CHAPTER

9

Phylogenetic Analysis*

OUTLINE

9.1 PHYLOGENETICS AND THE WIDESPREAD USE OF THE PHYLOGENETIC TREE

Phylogeny refers to the evolutionary history of species. **Phylogenetics** is the study of phylogenies—that is, the study of the evolutionary relationships of species. **Phylogenetic analysis** is the means of estimating the evolutionary relationships. In molecular phylogenetic analysis, the sequence of a common gene or protein can be used to assess the evolutionary relationship of species. The evolutionary relationship obtained from phylogenetic analysis is usually depicted as branching, treelike diagram—the **phylogenetic tree**. Historically, the use of phylogenetic trees was restricted more or less to the study of evolutionary biology, and to disciplines like systematics and taxonomy. However, with the advent of sequencing and the widespread use of cladistics, the use of phylogenetic trees has pervaded many branches of biology and beyond. Construction of phylogenetic/evolutionary trees is now widespread in many areas of study where evolutionary divergence can be studied and demonstrated; be it pathogens, biological macromolecules, or languages.

Phylogenetics also provides the basis for **comparative genomics**, which is a more recent term that came into existence in the age of genomics. Comparative genomics is the study of the interrelationships of genomes of different species. Comparative genomics helps identify regions of similarity and differences among genomes. The comparison can be made at different levels, such as comparison of whole-genome sequences, comparison of genome sequences involving blocks of conserved synteny, comparison of the number of protein-coding genes, comparison of regulatory sequences, or other focused comparisons. An important application of comparative genomics is gene finding. From the standpoint of evolutionary biology, comparative genomics helps understand the evolutionary relationships among genomes.

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

A resource for comparative genomic analysis is **VISTA**, which can be accessed at http://genome.lbl.gov/vista/index.shtml.

9.2 PHYLOGENETIC TREES

A phylogenetic tree or evolutionary tree is a diagrammatic representation of the evolutionary relationships among various taxa (Figure 9.1 A–D). It is a branching diagram composed of **nodes** and **branches**. The branching pattern of a tree is called the **topology** of the tree. The nodes represent taxonomic units, such as species (or higher taxa), populations, genes, or proteins. A branch is called an edge, and represents the time estimate of the evolutionary relationships among the taxonomic units. One branch can connect only two nodes. In a phylogenetic tree, the terminal nodes represent the **operational taxonomic units** (**OTU**s) or **leaves**. The OTUs are the actual objects—such as the species, populations, or gene or protein sequences—being compared, whereas the internal nodes represent **hypothetical taxonomic units** (**HTU**s). An HTU is an inferred unit and it represents the **last common ancestor** (**LCA**) to the nodes arising from this point. Descendants (taxa) that split from the same node form **sister groups**, and a taxon that falls outside the **clade**[a] is called an **outgroup**. For example, in Figure 9.1 B, T_2 and T_3 are sister groups, and T_1 is an outgroup to T_2 and T_3.

Phylogenetic trees can be **scaled** or **unscaled**. In a scaled tree, the branch length is proportional to the amount of evolutionary divergence (e.g. the number of nucleotide substitutions) that has occurred along that branch. In an unscaled tree, the branch length is not proportional to the amount of evolutionary divergence, but usually the actual number is indicated somewhere on the branch.

Phylogenetic trees can be **rooted** (Figure 9.1 A and B) or **unrooted** (Figure 9.1 C). A rooted tree has a node (the root) from which the rest of the tree diverges.

