); retractable (1)
9	Prostate gland	Small and simple (0); large and bilobed (1)
10	Kidney structure	Simple (0); conglomerate (1)
11	External ear (pinna)	Present (0); absent (1)
12	Testis position	Scrotal (0); abdominal (1)

matrix. Likewise, dogs, cats, and many other carnivorans have long tails so they also get state 0, whereas bears, seals, sea lions, and walruses have short tails and are assigned state 1.

The complete character state matrix for the 12 characters for 10 ingroups and 1 outgroup is given in Table 7.2. Notice that some taxa may be scored as unknown for certain characters (conventionally represented with '?'). This could be because we are ignorant as to the proper scoring (e.g., soft tissues in a fossil) or because it is impossible to score (e.g., toe number in snakes). While not present in this matrix, a taxon whose members may express different

TABLE 7.2 Morphological character state matrix for carnivorans, with creodont included as the outgroup

	Character state scoring											
Taxon	1	2	3	4	5	6	7	8	9	10	11	12
Creodont	0	0	0	0	0	0	0	0	?	?	?	?
Cat	0	1	0	1	0	0	1	1	1	0	0	0
Hyena	0	1	0	1	0	0	1	0	1	0	0	0
Civet	0	1	0	0	0	0	0	0	1	0	0	0
Dog	1	0	0	0	1	0	0	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1	0	0
Otter	1	0	0	0	1	0	0	0	0	1	0	0
Seal	1	0	1	0	1	1	0	0	0	1	1	1
Walrus	1	0	1	0	1	1	0	0	0	1	1	1
Sea lion	1	0	1	0	1	1	0	0	0	1	0	0

character states can be scored as ***polymorphic*** by listing multiple states within a cell.

As you might imagine, it can be difficult to find a large number of morphological traits that show appropriate levels of variation for reconstructing a phylogeny. In comparison, it has become quite easy to obtain large amounts of DNA sequence data. Table 7.3 shows some DNA sequence data for the carnivorans. Because DNA is unavailable for creodonts, an alternative outgroup, a mole, has been substituted. Like the creodonts, we can be sure that moles are outside the Carnivora clade.

Whereas for morphological data the character states were coded as 0's and 1's, the states of DNA are the 4 bases (A, C, G, and T). You may also observe *gaps* (marked with a hyphen) in the DNA matrix. These gaps arise when bases are inserted or deleted during the course of evolution. The process of ***sequence***

TABLE 7.3 The states for 15 consecutive positions in the transthyretin 2 gene, with mole included as the outgroup

	Positions in DNA sequence														
Taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Mole	G	T	T	A	A	-	C	T	T	C	T	C	A	C	T
Cat	G	T	T	G	A	-	C	C	T	C	T	T	A	C	T
Hyena	G	T	T	G	A	-	C	C	T	C	T	C	A	C	T
Civet	G	T	T	G	A	-	C	C	T	C	T	C	A	C	T
Dog	G	T	T	A	A	G	C	A	T	C	T	G	C	C	T
Raccoon	G	T	T	A	A	G	G	G	T	C	T	G	C	C	T
Bear	C	T	T	A	A	G	T	G	T	C	T	G	C	C	T
Otter	G	T	T	A	A	G	G	G	T	C	T	G	C	C	T
Seal	G	T	A	A	A	G	C	G	T	C	T	G	C	C	T
Walrus	G	T	A	A	A	G	C	G	T	C	T	G	C	C	T
Sea lion	G	T	A	A	A	G	C	G	T	C	T	G	C	C	T

alignment is concerned with establishing the correct position of gaps so that homologous sequence positions are aligned above one another in the data matrix. Sequence alignment will be discussed more fully later in this chapter.

HENNIGIAN INFERENCE

In the middle part of the twentieth century, the German entomologist Willi Hennig and colleagues developed the first formal method for phylogeny reconstruction. This was described in Hennig's 1966 book, *Phylogenetic Systematics*, which is generally credited with launching the modern field of phylogenetics. While Hennigian inference (or Hennigian "argumentation") is no longer used, we believe it is worth knowing about. The method played an important role in the historical development of phylogenetics. It is simple to understand and illustrates the general point that the distribution of trait variation among

taxa contains information about evolutionary relationships. Moreover, by understanding the problems with Hennigian inference, and the reasons why it has been replaced by other methods, one can gain a clearer appreciation of the distinction between algorithmic and optimality methods for phylogeny reconstruction.

Hennigian inference makes two major assumptions: (1) There is a strictly treelike phylogenetic history, and (2) there is no homoplasy—each character evolved from an ancestral (plesiomorphic) to a derived (apomorphic) state once, without subsequent reversal. In other words, Hennigian methods require that there be no back mutations and no independent forward mutations. Given these assumptions, a set of tips sharing an apomorphic state must be a clade.

Let us apply this principle to the morphological data for Carnivora. The first problem is to determine which character states are ancestral and which are derived. For example, did dewclaws evolve within the group, or were dewclaws lost? While many methods have been proposed for determining character polarity (Chapter 4), the most widely used is the outgroup method. If a state is variable in the ingroup, the state that occurs in the outgroup is the ancestral state for the ingroup. If no character states are shared between the ingroup and the outgroup, the ancestral state for the ingroup cannot be determined. In this case the outgroup creodonts have dewclaws (character 7, Tables 7.1 and 7.2), so "dewclaws present" is the plesiomorphic or ancestral character state.

You may notice that creodonts are scored as uncertain ('?') for the soft tissue traits: kidneys, prostate gland, ears, and testes. By examining additional living outgroups, we can determine that the ancestral state for these four characters was probably also state 0. A modified data matrix with a hypothetical outgroup having all ancestral states is shown in Table 7.4.

The Hennigian method of phylogenetic inference involves identifying sets of taxa that share a derived character state and inferring that they form a clade. For example, character 11 supports a seal + walrus clade, and character 4 supports a cat + hyena clade. Applying this method we can now draw a tree that contains all of the clades suggested by the 12 morphological characters (Figure 7.2). To make it easier to interpret, we have indicated the numbers of the characters that support each clade.

Prior to Hennig, scientists had lacked well-defined protocols for phylogeny reconstruction. Hennigian inference provided a clear, objective method for using observed trait variation to reconstruct evolutionary history. As a result, this method revolutionized evolutionary biology and stimulated the emergence

TABLE 7.4 Morphological data matrix for carnivorans

	1	2	3	4	5	6	7	8	9	10	11	12
Outgroup	0	0	0	0	0	0	0	0	0	0	0	0
Cat	0	1	0	1	0	0	1	1	1	0	0	0
Hyena	0	1	0	1	0	0	1	0	1	0	0	0
Civet	0	1	0	0	0	0	0	0	1	0	0	0
Dog	1	0	0	0	1	0	0	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1	0	0
Otter	1	0	0	0	1	0	0	0	0	1	0	0
Seal	1	0	1	0	1	1	0	0	0	1	1	1
Walrus	1	0	1	0	1	1	0	0	0	1	1	1
Sea lion	1	0	1	0	1	1	0	0	0	1	0	0

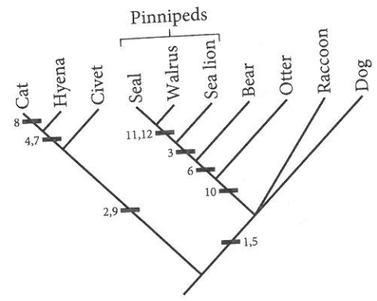


FIGURE 7.2 **Phylogeny deduced from the data in Table 7.4 using Hennigian inference.** Bars represent the origin of the derived character state of the character number(s) indicated.

of many other methods of phylogenetic inference. Although the Hennigian method described here was popular in the 1960s and 1970s, it is no longer in use. The core problem of Hennigian inference is that it makes unrealistic assumptions about trait evolution and provides no clear way of proceeding when those assumptions are not met.

The Hennigian method assumes that homoplasy is absent. But homoplasy does happen. Some characters secondarily evolve to closely resemble an ancestral state, and sometimes two indistinguishable traits evolve in parallel. Furthermore, even if evolution were strictly Hennigian, we should not expect all the traits, *as we score them*, to behave in a strictly Hennigian manner. We will sometimes make mistakes in character scoring, and we will sometimes make errors in the determination of character polarity.

If we examine additional carnivoran characters from the larger morphological matrix put together by Wyss and Flynn (1993), we can quickly find cases that serve to prove that Hennig's rules do not apply. Consider a thirteenth character: the presence/absence of lower premolar 1. This tooth occurs in the outgroup and all ingroups except cat, hyena, and otter. The absence of lower premolar 1 in cat, hyena, and otter suggests that these three form a clade. However, such a clade would require homoplasy in several other characters, for example, character 1 (branching of the turbinal bones). A tree that is fully consistent with character 13 would be inconsistent with character 1. Because there is no tree that is consistent with both characters 1 and 13, we know that Hennig's rules were broken.

If we allow the possibility of homoplasy, it becomes possible to reconcile any data matrix with any tree. But in that case Hennig's logic can no longer be used to deduce the true tree. Once the model is violated, the Hennigian deductive logic cannot tell us which tree is correct. Instead we need an ***optimality criterion***, a metric that can be used to decide, given some data, which trees are better and which trees are worse. Among the first optimality criterion proposed to replace Hennigian inference was ***maximum parsimony***.

