# 5

# Phylogenetic Trees

Phylogenetic analysis of DNA or protein sequences has become an important tool for studying the evolutionary history of organisms from bacteria to humans. Since the rate of sequence evolution varies extensively with gene or DNA segment (Wilson et al. 1977; Dayhoff et al. 1978), one can study the evolutionary relationships of virtually all levels of classification of organisms (e.g., kingdoms, phyla, families, genera, species, and intraspecific populations) by using different genes or DNA segments. Phylogenetic analysis is also important for clarifying the evolutionary pattern of multigene families (e.g., Atchley et al. 1994; Goodwin et al. 1996; Nei et al. 1997a) as well as for understanding the process of adaptive evolution at the molecular level (e.g., Jermann et al. 1995; Chandrasekharan et al. 1996; Zhang et al. 1998).

Reconstruction of phylogenetic trees by using statistical methods was initiated independently in numerical taxonomy for morphological characters (Sokal and Sneath 1963) and in population genetics for gene frequency data (Cavalli-Sforza and Edwards 1964). Some of the statistical methods developed for these purposes are still used for phylogenetic analysis of molecular data, but in recent years many new methods have been developed. In this book, we will discuss only methods that are useful for analyzing molecular data. For morphological data, the readers may consult Wiley et al. (1991), Maddison and Maddison (1992), and Swofford and Begle (1993). Before discussing tree-building methods, we first consider the types of phylogenetic trees in which molecular evolutionists are interested.

## 5.1. Types of Phylogenetic Trees

### *Rooted and Unrooted Trees*

Phylogenetic relationships of genes or organisms are usually presented in a tree-like form either with a root (Figure 5.1A) or without any root (Figure 5.1B). The former tree is called a **rooted tree** and the latter an **unrooted tree.** The branching pattern of a tree, whether rooted or unrooted, is called a **topology.** There are many possible rooted and unrooted tree

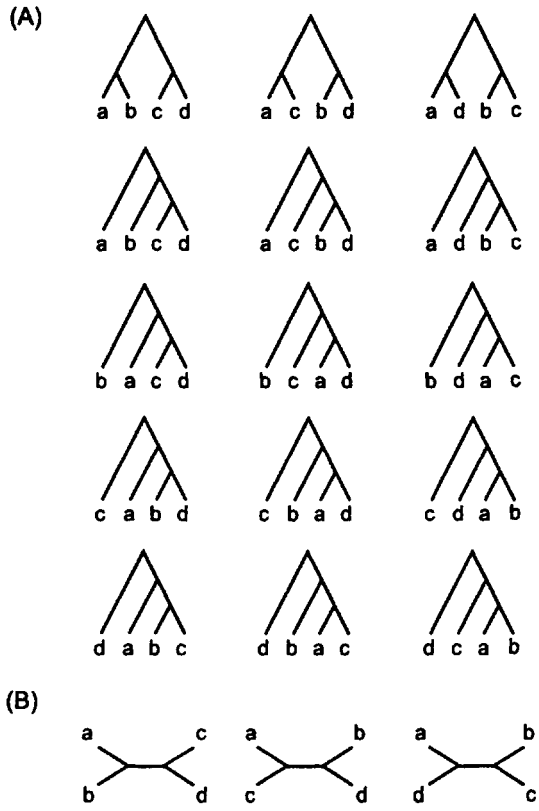


FIGURE 5.1. (A) Fifteen possible rooted trees and (B) three possible unrooted trees for four taxa.

topologies for a sizable number of **taxa** (any kind of taxonomic unit; families, species, populations, DNA sequences, etc.). If the number of taxa ($m$) is four, there are 15 possible rooted tree topologies and three possible unrooted tree topologies, as shown in Figure 5.1. The number of possible topologies rapidly increases with increasing $m$. In general, the number of possible topologies for a bifurcating rooted tree of $m$ taxa is given by

$$1 \cdot 3 \cdot 5 \cdots (2m - 3) = [(2m - 3)!] / [2^{m-2}(m - 2)!] \qquad (5.1)$$

for $m \geq 2$ (Cavalli-Sforza and Edwards 1967). This indicates that the numbers of topologies for $m = 2$, 3, 4, 5, and 6 are 1, 3, 15, 105, and 945, respectively. When $m = 10$, it becomes $1 \cdot 3 \cdot 5 \cdot 7 \cdot 9 \cdot 11 \cdot 13 \cdot 15 \cdot 17 = 34{,}459{,}425$. Only one of these topologies is the true tree. The number of possible topologies for a bifurcating unrooted tree of $m$ taxa is given by replacing $m$ by $m - 1$ in Equation (5.1). This becomes 2,027,025 for $m = 10$. In many cases, a majority of the possible topologies can be excluded from consideration because of obviously unlikely evolutionary relationships or because of other biological information. Nevertheless, it is a very difficult task to find the true tree topology when $m$ is large.

