7

Phylogenetic Inference: Maximum Parsimony Methods

Maximum parsimony (MP) methods were originally developed for morphological characters (Henning 1966), and there are many different versions (Wiley 1981; Felsenstein 1982; Wiley et al. 1991; Maddison and Maddison 1992; Swofford and Begle 1993). In this book, we consider only the methods that are useful for analyzing molecular data. Eck and Dayhoff (1966) seem to be the first to use an MP method for constructing trees from amino acid sequence data. Later, Fitch (1971) and Hartigan (1973) developed a more rigorous MP algorithm for nucleotide sequence data. In these MP methods, four or more aligned nucleotide (or amino acid) sequences ($m \geq 4$) are considered, and the nucleotides (amino acids) of ancestral taxa are inferred separately at each site for a given topology under the assumption that mutational changes occur in all directions among the four nucleotides (or 20 amino acids). The smallest number of nucleotide (or amino acid) substitutions that explain the entire evolutionary process for the topology is then computed. This computation is done for all potentially correct topologies, and the topology that requires the smallest number of substitutions is chosen to be the best tree. The theoretical basis of this method is William of Ockham's philosophical idea that the best hypothesis to explain a process is the one that requires the smallest number of assumptions. Sober (1988) states that the less we need to know about the evolutionary process to make a phylogenetic inference, the more confidence we can have in our conclusions. In this chapter, we are primarily concerned with nucleotide sequences, but the same approach can be used for amino acid sequences as well.

If there are no backward and no parallel substitutions (no homoplasy) at each nucleotide site and the number of nucleotides examined (n) is very large, MP methods are expected to produce the correct (realized) tree. In practice, however, nucleotide sequences are often subject to backward and parallel substitutions, and n is rather small. In this case, MP methods tend to give incorrect topologies (chapter 9). Furthermore, Felsenstein (1978) has shown that when the rate of nucleotide substitution varies extensively with evolutionary lineage, MP methods may generate incorrect topologies even if an infinite number of nucleotides are examined. Under certain conditions, this can happen even when the rate of substitution is constant for all lineages (Hendy and Penny 1989;

Zharkikh and Li 1993; Takezaki and Nei 1994; Kim 1996). In this case, long branches (or short branches) of the true tree tend to join together or attract each other in the reconstructed tree (chapter 9). Therefore, this phenomenon is often called **long-branch attraction** (Hendy and Penny 1989) or **short-branch attraction** (Nei 1996). In parsimony analysis, it is also difficult to treat the phylogenetic inference in a statistical framework, because there is no natural way to compute the means and variances of the minimum numbers of substitutions obtained by the parsimony criterion.

Nevertheless, MP methods have some advantages over other tree-building methods. First, they are relatively free from various assumptions that are required for nucleotide or amino acid substitution in distance or likelihood methods. Since any mathematical model currently used is a crude approximation to reality, model-free MP methods may give more reliable trees than other methods when the extent of sequence divergence is low (Miyamoto and Cracraft 1991). In fact, computer simulation has shown that when (1) the extent of sequence divergence is low ($d \leq 0.1$), (2) the rate of nucleotide substitution is more or less constant, and (3) the number of nucleotides examined is large, MP methods are often better than distance methods in obtaining the true topology (Sourdis and Nei 1988; Nei 1991). Furthermore, parsimony analysis is very useful for some types of molecular data such as insertion sequences and insertions/deletions, as will be discussed later.

There are many different versions of MP methods even just for molecular data, but they can be divided into **unweighted MP** and **weighted MP** methods. In unweighted MP methods, nucleotide or amino acid substitutions are assumed to occur in all directions with equal or nearly equal probability. In reality, however, certain substitutions (e.g., transitional changes) occur more often than other substitutions (e.g., transversional changes). It is therefore reasonable to give different weights to different types of substitutions when the minimum number of substitutions for a given topology is to be computed. MP methods incorporating this feature are weighted MP methods. In the following, we first consider unweighted MP methods.

7.1. Finding Maximum Parsimony (MP) Trees

Estimation of the Minimum Number of Substitutions

Let us now explain how to count or estimate the minimum number of substitutions for a given topology. We consider the topology of a rooted tree for six DNA sequences (1, 2, . . . , 6) given in Figure 7.1A and assume that the nucleotides at a given site for the six extant sequences are as given at the exterior nodes of the tree. There are one C, three T's and two A's. From these nucleotides, we can infer the nucleotides for the five ancestral taxa (nodes) *a, b, c, d,* and *e*. The nucleotide at node *a* must be either C or T if we consider the minimum possible number of substitutions. The nucleotide at node *b* is inferred to be T, whereas the nucleotide at

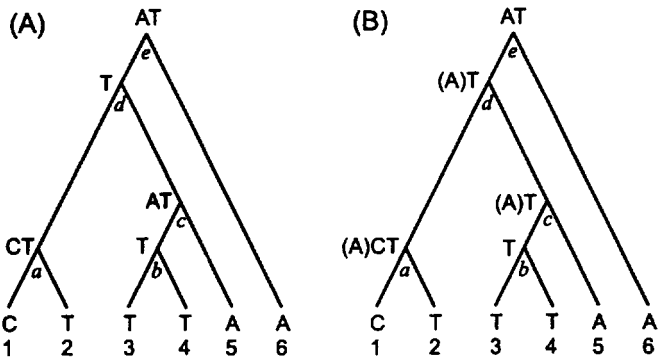

FIGURE 7.1. Nucleotides in six extant sequences and the possible nucleotides in five ancestral sequences.

node *c* must be A or T. Node *d* is expected to have T, because its immediate descendant nodes (*a* and *c*) both have T. Finally, we infer the nucleotide at node *e* to be either A or T. It is now clear that the minimum number of nucleotide substitutions for this set of taxa can be obtained by assuming that all the ancestral nodes had nucleotide T. The number is three. However, this set of nucleotides at the ancestral nodes (pathway) is not the only possible set that explains the evolutionary change of nucleotides.

If we assume that nodes *a, c, d,* and *e* all have A and node *b* has T, the number of substitutions required is again three (see Figure 7.1B). Actually, there are three more pathways that are possible with the same minimum number of substitutions. They are: (a − T, b − T, c − A, d − A, e − A), (a − C, b − T, c − A, d − A, e − A), and (a − T, b − T, c − T, d − T, e − A). These results show that the nucleotides at the ancestral nodes cannot always be determined uniquely, and all the nucleotides listed in Figure 7.1B are parsimonious ones. However, it is possible to count the minimum number of substitutions required. It is three for all of the above cases.

In the above example, we considered a rooted tree. However, the tree can be transformed into an unrooted tree by eliminating the apex node *e*. Elimination of this node does not change the minimum number of substitutions, but the number of possible pathways is reduced. For example, the two possibilities (e − T, d − T, c − T, b − T, a − T) and (e − A, d − T, c − T, b − T, a − T) are no longer distinguishable, because node *e* can be either T or A. In the present case, the total number of pathways for the unrooted tree is four. Because MP methods do not determine the root of the tree, unrooted trees are usually considered.

In the above example, the minimum number of substitutions was three, and there were four equally parsimonious pathways for the unrooted tree. Computation of these numbers was relatively easy in this case, but as the number of taxa increases, it becomes increasingly cumbersome. Therefore, all these computations are done by a computer using the above rule. The basic algorithm for these computations was developed by Fitch (1971) and Hartigan (1973).

Tree Lengths

In the above example, we considered only one topology, but in practice we have to consider all potentially correct topologies and determine the topology that requires the smallest minimum number of substitutions. Let us now consider this problem using the trees given in Figure 7.2 and assuming that taxa 1, 2, and 3 all have nucleotide A but taxa 4, 5, and 6 have G, G, and T, respectively. The trees given in the figure all consist of six taxa, but the topologies are not necessarily the same. We again consider a particular nucleotide site and compute the minimum number of substitutions required. In topology A, this number is obviously two. Topology B, in which taxa 3 and 4 are interchanged, requires at least three substitutions. Of course, there are several equally parsimonious pathways, and another pathway that requires three substitutions is given in tree C. Tree D has a different topology and requires at least three substitutions. However, in the case of six taxa, there are 105 different topologies, so we have to compute the minimum number of substitutions required for all topologies. If this computation is done for all sites and for all topologies, we can compute the sum of the minimum numbers of substitutions over all sites for each topology. This sum (L or TL) is called the **tree length.** The *maximum parsimony* (MP) tree is the topology that has the smallest tree length. In practice, it is possible that two or more different topologies have the same minimum number of substitutions. In this case, we cannot determine the final topology uniquely, and all equally parsimonious MP trees are considered as potentially correct topologies.

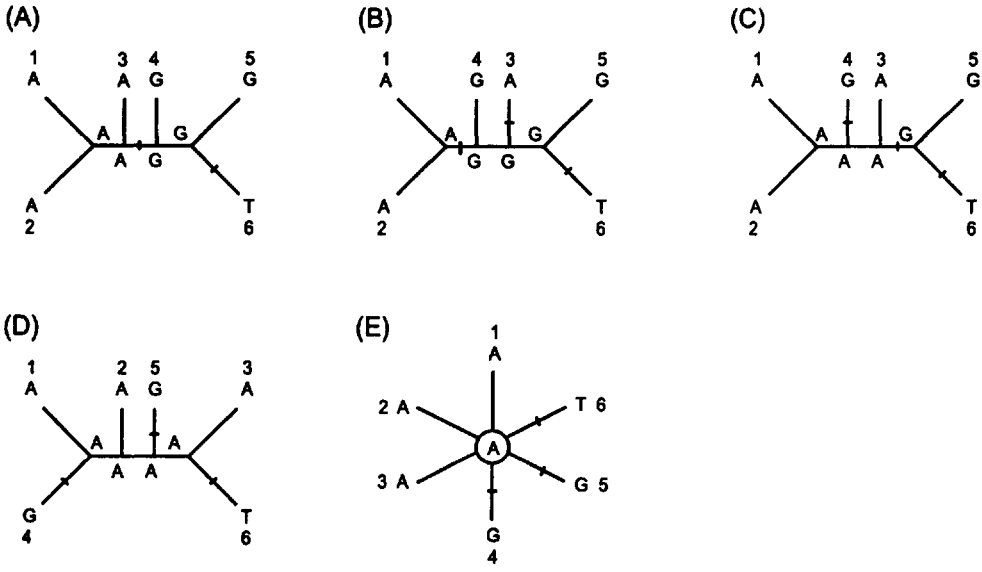

FIGURE 7.2. Assignment of mutations to different branches at a parsimony-informative site.

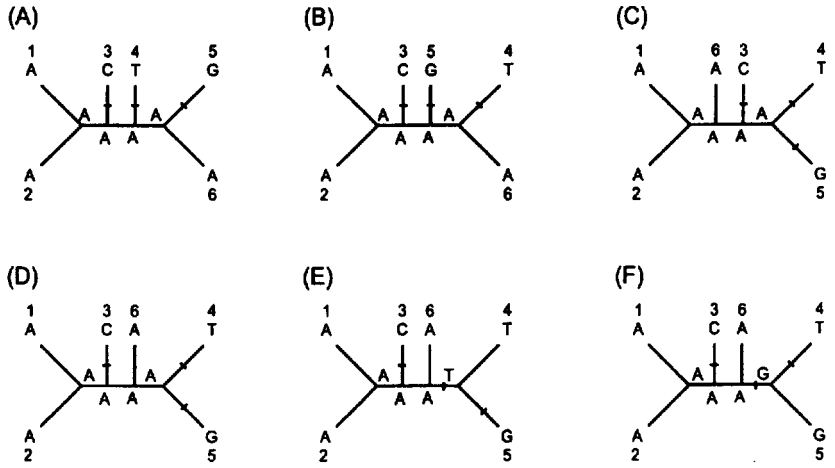

FIGURE 7.3. Assignment of mutations to different branches at a noninformative site.