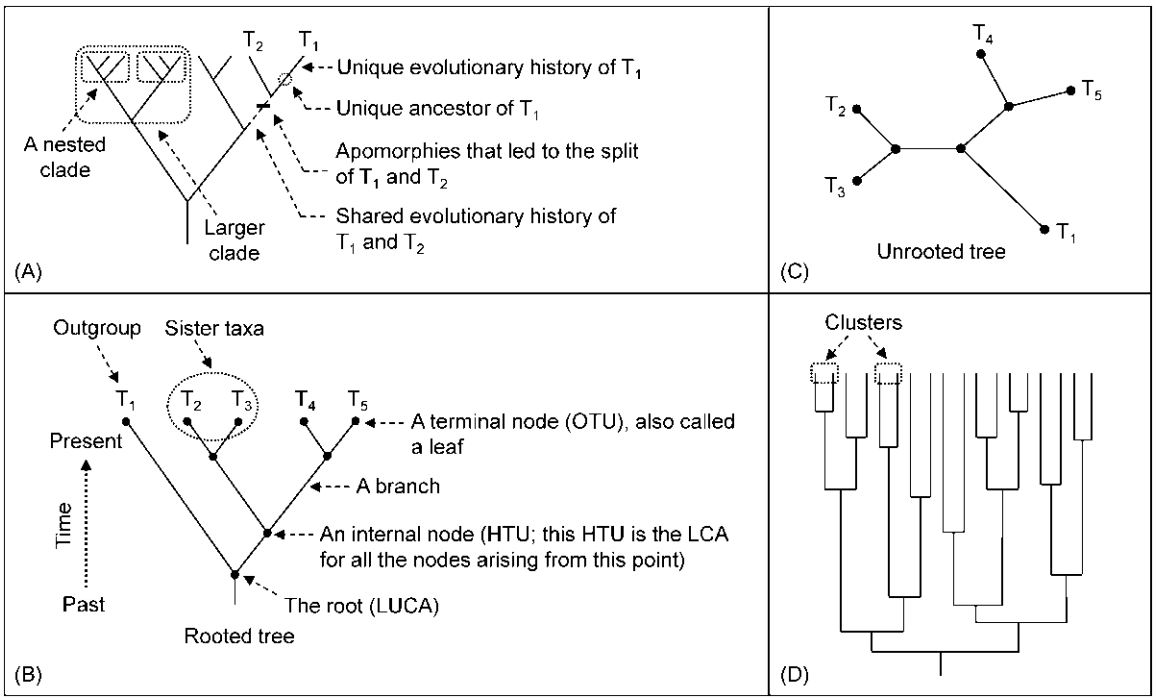

FIGURE 9.1 **Different forms of presentation of the phylogenetic tree.** The phylogenetic tree in D is a dendrogram derived from hierarchical clustering (see text). A, B, and D show rooted trees, while C shows an unrooted tree. Taxa that share specific derived characters are grouped into clades. (A) Smaller clades located within a larger clade are called nested clades. (B) The terminal nodes represent the operational taxonomic units, also called "leaves"; each terminal node could be a taxon (species or higher taxa), or a gene or protein sequence. The internal nodes represent hypothetical taxonomic units. An HTU represents the last common ancestor to the nodes arising from this point. Two descendants that split from the same node are called sister groups and a taxon that falls outside the clade is called an outgroup. Rooted trees have a node from which the rest of the tree diverges, frequently called the last universal common ancestor (LUCA).

[a]Taxa that share specific derived characters are grouped more closely together than those who do not. The groups are called **clades**; each clade consists of an ancestor and all of its descendants.

This root is frequently referred to as the **last universal common ancestor** (**LUCA**), from which the other taxonomic groups have descended and diverged over time. In molecular phylogenetics, the LUCA and LCA are represented by DNA or protein sequences. Obtaining a rooted tree is ideal, but most phylogenetic-tree-reconstruction algorithms produce unrooted trees.

9.2.1 Phylogenetic Trees, Phylograms, Cladograms, and Dendrograms

In the context of molecular phylogenetics, the expressions phylogenetic tree, phylogram, cladogram, and dendrogram are used interchangeably to mean the same thing—that is, a branching tree structure that represents the evolutionary relationships among the taxa (OTUs), which are gene/protein sequences. In the traditional evolutionary sense, the OTUs in the phylogenetic tree are represented by species. A **phylogram** is a scaled phylogenetic tree in which the branch lengths are proportional to the amount of evolutionary divergence. For example, a branch length may be determined by the number of nucleotide substitutions that have occurred between the connected branch points. A **cladogram** is a branching hierarchical tree that shows the relationships between clades; cladograms are unscaled. The word **dendrogram** means a hierarchical cluster arrangement where similar objects (based on some defined criteria) are grouped into clusters; hence, a dendrogram shows the relationships among various clusters (Figure 9.1 D). Dendrograms are also used outside the scope of phylogenetics and even outside of biology. Dendrograms are fequently used in computational molecular biology to illustrate the branching based on clustering of genes or proteins.

9.3 PHYLOGENETIC ANALYSIS TOOLS

The most convenient way to construct a phylogenetic tree is to use online tools. A good online phylogenetic analysis tool is available at **Phylogeny.fr** (http://www.phylogeny.fr/). This server provides "robust phylogenetic analysis for the non-specialist." The user can build a phylogenetic tree using the "**One Click**" option with all the default settings. Another tool for phylogenetic-tree construction is **MEGA** version 5[1] (as of October 2013). MEGA stands for **Molecular Evolutionary Genetics Analysis**, and it was developed by a group of well-known evolutionary biologists. MEGA can be downloaded from http://www.megasoftware.net/. MEGA is easy to operate, the toolbar is self-explanatory, and there are instructions provided. A recent publication by Hall[2] is also a good resource to understand MEGA. Another widely used and versatile downloadable software tool is **PHYLIP** (**Phylogenetics Inference Package**), which is a free package of programs for inferring phylogenies. It was developed by Joseph Felsenstein of the University of Washington (http://evolution.genetics.washington.edu/phylip.html). A widely used and affordable *commercial* software program for phylogenetic analysis is **PAUP** (**Phylogenetic Analysis Using Parsimony** (and Other Methods)), written by David Swofford. Another downloadable phylogenetic software tool is **MacClade** (http://macclade.org/macclade.html), written by David Maddison and Wayne Maddison. On the MacClade link, click on "Acquiring MacClade" or access the downloadable link directly at http://macclade.org/download.html.

There are several other phylogenetic analysis tools available on the web. Many of these require special formatting of data for entry, and they send the results through e-mail instead of providing real-time display of results. These tools can be checked out at the following link: http://molbiol-tools.ca/Phylogeny.htm.