In an unrooted bifurcating tree of $m$ taxa there are $2m - 3$ **branches.** Since there are $m$ **exterior branches** connecting to $m$ extant taxa, the number of **interior branches** is $m - 3$. The number of **interior nodes** is equal to $m - 2$. In a rooted tree, the numbers of interior branches and interior nodes are $m - 2$ and $m - 1$, respectively, and the total number of branches is $2m - 2$.

Theoretically, a DNA sequence splits into two descendant sequences at the time of speciation or gene duplication. Therefore, phylogenetic trees are usually **bifurcating.** However, when a relatively short sequence is considered, some interior branches may show no nucleotide substitution, so that a multifurcating node may appear. This type of tree is often called a **multifurcating tree.** Most tree-building methods are for constructing a bifurcating tree, but the tree obtained can be reduced to a multifurcating tree by eliminating any branch that has zero length. It is also possible that even if the true tree is bifurcating, the reconstructed tree becomes multifurcating because of statistical errors. In reality, it is difficult to distinguish between the two cases.

Phylogenetic relationships of different DNA sequences are sometimes presented in a form of network with some loops. Loops are required when recombination occurs within a sequence or when the resolving power of mutational differences is low (e.g., Bandelt et al. 1995; Fitch 1997; Saitou and Yamamoto 1997; Page and Holmes 1998). In the former case, phylogenetic relationships in network form are a natural representation. In the latter case, the use of an experimental technique with a higher resolving power often resolves the network and reduces it to a bifurcating tree. For example, Avise et al. (1987) obtained a network tree when the variation of mitochondrial DNA in the deer mouse *Peromyscus polionotus* was analyzed by using a single restriction enzyme, but they could produce a bifurcating tree when eight restriction enzymes were used. In practice, network trees are produced only occasionally, so they will not be considered in this book.

### *Gene Trees and Species Trees*

Evolutionists are often interested in a phylogenetic tree that represents the evolutionary history of a group of species or populations. This type of tree is called a **species** or **population tree.** In a species tree, the time of divergence between two species refers to the time when the two species were reproductively isolated. However, when a phylogenetic tree is constructed from one gene from each species, the tree obtained does not necessarily agree with the species tree. In the presence of polymorphic alleles at a locus, the times of divergence of genes sampled from different species are expected to be longer than the time of species divergence (Figure 5.2). The branching pattern of a tree constructed from genes may also be different from that of the species tree. To distinguish this tree from the species tree, we call it a **gene tree** (Nei 1986, 1987). Figure 5.3 shows three different possible relationships between species trees and gene trees for the case of three species. In relationships A and B, the topologies of the species and the gene trees are the same, but in relationship C they are different. If we use the gene genealogy theory in population ge-

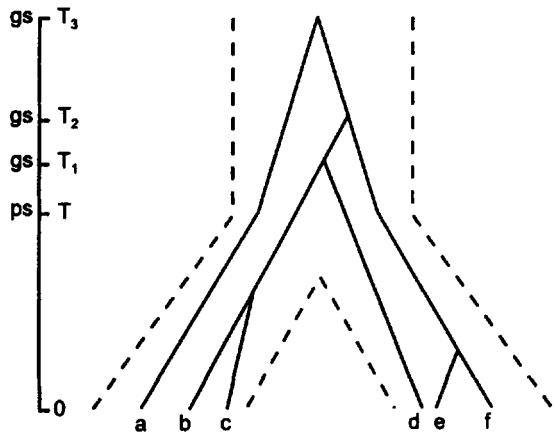


FIGURE 5.2. Diagram showing that the time of gene splitting (gs) is usually earlier than the time of population splitting (ps) when polymorphism exists. From Takahata and Nei (1985).

netics (Tajima 1983), it is possible to compute the probability of occurrence of events A, B, and C (Nei 1987; Pamilo and Nei 1988). The probability of occurrence of relationship C is quite high when the time interval between the first and second species splitting measured in terms of the number of generations ($T$) is short and the effective population size ($N$) is large.

