CHAPTER TEN

Phylogenetics Basics

Biological sequence analysis is founded on solid evolutionary principles (see Chapter 2). Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic trees. Thus, molecular phylogenetics is a fundamental aspect of bioinformatics. In this chapter, we focus on phylogenetic tree construction. Before discussing the methods of phylogenetic tree construction, some fundamental concepts and background terminology used in molecular phylogenetics need to be described. This is followed by discussion of the initial steps involved in phylogenetic tree construction.

MOLECULAR EVOLUTION AND MOLECULAR PHYLOGENETICS

To begin the phylogenetics discussion, we need to understand the basic question, "What is evolution?" Evolution can be defined in various ways under different contexts. In the biological context, evolution can be defined as the development of a biological form from other preexisting forms or its origin to the current existing form through natural selections and modifications. The driving force behind evolution is natural selection in which "unfit" forms are eliminated through changes of environmental conditions or sexual selection so that only the fittest are selected. The underlying mechanism of evolution is genetic mutations that occur spontaneously. The mutations on the genetic material provide the biological diversity within a population; hence, the variability of individuals within the population to survive successfully in a given environment. Genetic diversity thus provides the source of raw material for the natural selection to act on.

Phylogenetics is the study of the evolutionary history of living organisms using tree-like diagrams to represent pedigrees of these organisms. The tree branching patterns representing the evolutionary divergence are referred to as *phylogeny.* Phylogenetics can be studied in various ways. It is often studied using fossil records, which contain morphological information about ancestors of current species and the timeline of divergence. However, fossil records have many limitations; they may be available only for certain species. Existing fossil data can be fragmentary and their collection is often limited by abundance, habitat, geographic range, and other factors. The descriptions of morphological traits are often ambiguous, which are due to multiple genetic factors. Thus, using fossil records to determine phylogenetic relationships can often be biased. For microorganisms, fossils are essentially nonexistent, which makes it impossible to study phylogeny with this approach.

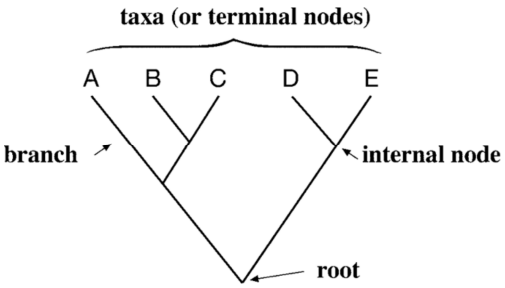


Figure 10.1: A typical bifurcating phylogenetic tree showing root, internal nodes, terminal nodes and branches.

Fortunately, molecular data that are in the form of DNA or protein sequences can also provide very useful evolutionary perspectives of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes. Because genes are the medium for recording the accumulated mutations, they can serve as *molecular fossils*. Through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

The advantage of using molecular data is obvious. Molecular data are more numerous than fossil records and easier to obtain. There is no sampling bias involved, which helps to mend the gaps in real fossil records. More clear-cut and robust phylogenetic trees can be constructed with the molecular data. Therefore, they have become favorite and sometimes the only information available for researchers to reconstruct evolutionary history. The advent of the genomic era with tremendous amounts of molecular sequence data has led to the rapid development of molecular phylogenetics.

The field of molecular phylogenetics can be defined as the study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules. Based on the sequence similarity of the molecules, evolutionary relationships between the organisms can often be inferred.

Major Assumptions

To use molecular data to reconstruct evolutionary history requires making a number of reasonable assumptions. The first is that the molecular sequences used in phylogenetic construction are homologous, meaning that they share a common origin and subsequently diverged through time. Phylogenetic divergence is assumed to be bifurcating, meaning that a parent branch splits into two daughter branches at any given point. Another assumption in phylogenetics is that each position in a sequence evolved independently. The variability among sequences is sufficiently informative for constructing unambiguous phylogenetic trees.

TERMINOLOGY

Before discussing methods for reconstruction of phylogenies, it is useful to define some frequently used terminology that characterizes a phylogenetic tree. A typical bifurcating phylogenetic tree is a graph shown in Figure 10.1. The lines in the tree are

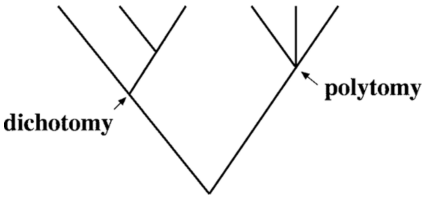


Figure 10.2: A phylogenetic tree showing an example of bifurcation and multifurcation. Multifurcation is normally a result of insufficient evidence to fully resolve the tree or a result of an evolutionary process known as *radiation*.

called *branches*. At the tips of the branches are present-day species or sequences known as *taxa* (the singular form is *taxon*) or operational taxonomic units. The connecting point where two adjacent branches join is called a *node*, which represents an inferred ancestor of extant taxa. The bifurcating point at the very bottom of the tree is the *root node*, which represents the common ancestor of all members of the tree.

A group of taxa descended from a single common ancestor is defined as a *clade* or *monophyletic group*. In a monophyletic group, two taxa share a unique common ancestor not shared by any other taxa. They are also referred to as *sister taxa* to each other (e.g., taxa B and C). The branch path depicting an ancestor–descendant relationship on a tree is called a *lineage*, which is often synonymous with a tree branch leading to a defined monophyletic group. When a number of taxa share more than one closest common ancestors, they do not fit the definition of a clade. In this case, they are referred to as *paraphyletic* (e.g., taxa B, C, and D).

The branching pattern in a tree is called *tree topology*. When all branches bifurcate on a phylogenetic tree, it is referred to as *dichotomy*. In this case, each ancestor divides and gives rise to two descendants. Sometimes, a branch point on a phylogenetic tree may have more than two descendents, resulting in a *multifurcating node*. The phylogeny with multifurcating branches is called *polytomy* (Fig. 10.2). A polytomy can be a result of either an ancestral taxon giving rise to more than two immediate descendants simultaneously during evolution, a process known as *radiation*, or an unresolved phylogeny in which the exact order of bifurcations cannot be determined precisely.

A phylogenetic tree can be either rooted or unrooted (Fig. 10.3). An *unrooted phylogenetic tree* does not assume knowledge of a common ancestor, but only positions the taxa to show their relative relationships. Because there is no indication of which node represents an ancestor, there is no direction of an evolutionary path in an

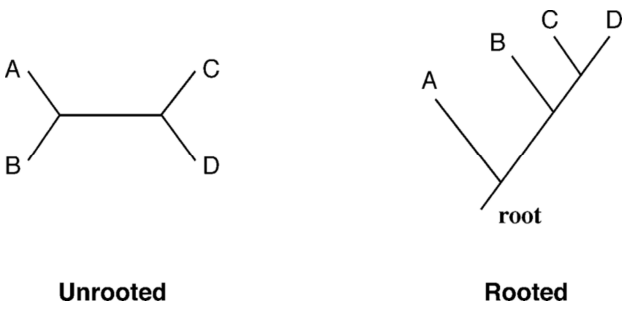


Figure 10.3: An illustration of rooted versus unrooted trees. A phylogenetic tree without definition of a root is unrooted (*left*). The tree with a root is rooted (*right*).

unrooted tree. To define the direction of an evolution path, a tree must be rooted. In a *rooted tree*, all the sequences under study have a common ancestor or root node from which a unique evolutionary path leads to all other nodes. Obviously, a rooted tree is more informative than an unrooted one. To convert an unrooted tree to a rooted tree, one needs to first determine where the root is.