Informative Sites and Homoplasy

In the search for MP trees, nucleotide or amino acids sites that have the same nucleotide for all taxa (**invariable sites**) are eliminated from the analysis, and only **variable sites** are used. However, not all variable sites are useful for finding an MP tree topology. Any nucleotide site at which only unique nucleotides (singletons) exist is not informative, because the nucleotide variation at the site can always be explained by the same number of substitutions in all topologies. Such a site is called a **singleton site.** For example, tree A in Figure 7.3 has three singleton substitutions C, T, and G and requires three substitutions, but the same number of substitutions is required for any other topology. This can be seen from trees B, C, and D in Figure 7.3. In all these trees (topologies), the three singleton substitutions can be assigned to exterior branches. In some topologies, however, we can assign singleton substitutions to both exterior and interior branches. In Figure 7.3, the topologies of trees D, E, and F are the same but have different assignments of mutational changes to different branches. However, the total number of substitutions is always three. Therefore, this site is not informative for identifying MP trees.

For a nucleotide site to be informative for constructing an MP tree, there must be at least two different kinds of nucleotides, each represented at least two times. These sites are called **informative sites** (Fitch 1977). In trees A, B, C, and D of Figure 7.2, the nucleotide site satisfies this condition, and thus it is useful for finding the topology with the minimum number of substitutions. However, note that singleton sites are informative for topology construction in other tree-building methods. Actually, even invariable sites have some phylogenetic information in distance and maximum likelihood methods. So, Fitch's terminology "phylogenetically informative sites" is not very appropriate. For this reason, we call these sites **parsimony-informative sites.**

In the construction of MP trees, it is sufficient to consider only parsimony-informative sites. However, some authors include singleton sub-

stitutions in the computation of tree lengths. This addition of singleton substitutions to the tree length for parsimony-informative sites does not affect the identification of the MP tree, because the number of singleton substitutions is the same for all topologies. Nevertheless, one should be cautious about the tree length of a published tree and should know whether it is based on only parsimony-informative sites or all variable sites.

Because only informative sites contribute to finding MP trees, it is important to have many informative sites to obtain reliable MP trees. However, when the extent of **homoplasy** (backward and parallel substitutions) is high, MP trees would not be reliable even if there are many informative sites available. For this reason, Kluge and Farris (1969) proposed a quantity called the **consistency index** to measure the extent of homoplasy. This index for a single nucleotide site (i-th site) is given by $c_i = m_i/s_i$, where m_i is the minimum possible number of substitutions at the site for any conceivable topology, and s_i is the minimum number of substitutions required for the topology under consideration. The minimum possible number of substitutions (m_i) is one fewer than the number of different kinds of nucleotides at the site, assuming that one of the observed nucleotides is ancestral. For example, there are three different nucleotides in tree A of Figure 7.2. Therefore, $m_i = 2$. For this topology, s_i is also equal to 2, so $c_i = 1$. This indicates that the nucleotide configuration at this site is supportive of tree A under the MP principle. By contrast, $s_i = 3$ for topologies B, C, and D, so $c_i = 2/3$. Therefore, these topologies are not well supported.

However, the lower bound of the consistency index is not 0, and c_i varies with topology. For this reason, Farris (1989) proposed two more quantities called the **retention index** (see also Archie 1989) and the **rescaled consistency index**. The retention index is given by $r_i = (g_i - s_i)/(g_i - m_i)$, where g_i is the maximum possible number of substitutions at the i-th site for any conceivable tree under the parsimony principle and is equal to the number of substitutions required for a star topology when the most frequent nucleotide is placed at the central node. Diagram E in Figure 7.2 shows such a tree, and in this tree $g_i = 3$. The retention index becomes 0 when the site is least informative for MP tree construction, that is, $s_i = g_i$. In the examples of Figure 7.2, we have $r_i = (3 - 2)/(3 - 2) = 1$ for tree A and $r_i = (3 - 3)/(3 - 2) = 0$ for trees B, C, and D. Therefore, the site under consideration is supportive of tree A but not of the other trees. By contrast, the rescaled consistency index (rc_i) is given by $r_i c_i$. That is,

$$rc_i = \frac{g_i - s_i}{g_i - m_i} \frac{m_i}{s_i} \tag{7.1}$$

This index also is 1 for tree A and 0 for trees B, C, and D. In the present case, therefore, rc_i is identical with r_i, but this is not always the case.

In the above discussion, we considered c_i, r_i, and rc_i for one site. In practice, however, these values are computed for all informative sites, and the **ensemble** or **overall consistency index (*CI*) overall retention index (*RI*)**, and **overall rescaled index (*RC*)** for all sites are considered.

These indices are defined as $CI = \Sigma_i m_i / \Sigma_i s_i$, $RI = (\Sigma_i g_i - \Sigma_i s_i)/(\Sigma_i g_i - \Sigma_i m_i)$, and $RC = CI \times RI$, respectively, where i refers to the i-th informative site. These indices should be computed only for informative sites, because for uninformative sites c_i becomes 1 and r_i and rc_i are undefinable. These indices are often used as a measure of accuracy of the topology obtained, particularly for an MP tree obtained from morphological characters. In systematics, $HI \equiv 1 - CI$ is called the **homoplasy index.** When there are no backward and no parallel substitutions, we have $CI = 1$ and $HI = 0$. In this case, the topology is uniquely determined.

Example 7.1. MP Trees for Five Hominoid Species

Let us again consider the DNA sequences given in Figure 6.1 and construct the MP tree. In this data set, if we exclude site 560, in which a deletion exists, there are 281 variable sites of which 90 are parsimony informative. Using these informative sites, we can compute the tree lengths (L) for all the topologies. Only three topologies (B(O(G(C, H)))), (B(O(H(G, C)))), and (B(O(C(G, H)))) have $L = 148$ or less, and all others have much larger L values (Brown et al. 1982). The L values for the three topologies are 147, 145, and 148, respectively, and therefore the topology (B(O(H(G, C)))) is the MP tree. The branch length estimates of this tree are given in Figure 7.4B. The topology of this tree is different from that of the trees obtained by distance methods (Figure 6.2). However, the difference between the two topologies is one branch interchange ($d_T = 2$), and the L value differs only by 2. Therefore, the difference is unlikely to be significant. When the entire sequences of mitochondrial DNA are used, we obtain the same topology as that of the trees in Figure 6.2 (Horai et al. 1995). Another reason why we obtained an erroneous tree seems to be that in this case the transition/transversion ratio is high. In fact, if we use a weighted parsimony method described later, we obtain the same topology as that of the distance trees (Figure 7.4D).

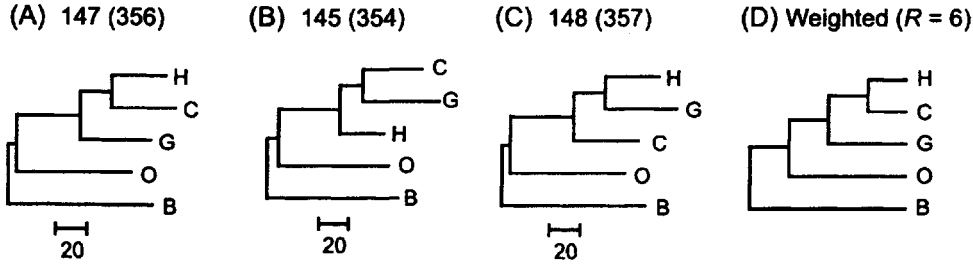

FIGURE 7.4. A–C. Three possible trees (topologies) for the human (H), chimpanzee (C), gorilla (G), orangutan (O), and gibbon (B). These trees were obtained from the DNA sequence data in Figure 6.1. The number given to each topology is the tree length for parsimony-informative sites. The number in parentheses refers to the tree length for all variable sites. D. Maximum parsimony tree obtained when transversions are given six times the weight of the transitional substitutions. Only the branching pattern is shown in D. The overall consistency index (CI) is 0.67, 0.68, and 0.67 for trees A, B, and C, respectively, whereas the overall retention index (*RI*) is 0.47, 0.49, and 0.46 for the three trees.

We have also computed the overall consistency index (*CI*) and the overall retention index (*RI*) for the three topologies. The *CI* values for trees A, B, and C were 0.67, 0.68, and 0.67, respectively. Therefore, the differences in *CI* between the three trees are very small. By contrast, the *RI* values for the three trees were 0.47, 0.49, and 0.46, respectively. These values are negatively correlated with the tree lengths.

7.2. Strategies of Searching for MP Trees

When the number of sequences or taxa (m) is small, say, $m < 10$, it is possible to compute the tree lengths of all possible trees and determine the MP tree. This type of search for MP trees is called the **exhaustive search.** As previously mentioned, the number of topologies rapidly increases as m increases (Equation [5.1]). Therefore, it is virtually impossible to examine all topologies if m is large. However, if we know clearly incorrect topologies, as in the case of the five hominoid species in Figure 6.1, we do not have to compute the L values for them. We can simply compute L's only for potentially correct trees. This type of search is called the **specific-tree search.**

There are two ways of obtaining MP trees when $m > 10$ and the specific-tree search is not applicable. One is to use the **branch-and-bound method** (Hendy and Penny 1982). In this method, the trees that obviously have a tree length longer than that of a previously examined tree are all ignored, and the MP tree is determined by evaluating the tree lengths for a group of trees that potentially have shorter tree lengths. This method guarantees finding of all MP trees, although it is not an exhaustive search. However, even this method becomes very time-consuming if m is about 20 or larger. In this case, one has to use another approach called the **heuristic search.** In this method, only a small portion of all possible trees is examined, and there is no guarantee that the MP tree will be found. However, it is possible to enhance the probability of obtaining the MP tree by using several algorithms.

Branch-and-Bound Search

After the branch-and-bound method was introduced by Hendy and Penny (1982) in parsimony analysis, several different versions were developed (Swofford and Begle 1993). The differences in algorithm and efficiency among them are rather small, and here we present Kumar et al.'s (1993) version. In this version of the branch-and-bound method, the search for an MP tree starts with an initial core tree of three taxa, which has only one unrooted tree (tree A in Figure 7.5). The remaining taxa are added to this core tree one by one according to a certain order, and the tree length of the new tree is computed at each stage of taxon addition. If the addition of a taxon to a particular branch of a core tree results in a tree length greater than a predetermined upper bound of tree length (L_U), this topology and all the subsequent topologies that can be generated by adding more taxa to this core tree are ignored from further consideration.