Suppose that the long-term effective population size ($N$) is 10,000 as in the case of some mammals (Nei and Graur 1984) and the interval between two speciation events is one million years. If the generation time of the organism under consideration is 5 years, $T$ becomes 200,000 generations. In this case, the probability [$P(C)$] of occurrence of relationship C is $(2/3)[\exp - (T/2N)] = 0.00003$, which is virtually 0. If $N$ is as large as 100,000 but the generation time is 1 year as in the case of some invertebrate organisms, $P(C)$ becomes 0.004, which is again negligibly small. Therefore, if we consider a group of organisms where speciation event has occurred every one or two million years, the probability that the gene tree is different from the species tree is very small.

By contrast, if $N = 10{,}000$, $T = 100{,}000$, and the generation time is 5 years, we obtain $P(C) = 0.245$, which is substantial. Therefore, for a group of closely related species or intraspecific populations, the chance that the gene tree does not agree with the species or population tree is quite high. This was indeed the case with the DNA sequences from several nuclear genetic loci in a group of recently generated cichlid fish species of Lake Victoria in Africa (Nagl et al. 1998) or with the mitochondrial DNA sequences from several different human populations (Vigilant et al. 1991). To obtain a reliable tree of intraspecific populations or closely related species, interpopulational genetic distances based on a large number of genes from independently evolving (unlinked) loci need to be used (Saitou and Nei 1986; Pamilo and Nei 1988).

It should also be noted that even if the actual pattern of gene splitting

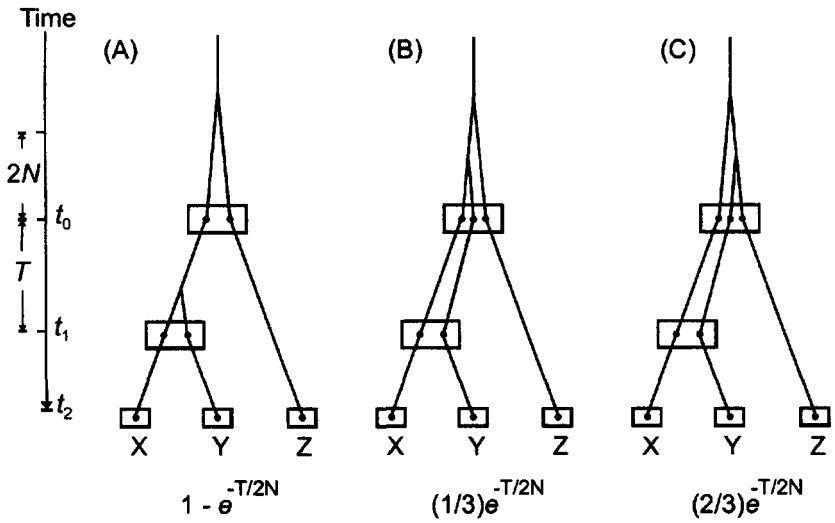


FIGURE 5.3. Three possible relationships between the species and gene trees for the case of three species in the presence of polymorphism. The times of the first and second species splitting are $t_0$ and $t_1$, respectively. The probability of occurrence of each tree is given underneath the tree. $T = t_1 - t_0$, and $N$ is the effective population size. From Nei (1987).

agrees with that of species splitting, the branching pattern of a reconstructed gene tree may not agree with that of the species tree if the number of nucleotides or amino acids examined is small. This is because nucleotide or amino acid substitution occurs stochastically, so that the number of substitutions in lineage $Z$ in Figure 5.3A or B may be smaller than that in lineage X or Y. To avoid this type of error, we must examine a large number of nucleotides or amino acids (Saitou and Nei 1986).

When the gene studied belongs to a multigene family, another problem may occur. Suppose that two related species, species 1 and 2, have two duplicate genes $a_1$ and $b_1$ and $a_2$ and $b_2$, respectively, and that the duplicate genes were generated by gene duplication that occurred before the divergence of the two species (Figure 5.4). In this case, genes $a_1$ and $a_2$ or $b_1$ and $b_2$ from the different species are called **orthologous genes,** whereas pairs of genes $a_1$ and $b_1$, $a_2$ and $b_2$, $a_1$ and $b_2$, and $a_2$ and $b_2$ are called **paralogous genes** (Fitch 1970). To construct a phylogenetic tree of different species, we should use orthologous genes rather than paralogous genes, because only orthologous genes represent speciation events. In practice, however, the distinction between orthologous and paralogous genes is not always easy, particularly when there are many copies of duplicate genes in the genome. We should, therefore, exercise great caution in the inference of species trees from gene trees.