9.4 PRINCIPLES OF PHYLOGENETIC-TREE CONSTRUCTION

Although a number of online resources have been mentioned above that can be used to construct/reconstruct phylogenetic trees, it is nevertheless important to understand the assumptions and steps involved in phylogenetic-tree construction for conceptual clarity.

There are certain assumptions behind making a phylogenetic tree, such as (1) the sequences are homologous—that is, the sequences share a common ancestry and they diverged through time as they evolved—and (2) each position evolved independently. The quality of multiple sequence alignment is the key to obtaining a reliable phylogenetic tree. *When using coding sequences, it is desirable to use the protein sequences to reconstruct the phylogenetic tree*.

Construction of a phylogenetic tree involves the following steps: (1) Selection of the appropriate molecular marker (genes/proteins/mitochondrial DNA), (2) Multiple sequence alignment, (3) Selection of a model of evolution, (4) Construction of the phylogenetic tree, (5) Assessment of the reliability of the tree.

9.4.1 Selection of the Appropriate Molecular Marker

The choice of nucleic acid or protein sequences as the appropriate marker depends on the need. A molecular marker in phylogenetic analysis is the

biological information that is used to infer the evolutionary relationships among taxa. In general, when coding sequences are used, it is desirable to use protein sequences to construct the phylogenetic tree. Some of the reasons why protein sequences are more appropriate are as follows:

1. There are more possible character states for amino acids (20) than nucleotides (4); the terminals may share a character state by chance simply because a given position can have only one of 4 possible character states (as opposed to 20 for amino acids).
2. Amino-acid-substitution matrices are more sophisticated than nucleotide-substitution matrices.
3. The existence of codon bias for the same amino acid in different species might artificially inflate the nucleotide sequence variation.

However, nucleotide sequences can also be used under certain circumstances to obtain a reliable tree, such as when comparing genes whose sequences are highly conserved among species, or comparing the evolution of genes in geographically separated populations within a species. Slowly evolving gene sequences can be used to assess the evolutionary relationship between distantly related species and, conversely, rapidly evolving gene sequences can be used for recently evolved species.

9.4.2 Multiple Sequence Alignment

Alignment of sequences is the most important step in constructing a reliable phylogenetic tree. Multiple sequence alignment identifies blocks of conserved residues. A good alignment should also have fewer gaps/long gaps. Gaps indicate sequences gained or lost (insertions–deletions) during evolution. The user may decide to use the entire alignment or use parts of it. There are no set rules regarding which sections of the alignment to remove; the user should apply judgment. If the alignment is ambiguous at the two ends, the ends can be removed. Such editing can also be done using **Gblocks**[3,4]. Gblocks eliminates poorly aligned positions and divergent regions of a DNA or protein alignment to make it more suitable for phylogenetic analysis. Gblocks can be accessed at http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=gblocks, or at http://molevol.cmima.csic.es/castresana/Gblocks_server.html. The former link provides an example of how to enter the alignment data. The latter link provides an example of an output file showing the blocks selected from a protein alignment.

The "One Click" link of Phylogeny.fr (http://www.phylogeny.fr/) provides the option to utilize Gblocks to eliminate poorly aligned positions and divergent regions. This option is selected as part of the default settings. The user may choose to uncheck this option in order to use the entire sequence instead of the edited sequence.

9.4.3 Selection of a Model of Evolution

An evolutionary model of sequence data is a model of nucleotide or amino-acid substitution and consequent divergence of sequences. The evolutionary (substitution) models play an important role in the analysis of molecular sequence data. These models filter the complexity of the biological mutation process into simpler patterns that can be described and predicted using a small number of parameters. Substitution models attempt to predict the rate of substitution for nucleotides or amino acids at a given site, and also the distribution of substitutions across the entire sequence. The differential rate of substitutions across the sequence is called the **rate heterogeneity**.

Multiple alignment is followed by the selection of an appropriate evolutionary model. There are many such models. All statistical models are based on certain assumptions. One assumption is that each position in the nucleic acid or protein evolves independently. In reality, that is not the case; there are hot spots of mutation, and also some mutations are more tolerated than others.

The simplest way to determine divergence is to count the number of substitutions. However, there are caveats in such a simplistic approach. For example, an observed substitution (e.g. A→G) may not be the original substitution, but may have involved an intermediate substitution (e.g. A→T→G). Likewise, the absence of substitution at a position may also mean that an original substitution has been reversed (reverse mutation) during evolution to restore the original residue (e.g. A→G→A). Substitution models are statistical models that are supposed to correct for these biases. *Note that these methods are based on general mathematical and statistical principles that have their own set of assumptions.* The simplest substitution model *for nucleotides* is the **Jukes–Cantor (JC) one-parameter model**, which assumes that all nucleotides occur in equal frequency (25%) and are substituted with equal probability. This model requires a single parameter denoting rate. However, it is well known that transition mutations are more common than transversion mutations. **Kimura's two-parameter model** accounts for this, and proposes that transition mutations provide a better estimate of evolutionary divergence than transversion mutations. This model requires two parameters denoting rate. Like the Jukes–Cantor model, Kimura's model also assumes that all nucleotides occur in equal