Strictly speaking, the root of the tree is not known; the common ancestor is already extinct. In practice, however, it is often desirable to define the root of a tree. There are two ways to define the root of a tree. One is to use an *outgroup*, which is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time. Outgroups are generally determined from independent sources of information. For example, a bird sequence can be used as a root for the phylogenetic analysis of mammals based on multiple lines of evidence that indicate that birds branched off prior to all mammalian taxa in the ingroup. Outgroups are required to be distinct from the ingroup sequences, but not too distant from the ingroup. Using too divergent sequences as an outgroup can lead to errors in tree construction. In the absence of a good outgroup, a tree can be rooted using the *midpoint rooting approach*, in which the midpoint of the two most divergent groups judged by overall branch lengths is assigned as the root. This type of rooting assumes that divergence from root to tips for both branches is equal and follows the "molecular clock" hypothesis.

Molecular clock is an assumption by which molecular sequences evolve at constant rates so that the amount of accumulated mutations is proportional to evolutionary time. Based on this hypothesis, branch lengths on a tree can be used to estimate divergence time. This assumption of uniformity of evolutionary rates, however, rarely holds true in reality.

GENE PHYLOGENY VERSUS SPECIES PHYLOGENY

One of the objectives of building phylogenetic trees based on molecular sequences is to reconstruct the evolutionary history of the species involved. However, strictly speaking, a gene phylogeny (phylogeny inferred from a gene or protein sequence) only describes the evolution of that particular gene or encoded protein. This sequence may evolve more or less rapidly than other genes in the genome or may have a different evolutionary history from the rest of the genome owing to horizontal gene transfer events (see Chapter 17). Thus, the evolution of a particular sequence does not necessarily correlate with the evolutionary path of the species. The species evolution is the combined result of evolution by multiple genes in a genome. In a species tree, the branching point at an internal node represents the speciation event whereas, in a gene tree, the internal node indicates a gene duplication event. The two events may or may not coincide. Thus, to obtain a species phylogeny, phylogenetic trees from a variety of gene families need to be constructed to give an overall assessment of the species evolution.

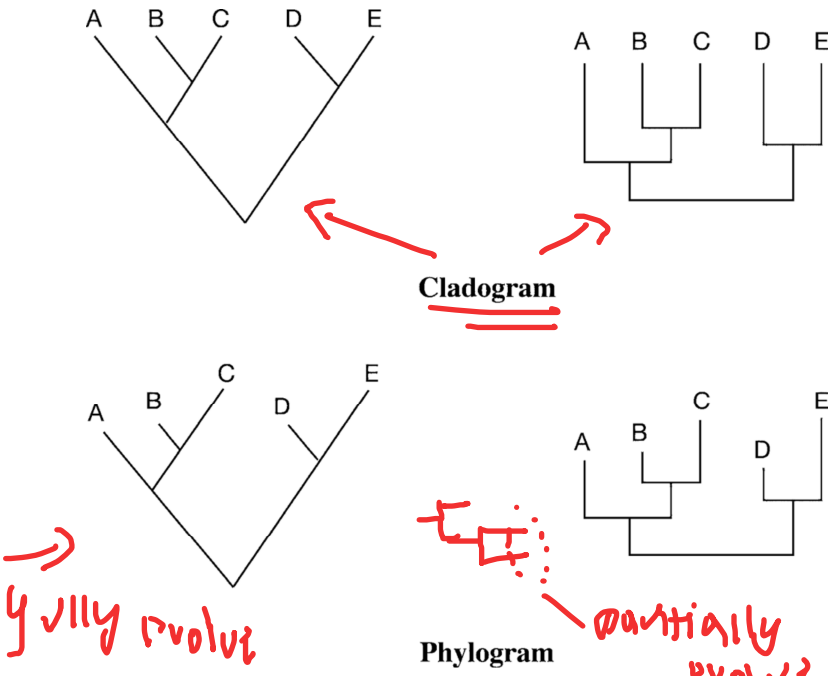


Figure 10.4: Phylogenetic trees drawn as cladograms (*top*) and phylograms (*bottom*). The branch lengths are unscaled in the cladograms and scaled in the phylograms. The trees can be drawn as angled form (*left*) or squared form (*right*).

FORMS OF TREE REPRESENTATION

The topology of branches in a tree defines the relationships between the taxa. The trees can be drawn in different ways, such as a cladogram or a phylogram (Fig. 10.4). In each of these tree representations, the branches of a tree can freely rotate without changing the relationships among the taxa.

In a *phylogram*, the branch lengths represent the amount of evolutionary divergence. Such trees are said to be scaled. The scaled trees have the advantage of showing both the evolutionary relationships and information about the relative divergence time of the branches. In a *cladogram*, however, the external taxa line up neatly in a row or column. Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning. In such an unscaled tree, only the topology of the tree matters, which shows the relative ordering of the taxa.

To provide information of tree topology to computer programs without having to draw the tree itself, a special text format known as the *Newick format* is developed. In this format, trees are represented by taxa included in nested parentheses. In this linear representation, each internal node is represented by a pair of parentheses that enclose all member of a monophyletic group separated by a comma. For a tree with scaled branch lengths, the branch lengths in arbitrary units are placed immediately after the name of the taxon separated by a colon. An example of using the Newick format to describe tree topology is shown in Figure 10.5.

Sometimes a tree-building method may result in several equally optimal trees. A consensus tree can be built by showing the commonly resolved bifurcating portions

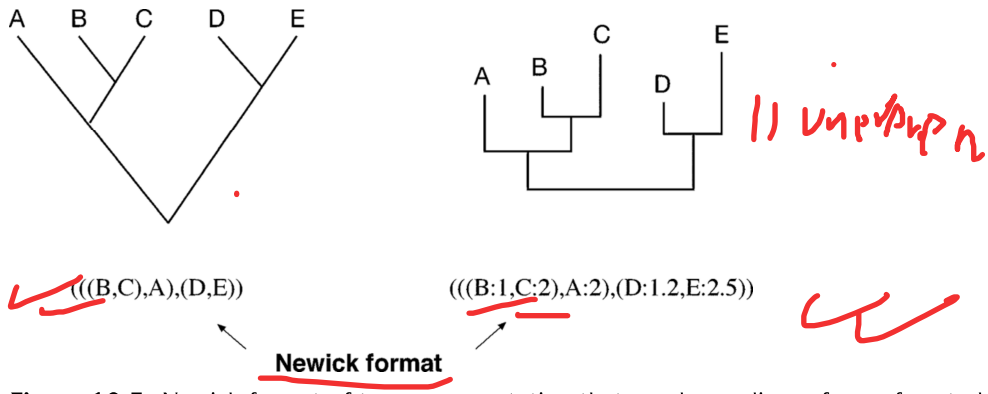


Figure 10.5: Newick format of tree representation that employs a linear form of nested parentheses within which taxa are separated by commas. If the tree is scaled, branch lengths are indicated immediately after the taxon name. The numbers are relative units that represent divergent times.

and collapsing the ones that disagree among the trees, which results in a polytomy. Combining the nodes can be done either by strict consensus or by majority rule. In a strict consensus tree, all conflicting nodes are collapsed into polytomies. In a consensus tree based on a majority rule, among the conflicting nodes, those that agree by more than 50% of the nodes are retained whereas the remaining nodes are collapsed into multifurcation (Fig. 10.6).