Of course, gene trees are not always produced just to infer species trees. In the study of evolution of multigene families, it is important to know the evolutionary history of member genes and the process of gene duplication. In this case, we must study gene trees.

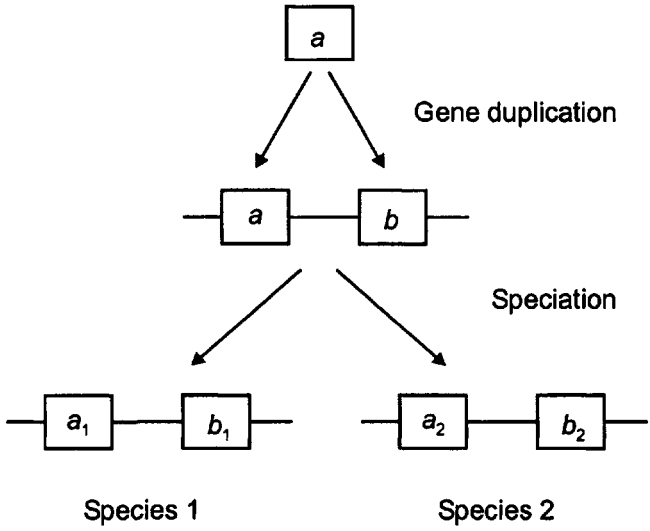


FIGURE 5.4. Duplicate genes from two different species. Genes $a_1$ and $a_2$ and $b_1$ and $b_2$ are orthologous, whereas pairs of genes $a_1$ and $b_1$, $a_1$ and $b_2$, $a_2$ and $b_1$, etc., are called paralogous.

### *Expected and Realized Trees*

In the theory of phylogenetic inference, it is often assumed that the DNA or protein sequences to be studied are very long (theoretically infinitely long) and that a large number of amino acids or nucleotides that represent a random sample from the long sequences are sampled. While this assumption simplifies the statistical analysis of DNA or protein sequences, investigators are often interested in reconstructing the evolutionary history of a short sequence. For example, if one wants to know the long-term evolution of homeobox genes, he or she must work with a sequence of about 60 codons, because this is the size of the highly conserved homeobox domain (Kappen et al. 1993; Duboule 1994).

If we consider a short gene or a short segment of DNA, the number of nucleotide or amino acid substitutions is subject to large stochastic errors. Therefore, even if the expected number of substitutions increases linearly with time, a phylogenetic tree representing the actual number of substitutions could be very different from what one might expect intuitively. In this case, even the topology of the tree could be different from that of the tree from long DNA sequences. A tree that can be constructed by using infinitely long sequences or the expected number of substitutions for each branch is called an **expected tree,** whereas a tree based on the actual number of substitutions is called a **realized tree** (Nei 1987; Kumar 1996b). Note that both expected and realized trees are often different from the tree reconstructed (**reconstructed** or **inferred tree**) from observed sequence data.

Figure 5.5 shows one example of the differences among the expected, realized, and reconstructed trees when the molecular clock is assumed to work. Tree A in this figure represents an expected tree with each