THE MAXIMUM PARSIMONY CRITERION

Once we acknowledge that traits sometimes show homoplasy, one logical way to proceed is to allow that some homoplasy occurred, but to minimize the amount of homoplasy. This is an application of the principle of parsimony. We introduced this principle in Chapter 4 as a method for reconstructing the evolutionary history of a character, given a tree. Here we are using it in a slightly different way. In this context, the maximum parsimony criterion holds that the best estimate of phylogeny is that tree which explains all of the observed data by invoking the least homoplasy, which is to say, the fewest character state

changes. Referring back to the consistency index (Chapter 4), parsimony selects the tree that maximizes the average consistency index of the characters in the matrix. The simplest implementation of parsimony proceeds in three steps:

1. For a single tree, we consider each character in turn and determine the minimum number of character state changes, or ***steps***, that are required to account for the distribution of states among tips (see Chapter 4).
2. We sum up the number of steps required by each character. The number of steps required to explain all of the characters' evolution is called the ***tree length***.
3. We repeat the preceding steps for all alternative trees and then identify the tree with the lowest tree length, which is the ***shortest*** or ***most parsimonious tree***.

Before applying parsimony to the Carnivora data set, let us consider a simple data matrix for four taxa (Table 7.5). If we assume that taxon O (the outgroup) is the sister taxon to the remaining species, that is, that taxa A–C form a clade, then three trees are possible (Figure 7.3).

We can start by considering tree 1 and seeing how we can explain each character in turn. If tree 1 were true, the simplest way to explain the first character is that all lineages began with state 0, but that a single change to state 1 occurred

TABLE 7.5 A simple morphological data matrix for four taxa

	1	2	3	4	5	6	7	8
O	0	0	0	0	0	0	0	0
A	0	1	0	0	0	1	1	0
B	1	1	0	1	1	1	1	1
C	0	0	1	1	0	0	0	0

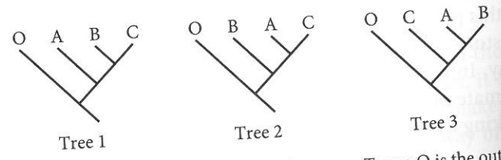


FIGURE 7.3 The three possible rooted trees for four taxa. Taxon O is the outgroup and taxa A, B, and C constitute the ingroup.

somewhere on the lineage leading to taxon B. Thus, we can explain this character with a single step: character 1 has length 1 on tree 1.

Character 2 is more difficult to map onto tree 1. There is no way to explain the distribution of states among the tips with only one change, but there are three ways to do it with two changes. These three equally parsimonious reconstructions are shown in Figure 7.4. The first scenario entails two independent transitions to state 1 (from state 0), the second entails two independent transitions to state 0 (from state 1), and the third entails one change to state 1 and one reversal back to state 0. When inferring a phylogeny, we do not need to know which of these reconstructions is correct. All that matters is that it takes a minimum of two changes to map character 2 onto tree 1: character 2 has length 2 on tree 1.

Using the same approach we can now map all eight characters onto tree 1. Characters 1, 3, 4, 5, and 8 each have only one most parsimonious reconstruction, whereas characters 2, 6, and 7 each have multiple, equally parsimonious reconstructions. For those characters, we have arbitrarily selected one of the most parsimonious mappings in Figure 7.5. In total, 11 steps are required to explain all the characters' evolution: tree 1 has a length of 11 for these data.

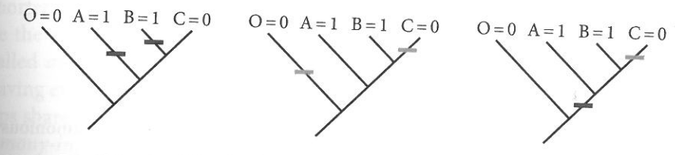


FIGURE 7.4 Alternative histories for character 1. Black bars: change from 0 to 1. Gray bars: change from 1 to 0.

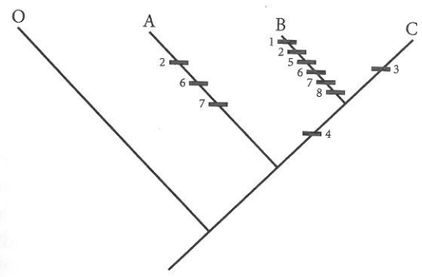


FIGURE 7.5 Tree 1 with all character state changes mapped.

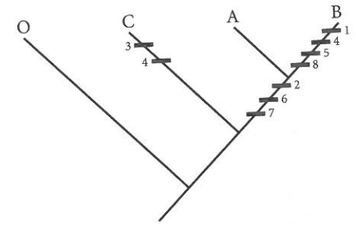


FIGURE 7.6 Tree 3 with all character state changes mapped.

TABLE 7.6 The length of each character on each of the possible trees.

	1	2	3	4	5	6	7	8	
O	0	0	1	0	1	1	0	0	
A	0	1	1	0	1	0	1	0	
B	1	1	1	1	0	0	1	1	
C	0	0	0	1	1	1	0	0	Total length
Length on tree 1	1	2	1	1	1	2	2	1	11
Length on tree 2	1	2	1	2	1	2	2	1	12
Length on tree 3	1	1	1	2	1	1	1	1	9 ← Most parsimonious

We can now apply the same procedures to the other trees. For these same data, tree 2 has a length of 12, whereas tree 3 has a length of 9 (Figure 7.6). This tells us that tree 3 is the most parsimonious tree and is the one that would be preferred under the maximum parsimony optimality criterion.

The length of each tree is a summation of the number of steps required to explain each character on that tree. As Table 7.6 shows, the tree length corresponds to the sum of the length of each of the eight characters.

It is worth highlighting that, although the optimal (most parsimonious) tree has the shortest length overall, it is not optimal for all characters. Character 4 has a length of two on the optimal tree (and on tree 2), but a length of one on tree 1. Character 4 can be said to support tree 1 over trees 2 and 3. However, the totality of the evidence still favors tree 3 over tree 1.

TABLE 7.7 Examples of informative and uninformative characters, with ? used to represent uncertain or missing character states.

Taxon	Parsimony-informative							Parsimony-uninformative						
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	1	0	?	1	0	1	0	0	0	1	0
3	0	1	1	2	1	0	2	0	1	1	0	0	2	0
4	1	1	1	1	?	1	2	0	1	0	1	1	2	1
5	1	1	2	3	1	1	0	0	1	2	?	2	2	2
6	1	0	2	2	1	?	1	0	1	0	?	3	2	?

A second observation can be drawn from this table. Characters 1, 3, 5, and 8 have the same length on each of the three trees. As a result, they do not contain information that helps choose among trees—they are ***parsimony-uninformative***. If these characters were deleted, each tree would be four steps shorter than shown, but the rank order and the difference in tree length would be the same. These characters are uniquely derived characters (sometimes called ***autapomorphies***). Because they can be parsimoniously explained as having evolved on a terminal branch of the tree, they do not help tell us which tips share more recent common ancestry. The other four characters are ***parsimony-informative*** in that their length varies among trees. Only parsimony-informative characters have the potential to influence which tree is optimal under the parsimony criterion.

You may wonder whether it is possible to know if a character is uninformative before looking at the length of all the possible trees. Here is a simple rule of thumb: a character is parsimony-informative if there are at least two states that each occur in two or more taxa. Some examples of parsimony-informative and -uninformative characters are shown in Table 7.7.

A JUSTIFICATION OF PARSIMONY

The aim of phylogenetics is to choose the tree that is most likely to be true given all of the observed trait data and our prior understanding of the evolutionary

process. Why would we expect more parsimonious (shorter) trees to be better estimates of the true tree than less parsimonious (longer) trees? Why does a set of data for which tree 1 is shorter than tree B suggest that tree A is a better hypothesis than tree 2?

In Chapter 4 we described a metaphor for the principle of parsimony, in which an emergency call center in a North American city receives two calls in the same day reporting a tiger on the loose. Because reports of tigers are rare, it is logical to assume that the two calls refer to the same tiger. Likewise, if characters change state relatively rarely, then when two tips share the same derived state, it is more likely that the trait evolved once in the tips' common ancestor than that it evolved twice. Thus, taxa sharing a derived character state are *a priori* likely to form a clade. This does not mean that the derived character state absolutely must have evolved once, just that this is the more likely explanation for this character's evolution, taken in isolation.

This principle can be extended to consider all of the characters in a data matrix. Let us start by assuming that the rate of evolution of all characters on the true tree is reasonably low. In that case, the true tree should be able to explain most of the characters' evolution by invoking few evolutionary events, so the length of the tree will be relatively short. An incorrect tree, in contrast, may be able to explain the evolution of some characters without invoking homoplasy, but for most characters we expect them to show some homoplasy on the incorrect tree. Some homoplasy is also expected on the true tree, but the amount of homoplasy should be less than on the incorrect tree. Thus, we expect the true tree to be shorter than incorrect trees.

But what if all traits evolve very rapidly? Should we still expect the true tree to be shorter than the incorrect tree? Because many squirrel sightings occur every day in a typical North American city, the fact that two squirrel sightings occur on the same day is not compelling evidence that the squirrels observed are one and the same. Similarly, if a trait has evolved rapidly, the shared occurrence of a derived character state is not compelling evidence that the taxa with the derived state form a clade. Nonetheless, unless there is a reason to expect different characters to show the same homoplasious pattern, we expect different rapidly evolving characters to support different trees.

If the rate of evolution for all characters in a data matrix is very high, the data should lack a consistent signal favoring one tree over another: that is, all trees should be about the same length. Some simple statistical methods are available to detect cases in which there is no phylogenetic signal in a data set (see Chapter 9).

But what if the rate of evolution is low for some characters but high for others? In this case the rapidly evolving characters should tend to yield a noisy pattern that will not strongly favor any one tree over the others. Still, provided that there are enough slowly evolving characters, these will tend to agree with one another and should collectively support trees that resemble the true tree. Thus, parsimony should still tend to point toward the true tree.