WHY FINDING A TRUE TREE IS DIFFICULT

The main objective of molecular phylogenetics is to correctly reconstruct the evolutionary history based on the observed sequence divergence between organisms. That means finding a correct tree topology with correct branch lengths. However, the search for a correct tree topology can sometimes be extremely difficult and computationally demanding. The reason is that the number of potential tree topologies can

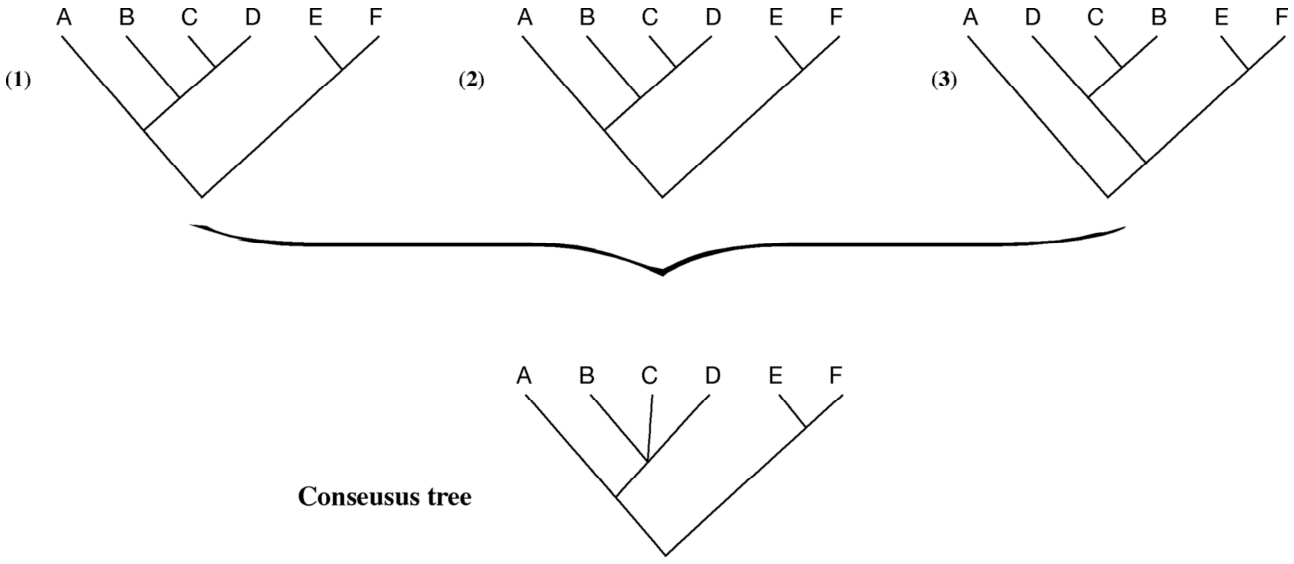


Figure 10.6: A consensus tree is derived from three individual inferred trees based on a majority rule. Conflicting nodes are represented by a multifurcating node in the consensus tree.

be enormously large even with a moderate number of taxa. The increase of possible tree topologies follows an exponential function. The number of rooted trees (N_R) for n taxa is determined by the following formula:

$$N_R = (2n-3)!/2^{n-2}(n-2)! \quad \text{(Eq. 10.1)}$$

In this formula, $(2n-3)!$ is a mathematical expression of factorial, which is the product of positive integers from 1 to $2n-3$. For example, $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$.

For unrooted trees, the number of unrooted tree topologies (N_U) is:

$$N_U = (2n-5)!/2^{n-3}(n-3)! \quad \text{(Eq. 10.2)}$$

An example of all possible rooted and unrooted tree topologies for three and four taxa is shown in Figure 10.7. For three taxa, there is only one possible unrooted tree but three different rooted trees. For four taxa, one can construct three possible unrooted trees and fifteen rooted ones. The number of possible topologies increases extremely rapidly with the number of taxa. According to Equation 10.1 and Equation 10.2, for six taxa, there are 105 unrooted trees and 945 rooted trees. If there are ten taxa, there can be 2,027,025 unrooted trees and 34,459,425 rooted ones. The exponential relationship between the number of tree topologies and the number of taxa is clearly represented in Figure 10.8. There can be an explosive increase in the possible tree topologies as the number of taxa increases. Therefore, it can be computationally very demanding to find a true phylogenetic tree when the number of sequences is large. Because the number of rooted topologies is much larger than that for unrooted ones, the search for a true phylogenetic tree can be simplified by calculating the unrooted trees first. Once an optimal tree is found, rooting the tree can be performed by designating a number of taxa in the data set as an outgroup based on external information to produce a rooted tree.

PROCEDURE

Molecular phylogenetic tree construction can be divided into five steps: (1) choosing molecular markers; (2) performing multiple sequence alignment; (3) choosing a model of evolution; (4) determining a tree building method; and (5) assessing tree reliability. Each of first three steps is discussed herein; steps 4 and 5 are discussed in Chapter 11.