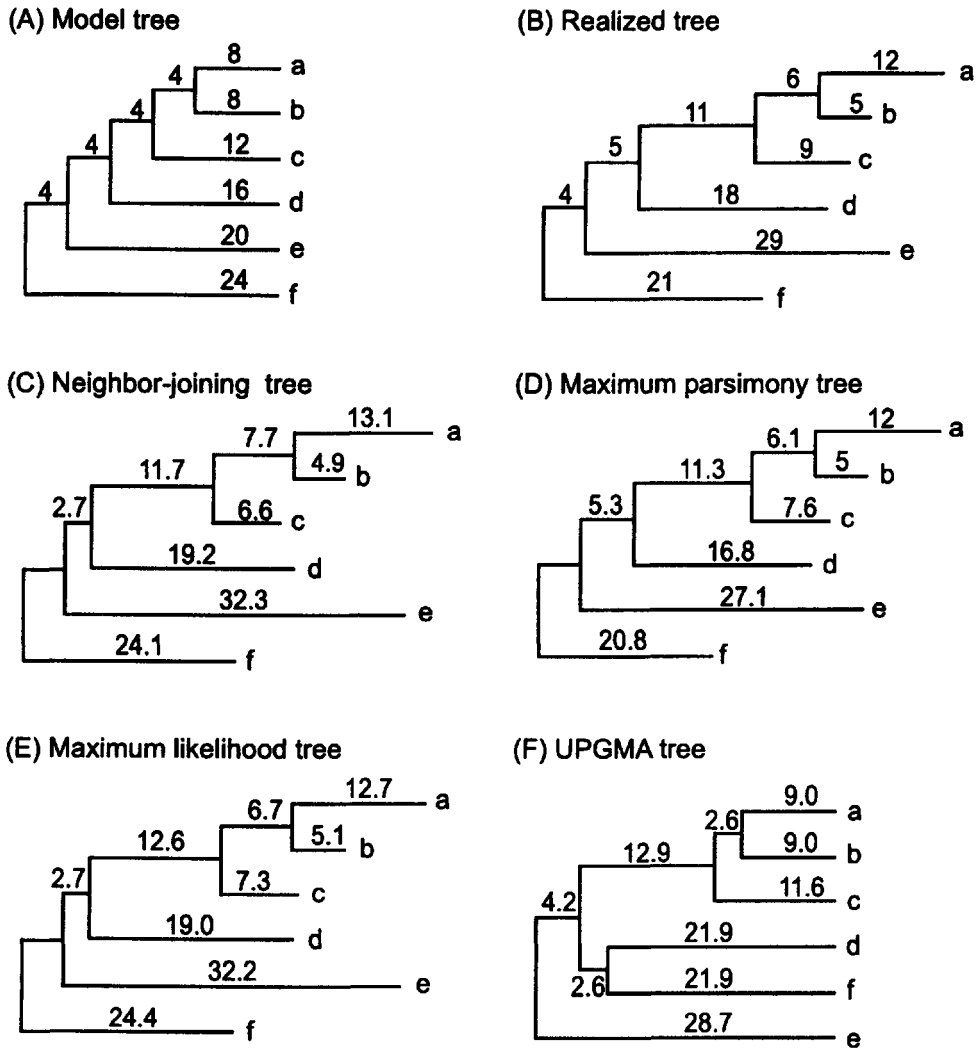


FIGURE 5.5. (A) Model tree, (B) realized tree, and (C–F) reconstructed trees. The neighbor-joining, maximum likelihood, and UPGMA trees were constructed with the Jukes-Cantor model. The branch lengths of the maximum parsimony tree were estimated by the average pathway method.

branch length equal to the expected number of nucleotide substitutions. In this case, the expected number of substitutions from the root to each terminal node is 0.12 per nucleotide site. Therefore, if a sequence of 200 nucleotides is used, the expected number of substitutions per sequence is 24. Tree B is a realized tree obtained in a replication of computer simulation under the assumption that the number of nucleotides used is 200 and nucleotide substitution occurs following the Jukes-Cantor model. The number given for each branch of tree B represents the number of substitutions that actually occurred for that branch. This number is considerably different from the expected value in tree A because of stochastic errors of nucleotide substitution.

Which tree does a tree-building method attempt to reconstruct, the ex-

pected tree or the realized tree? The answer to this question depends on the method of reconstruction, but most methods are intended to reconstruct realized trees. Figure 5.5C shows a tree reconstructed by the neighbor-joining method, which will be explained in chapter 6. The topology of the tree is identical with that of the expected (model) and the realized trees. However, the branch lengths of this tree are very different from those of the model tree and are close to those of the realized tree. This clearly indicates that the neighbor-joining method is intended to infer a realized tree rather than the model tree. Figures 5.5D and 5.5E show the trees constructed by the maximum parsimony and the maximum likelihood methods, respectively. The reconstructed trees are closer to the realized tree than to the model tree, indicating that they are also for inferring a realized tree.

By contrast, the topology of the tree (Figure 5.5F) obtained by the unweighted pair-group method with arithmetic mean (UPGMA) (see chapter 6) is different from that of both the model tree and the realized tree. Because of the incorrect topology, comparison of the branch lengths of the UPGMA tree with those of the model or the realized tree is not very meaningful, but the branch lengths for the correct part (sequences a, b, and c) of the topology are closer to those of the model tree rather than to those of the realized tree. In the present example, the expected number of nucleotide substitutions (4) for each interior branch was small, so that UPGMA could not produce the correct topology because of stochastic errors. However, if the number is two times higher, UPGMA would produce the correct topology with a high probability, and in this case, the branch lengths would have been closer to those of the model tree (Tateno et al. 1982). In other words, UPGMA is intended to infer the model tree or species tree, but unfortunately the topology of the UPGMA tree is disturbed by stochastic errors and other factors more easily than that of the trees obtained by other tree-building methods.