frequency (25%). There are other more complex models of nucleotide substitution, such as the **Felsenstein model** and the **Hasegawa–Kishono–Yano (HKY) model**, which assume that nucleotides occur at different frequencies, and that transitions and transversions occur at different rates. The **general time reversible (GTR) model**, also known as the **general reversible (REV) model** is even more complex and assumes different rates of substitution for each pair of nucleotides, in addition to assuming different frequencies of occurrence of nucleotides. For these models, the nucleotide frequencies are estimated by the observed frequencies in the alignment. Some *amino acid substitution models* are the **Dayhoff model (PAM)**, the **Bishop–Friday model**, the **Jones–Taylor–Thornton (JTT) model**, the **Whelan and Goldman (WAG) model**, and the **Le Gascuel (LG) model**. The simplest model is the Bishop–Friday model, which assumes that all amino acids occur at equal frequency and all substitutions occur at the same rate. All other models assume different amino-acid frequencies and different substitution rates, which are experimentally determined.

The substitution model utilized for a particular data set can be displayed by the software, such as **MEGA** version 5[1] (discussed above).

9.4.4 Construction of the Phylogenetic Tree

The choice of an appropriate tree-building method for a given data set is a crucial but complex issue. Many methods have been described for reconstructing phylogenetic trees; each one has its own merits and demerits[5]. This is a highly specialized area of computation and statistics. Therefore, only some overall principles are discussed here. The methods to construct phylogenetic trees can be classified into two major types: (1) **distance-based** and (2) **character-based**, also called the **discrete method**.

9.4.4.1 Distance-Based (Distance-Matrix) Methods

In distance-based methods, the distance between each pair of sequences is calculated, and a distance matrix is computed. This distance matrix is used for tree construction. Distance-based methods use substitution models; hence, they are model based. Figure 9.2 A shows a simple distance matrix of four 10-nt-long sequences that differ from one another by 1, 2, 3, or 4 nucleotides. These nucleotide differences are used to compute the evolutionary distances among these sequences. There are two popular distance-based methods, the **unweighted pair group method with arithmetic mean (UPGMA)** and **neighbor joining (NJ)**.

The **UPGMA** is the simplest distance-matrix method, and it employs sequential clustering to build a rooted phylogenetic tree. First, all sequences are compared through pairwise alignment to compute the distance matrix. Using this matrix, the two sequences with minimum distance are identified and clustered as a single pair. Next, the distance between this pair and all other sequences is recalculated to form a new matrix. Using this new matrix, the sequence that is closest to the first pair is identified and clustered. This process is repeated until all sequences have been incorporated in the cluster. Figure 9.2 B shows how an UPGMA tree is computed. *Because the process is "unweighted," all pairwise distance are assumed to contribute equally.*

The **neighbor-joining (NJ) method**[6] is the most widely used distance-matrix method. It starts with a star tree—that is, it is assumed that the branches leading to the respective OTUs (the sequences) radiate from one internal node forming a star-like pattern. Next, a pair of sequences is chosen at random, removed from the star, and attached to a second internal node which is connected by a branch to the center of the star-like pattern (Figure 9.3). The branch lengths are calculated. These two sequences are then returned to their original positions and another pair is selected to repeat the same operation. The goal of these repetitive operations until all possible pairs have been examined is to find out the combination of neighbors that minimizes the total length of the phylogenetic tree.

9.4.4.2 Character-Based Methods

In contrast to the distance-matrix methods, the character-based methods utilize the sequence itself rather than the pairwise distance obtained from the sequence features. A character is a site (position) in the alignment. There are two popular character-based methods, **maximum parsimony (MP) and maximum likelihood (ML)**.

The **maximum parsimony** method computes many trees from the given data set and assigns a cost to each tree. The assumption of maximum parsimony is that the simplest tree is the most plausible tree. The simplest tree is the one that requires the fewest number of changes to explain the data in the alignment. Thus, parsimony uses the data and does not attempt to use any model to estimate the total number of changes. The tree score is the sum of character lengths over all sites. If more than one tree with a smallest number of changes can be obtained, then the trees are said to be equally parsimonious. In maximum parsimony, the site (position of the sequence) that has *at least two different kinds of nucleotides (bases) represented in at least two of the sequences* is considered to be an informative site (Figure 9.4 A). Figure 9.4 B shows the principle of tree