Choice of Molecular Markers

For constructing molecular phylogenetic trees, one can use either nucleotide or protein sequence data. The choice of molecular markers is an important matter because it can make a major difference in obtaining a correct tree. The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study. For studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins, can be used. For example, for evolutionary analysis of different individuals within a population, noncoding regions of mitochondrial

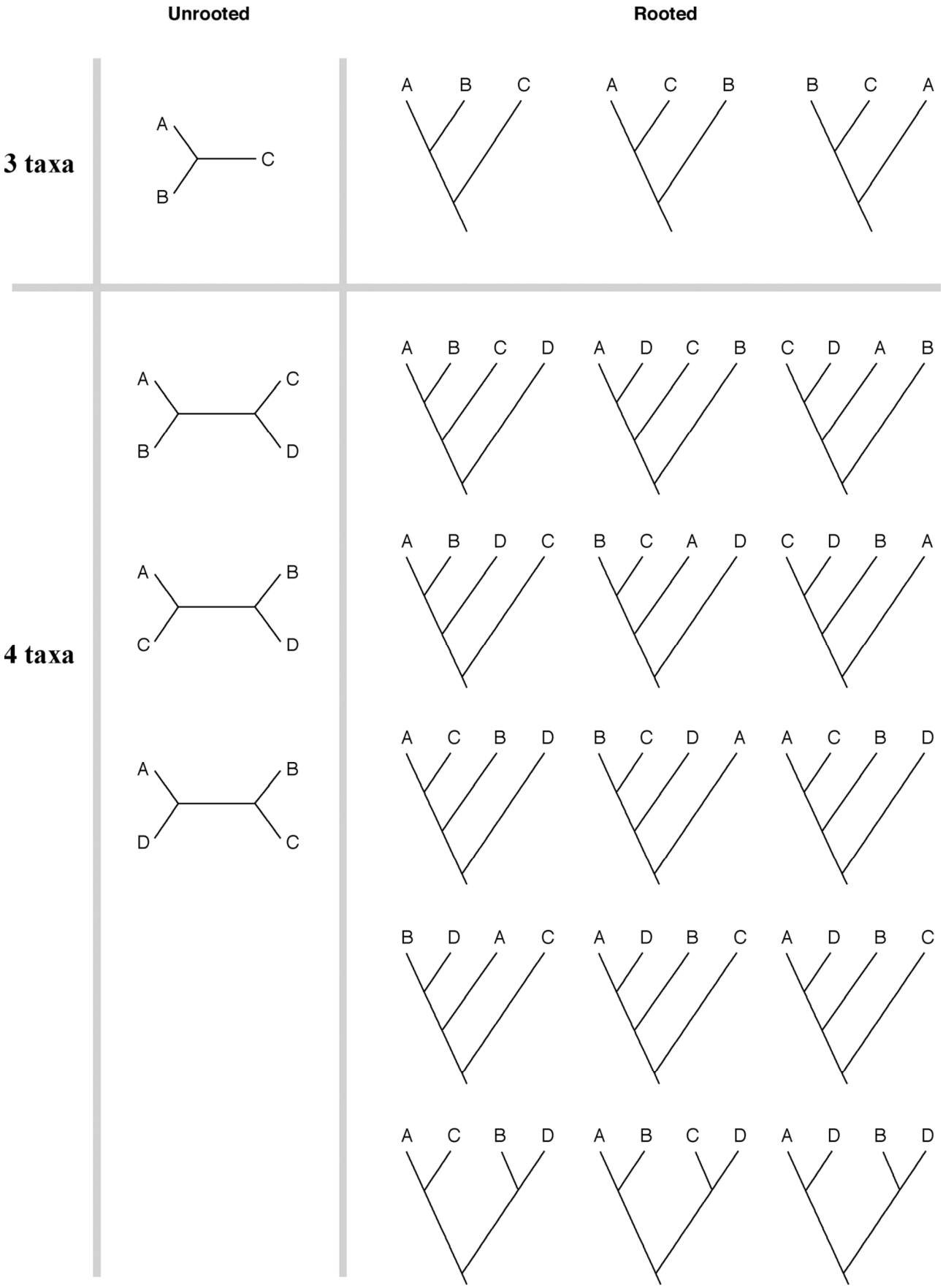


Figure 10.7: All possible tree topologies for three and four taxa. For three taxa, there are one unrooted and three rooted trees. For four taxa, there are three unrooted and fifteen rooted trees.

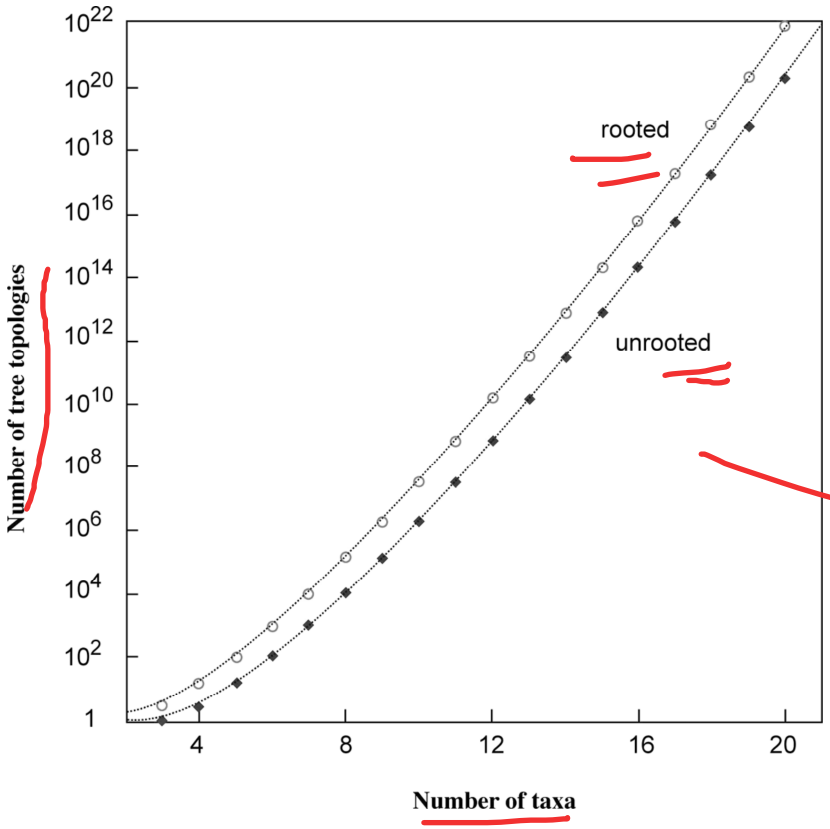


Figure 10.8: Total number of rooted (○) and unrooted (◆) tree topologies as a function of the number of taxa. The values in the *y*-axis are plotted in the log scale.

DNA are often used. For studying the evolution of more widely divergent groups of organisms, one may choose either slowly evolving nucleotide sequences, such as ribosomal RNA or protein sequences. If the phylogenetic relationships to be delineated are at the deepest level, such as between bacteria and eukaryotes, using conserved protein sequences makes more sense than using nucleotide sequences. The reason is explained in more detail next.

In many cases, protein sequences are preferable to nucleotide sequences because protein sequences are relatively more conserved as a result of the degeneracy of the genetic code in which sixty-one codons encode for twenty amino acids, meaning thereby a change in a codon may not result in a change in amino acid. Thus, protein sequences can remain the same while the corresponding DNA sequences have more room for variation, especially at the third codon position. The significant difference in evolutionary rates among the three nucleotide positions also violates one of the assumptions of tree-building. In contrast, the protein sequences do not suffer from this problem, even for divergent sequences.

DNA sequences are sometimes more biased than protein sequences because of preferential codon usage in different organisms. In this case, different codons for the same amino acid are used at different frequencies, leading to sequence variations not attributable to evolution. In addition, the genetic code of mitochondria varies from the standard genetic code. Therefore, for comparison of mitochondria protein-coding genes, it is necessary to translate the DNA sequences into protein sequences.