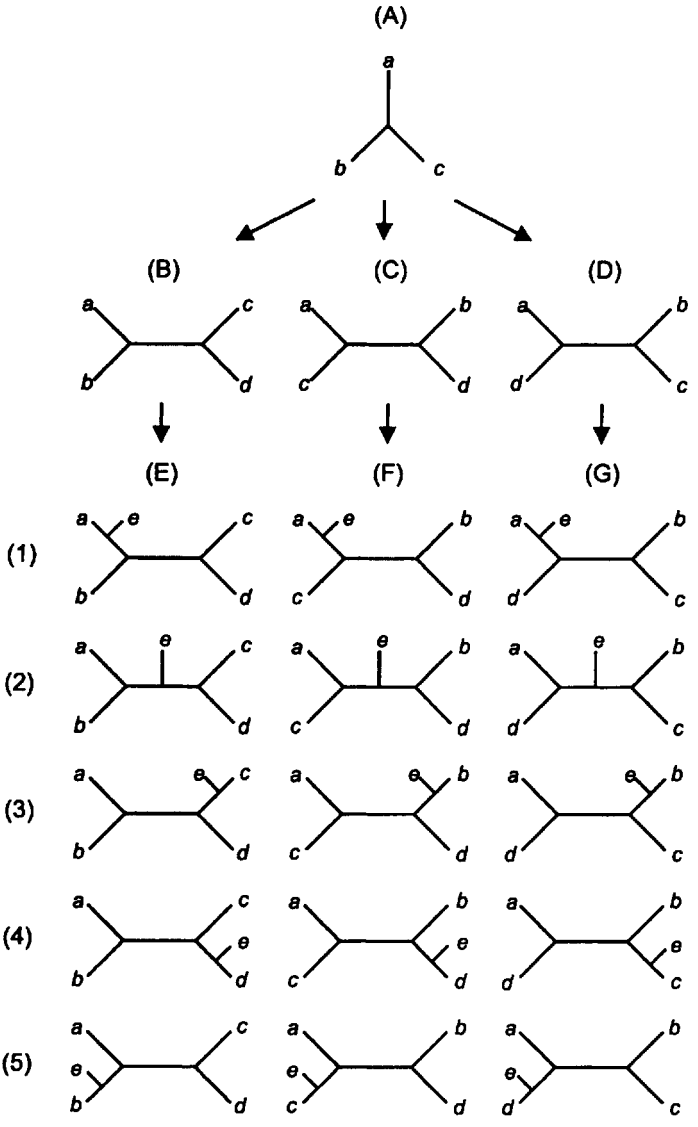

FIGURE 7.5. Diagrams showing the procedures of the branch-and-bound and the heuristic branch-and-bound-like searches.

Core Tree and Order of Taxon Addition

The initial core tree of three taxa is chosen such that the length (L) of the tree is largest or approximately largest among all possible three-taxon trees (Figure 7.5). This is to make L closer to the length (L_M) of the MP tree so that we can reach the MP tree faster. The next step is to determine the order of taxon addition that makes the search for the MP tree faster. To do this, we place one of the remaining taxa on one of the three branches of the initial core tree and compute the tree length by the MP procedure. We repeat this computation for the two remaining branches

and record the minimum value of the three tree lengths. We repeat this procedure for all remaining taxa. We then find the taxon that shows the maximum value of the minimum tree lengths. This taxon is the first to be added to the initial core tree. We call this procedure the maximum-of-the-minimum-values algorithm or simply the **max-mini algorithm.** To find the next taxon, we apply this max-mini algorithm for the remaining taxa using the tree for the first four taxa as the next core tree. In this case, of course, the number of minimum tree lengths to be computed for each taxon is five, because a four-taxon tree has five branches. We can then find a taxon that shows the maximum of the minimum tree lengths. This taxon will be the second one to be added to the initial core tree of three taxa. This process is repeated until the addition order of all taxa is determined. Since the maximum of the minimum values is closer to L_M than many other value (e.g., the minimum of the minimum values), this order of taxon addition is expected to speed up the search for the MP tree.

Search for MP Tree(s)

Once the initial core tree and the order of taxon addition are determined, we are in a position to search for the MP tree. Before starting this search, we must have a predetermined upperbound of tree length, that is, L_U for a **temporary MP tree.** This value is a temporary minimum number of substitutions, which is likely to be slightly larger than the real minimum number, L_M. We determine this value by running the heuristic search called the stepwise addition or the branch-and-bound-like algorithm.

Let us now explain the algorithm for finding the MP tree by using the diagrams in Figure 7.5. We start with the initial core tree in diagram A. In this example of five taxa, taxa *a, b,* and *c* form the initial core tree, and taxa *d* and *e* are added in this order. There are three ways of adding *d* to the core tree (trees B, C, and D). We first compute the tree length (L) for tree B. If this L is greater than L_U, we ignore all the subsequent trees that are generated by adding taxon *e* to this tree (five trees given in column E). If $L \leq L_U$, we add *e* to each of the five branches of tree B to form five different trees with five taxa. We again compute L for each of these five trees and find a tree (or trees) that shows the smallest L value. If this L is greater than L_U, then we move on to tree C. However, if L is equal to L_U for a tree, we save the tree as another potential MP tree and move on to tree C. If a tree (or trees) in column E has an L smaller than L_U, then this tree will become the next temporary MP tree, and L_U is now replaced by this new L value. We then move to tree C. We apply the same procedure to tree C and the trees generated by adding *e* to tree C. If all these trees are examined, we then move to tree D and its descendant trees. Since we adjust L_U whenever we find a tree with an L smaller than the previous L_U, we are assured of finding the MP tree. Of course, there may be two or more equally parsimonious trees, and in this case all these trees are identified by the present method. The same algorithm can be used for the case where the number of taxa (m) is greater than five. This algorithm saves computer time considerably, because many trees need not be examined

if L_U is sufficiently close to the tree length (L_M) of the true MP tree(s). However, even this method becomes time-consuming when $m \geq 20$.

Heuristic Search

Several algorithms of the heuristic search for MP trees are now available (see Maddison and Maddison 1992; Swofford and Begle 1993), but many of them are based on the same principle. In these algorithms, a provisional MP tree is first constructed by using a procedure called the **stepwise addition algorithm,** and this provisional MP tree is then subjected to some kind of **branch swapping** to find a more parsimonious tree. In the following, we first explain the principle of the stepwise addition algorithm and then branch swapping procedures. In addition, we present one more heuristic search algorithm whose principle is different from that of the traditional ones.

Stepwise Addition Algorithms

In this set of algorithms, an initial core tree of three taxa is first formed according to a certain rule, and each of the remaining taxa is then chosen for the next taxon addition. This taxon is connected to one of the three branches of the initial core tree, and the tree lengths of the three resulting trees are evaluated. After this evaluation, the tree of four taxa whose tree length is shortest is saved for the next step of taxon addition. The next taxon is then connected to each of the five branches of the four-taxon tree, and the five-taxon tree whose tree length is shortest is chosen. This process is continued until a tree of all taxa is produced. This final tree is the provisional MP tree. This provisional MP tree usually has a longer tree length than that (L_M) of the MP tree. Therefore, this tree is subjected to branch swapping procedures to find a tree that has a smaller L value. Application of several rounds of branch swapping usually produces a tree whose branch length is considerably shorter than that of the provisional tree, and this tree is regarded as the MP tree.

Swofford and Begle (1993) describe various ways of producing the provisional MP tree considering the order of taxon addition. The simplest one is the **"as is"** option, in which the initial core tree is produced by the first three taxa given in the data set, and the following taxon addition is done according to the taxon order in the data set. Usually this method is not very effective for finding a tree with a small L. The second simplest method is the **"random"** option, where pseudorandom numbers are used to determine the order of taxon addition, and this procedure is applied many times to obtain a provisional MP tree. Another one is called the **"closest"** option, in which the initial core tree is produced by examining all triplets of taxa and choosing the one that shows the smallest L, and in the following steps, a taxon whose addition to the previous core tree shows the smallest increase in L is chosen. The reader should refer to Swofford and Begle (1993) for details of these options. All of these options are included in the software PAUP*, whereas PHYLIP primarily uses the second (Jumble) option. Once a provisional MP tree is produced,

the tree is subjected to one or two of the following algorithms of branch swapping.

Branch Swapping

The most popular algorithms of branch swapping are (1) **nearest neighbor interchanges (NNI)**, (2) **subtree pruning regrafting (SPR)**, and (3) **tree bisection-reconnection (TBR)** (Swofford and Begle 1993). The first algorithm is the same as the examination of all trees that are different from the provisional MP tree by a topological distance of $d_T = 2$ (Figure 7.6A). For example, for the five-taxon tree in Figure 7.6A, there are two alternative trees (interchanges of taxa 2 and 3 and taxa 1 and 3) with a topological distance of $d_T = 2$ from the original tree when the interior branch a is considered. Two more alternative trees can be produced if we consider the interior branch b. This algorithm is obviously related to the *close neighbor interchange* (CNI) algorithm described in relation to the ME method (chapter 6). In the latter algorithm, the trees that are different from the provisional tree by $d_T = 2$ and 4 are examined, and this search is repeated until no tree with a smaller L is found. Therefore, this algorithm examines more trees than the NNI search.

In the SPR algorithm, a branch of a provisional tree is cut into two parts, a pruned subtree and the residual tree. The cutting point of the pruned subtree is then grafted onto each branch of the residual tree to produce a new topology. This is done for all branches of the residual tree to produce more trees to be examined. This is illustrated in Figure 7.6B. In this example, the exterior branch a was cut, and the pruned subtree consists of taxon 1 only, whereas the residual tree is composed of taxa 2, 3, 4, and 5. There are four ways of grafting the subtree to the residual trees in this case. If the interior branch b is cut instead, the subtree is grafted to two exterior branches (4 and 5) of the residual tree to produce two alternative trees. This procedure can be used for a tree of any number of taxa.

In the TBR search, a provisional tree is cut into two subtrees at a branch, and these two subtrees are then reconnected by joining two branches, one from each subtree, to generate a different topology (Figure 7.6C). This is tried for all possible pairs of branches of the two subtrees to generate many different topologies. In the SPR search, only the cutting point of a subtree was regrafted to a branch of the residual tree, whereas in the TBR search all combinations of branches from the two subtrees are considered for reconnection. Therefore, the number of topologies generated is larger than that generated by the SPR search when the number of taxa is greater than five.

Since the TBR search examines a larger number of trees than the NNI and SPR searches, many investigators use this method. However, even this method examines only a limited number of trees when the number of taxa is large (Maddison 1991). One way to increase the number of trees to be examined is to use the "random" option of stepwise addition and the TBR search repeatedly. If this approach is used a large number of times, the chance of finding the MP or a suboptimal MP tree is quite high.