Based on this reasoning, we can see that when the rate of evolution is low, at least for some characters, parsimony is a reasonable tool for inferring phylogenetic trees: shorter trees are more likely to be true than longer trees. Furthermore, tree length provides a crude measure of how much better one tree is than another. Thus, if tree 1 is one step shorter than tree 2 but 15 steps shorter than tree 3, we can say that the data argue against tree 3 more strongly than against tree 2. However, the magnitude of the length difference between two trees is dependent on the particular matrix of characters used. Without other analyses (such as those presented in Chapter 9), we cannot assert that one tree is "significantly" better than another. Although parsimony gives us valuable insights into the trees implied by our data, statistical methods must be applied to determine whether the data convincingly favor some topologies over others.

FINDING OPTIMAL TREES

In the example above, we dealt with a simple case involving just four taxa. With three ingroup taxa and one outgroup, there are only three possible fully resolved tree topologies. This made it easy to determine the parsimony score of each of these trees and to identify the most parsimonious tree.

Suppose instead that we have four, not three, ingroup taxa (five taxa in all). Because the fourth taxon could be added to any of five branches on the three possible trees for three ingroups, there are 15 possible rooted topologies (Figure 7.7). In turn, each of these 15 trees has seven places to add yet another ingroup taxon, meaning that there are $7 \times 15 = 105$ possible rooted trees for five ingroups, and so on. For the mathematically inclined, if we assume one outgroup and n ingroup taxa, the number of rooted tree topologies is $(2n-3) \times (2n-5) \times (2n-7) \times \ldots \times 3 \times 1$. This can also be written: $(2n-3)! / [2^{n-2} \times (n-2)!]$. Table 7.8 lists the numbers of tree topologies for cases with even more taxa. As you can see, the number of possible trees increases very rapidly as the number of taxa increases. When you get to 52

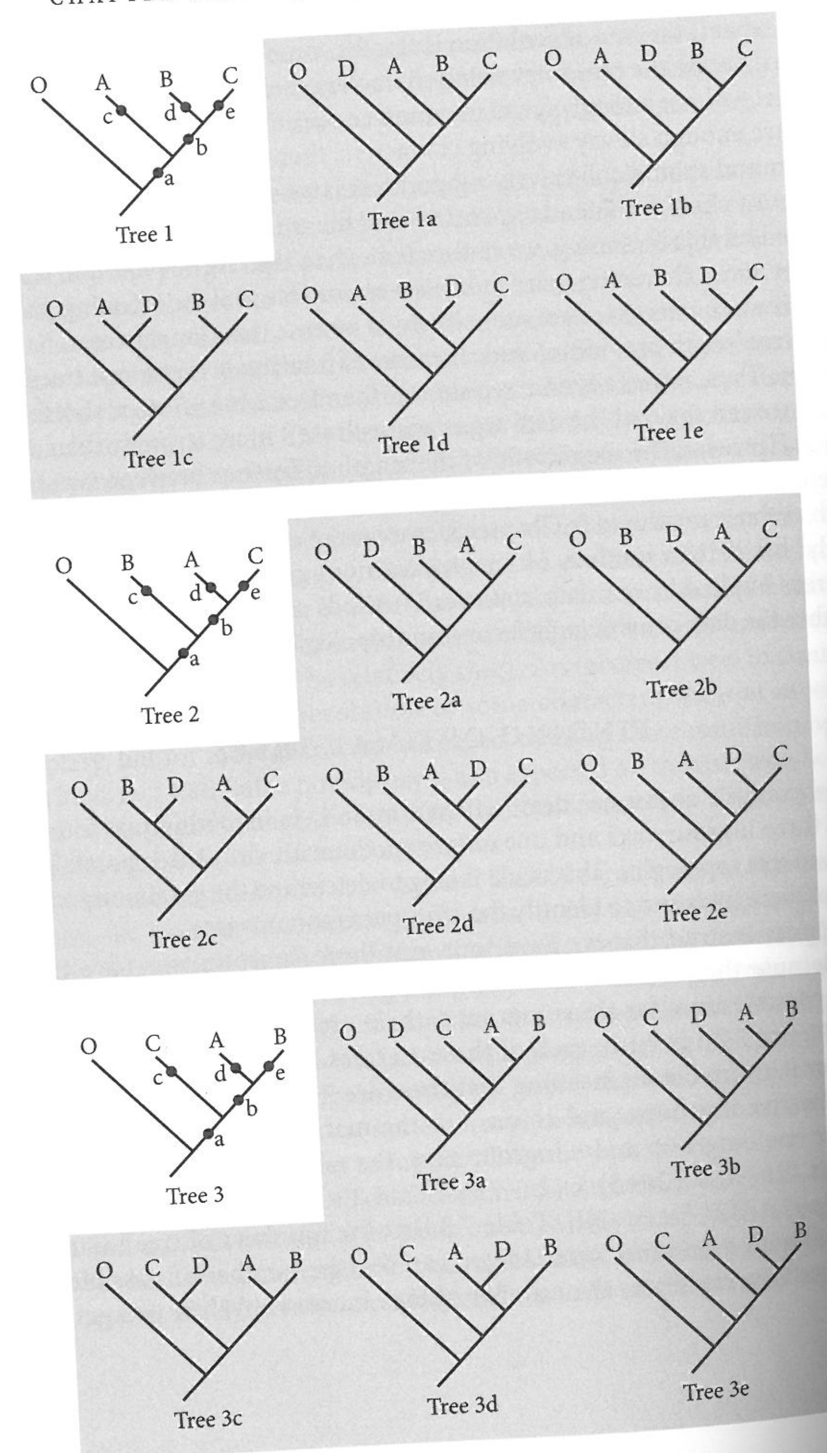


FIGURE 7.7 All possible rooted trees for five taxa.

TABLE 7.8 The rapid increase in the number of rooted tree topologies with increasing numbers of taxa

Number of ingroup taxa (with a single outgroup)	Number of possible topologies
2	1
3	3
4	15
5	105
6	945
10	34,459,425
20	~8,200,794,532,637,890,000,000
50	~27,529,213,532,835,600,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000
100	~33,499,038,543,080,200,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

taxa, the number of possible trees exceeds the estimated number of electrons in the entire universe!

Using a modern computer, it is possible to calculate the length of every possible tree for up to about 10 taxa. But when you consider that there are perhaps 2 million named species alive today, phylogenetics would be limited indeed if we could never study more than 10 taxa at one time. How then, for larger numbers of taxa, can we find most parsimonious trees?

For up to about 20 taxa, it is possible to use a method called branch-and-bound, which is guaranteed to find the optimal tree, yet does so without calculating a length for all trees. The key principle behind the branch-and-bound algorithm is that adding taxa to a tree can never decrease tree length. At best additional taxa can be added without introducing extra homoplasy, because adding taxa to a tree can never "undo" homoplasy that is already implied by the smaller tree. As a result, if a pruned tree is less parsimonious (longer) than the best full tree found so far (the "bound"), then we can be sure that all trees that derive from the pruned tree (i.e., match the pruned tree with the missing taxa grafted on) are also worse than the current bound.

For example, suppose we have 12 taxa in a data set and have already found one 12-taxon tree of length 712. If we calculate the length of a certain 10-taxon tree and find that it had a length of 713, we would know that any 12-taxon tree that could be pruned to yield this 10-taxon tree must be less parsimonious than the 12-taxon tree that we have already found. Therefore, all trees that are derived from the pruned tree can be excluded without even calculating their length. Depending on the structure in the data, branch-and-bound can usually find the most parsimonious tree while only calculating the length of a subset of trees. However, even branch-and-bound becomes impractical with more than 20 taxa. So what can be done for yet larger numbers of taxa?

Computer scientists have developed ***heuristic search*** algorithms that allow one to analyze indefinitely large data sets. These programs are not guaranteed to find the optimal tree, but they usually do so and even when they do not, the optimal tree overall is expected to be quite similar to the best trees found. To explain how heuristic algorithms work, it will be helpful to introduce the concept of tree space.

As discussed in Chapter 3, one measure of the similarity of two tree topologies is the number of rearrangements needed to convert one tree into the other. Among the several possible rearrangement methods, we introduced one: subtree pruning and regrafting (SPR).

Now imagine a space in which all possible trees are cleverly laid out such that each tree is placed adjacent to all those trees from which it is one SPR rearrangement away. This tree space is multidimensional: each dimension being the distance of one specific tree to all the other trees. However, for the sake of visualization let us pretend that tree space could be flattened into two dimensions.

Now imagine an additional dimension: a measure of tree quality. In a parsimony framework, tree length is a measure of how well a tree explains a set of data. Each point in tree space corresponds to a tree with a definite tree length for the data we are analyzing. Somewhere on this space there must be one or more trees that are shorter than all the rest—the most parsimonious trees. All other trees are a certain number of steps less parsimonious: some are one step longer, some are two steps longer, and so on. The worse the tree, the lower its "altitude" (Figure 7.8). Thus, our objective is to search through tree space to find the highest peak, which corresponds to the set of most parsimonious trees.