As mentioned in Chapter 4, protein sequences allow more sensitive alignment than DNA sequences because the former has twenty characters versus four in the latter. It has been shown that two randomly related DNA sequences can result in up to 50% sequence identity when gaps are allowed compared to only 10% for protein sequences. For moderately divergent sequences, it is almost impossible to use DNA sequences to obtain correct alignment. In addition, to align protein-coding DNA sequences, when gaps are introduced to maximize alignment scores, they almost always cause frameshift errors, making the alignment biologically meaningless. Protein sequences clearly have a higher signal-to-noise ratio when it comes to alignment and phylogenetic analysis. Thus, protein-based phylogeny in most cases may be more appropriate than DNA-based phylogeny.

Despite the advantages of using protein sequences in phylogenetic inference, DNA sequences can still be very informative in some cases, such as those for closely related sequences. In this case, faster evolutionary rates at the DNA level become an advantage. In addition, DNA sequences depict synonymous and nonsynonymous substitutions, which can be useful for revealing evidence of positive or negative selection events.

To understand positive or negative selection, it is necessary to make a distinction between synonymous substitutions and nonsynonymous substitutions. *Synonymous substitutions* are nucleotide changes in the coding sequence that do not result in amino acid sequence changes for the encoded protein. *Nonsynonymous substitutions* are nucleotide changes that result in alterations in the amino acid sequences.

Comparing the two types of substitution rates helps to understand an evolutionary process of a sequence. For example, if the nonsynonymous substitution rate is found to be significantly greater than the synonymous substitution rate, this means that certain parts of the protein are undergoing active mutations that may contribute to the evolution of new functions. This is described as *positive selection* or *adaptive evolution*. On the other hand, if the synonymous substitution rate is greater than the nonsynonymous substitution rate, this causes only neutral changes at the amino acid level, suggesting that the protein sequence is critical enough that changes at the amino acid sequence level are not tolerated. In this case, the sequence is said to be under *negative* or *purifying selection*.

Alignment

The second step in phylogenetic analysis is to construct sequence alignment. This is probably the most critical step in the procedure because it establishes positional correspondence in evolution. Only the correct alignment produces correct phylogenetic inference because aligned positions are assumed to be genealogically related. Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree. For that reason, it is essential that the sequences are correctly aligned. Multiple state-of-the-art alignment programs such as T-Coffee should be used. The alignment results from multiple sources should be inspected and compared carefully

to identify the most reasonable one. Automatic sequence alignments almost always contain errors and should be further edited or refined if necessary.

Manual editing is often critical in ensuring alignment quality. However, there is no firm rule on how to modify a sequence alignment. As a general guideline, a correct alignment should ensure the matching of key cofactor residues and residues of similar physicochemical properties. If secondary structure elements are known or can be predicted (see Chapter 14), they can serve to guide the alignment. One of the few alignment programs that incorporates protein secondary structure information is Praline (see Chapter 5).

It is also often necessary to decide whether to use the full alignment or to extract parts of it. Truly ambiguously aligned regions have to be removed from consideration prior to phylogenetic analysis. Which part of the alignment to remove is often at the discretion of the researcher. It is a rather subjective process. In extreme cases, some researchers like to remove all insertions and deletions (indels) and only use positions that are shared by all sequences in the dataset. The clear drawback of this practice is that many phylogenetic signals are lost. In fact, gap regions often belong to *signature indels* unique to identification of a subgroup of sequences and should to be retained for treeing purposes.

In addition, there is an automatic approach in improving alignment quality. Rascal and NorMD (see Chapter 5) can help to improve alignment by correcting alignment errors and removing potentially unrelated or highly divergent sequences. Furthermore, the program Gblocks (http://woody.embl-heidelberg.de/phylo/) can help to detect and eliminate the poorly aligned positions and divergent regions so to make the alignment more suitable for phylogenetic analysis.

Multiple Substitutions

A simple measure of the divergence between two sequences is to count the number of substitutions in an alignment. The proportion of substitutions defines the observed distance between the two sequences. However, the observed number of substitutions may not represent the true evolutionary events that actually occurred. When a mutation is observed as A replaced by C, the nucleotide may have actually undergone a number of intermediate steps to become C, such as A → T → G → C. Similarly, a back mutation could have occurred when a mutated nucleotide reverted back to the original nucleotide. This means that when the same nucleotide is observed, mutations like G → C → G may have actually occurred. Moreover, an identical nucleotide observed in the alignment could be due to parallel mutations when both sequences mutate into T, for instance.

Such multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between sequences. This effect is known as *homoplasy*, which, if not corrected, can lead to the generation of incorrect trees. To correct homoplasy, statistical models are needed to infer the true evolutionary distances between sequences.

Choosing Substitution Models

The statistical models used to correct homoplasy are called *substitution models* or *evolutionary models*. For constructing DNA phylogenies, there are a number of nucleotide substitution models available. These models differ in how multiple substitutions of each nucleotide are treated. The caveat of using these models is that if there are too many multiple substitutions at a particular position, which is often true for very divergent sequences, the position may become saturated. This means that the evolutionary divergence is beyond the ability of the statistical models to correct. In this case, true evolutionary distances cannot be derived. Therefore, only reasonably similar sequences are to be used in phylogenetic comparisons.

Jukes–Cantor Model

The simplest nucleotide substitution model is the Jukes–Cantor model, which assumes that all nucleotides are substituted with equal probability. A formula for deriving evolutionary distances that include hidden changes is introduced by using a logarithmic function.

$$d_{AB} = -(3/4)\ \ln[1 - (4/3)\, p_{AB}] \qquad \text{(Eq. 10.3)}$$

where d_{AB} is the evolutionary distance between sequences A and B and p_{AB} is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

For example, if an alignment of sequences A and B is twenty nucleotides long and six pairs are found to be different, the sequences differ by 30%, or have an observed distance 0.3. To correct for multiple substitutions using the Jukes–Cantor model, the corrected evolutionary distance based on Equation 10.3 is:

$$d_{AB} = -3/4 \ln[1 - (4/3 \times 0.3)] = 0.38$$

The Jukes–Cantor model can only handle reasonably closely related sequences. According to Equation 10.3, the normalized distance increases as the actual observed distance increases. For distantly related sequences, the correction can become too large to be reliable. If two DNA sequences have 25% similarity, p_{AB} is 0.75. This leads the log value to be infinitely large.