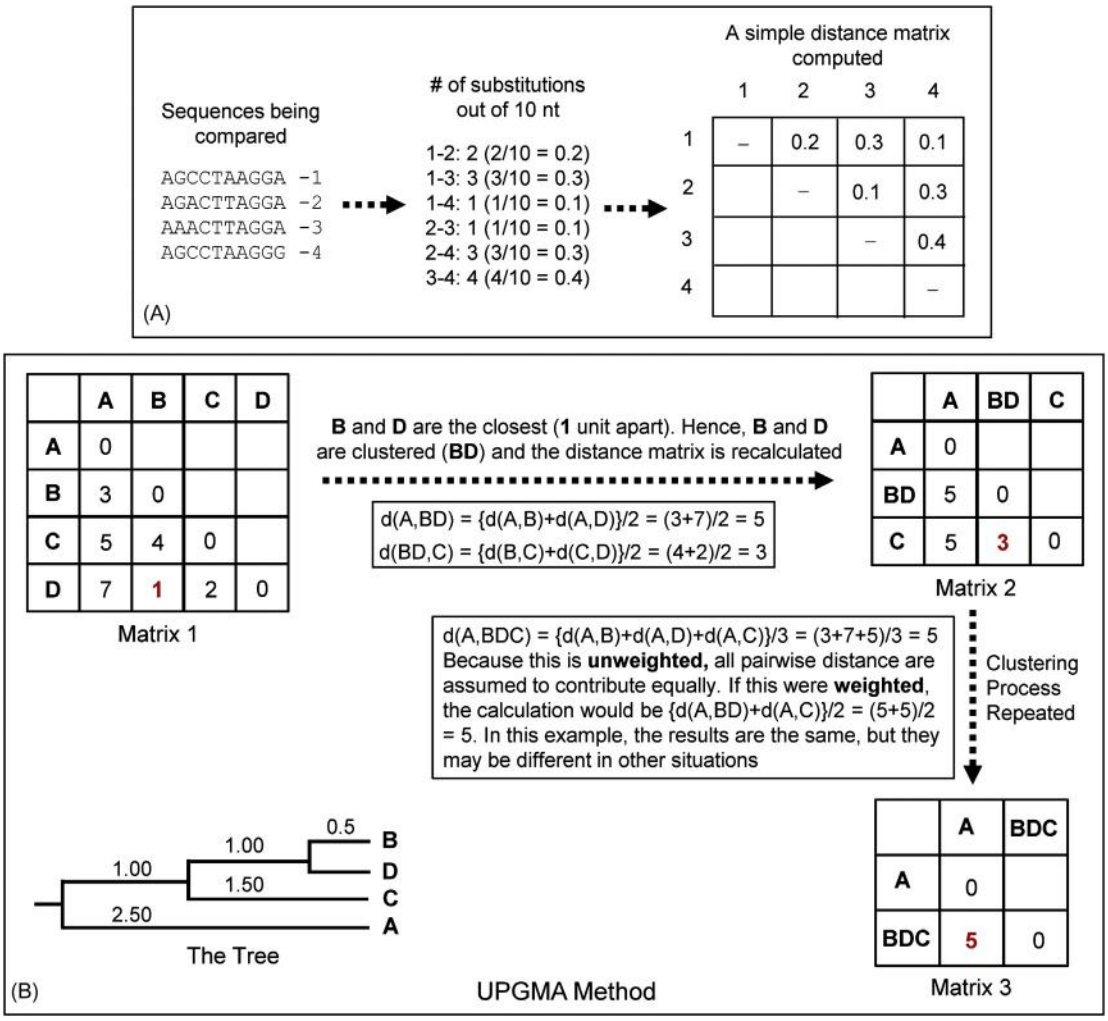

FIGURE 9.2 **Construction of phylogenetic tree using the distance-matrix method.** (A) A simple distance matrix of four 10-nt-long sequences is shown; the sequences differ from one another by 1, 2, 3, or 4 nucleotides. (B) The UPGMA method involves sequential clustering, with calculation of a new distance matrix at each step (see text).

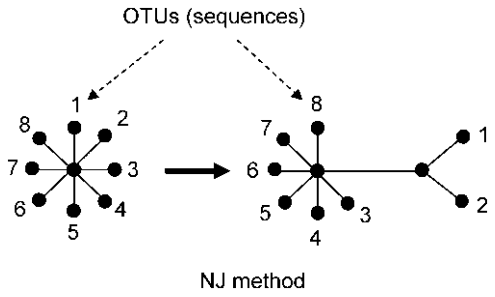

FIGURE 9.3 **Construction of phylogenetic tree using Saitou and Nei's neighbor-joining method.** See text for details.

construction by maximum parsimony using the informative sites (positions 7 and 9) of the sequences shown in Figure 9.4 A. The figure shows that tree 1 is the most parsimonious tree because its topology is based on the minimum number of mutations.

Maximum likelihood is a statistical method that estimates the unknown parameters of a probability model. The maximum-likelihood method is currently widely used for the construction of phylogenetic trees because of increased computational ability. Maximum likelihood evaluates the probability that the selected evolutionary model predicts the observed sequences. In other words, the topology of the phylogenetic trees constructed using maximum likelihood should yield the highest probability of producing the observed sequences.