Tree space is not infinitely rugged. Two adjacent trees cannot be very different in altitude (length) because a single rearrangement of a tree cannot greatly change the length of all characters. This means that the best (shortest) trees will

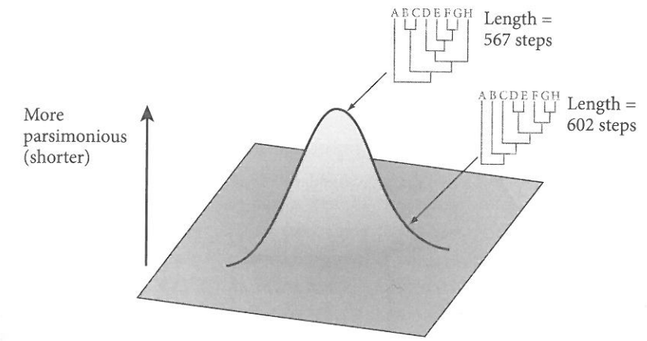


FIGURE 7.8 Visualization of tree space, where the best (shortest) trees sit at the peak.

tend to be adjacent to trees that are almost as good and the worst (longest) trees will generally be adjacent to other bad trees. This fact allows computer programs to search out peaks (short trees) without having to survey all the trees in tree space.

The approach that phylogenetic computer programs usually use when looking for the optimal trees is a hill-climbing algorithm. The program grabs a starting tree (there are clever ways to start searches on decent trees to speed up the analysis) and calculates its length. It then "visits" all the adjacent trees by making all the possible rearrangements to the starting tree, and for each adjacent tree it calculates the length. If the initial tree is shorter than all adjacent trees, it is a peak and the search stops. If it is not a peak, then the computer identifies the shortest of the adjacent trees and then "moves" to that tree (Figure 7.9). It then looks at all of *its* adjacent trees, and so on. By reiterating this procedure, the algorithm is guaranteed to identify a tree or set of related trees that are more parsimonious (= higher) than their surrounding trees. Thus, instead of calculating the number of steps on every possible tree, a heuristic search moves toward the most parsimonious tree or trees by wandering through tree space to successively shorter trees.

Heuristic searches can be adjusted in various ways to make them run faster and to have a higher probability of finding the global optimum in cases of "rugged" tree spaces that have local optima (the shorter peaks in Figure 7.9).

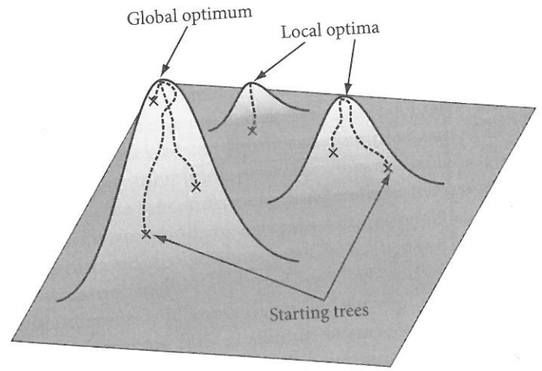


FIGURE 7.9 Searching a tree space with multiple optima.

For example, it is common to repeat heuristic searches hundreds or thousands of times initiating with a slightly different starting tree. This procedure would be analogous to finding the highest point on an island by parachuting explorers all over the island with instructions to walk uphill until they reached a peak and to then report the peak's altitude. If most of the explorers met on the same peak, you would be more confident that the island had a conical form and that the true peak had been found. If instead each explorer found a different peak, you should worry that the landscape is a jagged space whose global peak had not yet been found.

Through these and other procedures, computer programs have become able to apply the parsimony criterion to data sets with more than 1000 tips. In consequence, computer power is no longer a major impediment to conducting phylogenetic analysis using parsimony.

PARSIMONY ANALYSIS OF THE CARNIVORAN MORPHOLOGICAL DATA

Having introduced the principles of parsimony and the concept of tree space, we can return to the carnivoran data set to find the most parsimonious trees.

TABLE 7.9 Morphological data for Carnivora

	1 (4)	2 (21)	3 (32)	4 (45)	5 (52)	6 (54)	7 (56)	8 -	9 (59)	10 (60)	11 (61)	12 (62)	13 (40)	14 (50)	15 (51)	16 (1)	17 (2)	18 (3)	19 (24)	20 (26)
Outgroup	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cat	0	1	0	1	0	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0
Hyena	0	1	0	1	0	0	1	0	1	0	0	0	1	1	1	1	0	0	0	0
Civet	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0
Dog	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	1
Otter	1	0	0	0	1	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0
Seal	1	0	1	0	1	1	0	0	0	1	1	1	0	1	1	1	?	0	1	1
Walrus	1	0	1	0	1	1	0	0	0	1	1	1	0	0	1	1	0	1	1	1
Sea lion	1	0	1	0	1	1	0	0	0	1	0	0	0	1	1	1	0	1	1	1

The data matrix shown in Table 7.9 includes the 13 characters listed earlier plus 7 more characters from Wyss and Flynn (1993). In case you want to refer back to the original study, the numbers in parentheses correspond to the character numbers used in that study. Character 8, the presence or absence of retractile claws, was added to provide an example of a parsimony-uninformative character.

Using the branch-and-bound algorithm implemented in the computer program PAUP* (Swofford 2002), we find that there are two equally most parsimonious trees for these data, requiring 30 character state changes to explain the 20 morphological characters. The trees' consistency index (equal to the average CI of the 20 characters) is 20/30, or 0.67. As shown in Figure 7.10, these trees differ only in the resolution within the pinnipeds. The information common to both trees can be shown in a strict consensus tree (Chapter 3), a tree that contains only those clades present in all equally most parsimonious trees, as shown in Figure 7.11. The strict consensus tree has a polytomy where the two equally most parsimonious trees disagree.

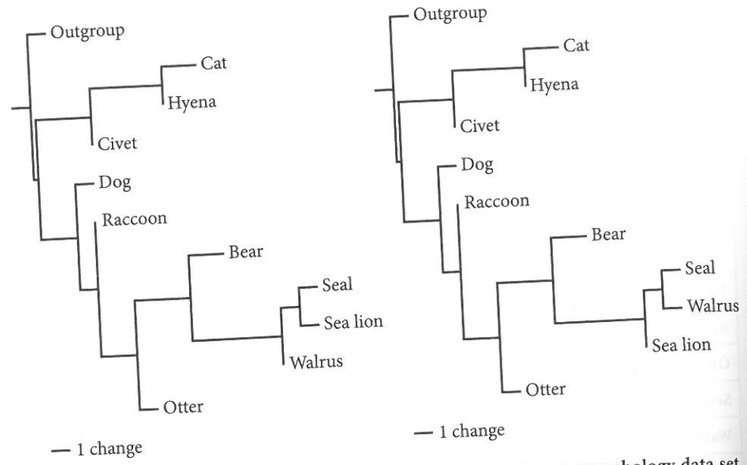


FIGURE 7.10 The two most parsimonious trees for the carnivoran morphology data set. Trees were rooted with the outgroup. These trees differ only in the resolution within the pinniped clade (seal, walrus, sea lion). To indicate the correct rooting, a short internal branch has been added at the base of the tree.

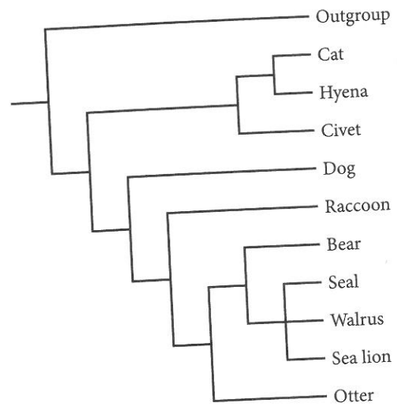


FIGURE 7.11 Strict consensus of the two most parsimonious trees for the carnivoran morphology data set. This shows that the only point of uncertainty is within the pinniped clade (seal, walrus, sea lion).

PARSIMONY ANALYSIS OF DNA SEQUENCE DATA

Once molecular sequence data became available in the 1980s, scientists began using these data for phylogenetic inference using maximum parsimony. As outlined in Chapter 8, most modern phylogenetic studies of molecular data utilize maximum likelihood or Bayesian inference methods in addition to or instead of maximum parsimony. Nonetheless, we believe it will be useful to describe parsimony analysis of DNA sequence data because (a) this accurately summarizes the historical development of the field of phylogenetics and (b) doing so will allow us to introduce some themes and concepts that will come into play when we introduce maximum likelihood approaches.

The raw data used for phylogenetic analysis are DNA sequences obtained for each taxon. Two such sequences are shown below. You will see that they are slightly different in length, having different numbers of bases. In order to proceed to use them for phylogenetic analysis, the first step is ***sequence alignment***: aligning sequences to one another so that a nucleotide position in one is matched with the homologous position in other sequences.