Kimura Model

Another model to correct evolutionary distances is called the Kimura two-parameter model. This is a more sophisticated model in which mutation rates for transitions and transversion are assumed to be different, which is more realistic. According to this model, transitions occur more frequently than transversions, which, therefore, provides a more realistic estimate of evolutionary distances. The Kimura model uses the following formula:

$$d_{AB} = -(1/2)\ \ln(1 - 2\,p_{ti} - p_{tv}) - (1/4)\ \ln(1 - 2\,p_{tv}) \qquad \text{(Eq. 10.4)}$$

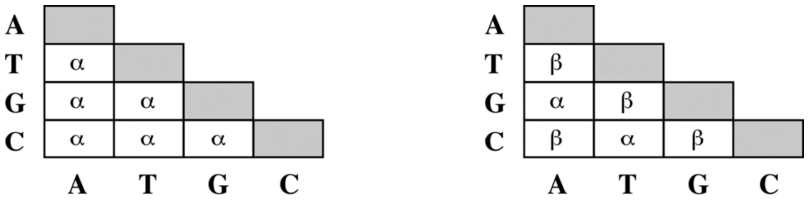


Figure 10.9: The Jukes–Cantor and Kimura models for DNA substitutions. In the Jukes–Cantor model, all nucleotides have equal substitution rates (α). In the Kimura model, there are unequal rates of transitions (α) and transversions (β). The probability values for identical matches are shaded because evolutionary distances only count different residue positions.

where d_{AB} is the evolutionary distance between sequences A and B, p_{ti} is the observed frequency for transition, and p_{tv} the frequency of transversion. Comparison of the Jukes–Cantor model and the Kimura model is graphically illustrated in Figure 10.9.

An example of using the Kimura model can be illustrated by the comparison of sequences A and B that differ by 30%. If 20% of changes are a result of transitions and 10% of changes are a result of transversions, the evolutionary distance can be calculated using Equation 10.4:

$$d_{AB} = -1/2 \ln(1 - 2 \times 0.2 - 0.1) - 1/4 \ln(1 - 2 \times 0.1) = 0.40$$

In addition to these models, there are more complex models, such as TN93, HKY, and GTR, that take many more parameters into consideration. However, these more complex models are normally not used in practice because the calculations are too complicated and the variance levels resulting from the formula are too high.

For protein sequences, the evolutionary distances from an alignment can be corrected using a PAM or JTT amino acid substitution matrix whose construction already takes into account the multiple substitutions (see Chapter 3). Alternatively, protein equivalents of Jukes–Cantor and Kimura models can be used to correct evolutionary distances. For example, the Kimura model for correcting multiple substitutions in protein distances is:

$$d = -\ln(1 - p - 0.2p^2) \qquad \text{(Eq. 10.5)}$$

whereas p is the observed pairwise distance between two sequences.

Among-Site Variations

In all these calculations, different positions in a sequence are assumed to be evolving at the same rate. However, this assumption may not hold up in reality. For example, in DNA sequences, the rates of substitution differ for different codon positions. The third codon mutates much faster than the other two. For protein sequences, some amino acids change much more rarely than others owing to functional constraints. This variation in evolutionary rates is the so-called among-site rate heterogeneity, which can also cause artifacts in tree construction.

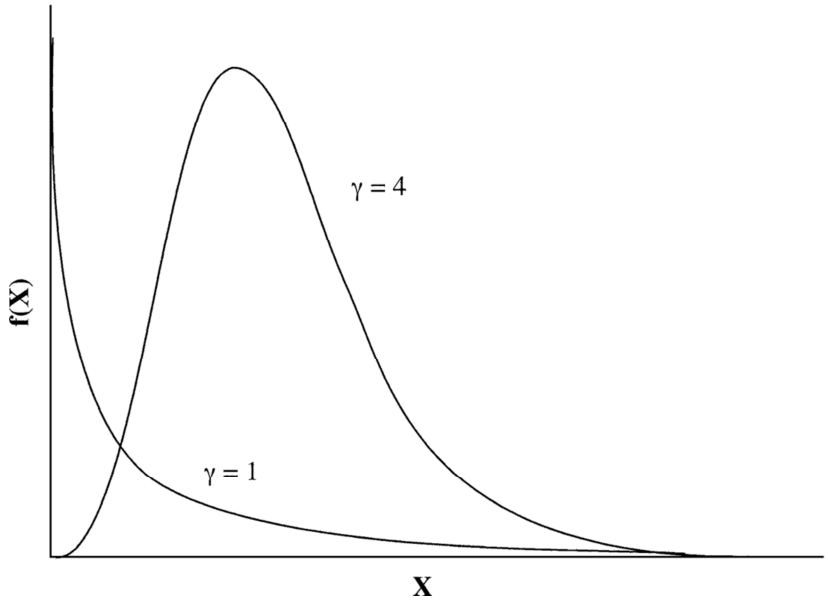


Figure 10.10: Probability curves of γ distribution. The mathematical function of the distribution is $f(x) + (x^{\gamma-1}\ e^{-x})/\Gamma(\gamma)$. The curves assume different shapes depending on the γ-shape parameter (γ).

It has been shown that there are always a proportion of positions in a sequence dataset that have invariant rates and a proportion that have more variable rates. The distribution of variant sites follows a γ distribution pattern. The γ distribution is a general probability function that has distribution curves with variable shapes depending on the values of the γ shape parameter (Fig. 10.10). Therefore, to account for site-dependent rate variation, a γ correction factor can be used. For the Jukes–Cantor model, the evolution distance can be adjusted with the following formula:

$$d_{AB} = (3/4)\alpha[(1 - 4/3\,p_{AB})^{-1/\alpha} - 1 \qquad \text{(Eq. 10.6)}$$

where α is the γ correction factor. For the Kimura model, the evolutionary distance with γ correction factor becomes

$$d_{AB} = (\alpha/2)[1 - 2p_{ti} - p_{tv})^{-1/\alpha} - (1/2)(1 - 2p_{tv})^{-1/\alpha} - 1/2] \qquad \text{(Eq. 10.7)}$$

Estimation of the value of the γ correction factor (α) is implemented in a number of tree-building programs.

SUMMARY

Molecular phylogenetics is the study of evolutionary relationships among living organisms using molecular data such as DNA and protein sequences. It operates on the basis of a number of assumptions – e. g., an evolutionary tree is always binary and all sequence positions evolve independently. The branches of a tree define its topology. The number of possible tree topologies depends on the number of taxa and increases extremely rapidly as the number taxa goes up. A tree based on gene sequences does not always correlate with the evolution of the species. Caution is

needed in extrapolation of phylogenetic results. A phylogenetic tree can be rooted or unrooted. The best way to root a tree is to use an outgroup, the selection of which relies on external knowledge. The first step in phylogenetic construction is to decide whether to use DNA sequences or protein sequences, each having merits and limitations. Protein sequences are preferable in most cases. However, for studying very recent evolution, DNA is the marker of choice.

The second step is to perform multiple sequence alignment. Obtaining accurate alignment is critical for phylogenetic tree construction. The unique aspect of multiple alignment for phylogenetic analysis is that it often requires manual truncation of ambiguously aligned regions. The next step is to select a proper substitution model that provides estimates of the true evolutionary event by taking into account multiple substitution events. Corrected evolutionary distances are used both in distance-based and likelihood-based tree-building methods. The commonly used nucleotide substitution models are the Jukes–Cantor and Kimura models. The commonly used amino acid substitution models are the PAM and JTT models. Other adjustments to improve the estimation of true evolutionary distances include the incorporation of rate heterogeneity among sites.