The use of **Bayesian** phylogenetic analysis is far more recent than the maximum-parsimony and maximum-likelihood methods. The Bayesian phylogenetic method has gained considerable ground ever since the use of Bayesian statistics in phylogenetics was proposed in the mid-1990s. The Bayesian method draws inference on the probability of an unknown event by deriving a "posterior probability." Unlike

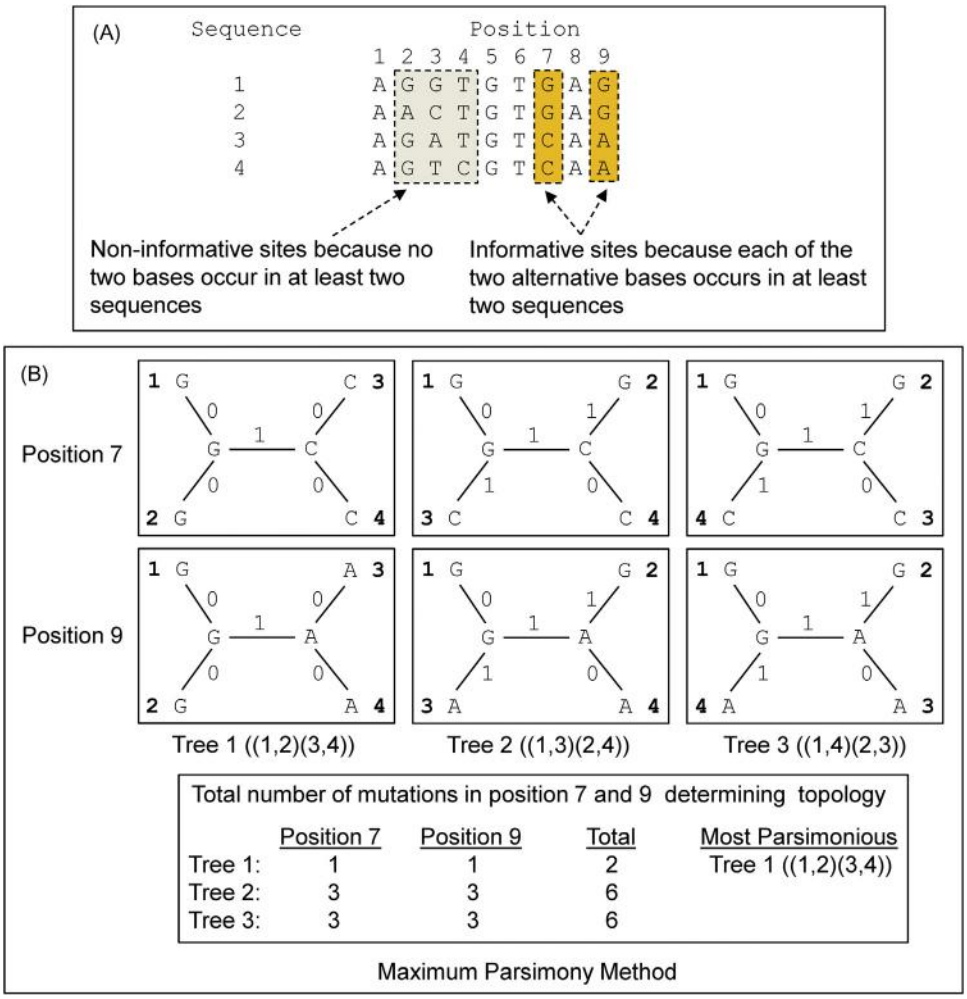

FIGURE 9.4 **The maximum parsimony method.** (A) Informative and non-informative sites considered in maximum parsimony. Non-informative sites do not have each of the alternative bases occurring in at least two sequences. In contrast, in an informative site, each of the alternative bases occurs in at least two sequences. (B) Principles of tree construction by the maximum parsimony method. Tree 1 is the most parsimonious tree because its topology is based on the minimum number of mutations (see text).

standard statistical tests, in which the existing data are used to test a hypothesis, Bayesian statistics uses prior knowledge, in addition to the existing data, to test a hypothesis. The prior knowledge/data provide an estimate of the prior probability of an event, whereas integrating the existing data with the prior probability helps estimate the posterior probability of the event. A prior probability might be derived based on a set of known principles or experimental results. Tree construction in the Bayesian method utilizes repetitive random sampling using a Markov chain Monte Carlo (MCMC) process, which seeks the tree topology with increasingly higher score with each repetitive sampling. Finally, the consensus tree with the highest posterior probability is built from a set of high-scoring tree topologies. The Bayesian method is faster than the ML method, and hence can handle large data sets. **MrBayes** is a Bayesian phylogenetic analysis tool. An online version is available at http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type = mrbayes. This link also shows the format of data entry. Alternatively, MrBayes can be downloaded from http://mrbayes.sourceforge.net/. MrBayes was written by John Huelsenbeck, Bret Larget, Paul van der Mark, Fredrik Ronquist, Donald Simon, and Maxim Teslenko (http://mrbayes.sourceforge.net/authors.php).

9.4.5 Assessment of the Reliability of a Phylogenetic Tree

Construction of a phylogenetic tree is followed by an assessment of the reliability of the tree. Determining the reliability of the tree means determining whether the topology of the tree is accurate or whether a better tree can be obtained. These questions are answered by **bootstrapping** the reconstructed tree.