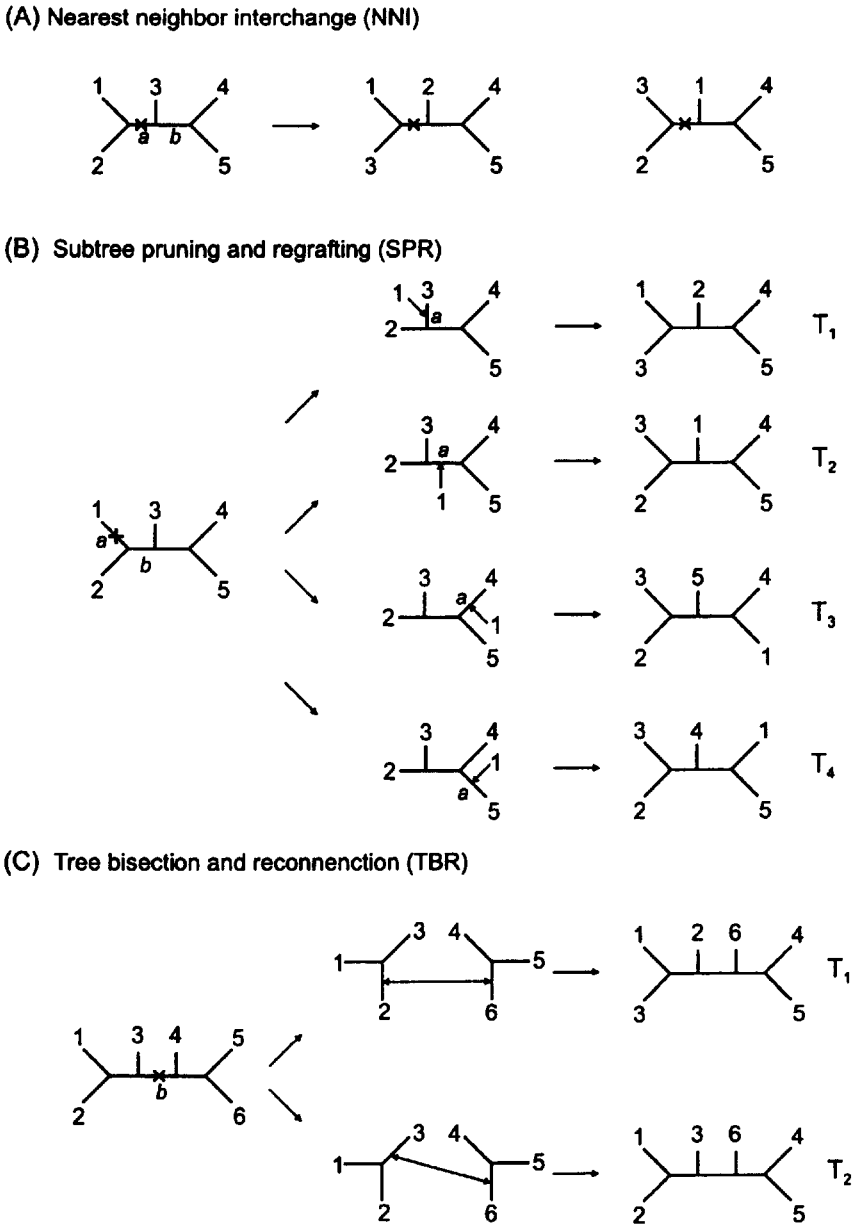

FIGURE 7.6. Three different methods of branch swapping for finding MP trees.

Branch-and-Bound-Like Algorithm

Kumar et al. (1993) proposed a heuristic search algorithm, which is conceptually different from the algorithms mentioned above but is similar to the branch-and-bound method. In this algorithm, we start with an initial core tree of three taxa that is determined as in the case of the branch-and-bound method. The order of taxon addition is also determined in a similar fashion except for the following. In the branch-and-bound

method, we computed the minimum numbers of substitutions for all taxa for each core tree (each step of taxon addition) and then chose the taxon that showed the maximum value among all the minimum values. For this heuristic search, which may be called the **min-mini algorithm,** we choose the minimum of all the minimum values, because we are not going to do a semiexhaustive search as in the case of the branch-and-bound method and want to reach the MP or a suboptimal MP tree relatively quickly.

The algorithm of searching for the MP tree is also similar to that of the branch-and-bound method. Let us again consider Figure 7.5 to explain this algorithm. As before, we start with the core tree A and first connect taxon *d* to branch *c* to produce tree B. We then compute the tree length (*L*) of this tree. We call this *L* value the **local upperbound** (L_1) for the first taxon addition and keep this value for future use. We then connect taxon *e* to branch *a* of tree B to produce tree E (1). We again compute the *L* value of this tree and call it the local upperbound (L_2) for the second taxon addition. If there is another taxon (*f*) to be added, we connect this taxon to branch *a* of tree E (1) and obtain tree E (1, 1) in Figure 7.7. If *f* is the last taxon to be added, we now compute the *L* value not only for tree E (1, 1) but also for all other six trees that can be derived from tree E (1). We then choose the tree that shows the smallest *L* value among the seven trees and call it a temporary MP tree. The *L* for this tree is the temporary upperbound (L_U) in this case.

The next step is to go back to tree E (2) in Figure 7.5 and compute the *L* value. If this *L* is greater than L_2, we neglect all trees that can be generated by adding *f* to this tree. If $L = L_2$, we compute *L* for all the descendant trees. If any of the descendant trees show an *L* equal to L_U, the tree is saved as another potential MP tree. If there is any tree showing an *L* less than L_U, this tree now becomes a new temporary MP tree, and the previous L_U is replaced by this *L*. By contrast, if tree E (2) shows an *L* less than L_2, L_2 is replaced by this *L*. The *L* values for all descendant trees are then computed, and a new potential MP tree or a new temporary MP tree is searched for. This procedure is applied to the remaining three trees E (3), E (4), and E (5) of five taxa, and the temporary MP tree (or trees) that shows the smallest *L* value among the 35 ($= 5 \times 7$) trees derived from tree B is determined.

If the above computation is completed, we now move on to tree C (and tree D) in Figure 7.5 and apply the same procedure to all trees that can be derived from these trees. When this is completed, we have the final tree or trees. When there are more than six taxa, essentially the same algorithm is applied. The only difference is that there are many steps of taxon addition and that at each step of taxon addition the local upperbound (L_1, L_2, L_3, . . . , L_{m-3}, or L_U) is computed, where *m* is the number of taxa. L_1, L_2, L_3, . . . , L_{m-4}, and L_U are then used to determine whether a group of descendant trees should be ignored or not in later computations.

In this algorithm, many trees that are unlikely to have a small *L* value are ignored, and thus the algorithm speeds up the search for the MP tree. However, the final tree or trees obtained by this algorithm may not be the true MP tree(s), because the upperbounds of the *L* values used here are

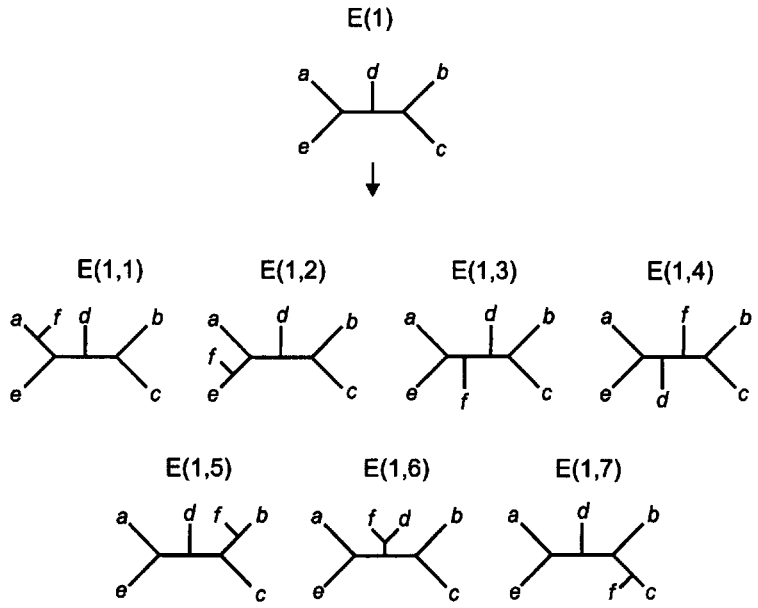

FIGURE 7.7. All possible trees that can be generated by adding taxon *f* to tree E(1) in Figure 7.5.

local upperbounds rather than the global upperbound as used in the branch-and-bound method, and the tree with the global minimum value of L may not have been obtained.

There is a way to improve the efficiency of finding the MP tree. It is to increment the local upperbound at each step of taxon addition. If the local upperbound is large, the number of trees to be examined automatically increases. In the above algorithm, the local upperbound at the i-th step of taxon addition was L_i except for the first step. We now increase L_i by x_i so that the upperbound is given by $L_i' = L_i + x_i$. If x_i is large for all i's, a large number of topologies will be examined. In this case, however, the computational time will be prohibitively large. We call x_i a **search factor.** In the default option of MEGA, $x_i = 2$ is used for all steps, but the user may change it as desired. MEGA2 has another option, in which x_i is defined as $x_i = p(L_{i+1} - L_i)$, where L_i and L_{i+1} are the upperbound of L for the i-th and the $(i + 1)$-th steps, and p is the fraction of $L_{i+1} - L_i$ that one wishes to use. Suppose $L_{i+1} = 200$ and $L_i = 180$, and one wishes to use $p = 0.1$. Then, x_i becomes 2. We call the p value a **proportional search factor.** The optimum p value varies with data set, and the user of MEGA2 may find it by trial and error.

Some Remarks

As mentioned earlier, MP methods tend to give incorrect topologies when the number of sequences used (m) is large and the number of nucleotides used (n) is small. In this case, one may choose an incorrect topology by making an excessive effort to find the real MP tree. When m

is large, some parts of the MP tree (or any other tree) are likely to be incorrect, and a submaximum parsimony tree may be as good as the MP tree in finding the true topology. For this reason, Nei et al. (1998) suggested that a relatively crude method of finding MP or potential MP trees gives essentially the same conclusion about phylogenetic inference as the exhaustive search when the accuracy of the tree obtained is examined by the bootstrap test. In fact, our computer simulation (Takahashi and Nei 2000) has shown that for randomly generated model trees of 48 sequences with n = 1000 the NNI search of MP trees is as efficient as the TBR search in inferring the true tree. This indicates that MP trees are often incorrect and that there is no need to spend an enormous amount of computer time for obtaining MP trees.

7.3. Consensus Trees

Strict and Majority-Rule Consensus Trees

As mentioned above, MP methods often produce several equally parsimonious trees. In this case, it is difficult to present all the trees for publication. One way to solve this problem is to make a composite tree that represents all the trees. Such a composite tree is called a consensus tree.

There are several different types of consensus trees (Swofford and Begle 1993), but the most commonly used ones are the **strict consensus trees** and the **majority-rule consensus trees.** Let us explain these trees using the examples given in Figure 7.8. Suppose that trees A, B, and C are three equally parsimonious trees obtained by an MP method. In a strict consensus tree, any conflicting branching patterns for a set of sequences among the rival trees are resolved by forming a multifurcating branching pattern. Thus, the strict consensus tree for trees A, B, and C is given by tree D. Among the majority-rule consensus trees, the most commonly used is the 50% majority-rule consensus tree. In this tree, a branching

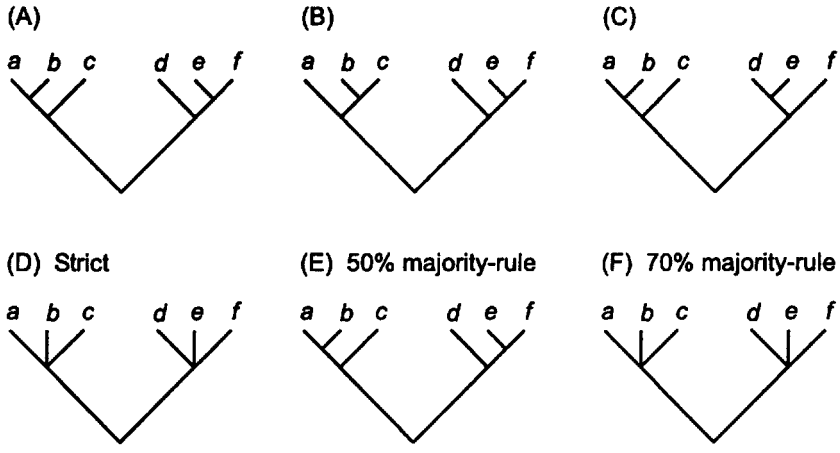

FIGURE 7.8. Examples of consensus trees.

pattern that occurs with a frequency of 50% or more is adopted. In the present example, the branching pattern ((*a, b*) *c*) for taxa *a, b,* and *c* occurs two times among the three rival trees, so this pattern is adopted. Similarly, branching pattern ((*e, f*) *d*) occurs two times among the three trees. Therefore, the 50% majority-rule consensus tree is given by tree E. It is possible to increase the majority-rule percentage. For example, if we use 70%, none of the branching patterns of the two three-taxon clusters reaches 70%. Therefore, the 70% majority-rule consensus tree (tree F) is identical with the strict consensus tree. Note that the 100% majority-rule consensus tree is always identical with the strict consensus tree.