One might argue that what we really want to know is the expected or true tree rather than the realized tree. This is surely the case when a phylogenetic tree for a group of species is to be constructed. In practice, however, it is easier to reconstruct a realized tree than an expected tree, because the sequence data available refer to the realized tree. Note also that the topology of a realized tree is the same as that of the expected tree, unless a realized tree becomes a multifurcating tree because of stochastic errors. A realized tree may become a multifurcating tree when no nucleotide substitution occurs for one or more interior branches of the model tree by chance. As the number of nucleotides examined ($n$) increases, the realized tree is expected to approach the expected tree.

When one is interested in constructing species or population trees, the expected tree must have branch lengths proportional to evolutionary times, and two evolutionary lineages descendent from an interior node must have the same branch lengths. In practice, it is not easy to reconstruct a species tree defined in this way. Since the evolutionary change of genes is subject to stochastic errors as well as some kinds of selection, even a tree based on many genes could be different from the true species tree. At the present time, many investigators seem to be satisfied if they can reconstruct the correct or nearly correct topology even though the

branch length estimates are not proportional to evolutionary time. In estimating species or population trees, it is important to use as many genes as possible (e.g., Kidd et al. 1974; Nei and Roychoudhury 1974; Doolittle et al. 1996).

## 5.2. Topological Differences

### *Topological Distance*

Although the true topology is generally unknown in actual data analysis, it is often useful to measure the extent of topological differences between two trees. For example, when one wants to know alternative trees that are closely related to a reconstructed tree, it is necessary to measure the topological distances of the alternative trees from the reconstructed tree. In the measurement of topological distance, it is customary to give no consideration to branch length differences.

The **topological distance** between two different trees is commonly measured by using Penny and Hendy's (1985) method of sequence partitioning. This distance gives the same numerical values as those obtained by Robinson and Foulds' (1981) method but is simpler to compute. For unrooted bifurcating trees, this distance is twice the number of interior branches at which sequence partition is different between the two trees compared. As an example, consider unrooted trees A and B in Figure 5.6. Both trees are for eight sequences and have five interior branches. It is possible to cut the tree at any interior branch and divide the sequences into two groups. Cutting at some interior branch results in the same partition of sequences in both trees A and B but not at other branches. For example, a cut at branch *a* produces two sequence groups (1,2) and (3, 4, 5, 6, 7, 8) in both trees. A cut at branch *c*, however, produces different partitions in trees A and B. That is, the two groups produced by this cut are (1, 2, 3, 4) and (5, 6, 7, 8) in tree A but (1, 2, 3, 5) and (4, 6, 7, 8) in tree B. Similarly, a cut at branch *d* produces different partitions in the two trees. In the present example, only these two cuts result in different partitions. Therefore, the topological distance between the two trees ($d_T$) is $2 \times 2 = 4$.

In general, if two trees for eight sequences have the same topology, $d_T = 0$, and if all interior branches produce different partitions, $d_T = 10$. However, if the two trees compared have multifurcating nodes, the above rule does not work. In this case, we can use Rzhetsky and Nei's (1992a)

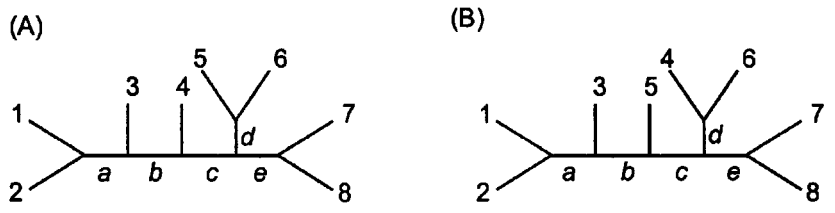


FIGURE 5.6. Two unrooted trees for eight sequences.

general formula for computing $d_T$ for a pair of arbitrary trees for $m$ sequences.

$$d_T = 2[\text{Min}(q_1,q_2) - p] + |q_1 - q_2| \quad (5.2)$$

where $q_1$ and $q_2$ are the total numbers of possible partitions (interior branches) for trees 1 and 2, respectively, and $p$ is the number of partitions that are identical for the two trees. $q_1$ and $q_2$ may not be the same when multifurcating nodes are involved, and $\text{Min}(q_1,q_2)$ means the smaller value of $q_1$ and $q_2$. For bifurcating trees, however, $q_1$ and $q_2$ are always the same, and $d_T$ takes only even numbers. In general, an unrooted bifurcating tree for $m$ sequences has $m - 3$ interior branches, so that the maximum possible value of $d_T$ is $2(m - 3)$.