FURTHER READING

Graur, D., and Li, W.-H. 2000. *Fundamentals of Molecular Evolution.* Sunderland, MA: Sinauer Associates.

Higgins, D. G. 2000. Amino acid-based phylogeny and alignment. *Adv. Protein Chem.* 54:99–135.

Nei, M., and Kumar, S. 2000. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Salemi, M., and Vandamme, A.-M. 2003. *The Phylogenetics Handbook – A Practical Approach to DNA and Protein Phylogeny*. Cambridge, UK: Cambridge University Press.

Thornton, J. W., and DeSalle, R. 2000. Gene family evolution and homology: Genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet.* 1:41–73.

Whelan, S., Lio, P., and Goldman, N. 2001. Molecular phylogenetics: State of the art methods for looking into the past. *Trends Genet.* 17:262–72.

CHAPTER ELEVEN

Phylogenetic Tree Construction Methods and Programs

To continue discussion of molecular phylogenetics from Chapter 10, this chapter introduces the theory behind various phylogenetic tree construction methods along with the strategies used for executing the tree construction.

There are currently two main categories of tree-building methods, each having advantages and limitations. The first category is based on discrete characters, which are molecular sequences from individual taxa. The basic assumption is that characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved. Therefore, the character states of the common ancestor can be traced from this dataset. Another assumption is that each character evolves independently and is therefore treated as an individual evolutionary unit. The second category of phylogenetic methods is based on distance, which is the amount of dissimilarity between pairs of sequences, computed on the basis of sequence alignment. The distance-based methods assume that all sequences involved are homologous and that tree branches are additive, meaning that the distance between two taxa equals the sum of all branch lengths connecting them. More details on procedures and assumptions for each type of phylogenetic method are described.

DISTANCE-BASED METHODS

As mentioned in Chapter 10, true evolutionary distances between sequences can be calculated from observed distances after correction using a variety of evolutionary models. The computed evolutionary distances can be used to construct a matrix of distances between all individual pairs of taxa. Based on the pairwise distance scores in the matrix, a phylogenetic tree can be constructed for all the taxa involved. The algorithms for the distance-based tree-building method can be subdivided into either clustering based or optimality based. The clustering-type algorithms compute a tree based on a distance matrix starting from the most similar sequence pairs. These algorithms include an unweighted pair group method using arithmetic average (UPGMA) and neighbor joining. The optimality-based algorithms compare many alternative tree topologies and select one that has the best fit between estimated distances in the tree and the actual evolutionary distances. This category includes the Fitch–Margoliash and minimum evolution algorithms.

Clustering-Based Methods

Unweighted Pair Group Method Using Arithmetic Average

The simplest clustering method is UPGMA, which builds a tree by a sequential clustering method. Given a distance matrix, it starts by grouping two taxa with the smallest pairwise distance in the distance matrix. A node is placed at the midpoint or half distance between them. It then creates a reduced matrix by treating the new cluster as a single taxon. The distances between this new composite taxon and all remaining taxa are calculated to create a reduced matrix. The same grouping process is repeated and another newly reduced matrix is created. The iteration continues until all taxa are placed on the tree (see Box 11.1). The last taxon added is considered the outgroup producing a rooted tree.

The basic assumption of the UPGMA method is that all taxa evolve at a constant rate and that they are equally distant from the root, implying that a molecular clock (see Chapter 10) is in effect. However, real data rarely meet this assumption. Thus, UPGMA often produces erroneous tree topologies. However, owing to its fast speed of calculation, it has found extensive usage in clustering analysis of DNA microarray data (see Chapter 17).

Neighbor Joining

The UPGMA method uses unweighted distances and assumes that all taxa have constant evolutionary rates. Since this molecular clock assumption is often not met in biological sequences, to build a more accurate phylogenetic trees, the neighbor-joining (NJ) method can be used, which is somewhat similar to UPGMA in that it builds a tree by using stepwise reduced distance matrices. However, the NJ method does not assume the taxa to be equidistant from the root. It corrects for unequal evolutionary rates between sequences by using a conversion step. This conversion requires the calculations of "*r*-values" and "transformed *r*-values" using the following formula:

$$d'_{AB} = d_{AB} - 1/2 \times (r_A + r_B) \qquad \text{(Eq. 11.1)}$$

where d'_{AB} is the converted distance between A and B and d_{AB} is the actual evolutionary distance between A and B. The value of r_A (or r_B) is the sum of distances of A (or B) to all other taxa. A generalized expression of the *r*-value is r_i calculated based on the following formula:

$$r_i = \Sigma d_{ij} \qquad \text{(Eq. 11.2)}$$

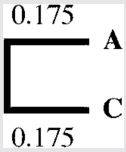
where *i* and *j* are two different taxa. The *r*-values are needed to create a modified distance matrix. The transformed *r*-values (r') are used to determine the distances of an individual taxon to the nearest node.

$$r'_i = r_i/n-2 \qquad \text{(Eq. 11.3)}$$

Box 11.1 An Example of Phylogenetic Tree Construction Using the UPGMA Method

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

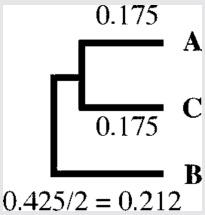
1. Using a distance matrix involving four taxa, A, B, C, and D, the UPGMA method first joins two closest taxa together which are A and C (**0.35** in grey). Because all taxa are equidistant from the node, the branch length for A to the node is AC/2 = 0.35/2 = 0.175.



2. Because A and C are joined into a cluster, they are treated as one new composite taxon, which is used to create a reduced matrix. The distance of A-C cluster to every other taxa is one half of a taxon to A and C, respectively. That means that the distance of B to A-C is (AB + BC)/2; and that of D to A-C is (AD + CD)/2.

	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70

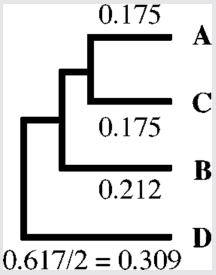
3. In the newly reduced-distance matrix, the smallest distance is between B and A-C (in grey), which allows the grouping of B and A-C to create a three-taxon cluster. The branch length for the B is one half of B to the A-C cluster.



4. When B and A-C are grouped and treated as a single taxon, this allows the matrix to reduce further into only two taxa, D and B-A-C. The distance of D to the composite taxon is the average of D to every single component which is (BD + AD + CD)/3.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

5. D is the last branch to add to the tree, whose branch length is one half of D to B-A-C.



6. Because distance trees allow branches to be additive, the resulting distances between taxa from the tree path can be used to create a distance matrix. Obviously, the estimated distances do not match the actual evolutionary distances shown, which illustrates the failure of UPGMA to precisely reflect the experimental observation.