Bootstrap Consensus Trees

One of the effective ways of testing the reliability of an MP tree is to use the bootstrap test, which will be discussed in chapter 9. In this test, the reliability of an inferred tree is examined by using Efron's bootstrap resampling techniques. A set of nucleotide sites is randomly sampled with replacement from the original set, and this random set that has the same number of nucleotide sites as that of the original set is used for constructing a new tree. The topology of this tree may be or may not be the same as that of the inferred tree. This process is repeated many times (over 100 times), and the reliability of the inferred tree is evaluated by the percentage of times in which each branching pattern (sequence partition) is found among all the replicate bootstrap trees (see chapter 9 for details).

Felsenstein (1985) proposed to construct a consensus tree from the replicate bootstrap trees and use it as a new inferred tree. This new inferred tree may be different from the original inferred tree, but since it is an "average" tree of many bootstrap trees, it may be more reliable than the original one, though there is no proof. In the case of MP trees, this procedure also has an advantage to avoid multifurcating trees by producing a low-percentage majority-rule consensus tree. In Figure 7.8, we saw that the 50% majority-rule tree for trees A, B, and C is a bifurcating tree. If we have several hundred bootstrap trees and if we make a 5% majority rule consensus tree, the tree will be almost always a bifurcating tree.

7.4. Estimation of Branch Lengths

MP methods are often used to construct a tree topology without branch lengths. However, it is possible to estimate the branch lengths of a reconstructed tree under certain assumptions, and these estimates should be presented for an MP tree as much as possible.

The branch lengths of an MP tree are estimated by considering all evolutionary pathways at each variable site and computing the average number of substitutions for each exterior or interior branch. When there is only one singleton substitution at a site, this substitution can always be assigned to the exterior branch leading to the taxon that has the substitution. When there are two or more singleton substitutions, there are sev-

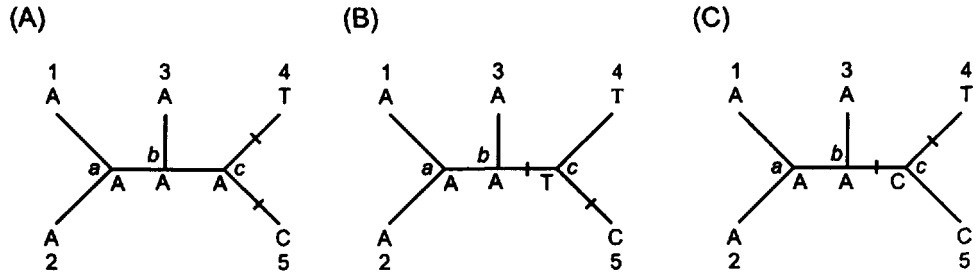

FIGURE 7.9. Assignment of substitutions to different branches when there are two or more singleton substitutions.

eral ways of assigning the substitutions (evolutionary pathways). For example, Figure 7.9 shows a case of two singleton substitutions, in which there are three different evolutionary pathways, and the average number of substitutions for each of the branches $b - c$, $c - 4$, and $c - 5$ is 2/3. In the case of trees D–F in Figure 7.3, there are three singleton substitutions for the same topology, and the number of substitutions for the branch leading to taxon 3 is 1, whereas the number for the branch leading to taxa 4 and 5 is 2/3. In addition, one interior branch has the average number of 2/3 substitutions. Therefore, if we know all the pathways, we can compute the average number of substitutions for each branch.

This is also true for parsimony-informative sites. Let us consider this problem using the tree given in Figure 7.1. We have seen that the nucleotides observed in the six extant taxa generate four equally parsimonious pathways in the unrooted topology of this tree. They are (*a* - T, *b* - T, *c* - T, *d* - T), (*a* - T, *b* - T, *c* - A, *d* - A), (*a* - C, *b* - T, *c* - A, *d* - A), and (*a* - A, *b* - T, *c* - A, *d* - A). The first pathway requires one substitution for each of the branches 1-*a*, 5-*c*, and 6-*d*, and the second requires one substitution for each of the branches 1-*a*, *b*-*c*, and *a*-*d*. Similarly, the third and fourth pathways require one substitution for branches 2-*a*, *b*-*c*, *a*-*d*, and 1-*a*, 2-*a*, *b*-*c*, respectively. We can therefore compute the average numbers of substitutions for all branches. We obtain the numbers 3/4, 2/4, 0, 0, 1/4, 1/4, 2/4, 3/4, and 0 for branches 1-*a*, 2-*a*, 3-*b*, 4-*b*, 5-*c*, 6-*d*, *a*-*d*, *b*-*c*, and *c*-*d*, respectively. The length of a branch can then be obtained by adding all substitutions for that branch at both singleton and informative sites. We call this way of estimating branch lengths the **average pathway method.**

Maddison and Maddison (1992) and Swofford and Begle (1993) also estimate the branch lengths by using two algorithms: **Acctran** and **Deltran.** In Acctran, evolutionary changes of nucleotides are assumed to occur as soon as possible from the root, whereas in Deltran the changes are assumed to occur as late as possible from the root (Swofford and Maddison 1987). As an example, let us consider tree B of Figure 7.1 and assume that the nucleotide (A) for sequence 6 represents the ancestral nucleotide at node *e*. In the Acctran algorithm, A is assumed to change to T at the earliest node *d*, and then the minimum number of nucleotide changes is considered. Therefore, nodes *a*, *b*, and *c* are also assumed to be T. In the Deltran algorithm, however, nucleotide changes are delayed

as much as possible. Therefore, the nucleotides at nodes *a*, *b*, *c*, and *d* are assumed to be A, T, A, and A, respectively. (Node *e* is not considered.) This indicates that the nucleotide assignments for the ancestral nodes are considerably different between Acctran and Deltran, and therefore the estimates of branch lengths will also be different. When closely related sequences are examined, however, the difference in branch length estimates between the two methods is not as large as one might suspect.

In general, the estimates of branch lengths obtained by parsimony methods tend to be smaller than the actual values, particularly when sequence divergence is high. One way to avoid this underestimation of branch lengths is to use the least squares or the ML method after the topology of the tree is determined by MP methods. Under certain conditions, MP methods seem to be superior to distance or ML methods for finding the correct topology (see chapter 9). Therefore, this approach may give a more reliable topology and more reliable estimates of branch lengths.

7.5. Weighted Parsimony

As mentioned earlier, MP methods are expected to produce more reliable trees when the number of backward and parallel substitutions (extent of homoplasy) is small than when it is large. Therefore, if a set of sequences used for phylogenetic analysis includes fast-evolving and slow-evolving sites, one would expect that the latter sites are more useful than the former sites for constructing MP trees when distantly related sequences are studied. Therefore, if we give more weight to slow-evolving sites than to fast-evolving sites, a more reliable tree may be obtained than when they are equally weighted (Farris 1969; Swofford et al. 1996).

For example, the nucleotides at the first, second, and third codon positions of protein-coding genes are known to evolve at different rates, those at the third positions evolving fastest and those at the second positions slowest (see Table 3.4). Therefore, one may give such weights as $w_1 = 3$, $w_2 = 5$, and $w_3 = 1$ for the first, second, and third positions, respectively. These weights would of course vary from gene to gene. Generally speaking, functionally less important parts of a gene are known to evolve faster than more important parts (Dickerson 1971; Kimura 1983). Therefore, one may give the former a lower weight than the latter when distantly related sequences are studied.

Weighted parsimony also allows different weights to be given to different types of substitutions at a given site. For example, transitional nucleotide substitutions generally occur more frequently than transversional substitutions, as mentioned earlier. In this case, it is convenient to use a **substitution weight matrix** as given in Figure 7.10A. A weight matrix is sometimes called a **step matrix** (Swofford and Begle 1993). If transitions occur twice as frequently as transversions, we may give $w = 2$. The matrix in Figure 7.10B gives a weight of 0 to all transitional changes and 1 to all transversional changes. Therefore, transitional changes are completely ignored, and only transversional changes are considered. This type of MP method is called **transversion parsimony.**

(A) Weighted parsimony

	A	T	C	G
A		*w*		1
T	*w*			*w*
C	*w*	1		*w*
G	1	*w*	*w*	

(B) Transversion parsimony

	A	T	C	G
A		1	1	0
T	1		0	1
C	1	0		1
G	0	1	1	

FIGURE 7.10. (A) Weight matrix for transitional and transversional substitutions. (B) Weight matrix for transversion parsimony.

Figure 7.4D shows a weighted parsimony tree obtained from the DNA sequence data for the five hominoid species considered earlier (Figure 6.1). Kimura's formula for estimating the transition/transversion ratio (Equation [3.18]) gave $\hat{R} = 6$ for this set of data. This suggests that the transition rate is about six times higher than the transversion rate. We therefore used $w = 6$ in constructing the weighted parsimony tree. Interestingly, the topology of this tree is the same as that of the trees obtained by distance methods (Figure 6.2).

However, there are some problems with weighted parsimony. First, we usually do not know the appropriate weights to be used in actual data analysis. In some cases, information from previous studies can be used, but there is no guarantee that they are appropriate for the data set under consideration. For this reason, a number of authors (Farris 1969; Sankoff and Cedergren 1983; Williams and Fitch 1990) proposed a method called **dynamically weighted parsimony.** In this method, a set of weight parameters that appear to be appropriate a priori are first used to construct an MP tree, and this tree is now used to obtain an improved set of weight parameters. These new parameters are then used to construct a new MP tree. This process is repeated until a stable tree (or trees) is obtained. This is a time-consuming method and does not guarantee convergence to a stable tree. Nevertheless, computer simulation has shown that this method improves the probability of obtaining the correct tree under certain conditions (Fitch and Ye 1991). Second, slowly evolving sites or slowly changing substitution types are informative only when distantly related sequences are studied. When closely related sequences are used, the fast-evolving sites or types of substitutions are obviously more informative. However, actual data often include both distantly related and closely related sequences, and in this case, it is not clear how useful weighted parsimony is. This problem needs more investigation.