Taxon 1:

CGTTTATGGTGACGGAGCCGGGGGAGGTAGCACGTGGCAAAAAGAACGGCCTCGATTATCTCTTCCAT
CTTTACGAACAGTGCCGGGAGTTCTTGATTCAAGTCCAAAACATCGCCAAGGAGCGCGGCGAAAAATG
CCCCACCAAGGTAACAATAGAAACAAATCTATTTTTAATGTTTCTTAAGTAAAATTTTGAATTCAAGC
TCCGTAAATGAATGAAAATATGAGAAATATCCTGTTTTTGATCCGATTCTCATGGAAAAATATGAAAC
TAGGATAGTTTTTGCATGGTGCACGAGGTTTGACACGGGACTAGCTGTAAAAACAAGGCTGTCTCTGT
TAGAATCTTAGAACTGGACCAGCCCTCCCATTAAAGCTAGGGTTTCTAGCCCATGAAAATGTGACAAC
TCAGGTACGGGGAGGAATGGAGTCTGAAAACTTGGGACATGTATGTCTAAATTTTTGCAGAGTAAGGT
CCCCTCCGCCCCAAAAGGTTGTACTTTTGTCTTTAAAGACTTTACTGTCTTCCTTTCTGAAGCCTCGT
TTTCCCTGTCCGGTTTAGCTGAGGTGGCGTGACCCTAATACGACAGCTCCACCAYTTTTGGATCCTAA
TCTTATTGCTTATACAGGTGACCAACCAAGTTTTCAGATATGCTAAGAAGGCTGGGGCGAGCTACATT
AACAARCCCAAAATGMGCCATTACGTCGGCAGGA

Taxon 2:

TCACCCACGACCGTTCATGGTGACGGAGCCGGGGGAGGTAGCAAGTGGCAAAAAGAACGGCCTCGATT
ATCTCTTCCATCTTTACGAGCAGTGCAGGGAGTTCTTGATTCAAGTCCAAAACATCGCCAAGGAACGC
GGCGAAAAATGCCCCACGAAGGTAACAATAGAAACAAATCTATTTTTAATGATTCTTAAGTAAAATTT
TGAATTCAAGCGTAAATGAATGAAATATGAGAAATATCCTGTTTTTGATCCGATTCTCATGGAAAAAT
ATGAAACTAGGATAGTTTTTGCATGGTGCACGAGGTTTGACACGTGACTAGCTGTAAAAACAAGGCTG
TCTCTGTTAGAATCTTAGAACTGGACCAACCCTCCCATTAAAGCTAGGGTTTCTAGCCCATGAAAATG
TGACAACTCAGGTACGGGGAGGAATGGAGTCTGAAAACTTGGGACATGTATGTCTAAATTTTTGCAGA
GTAAGGTCCCCTCCGCCCCAAAAGGTTGTACTTTTTGTCTTTAAACACTTTACTGTCCTCCTTTCTGA
AGCCTCGTTTTCCCCTGTCCGGTTGAGCTGAGGTGGCGTGACCCTAATACGACAGCTCCATTGGATCC
TAACCTTGTTACTTATACAGGTGACCAACCAAGTCCTCAGATATGCTAAGAAGGCTGGGGCGAGCTAC
ATTAACAAACCCAAAATGCGCCACTATGTC

Recall that a data matrix is composed of characters that are shared by taxa but potentially differ in state. For example, the character hair may adopt such character states as white, brown, or black. For DNA sequences, the character is the nucleotide position (numbered 1, 2, 3, etc.) and the character states are the nucleotides (A, C, G, and T). It is critical, therefore, that nucleotides in each taxon be assigned to the correct positions. Sequence alignment involves sliding the sequences over one another and inserting gaps, guided by the sequences themselves. Sequence alignment, done properly, can pose major computational challenges and has become a very technical subject. Here, we will just summarize the underlying issues and point to some additional resources.

A DNA molecule is a physical structure with nucleotides in a specific linear order. In the simple case in which the only kind of mutations are base substitutions, each nucleotide position in one taxon would be homologous to a nucleotide position at the same place in the sequence of another taxon: position 1 in taxon A will be homologous to position 1 in taxon B, position 2 to position 2, and so on. If we write out the two sequences, the homologous positions are aligned above one another. In the example below, the sequences from two closely related taxa are the same length but have some differences (shaded) due to base substitutions.

```
Taxon A: GTATTGACCACTGACTAGCAT
         |||||||||||||||||||||
Taxon B: GCATTAACCATTGTCTAGCAA
```

If the only kind of mutations were base substitutions, having found the homologous genes you would merely need to line up one homologous position and the rest of the alignment would be trivial. However, sequences are subject to additional kinds of mutation: deletions, insertions, inversions, and translocations. Of these, insertions and deletions appear to be the most common.

A ***deletion*** involves the removal of one or multiple continuous bases. Deletions may be due to errors during DNA replication, but can also happen due to imperfect DNA repair following damage, unequal crossing-over during recombination, or the action of mobile genetic elements. Deletions can be as short as one base pair or as long as thousands. When deletions happen, nucleotide positions in one sequence may lack any homolog in another sequence. The missing positions can be marked with a dash. For example, here is a case where a sequence experienced a five base-pair deletion relative to its ancestor.

```
Ancestor   GTATTGACCACTGACTAGCAT
           |||||||||||||||||||||
Descendant GCATT-----TTGTCTAGCAA
```


The same mechanisms that cause deletions (errors during replication and recombination, DNA damage, and mobile genetic elements) can cause the ***insertion*** of DNA sequences into a strand. When insertions happen, new bases emerge that have no homologs in the ancestral sequence (they may be copied from somewhere else in the genome, but we rarely know this). The lack of homologs in one taxon can be indicated, again, with a dash. In the following example, a sequence has experienced a three base-pair insertion relative to the ancestral sequence.

```
Ancestor   GTATTGACC---ACTGACTAGCAT
           |||||||||||||||||||||||||
Descendant GCATTAACCACCATTGTCTAGCAA
```

In practice we generally do not have access to ancestral sequences. When we find a gap in one sequence relative to another sequence, we do not know whether there was an insertion or deletion. In light of this ambiguity, the processes that generate gaps are often called insertion/deletion events, or indels (see also Chapter 4).

The process of sequence alignment aims to align homologous positions based on the true history of sequence evolution. Alignment is, thus, properly viewed as a problem of historical inference. Furthermore, because base substitutions and indels occurred along the branches of the gene tree, sequence alignment and tree inference are really two aspects of the same problem. Therefore, in the ideal world, we would have computer programs that could take raw, unaligned sequences and search for trees that could simultaneously account for the bases in the sequences and their indels. A few programs do conduct such combined alignment and phylogenetic inference. However, the problem is so computationally challenging that most phylogenetic analyses separate the two problems: first generating an alignment and then provisionally accepting that alignment as the basis for phylogenetic inference. Combined alignment and tree inference is, however, likely to become more common over the next decade.

To get a feel for how sequence alignment can be conducted free of a phylogeny, see if you can align the following pair of sequences.

```
ATGACCTGGCGGCTTTA
ATGTGGATATGGCATTA
```

You might conclude that these sequences are already well aligned, that there were seven substitutions affecting the shaded positions.

While this might be the best alignment, it is worth considering alternatives that can also explain these data through the addition of indel events. For

example, you could align these same two sequences by invoking five indels and no substitutions, or two substitutions and two indels, as shown below. Remember that an indel event can involve any number of bases, so whether it is one dash or four, it still counts as one indel.

Five indels:

```
A T G A C C T G G - - - - C G G C T - T T A
A T G - - - T G G A T A T - G G C - A T T A
```

Two substitutions and two indels:

```
A T G A C C T G G - - - C G G C T T T A
A T G - - - T G G A T A T G G C A T T A
```

To choose between these alignments, we need to ask ourselves whether it is more likely that there were five indels, or two substitutions and two indels. Data from molecular biology would say that base substitutions are generally more frequent than indels (especially in coding genes). Thus, we would tend to reject the first alignment because it invokes more indels than substitutions in addition to invoking more total events (five versus four).

This allows us to state a rule that is applied in almost all sequence alignment programs: invoke indels only when the number of base substitutions avoided is greater than the number of indel events implied. Indeed it is normal to set the ***gap penalty***, the threshold for the number of base substitutions avoided before an indel is inferred, higher still: gap penalties of 3 to 20 are common. Additionally, most computer programs impose an extra cost for longer gaps or gaps at the ends to avoid useless alignments such as the following, which invokes no base substitutions at the "cost" of two indels:

```
- - - - - - - - - - - - - - - - - - A T G A C C A G T A C G G C T T T A
A T G A T C G A T A T G G C A T T A - - - - - - - - - - - - - - - - - -
```

It is probably clear that alignment is easiest and most certain when both base substitutions and indels are rare. This is because conserved parts of the sequence provide a framework for identifying the position and size of indel events. For example, below are two true alignments. Which do you think we would be more likely to correctly infer?

```
A T G A - - - T G C A G C T T T A G G T A
A C A A C A G T A C G A - - C T A C - C A

A T G A C C A G T A C A G - T T T A G T T
A C G A C C - - T A C G G C T T C A G T A
```


The answer is the second one. While the number of indels is similar, the many extra base substitutions in the top case would make it very hard to identify the true alignment with confidence.

Pairwise alignment considers just two sequences at a time, whereas multiple alignment includes sequences from many taxa to obtain an entire aligned data matrix. A pairwise alignment is relatively simple for a computer to determine, even when a complex set of penalties is implemented. Multiple alignments are, however, disproportionately more difficult. As the number of sequences being aligned increases, the number of possible alignments goes up exponentially. Multiple alignment algorithms should allow the placement of gaps in one sequence to influence the placement of gaps in other sequences. This is because, when gaps in two species are in the same position, they can be attributed to a single indel occurring somewhere on the gene tree. However, detecting shared indels is not always easy for a computer program. As a result, although multiple alignment programs (e.g., CLUSTAL, MAAFT, MUSCLE, TCOFFEE, FAS) provide a good starting point, they usually need to be examined and adjusted by eye. To illustrate this, Figure 7.12 shows a problematic portion of an alignment that was returned by CLUSTAL and a manually edited version of the same. You

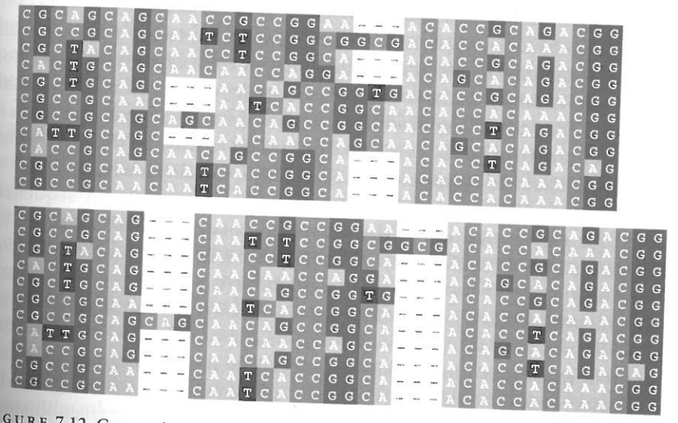


FIGURE 7.12 Comparison of a computer and manual ("eyeball") alignment of the same data set. The upper alignment was generated by a commonly used multiple alignment program, whereas the lower alignment was generated by hand.

might notice that not only does the second alignment imply a simpler mutational history, but also the human editor could take account of the codon structure of this gene (something that few computer programs keep track of) so as to keep both gaps in the same reading frame.