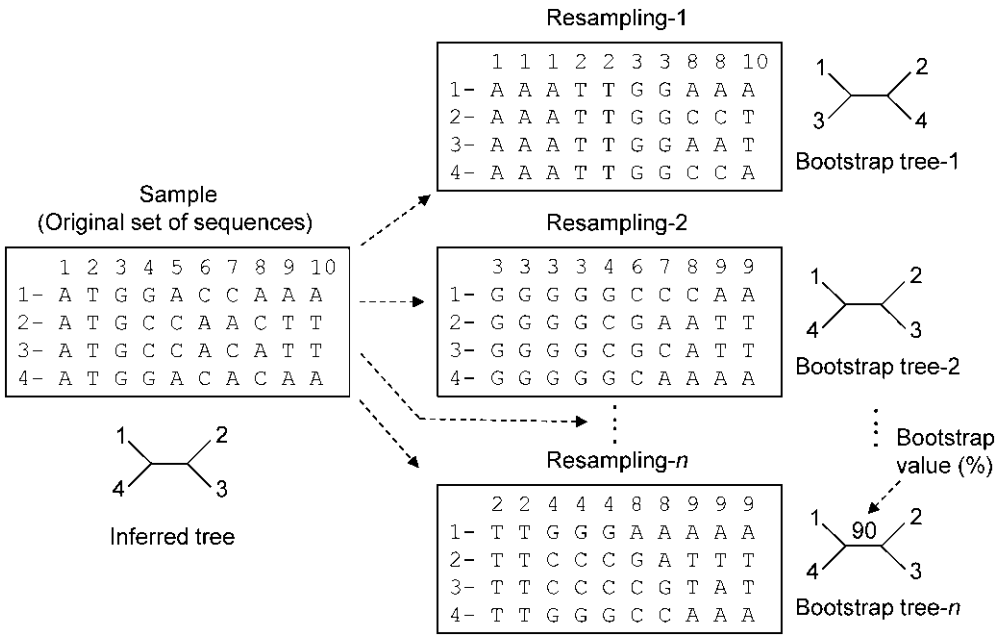

FIGURE 9.5 **Principles of bootstrapping the phylogenetic tree.** The bootstrap method involves repeated resampling (with replacement) from the original sample to create many new subsets of pseudosamples that are subjected to the same analysis as the original sample to obtain many bootstrap trees. The topology of these bootstrap trees is compared with that of the original tree to statistically assess the reliability of the original phylogenetic tree.

Felsenstein[7] first applied the bootstrap method to phylogenetic analysis to assess the reliability of the tree. (Phylogenetic) tree bootstrapping is a computationally performed statistical analysis, which is based on Efron's original bootstrap technique of resampling one's own data to infer the variability of the estimate. The bootstrap method involves repeated resampling (with replacement) from the original samples to create many new subsets of pseudosamples that are subjected to the same analysis as the original samples. The resampling with replacement means that some of the characters/data of the original samples will be in the bootstrap sample multiple times, whereas others will not appear at all. The statistical concept behind such resampling is that if a parameter can be estimated from samples drawn from a population, then the reliability of the estimate of that parameter can be verified by drawing new samples from the same population. The higher the number of resamplings, the greater is the confidence level of the estimate.

In the case of the bootstrap method using sequences, once the phylogenetic tree is constructed after aligning the original set of sequences, the sequences are repeatedly resampled to create many new subsets of derived sequences, i.e. the bootstrap samples. Each round of resampling (with replacement) of the original set of sequences creates a new subset of bootstrap samples of derived sequences. In each derived sequence, some of the bases from the original sequence will be represented multiple times, whereas other bases will not appear at all. One bootstrapping may perform 500−1000 such resamplings from the original sequences.

The derived sequences of each subset are then aligned and a new phylogenetic tree (bootstrap tree) is constructed using the same tree-construction method used to construct the original tree (e.g. neighbor-joining method, maximum-parsimony method, etc.). When the splitting pattern of an interior branch (branch topology) in the original tree is reproduced in the bootstrap tree, that branch is given a value of 1 (identity value). In other words, when an interior branch is given a value of 1, it is assumed to accurately predict the clade and the sister taxa, as reflected not only in the original tree but also in the bootstrap tree. Conversely, when the splitting pattern of an interior branch in the original tree is not reproduced in the bootstrap tree, that branch is given a value of 0. This process is repeated hundreds of times, and the percentage of times each interior branch is given a value of 1 is computed. This is known as the **bootstrap value** or **bootstrap confidence value**. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered accurate. Bootstrap values, expressed as percentages, are indicated on the branches. Therefore, a bootstrap value of 95 indicated on a branch means that 95% of the bootstrap trees support the topology at the branch obtained in the original phylogenetic tree. Figure 9.5 shows the principle of bootstrapping.

It should be remembered that, despite the rigor, the construction of phylogenetic trees is not exact and it involves general mathematical and statistical principles that have their own set of assumptions. As a result, many phylogenetic trees reconstructed from molecular sequences may conflict with common sense; they may be partially correct or even be incorrect[8].

9.5 MONOPHYLY, POLYPHYLY, AND PARAPHYLY

This concept relates to the groupings of organisms. If the classification is performed based on synapomorphic characters (shared derived characters), monophyletic groups are obtained. A monophyletic group includes the last common ancestor (LCA) plus all the descendants of the LCA. Monophyly can be assigned based on nodes as well as apomorphies (Figure 9.6). For example, mammals form a monophyletic group; so do birds, fish, etc. Monophyletic groups form clades and provide accurate information about the evolutionary history.