Example 7.2. Unweighted and Weighted Trees for Simulated Sequence Data

One of the virtues of MP methods is that when there are no backward or parallel substitutions and there are a sufficiently large number of informative sites, they are able to reconstruct the true tree irrespective of the pattern of nucleotide substitution. This suggests that they will produce a highly reliable tree when the extent of sequence divergence is low. To see whether this is the case or not, we constructed MP trees using the

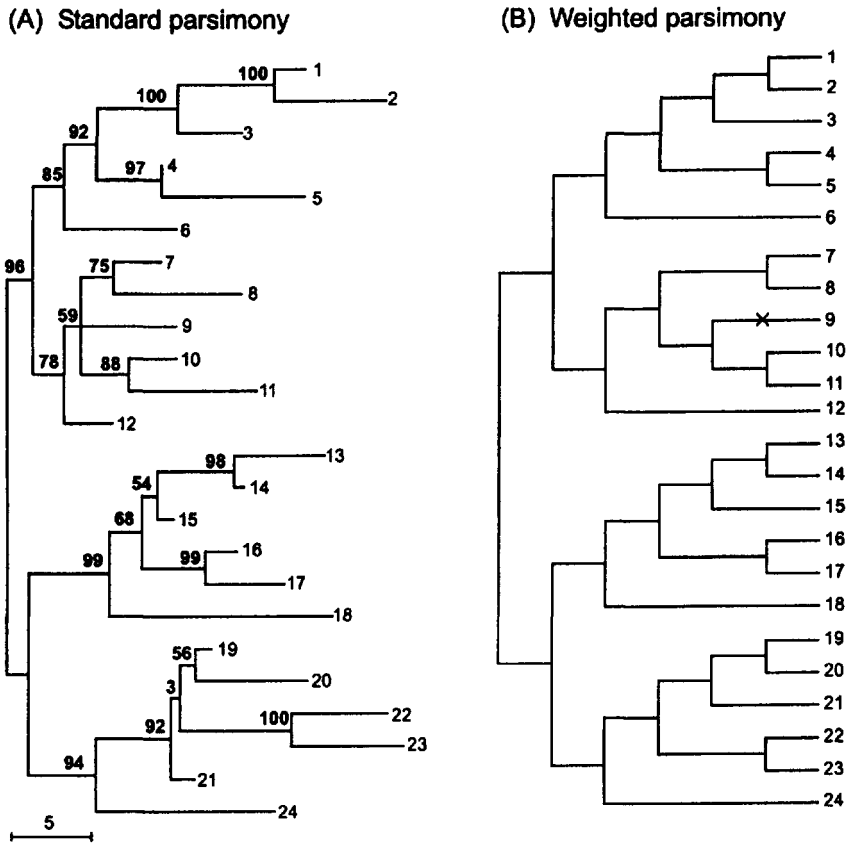

FIGURE 7.11. (A) Standard parsimony tree for simulated sequence data as inferred by the stepwise-addition with "closest" option in PAUP* (no branch swapping). Bootstrap values are shown in boldface (100 replications). (B) Weighted parsimony for the same data set ($w = 5$ was used). The branching pattern of sequence 9 is incorrect (see Figure 6.8A).

24 DNA sequences discussed in chapter 6 (Figure 6.8). We first used the stepwise addition algorithm with the "closest" option in PAUP* to construct the MP tree and the average pathway method to estimate the branch lengths. The tree obtained is presented in Figure 7.11A. Comparison of this tree with that in Figure 6.8B indicates that the MP tree is virtually the same as the true (realized) tree. The only topological difference observed is the interchange of sequence 21 with the cluster of sequences 22 and 23. The branch lengths are also very close to those of the true tree, though there is some tendency for the interior branches of the MP tree to be underestimated because of homoplasy. When we tried a heuristic search with 50 replications of TBR branch swapping, we found two more MP trees, which were different with respect to the splitting pattern of sequences 21, 22, and 23.

These results indicate that when sequence divergence is low, MP trees are very close to the true tree. They are also similar to the NJ tree in Figure 6.8C. It is interesting to note that NJ resolved the branching pattern of sequences 21, 22, and 23 correctly but did not produce a trifurcating node

for sequences 8, 9, and 10. By contrast, MP had no trouble identifying the trifurcating node but did not produce the correct branching pattern for sequences 21, 22, and 23.

We also constructed a weighted MP tree with a transition/transversion ratio (w) of 5. (Note that the DNA sequences used were generated with a transition/transversion ratio [R] of 5.) The tree is presented in Figure 7.11B. The topology of this tree is identical with that of the realized tree B in Figure 6.8 except for one topological error that occurred with respect to the branching pattern of sequence 9. The branch lengths of this tree do not have any biological meaning.

Example 7.3. Origin of Whales

Whales are the largest animals that have ever lived on Earth. They belong to the mammalian order Cetacea that includes whales, dolphins, and porpoises, which are all adapted to aquatic life. The evolutionary origin of cetaceans has been a mystery over a century. In recent years, however, their evolutionary relationships with other mammalian orders are being clarified thanks to molecular data. Although whales were once believed to be related to horses, elephants, or some other mammalian order, it is now generally agreed that they are most closely related with artiodactyls. The order Artiodactyla was traditionally divided into three suborders, Ruminantia (e.g., deer, giraffes, cows, sheep, chevrotains), Tylopoda (e.g., camels), and Suiformes (e.g., pigs, peccaries, and hippopotamuses), and each of these suborders had been considered to be monophyletic. Recent molecular data, however, suggest that the order Cetacea is most closely related to Ruminantia, and therefore Cetacea is included inside the order Artiodactyla (Graur and Higgins 1994; Gatesy 1997; Shimamura et al. 1997).

Here we construct a phylogenetic tree using DNA sequences of the blood-clotting protein γ-fibrinogen gene. This gene consists of 10 exons and spans an 8 kb region of nuclear DNA. Gatesy (1997) sequenced a 523–581 bp fragment of the gene for six species of artiodactyls, three species of cetaceans, two species (horse and Asiatic tapir) of Perissodactyla (odd-toes ungulates), and two species (spotted hyena and coyote) of Carnivora. Adding the human sequence available, Gatesy constructed an MP tree using PAUP* with 50 random taxon replicates and TBR branch swapping. Here we used the branch-and-bound method to construct the MP tree. The number of nucleotides used was 433 after elimination of all alignment gaps. The branch-and-bound method produced three equally parsimonious trees, one of which is presented in Figure 7.12A. The other two trees had different branching patterns for sheep, giraffe, and moose, that is, ([sheep, giraffe] moose) and ([sheep, moose] giraffe) instead of ([moose, giraffe] sheep). All three trees had a tree length of 485 substitutions. When a 30% bootstrap majority-rule consensus tree was constructed from 500 replications, we obtained the same tree as that of Figure 7.12A. However, the bootstrap value of the giraffe-moose cluster is so low that the branching pattern for sheep, giraffe, and moose remains unresolved. We also constructed the NJ and ME trees using Kimura distance for this data set. The NJ tree was identical with

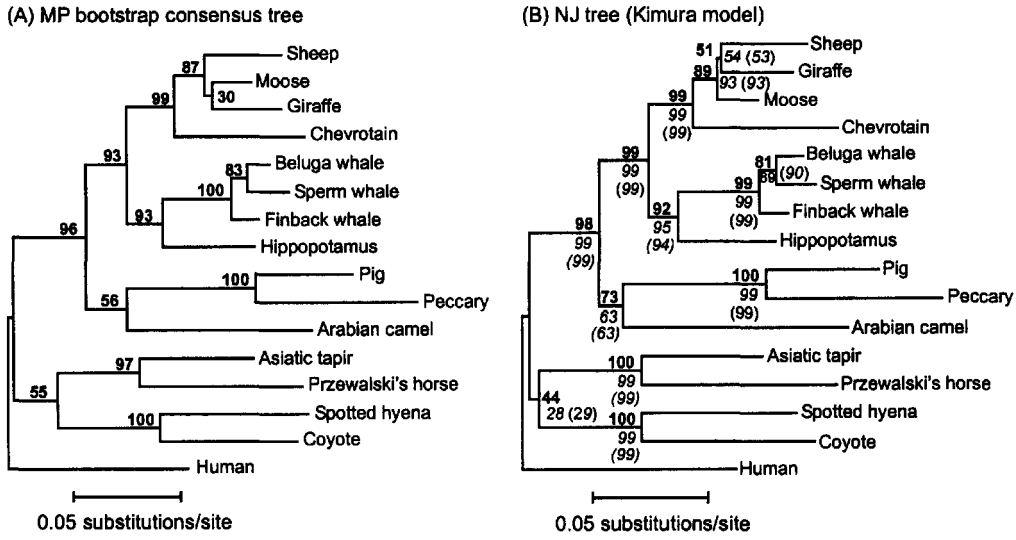

FIGURE 7.12. (A) MP bootstrap consensus tree obtained by the fast-heuristic search of PAUP*. (B) NJ tree. Bootstrap values are shown in boldface (1000 replications), and the PC values (for NJ tree) are shown in italics. The PC values obtained by Dopazo's method are shown in parentheses.

the ME tree and is given in Figure 7.12B. This tree has the same topology as that of tree A except for the branching pattern of sheep, giraffes, and moose, which again has low bootstrap values. Therefore, both MP and NJ trees give essentially the same conclusion as to the phylogenetic relationships of the organisms studied.

Both trees A and B show a branching pattern that is unexpected from the classical taxonomy. That is, cetaceans are close relatives of ruminants, and the other two suborders of Artiodactyla, i.e., Tylopoda and Suiformes, are outgroups of ruminants and cetaceans. This indicates that the order Artiodactyla is not monophyletic but **paraphyletic** (Wiley et al. 1991), because it does not include Cetacea, which is a close relative of the suborder Ruminantia. Since this classification is unnatural, Montgelard et al. (1997) proposed a new mammalian order named Cetartiodactyla that includes both Artiodactyla and Cetacea. This conclusion has been supported by Gatesy et al.'s (1999) further study using a larger set of DNA sequences. It is also supported by the work of Nikaido et al. (1999), who used an entirely different approach (see section 7.7).

7.6. MP Methods for Protein Data

Eck and Dayhoff (1966) used an MP method for protein sequence data. They considered 20 different amino acids as character states and constructed an MP tree, assuming that the evolutionary change can occur in all directions among the 20 amino acids. Dayhoff and her collaborators (Dayhoff 1972) used this method extensively and obtained quite reason-

able trees for various protein sequences. A computer program incorporating this method is available in PAUP* and MEGA2.

Theoretically, this approach is approximate, because some amino acid changes require two or three nucleotide substitutions, whereas other changes can be explained by one substitution. Furthermore, some amino acids are biochemically similar to one another, and substitution occurs more often within each group of similar amino acids than between groups. For this reason, a number of authors (Moore et al. 1973; Fitch and Farris 1974; Sankoff and Rousseau 1975; Felsenstein 1988) have developed various protein parsimony algorithms, taking into account the minimum number of nucleotide substitutions between any pair of amino acids and using the sum of these numbers to compute the tree length. Felsenstein's program PROTPARS in PHYLIP uses one of these algorithms (Felsenstein 1995). However, these algorithms are quite elaborate and depend on a number of simplifying assumptions. Therefore, it is not clear whether they are superior to Eck and Dayhoff's original version. Russo et al.'s (1996) empirical study suggests that Eck and Dayhoff's method is quite efficient in obtaining the true tree.