Given that sequence alignments will vary depending on how they were generated (which algorithm, what penalties, and whether they were manually edited), you might worry that phylogenetic analysis of DNA sequences is invalid. Actually, the problems are less than they may seem. Even if some part of a sequence is hard to align unambiguously, many regions can often be aligned confidently. The regions that are aligned correctly will tend to be composed of characters that provide consistent support for the same tree, whereas characters in regions that have been misaligned will tend to conflict with one another—they will constitute phylogenetic noise, similar to very rapidly evolving traits. This means that reasonable phylogenetic conclusions can often be obtained even when the alignment is imperfect. Nonetheless, it is wise to obtain a sense of how one's conclusions depend on the alignment. This is typically done by reconstructing the phylogeny using several different alignment schemes to see whether the major phylogenetic conclusions change. Finally, it is worth considering the use of one of several computer programs that treat sequence alignment and tree estimation as a single problem.

Having aligned DNA sequences, it is straightforward to analyze them using maximum parsimony (or with the methods described in Chapter 8). In its most basic form, parsimony analysis of DNA sequences counts all character state changes the same, regardless of which character is involved and what kinds of substitutions are invoked. Gaps are usually treated as missing data, but sometimes the inferred indel events are treated as additional characters.

To provide a concrete example, let us consider some DNA sequences obtained for representatives of the Carnivora. Sequences are available for a 1116 base-pair region of the transthyretin 2 gene (Flynn and Nedbal 1998) for the living species (Table 7.4). These sequences were aligned by eye and were then entered into a computer program, PAUP* (Swofford 2002), which searched for the most parsimonious tree. A branch-and-bound search yielded two trees of length 790 steps, which differed only in the placement of hyena (sister to civet or to cat). The strict consensus of the two is shown in Figure 7.13.

If you compare Figures 7.13 and 7.11, you will see that the trees obtained from morphological and molecular data are similar. Both data sets support a monophyletic pinniped group (seals, walruses, sea lions) and a division of the other living carnivores into two major subclades: feliforms (cats, hyenas, civets)

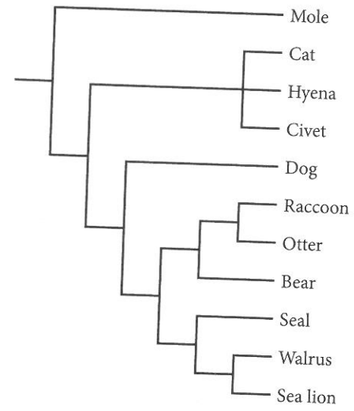


FIGURE 7.13 Consensus of the two most parsimonious trees for the carnivoran molecular data. Trees were rooted with the outgroup, mole.

and caniforms (dogs, raccoons, bears, otters, and pinnipeds), with dogs being the sister taxon to a clade comprising the remaining caniforms. This agreement is significant. If these data were not the result of evolution along a tree, the probability that the two data sets would yield trees this similar is less than 1 in 2000 (0.000375). The fact that both kinds of data yield such similar trees provides concrete evidence in support of the claim that these data are the result of descent along the same evolutionary tree. As discussed in Chapter 2, the agreement among independent phylogenetic data sets supports the hypothesis of common ancestry.

Although the trees inferred from morphological and molecular data are remarkably similar, they also have some differences. One difference is the sister group of the pinnipeds. The morphological data suggest that the bear lineage is the sister group (Figure 7.10), whereas the DNA sequences suggest that a clade composed of bear, otter, and raccoon is in this position. Such differences are best understood as being due to imperfect phylogenetic inference. One or the other or both trees are presumably incorrect in some details. While some conclusions, such as the fact that pinnipeds are embedded within the caniforms, are well demonstrated by these analyses, other results would have to remain uncertain pending the collection of more data. In fact, abundant

additional data have been collected on carnivoran phylogeny. If you want to know more, you can gain access to the literature by consulting Agnarsson et al. (2010).

CHARACTER AND CHARACTER STATE WEIGHTING

When we do phylogenetic analysis using parsimony (or another method), each character provides an independent piece of evidence of the actual evolutionary past. Drawing an analogy to forensics, another historical science, each character can be equated with a different piece of evidence collected at a murder scene. For example, character 1 might be the position of the body, character 2 might be the location of a bullet hole in the wall, and character 3 might be a scrap of paper in the victim's hand.

Until now, we have made the assumption that all changes are counted equally when deciding which trees are most parsimonious. This approach is called ***equally weighted parsimony*** or ***Fitch parsimony***, because it resembles a model proposed by Walter Fitch (Fitch 1971). Fitch parsimony is analogous to a forensic investigation in which all pieces of information are assigned equal weight in testing the innocence of a defendant. However, while many pieces of forensic evidence might influence one's belief in the guilt or innocence of a defendant, some pieces of evidence could be more compelling than others. For example, we are more likely to return a guilty verdict if the defendant's fingerprints were on the murder weapon than if a car matching the defendant's was seen near the site of the crime.

Applying this reasoning to phylogenetic inference, some characters ought to provide more reliable information about phylogeny. If some characters or kinds of character state change are less likely to show homoplasy than others, they are more likely to be consistent with the true tree and should provide more reliable phylogenetic information. What we need to use is a more flexible version of parsimony, called ***generalized (or weighted) parsimony***, in which we give more weight to those characters that we expect to provide more reliable information on phylogeny. Generalized parsimony allows more detailed prior knowledge of trait evolution to yield a more accurate assessment of the phylogeny. It is worth exploring the basics of generalized parsimony as a way to become more familiar with the logic of phylogenetic analysis, and as a useful lead-in to phylogenetic analysis by maximum likelihood.

The main reason why characters might differ in their tendency to show homoplasy is because they evolve at different rates. Because parsimony is most effective when rates of evolution are low (see A Justification of Parsimony earlier in this chapter), traits evolving more slowly provide more reliable phylogenetic evidence. Different kinds of characters are often expected to evolve at different rates. For example, gene sequences often include both slowly evolving regions (e.g., conserved domains, coding regions) and more rapidly evolving regions (e.g., introns, third codon positions). To assign all characters equal weight even when some provide more reliable evidence than others could give weak characters more weight than they deserve.

Returning to the small, hypothetical data set shown in Table 7.5, suppose that we judged character 4 to be five times as informative as the other characters in the matrix. To reflect this, we could count any change of character 4 as equivalent to five changes of the other characters. In generalized parsimony, the score of a tree is no longer simply the number of changes needed to explain the data but a sum of the cost of each character's evolution, where cost is the product of the character's length (number of steps) and its weight. Table 7.10 shows

TABLE 7.10 Tree scores when character 4 is assigned a weight of 5

	1	2	3	4	5	6	7	8			
O	0	0	1	0	1	1	0	0			
A	0	1	1	0	1	0	1	0			
B	1	1	1	1	0	0	1	1			
C	0	0	0	1	1	1	0	0			
Weight	1	1	1	5	1	1	1	1	Total length	Total cost	
Cost of tree 1	1	2	1	5	1	2	2	1	11	15	←Most parsimonious
Cost of tree 2	1	2	1	10	1	2	2	1	12	20	
Cost of tree 3	1	1	1	10	1	1	1	1	9	17	

the length and cost for each of the three possible trees (see Figure 7.3). The scores of all trees have gone up relative to Fitch (equally weighted) parsimony. However, whereas the scores of trees 2 and 3 have increased by eight (because two changes of character 4 are needed) the score of tree 1 has increased only by four. As a result, tree 1 is now the most parsimonious.

To provide a concrete example of generalized parsimony, let us consider two weighting schemes that we might consider applying to the carnivoran morphological data. First, imagine that you believed that gaining or losing external ears (pinnas) during evolution is rare and decided to reflect this by assigning a weight of 2 to character 11, while all other characters had a weight of 1. Rerunning the parsimony analysis results in finding just one most parsimonious tree with a length of 31. This tree resembles one of the two optimal trees from the flat-weighted analysis (Figure 7.10), the one in which the seal and walrus, which both lack external ears, form a clade.

Now, let us consider an alternative weighting scheme where we double the weight of all tooth characters (characters 1, 3, 4, 13, 14, 15, and 18). In this case there is again a single most parsimonious tree with a length of 42. This tree corresponds to the other tree that was found in the flat-weighted analysis (Figure 7.10), the one in which seals and sea lions form a clade. This illustrates that changing the weight of characters can change the conclusions, although in this case the impact is relatively minor.

Generalized parsimony analysis is flexible enough to accommodate another kind of differential weighting, called ***character state weighting***. Instead of assigning an elevated or lowered cost to all evolutionary changes occurring in a character, one applies different weights to particular character state transitions. Character state weighting is best represented with a step matrix, which shows the cost of transitions between each possible pair of states. For example, Table 7.11 shows a step matrix that corresponds to equally weighted (or Fitch) parsimony for DNA sequences. As you can see, all state changes are assigned the same cost.