If the classification is performed based on homoplastic characters (similar characters that evolved independently in different groups through convergent evolution), polyphyletic groups are obtained. A polyphyletic group includes the descendants only and excludes the LCA, and the taxa are grouped based on superficial similarities (Figure 9.6). Thus, polyphyletic taxa could be evolutionarily very distant but linked by homoplasy. Polyphyletic groups do not provide any accurate information about the evolutionary history. In fact, once it is realized that a group of taxa are polyphyletic, they are reclassified. For example, birds and bats could form a polyphyletic group based on homeothermy and the ability to fly. Similarly, sharks and dolphins could form a polyphyletic group based on the ability to swim and other aquatic adaptations.

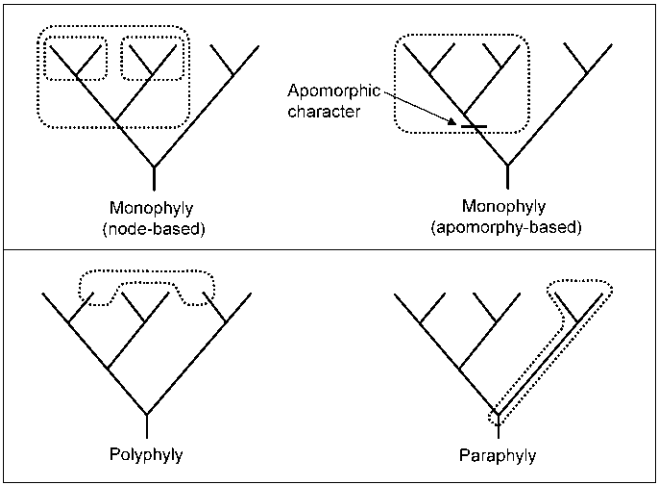

FIGURE 9.6 **Character-based classification to obtain monophyletic, polyphyletic, and paraphyletic groups.** A monophyletic group includes the last common ancestor (LCA) plus all the descendants of the LCA. A polyphyletic group includes the descendants only and excludes the LCA. A paraphyletic group includes the LCA but does not include one or more descendants.

If the classification is performed based on symplesiomorphic characters (shared ancestral characters), paraphyletic groups are obtained. A paraphyletic group includes the LCA but does not include one or more descendants. Therefore, a paraphyletic group is an incomplete clade and does not provide much information about the recent evolutionary history of the taxa concerned (Figure 9.6).

The terms polyphyly and paraphyly are of academic and historical interest. From the phylogenetic perspective, only monophyletic groups are important.

9.6 SPECIES TREES VERSUS GENE TREES

Phylogenetic trees can be constructed to depict the evolutionary history of species/populations or genes. A phylogenetic tree that shows the evolutionary history of species/populations is called a **species tree**. **Speciation** involves the splitting of an ancestral population into two populations that diverge and become reproductively isolated, giving rise to two species. Therefore, the branching in a species tree shows the time when the two species descended from the ancestral population and became reproductively isolated.

In contrast, when the phylogenetic tree is constructed based on a group of homologous gene sequences, where each sequence is sampled from a different species, then a gene tree is obtained. The general assumption is that gene trees are less ambiguous than species trees because gene trees are constructed based on definitive molecular data. However, the event that drives divergence between two populations leading to speciation is reproductive isolation, whereas the event that drives divergence between two homologous gene sequences is mutation. Mutations in genes and speciation do not necessarily happen at the same rate. Genetic polymorphism and multigene families add additional twists to the problem of gene tree to species tree extrapolation. When there is allelic polymorphism within species, a gene tree constructed from DNA sequences for a given gene can be quite different from the species tree, and this is particularly so when the time of divergence between different species is

short[9]. When the gene whose evolutionary history is being studied belongs to a multigene family, it may be difficult to correctly assign the homology of the sequences under study.

Therefore, inferring species trees from gene trees requires a great deal of caution. In general, gene trees are useful in studying the evolutionary history of the members a gene family, and inferring the evolutionary relatedness of the species from which the genes are obtained.

References

1. Tamura K, et al. *Mol Biol Evol* 2011;**28**:2731–9.
2. Hall BG. *Mol Biol Evol* 2013;**30** 1229-1135
3. Castresana J. *Mol Biol Evol* 2000;**17**:540–52.
4. Talavera G, Castresana J. *Syst Biol* 2007;**56**:564–77.
5. Yang Z, Rannala B. *Nat Rev Genet* 2012;**13**:303–14.
6. Saitou N, Nei M. *Mol Biol Evol* 1987;**4**:406–25.
7. Felsenstein J. *Evolution* 1985;**39**:783–91.
8. Lake JA, Moore JE. Trends guide to bioinformatics. *Trends J Suppl* 1998;**1998**:22–3.
9. Pamilo P, Nei M. *Mol Biol Evol* 1988;**5**:568–83.