When DNA sequence data became available in the 1980s, many investigators started to use them for phylogenetic inference. However, it gradually became clear that the evolutionary pattern of DNA sequences is so complex that they are not necessarily better than protein sequences. One problem with DNA sequences is that the substitution pattern is not the same for all nucleotide positions within codons and the GC content in third nucleotide positions often varies with species. For example, the rate of nucleotide substitution at the third codon positions in hominoid mitochondrial genes is so high that the proportion of different nucleotides reaches the saturation level rather quickly (Ruvolo et al. 1994). For these reasons, protein sequences are now again used for phylogenetic reconstruction, and Eck and Dayhoff's simple MP method often gives better results than DNA MP methods.

Example 7.4. MP and NJ Trees from the Cytochrome *b* Gene

The mitochondrial DNA (mtDNA) in vertebrates contains 13 protein coding genes, and the entire sequence of mtDNA is available for a substantial number of organisms. Russo et al. (1996) chose 11 organisms of which the evolutionary relationships are known from paleontological and morphological data and for which complete mtDNA sequences are available and then examined the ability of each gene to reconstruct the correct phylogeny. Here we consider only the cytochrome *b* gene, which is often used for phylogenetic inference. We first constructed the MP tree using amino acid sequence data for cytochrome *b,* which is composed of 377 amino acids. When the 11 species given in Figure 7.13A were used, there were 121 informative sites and 44 uninformative variable sites. The branch-and-bound search produced a single MP tree, which is presented in Figure 7.13A. The topology of this tree is identical with the biological tree we already know. An interesting observation about this tree is that the opossum, chicken, Xenopus, and fish sequences show considerably shorter branch lengths compared with what one would expect under the

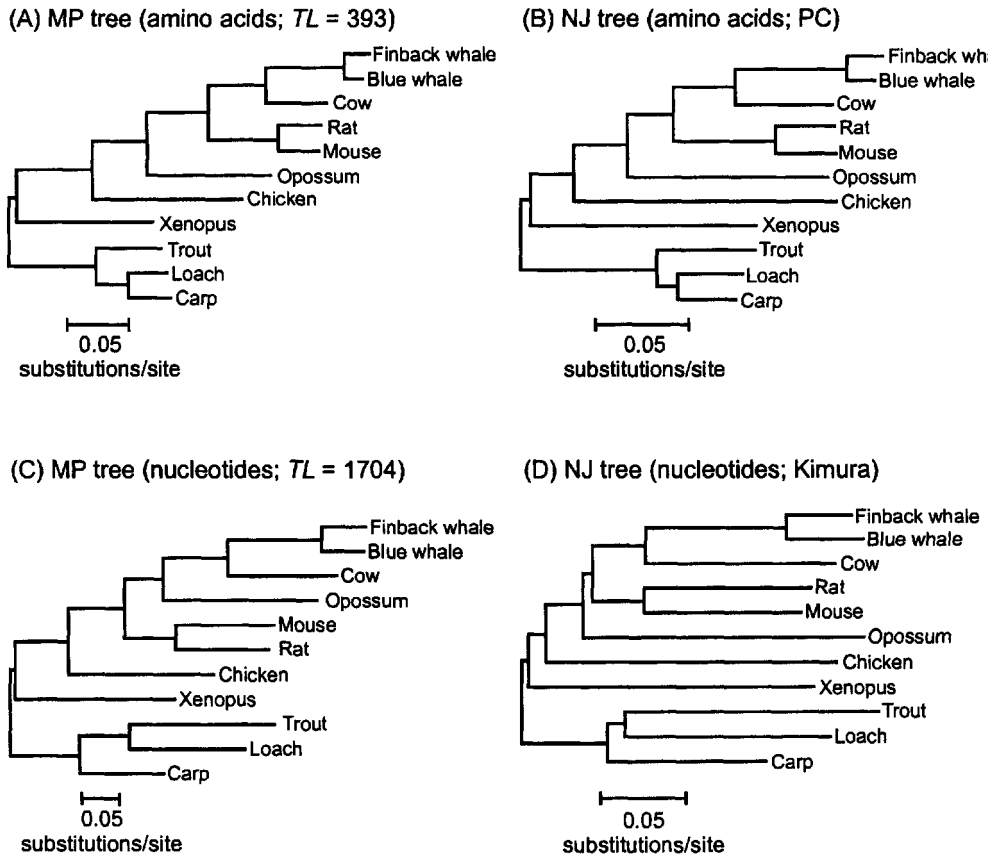

FIGURE 7.13. Inference of a "known" phylogeny of 11 vertebrates using the nucleotide and amino acid sequences of the mitochondrial cytochrome *b* gene. *TL*: Tree length. PC: Poisson correction distance. Kimura: Kimura distance.

molecular clock. Figure 7.13B shows the NJ (and ME) tree obtained by using the Poisson-correction distance. This tree also shows the correct topology, but the branch lengths for the opossum, chicken, and Xenopus are much longer than those of the MP tree.

Figure 7.13C shows the branch-and-bound MP tree for the nucleotide sequence data (1,131 bp long). In this case, there were 494 informative sites and 111 uninformative variable sites. Yet, the topology of the tree is wrong with two branch switches. That is, the opossum should be between chicken and the rodents, and the loach should be closer to the carp rather than to the trout. This indicates that a large number of informative sites alone does not necessarily produce a better tree. The wrong topology of the DNA tree appears to be caused by the fact that nucleotide differences at third codon positions have reached the saturation level, and they introduced noise in phylogenetic construction. However, even when we used only first and second codon position data, the topology was still incorrect with respect to the branching pattern of the three fish species. We also constructed the NJ and ME trees using the Kimura distance for all three codon position data. These trees were identical with

each other but showed one topological error with respect to the three fish species (Figure 7.13D). These results suggest that protein sequences are better than DNA sequences at least in this case. Russo et al. (1996) examined all the 13 protein-coding genes and found that the above conclusion is generally true.

7.7. Shared Derived Characters

Irreversible Shared Derived Characters

If a group of species share a unique and irreversible mutation (mutant character), they must be derived from the same common ancestral species in which this mutation occurred. We call this type of mutations **irreversible shared derived characters.** These characters are very useful for phylogenetic construction (Hennig 1950, 1966). For example, Figure 7.14 shows a phylogenetic tree for four species (1, 2, 3, and 4), in which one mutation ($a \rightarrow a'$, $b \rightarrow b'$, or $c \rightarrow c'$) has occurred in each of the three interior branches. Because these mutations are assumed to be unique and irreversible and each mutation defines a **clade** (a monophyletic group of species), they define the tree unambiguously. Furthermore, since the mutations are directional and the ancestral characters (a, b, and c) are known, we can infer a rooted tree without outgroup species. When the number of species is small, the phylogenetic tree can be easily determined from the distribution of character states among the species. However, when the number of species and the number of characters used are large, the topology and the assignment of mutations for each branch can be cumbersome. In this case, we can use the computer program incorporated in PAUP*.

In **cladistic parsimony,** where the clarification of the evolutionary changes of characters in the phylogeny is emphasized, only shared derived characters or **synapomorphies** are considered to be useful for con-

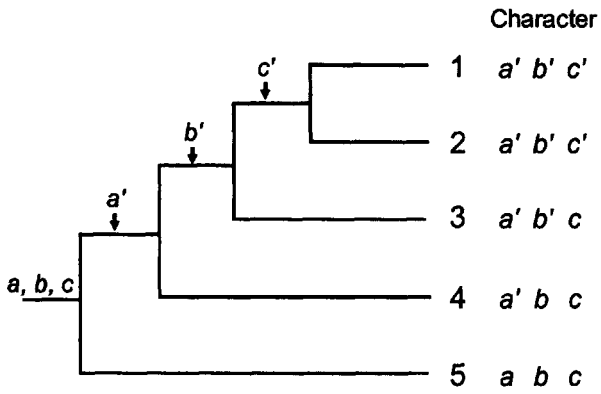

FIGURE 7.14. Phylogenetic tree for five species (1, 2, 3, 4, and 5), which is determined by irreversible mutations ($a \rightarrow a'$, $b \rightarrow b'$, and $c \rightarrow c'$). a, b, and c are the ancestral characters, and a', b', and c' are the derived characters. In this case, the root of the tree can also be determined.

structing phylogenetic trees (Henning 1966; Eldredge and Cracraft 1980; Sober 1988; Wiley et al. 1991). However, since the evolutionary changes of morphological characters and the nucleotide substitutions in DNA sequences are usually reversible, most parsimony analyses allow the reversibility of character states. Exceptions are Henning's (1950) strict cladistic analysis and Camin and Sokal's (1965) parsimony. The latter method is different from the Hennigian parsimony in that the same mutation may occur independently in different evolutionary lineages. In the past, however, these methods have rarely been used in actual data analysis because of the apparently unrealistic assumption of irreversibility (Wiley et al. 1991; Swofford et al. 1996).

SINEs and LINEs

However, recent molecular studies have shown that the genomes of higher organisms contain many unique shared derived characters that are apparently irreversible. Among the most well studied are **short interspersed repetitive elements (SINEs)** and **long interspersed repetitive elements (LINEs)** (Singer 1982; Jurka et al. 1988; Britten et al. 1988). SINEs are short sequences of 80–400 nucleotides, whereas LINEs are usually repeats of a few hundred to a few thousand nucleotides. Both SINEs and LINEs are retropseudogenes but are capable of self-replication. Replicated repeat elements are inserted at different locations of the genome, and once they are inserted, they are almost never excised unless they are eliminated by a rare event of large-scale DNA deletion (Hamdi et al.1999; Nikaido et al. 1999). These repeat elements are subject to mutation and minor insertions/deletions and lose their identity in the long run. However, if one is interested in constructing a phylogenetic tree for relatively closely related species (divergence times of up to about 50 million years), these repeat elements can be used as shared derived characters (e.g., Ryan and Dugaiczyk 1989; Okada 1991; Murata et al. 1993; Furano et al. 1994; Verneau et al. 1997).

SINEs and LINEs are identified by appropriate primer DNA sequences, but if they accumulate a substantial number of mutations, the primers may not be able to detect the repeat elements. If this happens for some of the species examined, they are treated as missing characters (Nikaido et al. 1999). When the elements are eliminated by rare deletion events, they are also treated as missing characters. However, since SINEs and LINEs are irreversible mutations, they are still very useful for phylogenetic analysis.

The most well-known family of SINEs is the *Alu* family, which has about 300,000 members in the human and the ape genomes. The members of this family are pseudogenes originally derived from 7SL RNA, one component of the signal recognition particle (Ullu and Tschudi 1984). The SINE families in other organisms are usually pseudogenes derived from various types of t-RNAs rather than 7SL RNA (Kido et al. 1991; Okada et al. 1997). Therefore, there are many different SINE families even in a single species. For example, the salmonid fish have at least three SINE families, and they are confined in this group of fish (Takasaki et al. 1994).