With DNA sequence data, bases A and G are chemicals called purines, and C and T are pyrimidines. This matters because mutations within base-types, called ***transitions***, happen more frequently than do mutations between base-types, called ***transversions***. Because transversions occur less frequently, they should be less prone to homoplasy and should be assigned a higher weight. Table 7.12 shows a step matrix that assigns twice the cost to transversions as to transitions. This means that homoplastic transversions exert a greater cost for

TABLE 7.11 Step matrix corresponding to Fitch parsimony with all character state changes receiving the same weight

From: \ To:	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

TABLE 7.12 Step matrix for 2:1 upweighting of transversions relative to transitions

From: \ To:	A	C	G	T
A	0	2	1	2
C	2	0	2	1
G	1	2	0	2
T	2	1	2	0

parsimony than do homoplastic transitions. Application of this step matrix to the carnivoran molecular data results in a single most parsimonious tree, which is identical to one of the two from the equally weighted analysis (Figure 7.10). Thus, the general conclusions in this case do not appear to be highly sensitive to changing parsimony costs.

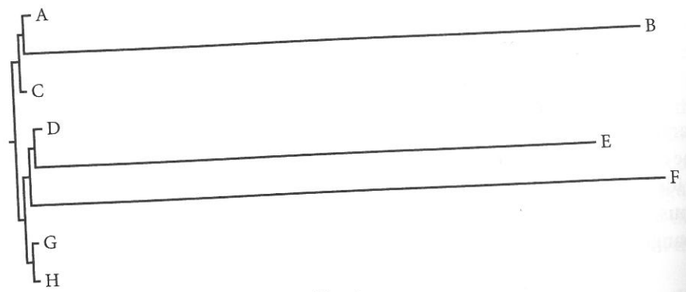
PROBLEMS OF PARSIMONY

As one of the first and most widely used methods for phylogenetic inference, maximum parsimony has led to many phylogenetic discoveries, such as identification of the major branches in the history of flowering plants and resolution of the relationships among humans and other primates. The method has also been tested in the laboratory by generating cultures of viruses with a known phylogenetic history (due to multiplying and periodically subdividing cultures) and using parsimony to infer that history from the viral DNA sequences (Hillis et al. 1992). Nonethless, while we know that parsimony works, it possesses several disadvantages that have led to the increasing use of alternative methods.

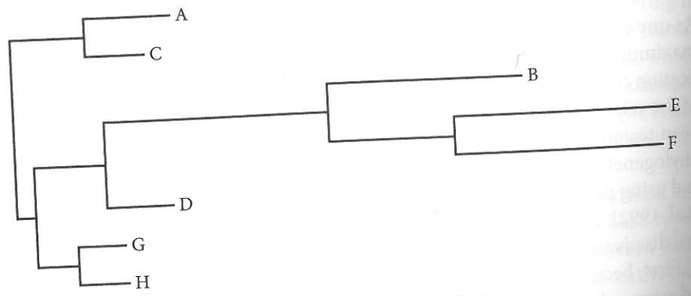
First, because parsimony does not take account of branch lengths, it can be led astray when the rate of evolution is high and the true tree has branches that have different lengths (Felsenstein 1978). If the true tree has some very long branches (e.g., due to rapid molecular evolution) and some very short branches

(e.g., due to slower molecular evolution), parsimony will tend to find an incorrect tree. Specifically, parsimony will typically yield a tree that clusters the long branches together. This problem with parsimony is usually called ***long-branch attraction***. To make matters worse, parsimony will tend to support the wrong tree more and more strongly as additional data are collected.

For example, if the tree in the upper panel of Figure 7.14 were true, the resulting data analyzed with parsimony would tend to yield a tree like that shown in the lower panel. Long-branch attraction arises because parsimony



True tree



Unrooted tree estimated by parsimony

FIGURE 7.14 Long-branch attraction. If the top tree is true, then a typical data set evolving along that tree would, if subjected to parsimony analysis, yield the lower tree.

does not allow information to flow from one character to another about the rate of evolution on different branches of the tree. If a large number of traits are inferred to change on a particular branch, then we ought to allow an increased probability of mapping additional homoplasy to those long branches. However, parsimony assigns changes based only on tree topology, making it unable to detect unusually long or short branches.

Second, parsimony requires us to select a weighting scheme (even if that scheme is equal weighting), and different weighting schemes often affect our conclusions. The problem is that there is no formal method for identifying the most appropriate weighting scheme for a given data set. The best we can do is to examine a range of plausible weighting schemes to assess how robust our conclusions are to the scheme selected. However, this approach reduces our ability to extract all the information in the data, sometimes resulting in poorly resolved estimates of phylogenetic history.

Most scientific journals accept analyses based on parsimony for morphological data, partly because formal mathematical models of morphological evolution are still poorly developed. However, the scientific community generally expects researchers to use maximum likelihood or Bayesian methods (Chapter 8) when analyzing molecular sequence data. Nonetheless, because parsimony is effective for many data sets, is less computationally demanding, and is easier to understand, it is still widely used in educational contexts and for preliminary data exploration.

FURTHER READING

Hennigian inference: Hennig 1966; Felsenstein 1978; Wiley 1981; Brooks et al. 1994
Parsimony/generalized parsimony: Wiley et al. 1991; Swofford et al. 1996
Justification of parsimony: Felsenstein 1981b; Farris 1983, 2000
Tree searching: Maddison 1991; Swofford et al. 1996; Nixon 1999; Quicke et al. 2001
Sequence alignment: Wheeler 1996, 2001; Liu et al. 2009; Morrison 2009
Long-branch attraction: Felsenstein 1978, 1983; Siddall and Whiting 1999; Sanderson and Kim 2000

CHAPTER 7 QUIZ

Questions 1–4. Refer to this data matrix. Assume that taxon H is the outgroup.

	1	2	3	4	5	6	7
A	0	0	0	0	0	0	0
B	0	1	1	0	1	1	0
C	1	1	0	1	1	0	1
D	1	1	0	1	1	0	1
E	1	1	0	0	0	0	0
F	0	1	1	0	0	1	0
G	0	1	1	1	1	1	0
H	0	1	1	1	0	1	0

1. For how many of the scored characters do taxa A and E share the same state?
   a. 0 b. 2 c. 4 d. 5 e. 7
2. For how many of the scored characters do taxa C and E share the same state?
   a. 0 b. 2 c. 4 d. 5 e. 7
3. Applying the principles of Hennigian inference to characters in isolation (i.e., ignoring the rest of the matrix), which of the following characters would suggest that C is more closely related to E than to A?
   a. 1 b. 3 c. 4 d. 5 e. 6
4. Which of the following characters directly contradicts the claim that C and E are more closely related to each other than either is to A?
   a. 1 b. 3 c. 4 d. 5 e. 6

Questions 5–7. Here is a small molecular data matrix. Assume that taxon A is the outgroup.

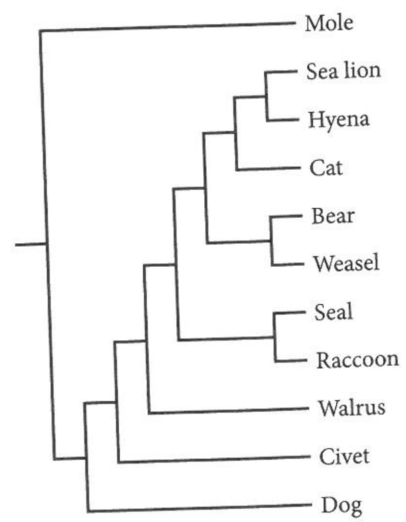
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	A	T	G	C	G	G	G	T	C	-	-	-	A	T	A	G	A	T	C	C	A
B	A	T	G	G	G	A	A	C	T	A	G	A	A	T	A	G	A	G	C	C	A
C	A	T	G	A	G	A	A	G	T	A	G	A	A	G	A	G	A	T	C	C	A
D	A	T	G	T	C	A	A	G	A	-	-	-	A	T	A	G	A	G	C	C	A
E	A	T	G	T	C	T	A	T	A	A	G	A	A	T	A	G	A	G	C	C	T
F	A	T	G	T	C	T	T	T	A	A	G	T	A	G	A	G	A	T	C	C	T
G	A	T	G	T	C	A	A	G	G	-	-	-	A	T	G	G	A	A	C	C	A
H	A	T	G	A	C	A	A	G	G	-	-	-	A	G	A	G	A	T	C	C	T

5. What do the dashes in columns 10–12 most likely represent?
   a. Positions occupied by a nonconventional base (neither A, C, G, nor T)
   b. Bases deleted during sequence evolution
   c. Positions where bases were inserted in other sequences
   d. Parsimony-uninformative character states
   e. Sequencing errors
6. Which of the following positions is parsimony-informative?
   a. 3 b. 5 c. 13 d. 15 e. Two of the other answers are correct
7. How would a parsimony analysis be affected by removing the first three positions from the data matrix?
   a. The optimal tree topology will definitely not change; the optimal tree will be three steps longer.
   b. The optimal tree topology will definitely not change; the optimal tree will be three steps shorter.
   c. The optimal tree topology will definitely not change, and neither will its length.
   d. The optimal tree topology might or might not change; the optimal tree length could be longer or shorter.
   e. The optimal tree topology will certainly change; the optimal tree will be shorter.