In some methods of phylogenetic inference (see chapter 6) all trees that are different from an initial tree with topological distances $d_T = 2$ and 4 are examined in the search for the most likely tree. (Actually, this process is repeated.) Rzhetsky and Nei (1992a) have shown that the number of trees (topologies) that are different from a given topology by $d_T = 2$ is given by $f(d_T = 2) = 2(m - 3)$, whereas the number for $d_T = 4$ is $f(d_T = 4) = 2(m^2 - 4m + 3m' - 6)$, where $m'$ is the number of tree nodes that are connected to one interior branch and two exterior branches ($m \geq 4$). For example, in tree A of Figure 5.6, $m = 8$ and $m' = 3$. Therefore, $f(d_T = 4) = 70$. This number is fairly large but represents a small portion of the total number of possible topologies (10,395). Therefore, it is much easier to examine the topologies with $d_T = 2$ and 4 rather than all topologies once a plausible tree is found.

### *Symbolic Expression of Topologies*

Although a bifurcating or multifurcating tree can generally be drawn in a two-dimensional space, it is often convenient to use symbolic expressions to represent different tree topologies. Actually, any bifurcating or multifurcating tree can be expressed by a simple symbolic expression. For example, the topology of trees A–E in Figure 5.5 can be expressed as (f(e(d(c(b, a))))) and that of tree F as (e((d, f)(c(b, a)))). A multifurcating tree can also be expressed in the same way. Suppose that taxa a, b, and c are derived from one trifurcating node rather than two bifurcating nodes in trees A–E. The topology of the tree can then be expressed as (f(e(d(c, b, a)))).

In the case of unrooted trees, there are several different ways of describing a topology. One simple method is to subdivide all the taxa into three subgroups of taxa that join at an interior node and then decompose each subgroup consisting of three or more taxa into further subgroups of taxa. For example, in the case of tree A of Figure 5.6, we can first consider three subgroups of taxa (1, 2), 3, and (4, 5, 6, 7, 8). This forms only one topology, but we can further decompose subgroup (4, 5, 6, 7, 8) and write the topology of the entire tree as ((1, 2) 3 (4 ((5, 6), (7, 8)))). When there are multifurcating nodes, we have to use a slightly different expression. Suppose that taxa 5, 6, 7, and 8 are connected through one mul-

tifurcating node in tree A. Then the topology can be written as ((1, 2) 3 (4 (5, 6, 7, 8))) or (((1, 2) 3) 4 (5, 6, 7, 8)).

If we use the above symbolic expressions, all tree topologies can be distinguished from one another. This method of distinction is important in the examination of many different topologies to find the most likely tree, which will be discussed later.

## 5.3. Tree-Building Methods

There are many statistical methods that can be used for reconstructing phylogenetic trees from molecular data. Commonly used methods are classified into three major groups: (1) distance methods, (2) parsimony methods, and (3) likelihood methods. Details of these methods will be discussed in the next three chapters. Recently, Hendy and his colleagues (Hendy and Charleston 1993; Hendy and Penny 1993; Hendy et al. 1994) proposed the use of Hadamard conjugation for phylogenetic reconstruction (closest tree method). Dopazo and Carazo (1996) also proposed a neural network method of phylogenetic reconstruction. However, the practical utility of these methods has not yet been examined. Therefore, these methods will not be discussed.

It is now customary to consider the reconstruction of a phylogenetic tree as a statistical inference of a true phylogenetic tree, which is unknown. There are two processes involved in this inference: "estimation" of the topology and estimation of branch lengths for a given topology. When the topology is known, estimation of branch lengths is relatively simple, and there are several statistical methods one may use (e.g., least squares and maximum likelihood methods). The problem is the estimation or reconstruction of a topology. When there are a sizable number of DNA or protein sequences (say, 20), the number of possible topologies is enormously large as mentioned above, so that it is generally very difficult to find the true topology among them.

In phylogenetic inference, a certain optimization principle such as the maximum likelihood or the minimum evolution principle is often used for choosing the most likely topology. The theoretical basis of this procedure is not well understood, as will be discussed later, but computer simulations have shown that the optimization principles currently used generally work quite well if the number of nucleotides or amino acids used ($n$) is large. When this number is small and the number of sequences used is large, the optimization principle tends to give incorrect topologies, as will be discussed in chapter 9.