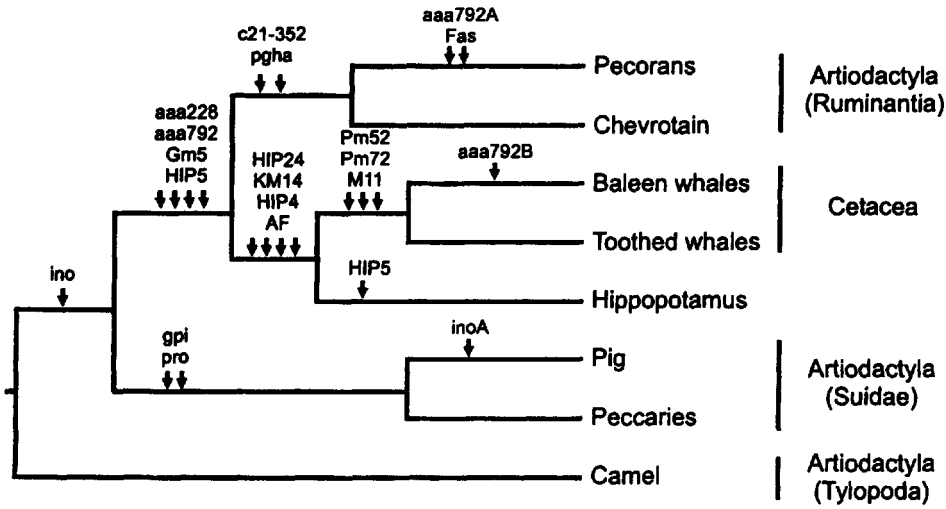

FIGURE 7.15. Evolutionary relationships among cetaceans and artiodactyls as inferred by the presence or absence of 21 different SINE elements. Arrows mark the insertions of SINEs (see Nikaido et al. 1999 for details).

Shimamura et al. (1997) and Nikaido et al. (1999) used mammalian SINE families to study the evolutionary relationships of whales, ruminants, pigs, and camels. The results obtained are presented in Figure 7.15. In this figure, the locations of branches, in which 21 different SINEs were inserted, are presented. As in the case of Figure 7.12, the pattern of insertions of SINEs indicates that whales are a sister group of ruminants and hippopotamuses and are a clade (monophyletic group) within the order Artiodactyla. These results strengthen the conclusion obtained by using the γ-fibrinogen gene (Figure 7.12). They are particularly significant because the SINEs used here are unique and largely irreversible genetic markers, and the tree based on them is unaffected by the error caused by short-branch (or long-branch) attraction (chapter 9). The Hennigian parsimony analysis of the SINE data has shown that this is the most parsimonious tree and that the consistency index (CI) is 1 and the homoplasy index (HI) is 0 (Nidaido et al. 1999). Therefore, the topology given in Figure 7.15 is likely to be correct. SINEs have also been used successfully in clarifying the evolutionary relationships of salmonid species (Murata et al. 1993; Takasaki et al. 1994) and some groups of primate species (Hamdi et al. 1999).

Some might wonder whether the phylogenetic trees based on SINEs are affected by the polymorphism of the presence and absence of SINE insertions in ancestral species (lineage sorting). Theoretically, this polymorphism may generate incongruent phylogenies among different SINE insertions, as in the case of polymorphic DNA sequences in Figure 5.3. This can be seen from the diagrams given in Figure 7.16. In this figure "+" stands for the allele or the genome having a SINE element at a given locus and "−" for the allele lacking it, and the arrow sign indicates the time at which the element was inserted. Here we consider only the cases

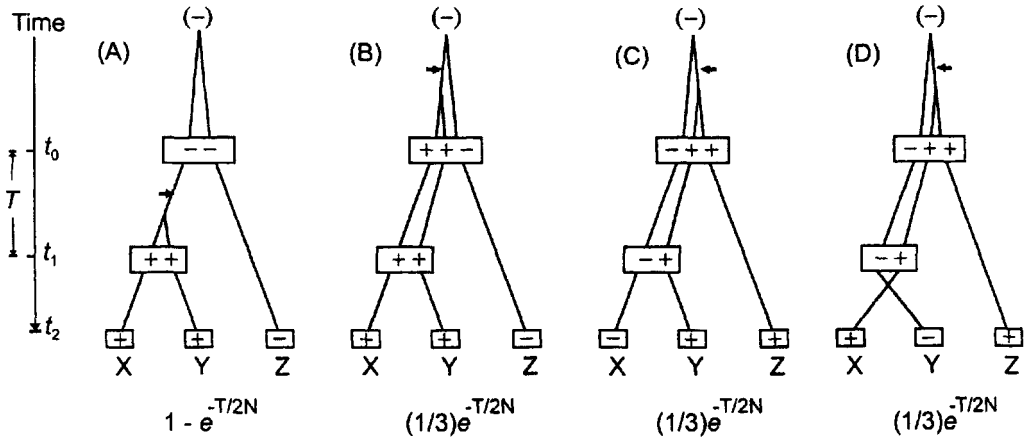

FIGURE 7.16. Four different evolutionary relationships of three species (X, Y, and Z) that may be inferred from SINE element polymorphism (+ and −) exists in ancestral species. The probability of occurrence of each relationship is given underneath the diagram of the relationship. The arrow sign indicates the time of occurrence of SINE insertion. The "−" and "+" are the ancestral and derived character states, respectively. The species tree is the same as that of Figure 5.3.

where two of the three species considered have the SINE element. In diagram (A), the insertion occurred between the times of occurrence of two speciation events (t_0 and t_1), and species X and Y form a clade, which is congruent with the species tree. If we assume that SINE insertion occurs with the same probability for all generations, it can be shown, by using the same mathematical method (coalescence theory) as used by Nei (1987, pp. 401–402), that the probability of occurrence of this event is $1 - \exp(-T/2N)$, where $T = t_1 - t_0$ and N is the effective population size. This probability is the same as that of having evolutionary relationship (A) for DNA sequences in Figure 5.3.

Diagrams (B)–(D) in Figure 7.16 show the cases where the SINE insertion occurred before the first speciation event (time t_0). In this case, the ancestral species can be polymorphic with respect to alleles "+" and "−", and this polymorphism may generate evolutionary relationships that are incongruent with the species tree. Relationship (B) is congruent with the species tree, but relationships (C) and (D) are incongruent. The probability of occurrence of each of these events is given underneath the diagram. Note that if SINE insertion occurs after the second speciation event (time t_1), the SINE element is not informative because it generates a singleton mutation.

Therefore, the effect of ancestral polymorphism of SINE elements is the same as that for DNA sequences discussed in chapter 5, and we have already shown that the probability of occurrence of incongruent relationships is generally very small if T is one million years or greater. (Tachida and Iizuka [1993] studied this problem under different assumptions, but their formulation appears to give an underestimation of the probability of occurrence of incongruent relationships.) Since T is likely to be greater than one million years when different families and genera in mammals

are studied, we can probably dismiss the effect of polymorphism in ancestral species in the case of the phylogenetic tree in Figure 7.15. In fact, all SINE insertions examined in Figure 7.15 are consistent with the topology presented, and there is no indication of incongruent phylogenies for different SINE insertions.

If SINE insertions are irreversible and the effect of ancestral polymorphism is negligible, the tree in Figure 7.15 can be regarded as established without further statistical tests. Application of a bootstrap test (Hillis 1999) to this tree is inappropriate, because every SINE defines a particular clade and exclusion of some SINE insertions (e.g., **ino**) in bootstrap pseudoresamples will make the tree look superficially unreliable. Note that theoretically a single SINE insertion for each interior branch is sufficient to support the topology obtained (Sober 1988).

Of course, this does not mean that no statistical test is needed for actual SINE data analysis. Although homoplasy appears virtually absent in the data of Figure 7.15, it is still too early to exclude homoplasy altogether, because the same SINE insertion may occur independently in different lineages though the probability appears to be very small. In some data sets, the effect of ancestral polymorphism may also generate incongruent phylogenies for different SINEs. In this case, we need some type of statistical test based on the special property of evolution of SINEs. At the present time, we are not sure how to test the reliability of SINE-based trees efficiently, but it is unlikely that the reliability of the tree in Figure 7.15 is questionably low. As a general strategy, it is important to have two or more SINE insertions for each interior branch.

LINEs are longer than SINEs and vary in size rather extensively from copy to copy. Therefore, it is harder to work with LINEs than with SINEs. However, LINEs can be used not only as shared derived characters but also for estimating the time of divergence between species, because LINEs diverge as mutations accumulate and are long enough to give reliable estimates of sequence divergence. Verneau et al. (1997, 1998) used the rodent LINE L1 family (45 rat L1 subfamilies) to clarify the evolutionary relationships and the times of speciation events among 26 rat species of the rodent subfamily Murinae. They showed that these species arose 5–6 million years ago and subsequently underwent different episodes of speciation, the first one occurring about 2.7 million years ago and the second one about 1.2 million years ago.

Other Shared Derived Characters

In addition to SINEs and LINEs, there are a variety of genetic markers that can be used for distinguishing between different groups of organisms. For example, the CMTIA-REP repeat in humans consists of two highly homologous 24 kb sequences (the proximal and the distal CMTIA-REP elements), and some forms of mutation of this repeat result in genetic diseases. Interestingly, this repeat exists only in humans and chimpanzees. However, the distal element of the repeat appears to be present in all primate species but is absent in other mammalian orders (Keller et al. 1999). Apparently, the distal element evolved in the common ancestor of primates, and the proximal element evolved as a result of duplication of the

distal element in the common ancestor of humans and chimpanzees. It is clear that if we find a large number of these shared derived characters they will be very useful for phylogenetic analysis.

Another important class of genetic markers is large-scale insertions or deletions of DNA sequences. For example, the human genome contains a duplication of a large DNA region encompassing about 25 immunoglobulin kappa variable region genes compared with other ape genomes (Zachau 1995). If this DNA duplication had occurred in the ancestor of African apes, it would have been a very useful phylogenetic marker. The class I MHC (HLA) C locus, which is known to be highly polymorphic in humans and chimpanzees, is present only in humans and African apes, so that this locus was apparently generated by gene duplication in the ancestor of these species (Chen et al. 1992). The DY/DI gene clusters in the cattle and pig class II MHC are also apparently confined only to a group of artiodactyl species (Trowsdale 1995). As the genomic structures of many different organisms are studied, we will find many such cladistic characters, and they will be important sources of phylogenetic analysis in the future.

In the study of evolutionary relationships of distantly related organisms, the presence and absence of introns in protein-coding genes will be useful. Although the debate over the intron-early and the intron-late hypotheses is still going on, it is now clear that at least in higher organisms introns are occasionally inserted or deleted, and therefore the presence or absence of introns can be used as cladistic characters. For example, the intron in the protamine P1 gene exists apparently in all mammals but not in other vertebrate species (Rooney et al. 2000). Venkatesh et al. (1999) used information on the presence or absence of introns in seven protein-coding genes for constructing a phylogenetic tree of fish species under the assumption that the probability of independent intron insertion in different lineages is negligibly small. They clarified the difficult-to-ascertain phylogenetic relationships of some ray-finned fishes.

Insertion of an intron is a rare event, and the same intron is almost never inserted twice in the same genomic location. However, the loss of an intron may occur independently in different evolutionary lineages. This property satisfies the conditions required for Farris's (1977) **Dollo parsimony** analysis. In this method, a new mutation (shared derived character) is assumed to be unique, but the loss of the mutation may occur independently in the descendant lineages. Therefore, intron insertion/deletion data can be analyzed by the Dollo parsimony, which is incorporated in PAUP*. However, note that the bootstrap test should not be applied to this type of data, because each intron insertion is a unique and unambiguous event.

This page intentionally left blank