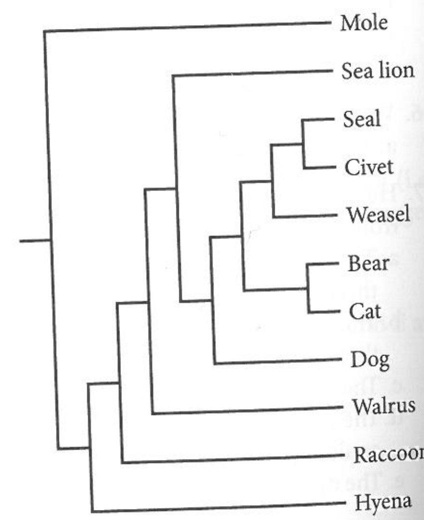
Questions 8–10. Below is a subset of the molecular data that Nedbal and Flynn collected for a phylogenetic study of Carnivora. The sequences come from the transthyretin Intron I. The mole is included as the outgroup.

Mole	C	G	T	C	A	A	C	A	G	T	C	G	A	C	A	T	T	C	A	T	G	C	T	T	T	T	C	T	T	T	T	T	G	T	T	T	G	C
Sea lion	T	A	T	T	C	C	C	G	C	T	C	C	C	T	G	T	T	T	G	T	C	T	A	G	G	C	G	A	T	T	T	A	G	A	G	C	G	C
Walrus	T	A	T	T	C	C	C	A	C	T	C	C	C	T	G	T	T	T	G	T	C	T	G	G	G	C	G	A	T	T	T	A	G	A	G	C	G	T
Seal	C	A	T	T	C	C	C	G	C	T	C	C	C	T	G	T	T	T	G	T	C	T	G	G	G	C	A	A	T	T	T	A	G	A	G	C	G	T
Bear	C	A	C	T	A	A	C	G	C	T	A	T	C	T	G	T	T	T	G	T	C	C	T	G	A	C	G	A	T	T	T	C	C	G	G	C	G	C
Raccoon	T	A	T	T	G	G	T	G	C	T	A	T	C	T	A	T	G	T	G	C	C	T	T	G	G	C	G	G	T	C	T	C	G	G	G	T	G	C
Weasel	T	A	T	T	A	A	C	G	C	T	A	G	C	T	A	T	G	T	G	C	C	T	T	G	G	T	A	G	T	C	C	C	G	G	G	C	G	A
Dog	T	A	T	T	A	A	T	G	G	T	G	T	A	C	C	T	T	T	A	T	C	T	T	T	T	C	G	C	T	T	T	C	C	A	T	C	G	C
Civet	T	C	C	T	A	A	C	A	G	G	A	T	A	T	A	C	T	G	A	T	G	T	T	T	T	C	A	A	A	T	C	A	G	C	T	C	A	C
Hyena	T	C	C	T	A	A	C	G	G	G	A	T	A	T	A	C	G	G	A	T	G	T	T	T	T	C	A	A	A	T	C	A	G	C	T	C	A	C
Cat	T	C	C	C	A	A	C	A	G	G	A	T	A	T	A	C	T	G	A	T	G	T	T	T	T	C	A	A	A	T	T	A	G	G	T	C	G	C

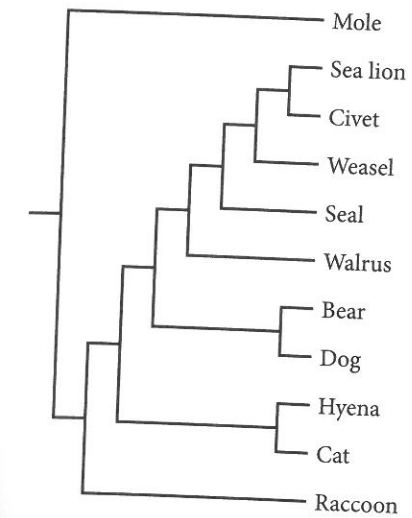
8. Which of the following five random trees is most compatible with the first nucleotide position (shaded), considered in isolation?



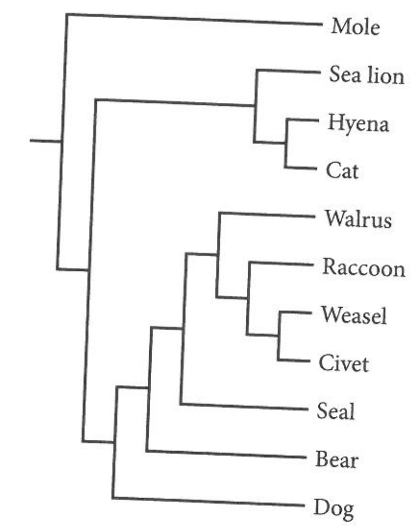
a



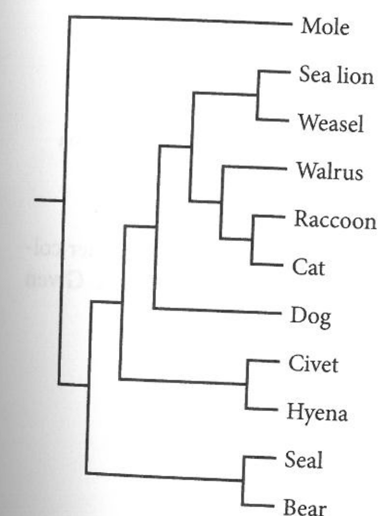
b



c



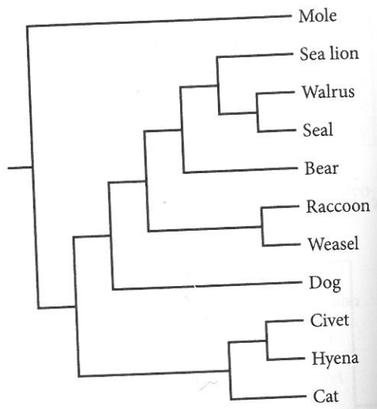
d



e

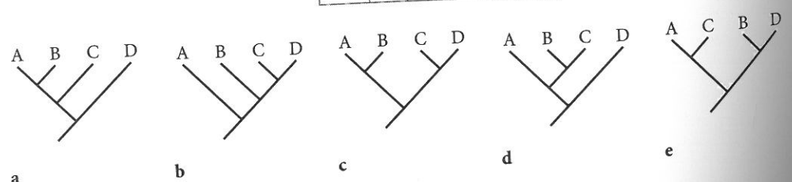
9. The matrix yields a single most parsimonious tree of length 72. What does the number 72 refer to?
   a. It is the number of characters that were analyzed.
   b. It is the number of character states that were analyzed.
   c. It is the number of character state changes needed to explain all the data on this tree.
   d. It is the number of clades in the most parsimonious tree.
   e. It is the number of trees that were considered during the heuristic search procedure before the optimal tree was found.

10. Under the assumption of parsimony, how many changes/steps does one need to explain the first nucleotide position (shaded) in the matrix on this tree?
    a. 1 b. 2 c. 3
    d. 4 e. >4



11. Here is a small data matrix. Taxa are labeled A–D in column 1, and the other columns show bases present at five positions in the aligned DNA sequences. Given these data, which of the five trees is most parsimonious?

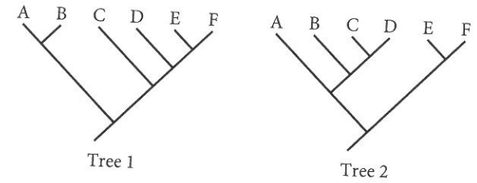
A	C	A	G	C	G
B	C	T	A	C	A
C	C	A	A	T	G
D	A	T	A	T	A



a **b** **c** **d** **e**

12. Below is a small data matrix and four alternative trees. For each character, calculate the number of steps needed to account for its evolution on each of the four trees. Sum across the characters to determine the overall tree length. Which of the four trees is the most parsimonious?

	Character									
Taxon	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
A	T	G	T	G	A	A	C	A	A	
B	T	G	T	G	A	C	C	A	A	
C	T	G	C	G	G	C	C	T	A	
D	A	G	C	G	G	C	G	T	A	
E	A	A	C	T	A	A	G	T	G	
F	A	A	C	T	A	A	G	C	G	**Total Length**
Steps on tree 1										
Steps on tree 2										
Steps on tree 3										
Steps on tree 4										



13. Generate (from your imagination) a DNA sequence matrix comprising 10 characters and seven taxa. Design the matrix to have the following features:
    a. All four bases A, C, G, T should be used.
    b. Two characters should include gaps.
    c. Eight characters should have one or two states, two should have three or four states.
    d. Eight of the characters should be parsimony-informative, two should be uninformative.
    e. Six informative characters should be consistent with one another. The other two should conflict with one or more of the other six characters.

Taxon	1	2	3	4	5	6	7	8	9	10
A										
B										
C										
D										
E										
F										
Outgroup										

    Draw two rooted trees: one should be the most parsimonious tree for the data matrix and the other should be a less parsimonious tree. For each tree indicate its length (treating indels as missing data).

14. Impose a character or character state weighting scheme to the data matrix in the preceding question and predict how it will change the topology of the most parsimonious tree.

15. Below is a small data matrix. Assuming that taxon A is the outgroup, infer the phylogeny using either Hennigian inference or parsimony (they yield the same result). Based on this tree, is taxon C more closely related to B or D? Based on the data matrix, does C share more traits in common with B or D? How do you explain this discrepancy?

	1	2	3	4	5	6	7
A	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0
C	1	1	1	0	0	0	0
D	1	1	1	1	1	1	1

16. What attributes must characters have or not have in order to be useful for inferring phylogenies using parsimony?

17. The Hennigian method allows one to conclude that some tree topologies are definitely false. Why is this not the case with parsimony?

18. Suppose you are studying a 100-taxon data set. Given the impossibility of calculating the length of every possible tree, can you hope to ever find the most parsimonious tree?

19. It has been argued that equally weighted parsimony is preferable to generalized parsimony because the former does not make assumptions about how characters evolve. What is wrong with this argument?

20. Is it true that parsimony assumes that few if any characters show homoplasy?