Some authors (e.g., Felsenstein 1978, 1988) have considered a tree topology as a **parameter** in statistical estimation and regarded a tree-building method as a **statistic** (or **estimator**) for estimating the parameter, as in the case of estimation of the mean of a statistical distribution. Therefore, Felsenstein (1978) used the concept of **inconsistency** in statistics to argue the inferiority of parsimony methods to likelihood methods under certain conditions. In statistical theory, if a statistic approaches the true parameter as the sample size (number of nucleotides

or amino acids in the present case) increases to infinity, the statistic is called a **consistent estimator.** In phylogenetic inference, a tree-building method does not represent any numerical quantity, so it is not a statistic as used in standard statistical theory. Nevertheless, inconsistency is a convenient way of describing a property of a tree-building method, so it is often used in phylogenetics.

If we are allowed to use a similar statistical concept for estimating a topology, we can consider whether a tree-building method gives the correct topology when the same evolutionary process is repeated an infinite number of times with a finite value of $n$. If a tree-building method gives the correct topology in this case, one may say that the tree-building method is an "unbiased estimator." If we use this definition, it can be shown that all tree-building methods based on the optimization principle are not "unbiased estimators" of the true topology and therefore tend to give incorrect topologies (Nei et al. 1998). In the case of maximum likelihood methods, some authors considered topologies as random variables and attempted to estimate the topology under this statistical framework (Cavalli-Sforza and Edwards 1967; Rannala and Yang 1996). In this case, one has to use a mathematical model for the pattern of species splitting. Rannala and Yang (1996) used the birth-and-death process in probability theory for this purpose, but since the real pattern of species splitting is very complicated, it is unclear how well this approach performs in real data analysis.

At present, the methodology of phylogenetic reconstruction is quite controversial. There seem to be at least three reasons for this controversy, putting aside personal preference. First, some workers were originally trained as systematicists using morphological characters, and they tend to be suspicious about any methods that are based on mathematical models of evolutionary changes, because the evolutionary change of morphological characters is so complex that it does not obey any simple rule. They therefore prefer parsimony methods, which require a minimum number of assumptions. Another group of workers has been trained as geneticists or molecular biologists and tends to prefer using analytical approaches but does not trust highly sophisticated mathematical models. A third group of workers is primarily trained as mathematicians or statisticians and tries to understand the construction of phylogenetic trees as a mathematical problem rather than a practical problem, using abstract mathematical concepts. Since the approaches used by the three groups of workers are quite different, controversies naturally occur.

Second, some scientists are primarily interested in short-term evolution within species or between closely related species, whereas others are interested in long-term evolution dealing with different orders, phyla, or kingdoms. The methodologies used by these two groups of scientists are quite different, and one group tends to feel wary of the approach used by the other group.

Third, in phylogenetic analysis, the true tree is almost always unknown, and it is difficult to test the accuracy of the trees obtained by different tree-building methods. Currently, there are several statistical criteria for evaluating the accuracy, but all of them depend on a number of simplifying assumptions. Therefore, none of them is perfect. Fur-

thermore, the theoretical basis of the statistical methods currently used for phylogenetic reconstruction is not well established, as mentioned above. The mathematical models used for describing sequence evolution are crude approximations to reality, and a sophisticated model does not necessarily give better results. Therefore, there is plenty of room for controversy.

Molecular phylogenetics is still a young scientific discipline, and it is important to realize that every statistical method has some strengths and some weaknesses, and none of the methods is almighty. One cannot reject a method simply because it did not work in a particular study or a particular computer simulation. The overall assessment of superiority of one method over the other should come from broad theoretical and experimental studies. As mentioned above, the evolutionary change of DNA or proteins is so complicated that the mathematical model used is necessarily approximate. Unlike the case in physics, the predictive power of a model in biology is quite low. It seems to us that if the prediction (e.g., a phylogenetic tree reconstructed) of a model is correct in 80% of the cases, it is a good model at least at the present time. In the case of molecular phylogenetics, one can study the phylogeny of a group of organisms using a large number of genes, and this comprehensive study will eventually clarify the evolutionary relationships of organisms.

In the following three chapters, we will discuss various tree-building methods without going into mathematical details and cover only methods that have proved to be useful for practical data analysis. However, the theoretical basis of each method will be discussed with minimum mathematics and verbal arguments as much as possible. In these chapters, we assume that the number of nucleotides or amino acids used ($n$) is sufficiently large so that phylogenetic inference based on optimization criteria works well. The performance of optimization criteria when $n$ is small will be discussed in chapter 9.

*This page intentionally left blank*