	A	B	C
B	0.42		
C	0.35	0.42	
D	0.62	0.62	0.62

where n is the total number of taxa. For example, assuming A and B form a node called U, the distance A to U is determined by the following formula:

$$d_{AU} = [d_{AB} + (r'_A - r'_B)]/2 \qquad \text{(Eq. 11.4)}$$

An example of this distance conversion and NJ tree building is shown in Box 11.2. The tree construction process is somewhat opposite to that used UPGMA. Rather than building trees from the closest pair of branches and progressing to the entire tree, the NJ tree method begins with a completely unresolved star tree by joining all taxa onto a single node and progressively decomposes the tree by selecting pairs of taxa based on the above modified pairwise distances. This allows the taxa with the shortest corrected distances to be joined first as a node. After the first node is constructed, the newly created cluster reduces the matrix by one taxon and allows the next most closely related taxon to be joined next to the first node. The cycle is repeated until all internal nodes are resolved. This process is called *star decomposition*. Unlike UPGMA, NJ and most other phylogenetic methods produce unrooted trees. The outgroup has to be determined based on external knowledge (see Chapter 10).

Generalized Neighbor Joining

One of the disadvantages of the NJ method is that it generates only one tree and does not test other possible tree topologies. This can be problematic because, in many cases, in the initial step of NJ, there may be more than one equally close pair

Box 11.2 Phylogenetic Tree Construction Using the Neighbor Joining Method

	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

1. The NJ method is similar to UPGMA, but uses an evolutionary rate correction step before tree building. Using the same distance matrix as in the UPGMA tree building (see Box 11.1), the first step of the NJ method is r-value and r'-value calculation. According to Eq. 11.1 and 11.2, r and r' for each taxon are calculated as follows:

$$r_A = AB+AC+AD = 0.4+0.35+0.6 = 1.35$$
$$r'_A = r_A/(4-2) = 1.35/2 = 0.675$$
$$r_B = BA+BC+BD = 0.4+0.45+0.7 = 1.55$$
$$r'_B = r_B/(4-2) = 1.55/2 = 0.775$$
$$r_C = CA+CB+CD = 0.35+0.45+0.55 = 1.35$$
$$r'_C = r_C/(4-2) = 1.35/2 = 0.675$$
$$r_D = DA+DB+DC = 0.6+0.7+0.55 = 1.85$$
$$r'_D = r_D/(4-2) = 1.85/2 = 0.925$$

2. Based on Eq. 11.4 and the above r-values, the corrected distances are obtained as follows:

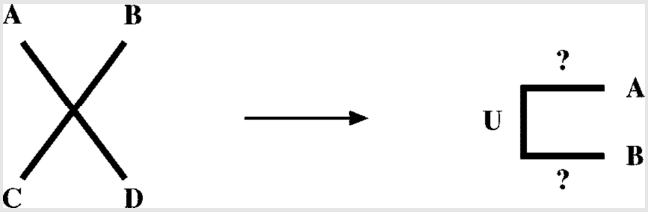
$$d'_{AB} = d_{AB} - 1/2 * (r_A+r_B) = 0.4 - (1.35+1.55)/2 = -1.05$$
$$d'_{AC} = d_{AC} - 1/2 * (r_A+r_C) = 0.35 - (1.35+1.35)/2 = -1$$
$$d'_{AD} = d_{AD} - 1/2 * (r_A+r_D) = 0.6 - (1.35+1.85)/2 = -1$$
$$d'_{BC} = d_{BC} - 1/2 * (r_B+r_C) = 0.45 - (1.55+1.35)/2 = -1$$
$$d'_{BD} = d_{BD} - 1/2 * (r_B+r_D) = 0.7 - (1.55+1.85)/2 = -1$$
$$d'_{CD} = d_{CD} - 1/2 * (r_C+r_D) = 0.55 - (1.35+1.85)/2 = -1.05$$

3. The rate-corrected distances allow the construction of a new distance matrix.

	A	B	C
B	-1.05		
C	-1	-1	
D	-1	-1	-1.05

4. Before tree construction, all possible nodes are collapsed into a star tree. The pair of taxa with the shortest distances in the new

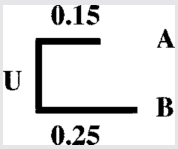
matrix are separated from the star tree first, according to the corrected distances. In this case, A and B as well as C and D are the shortest (−1.05, in grey). Therefore, the first node to be built can be either A-B or C-D. Choosing either pair first will give the same result. Let's choose A and B first and name the node U.



5. The branch lengths for A and B to the node U are calculated according to Eq. 11.4.

$$d_{AU} = [d_{AB} + (r'_A - r'_B)]/2 = [0.4 + (0.675 - 0.775)]/2 = 0.15$$

$$d_{BU} = [d_{AB} + (r'_B - r'_A)]/2 = [0.4 + (0.775 - 0.675)]/2 = 0.25$$



6. The new cluster allows the construction of a reduced matrix. This starts with actual distances. Unlike in UPGMA, the distance from a taxon to a node is the average of the original distances to each of the components of the composite taxon, subtracted from the inferred branch lengths.

$$d_{CU} = [(d_{AC} - d_{UA}) + (d_{BC} - d_{UB})]/2 = [(0.35 - 0.15) + (0.45 - 0.25)]/2 = 0.2$$

$$d_{DU} = [(d_{AD} - d_{UA}) + (d_{BD} - d_{UB})]/2 = [(0.6 - 0.15) + (0.7 - 0.25)]/2 = 0.45$$

	U	B
C	0.20	
D	0.45	0.55

7. Based on the reduced distance matrix, a new set of r- and r'-values are calculated.

$$r_C = CU + CD = 0.2 + 0.55 = 0.75$$

$$r'_C = r_C/(3-2) = 0.75/1 = 0.75$$

$$r_D = DU + CD = 0.45 + 0.55 = 1$$

$$r'_C = r_C/(3-2) = 1/1 = 1$$

$$r_U = CU + DU = 0.2 + 0.45 = 0.65$$

$$r'_U = r_U/(3-2) = 0.65/1 = 0.65$$

Box 11.2 (continued)

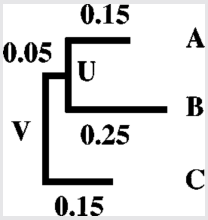
8. The new r- and r′-values allow construction of the corrected distance matrix.

$$d'_{CU} = d_{CU} - 1/2 * (r_C + r_U) = 0.2 - (0.75 + 0.65)/2 = -0.5$$
$$d'_{DU} = d_{DU} - 1/2 * (r_D + r_U) = 0.45 - (1 + 0.65)/2 = -0.375$$
$$d'_{CD} = d_{CD} - 1/2 * (r_C + r_D) = 0.55 - (0.75 + 1)/2 = -0.325$$

	U	B
C	-0.5	
D	-0.375	-0.325

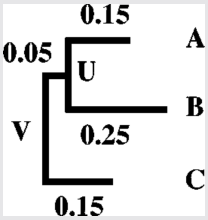
9. In the corrected distance matrix, C to node U has the shortest distance (−0.5, in grey). This allows creation of the second node named *V*. The branch length is calculated as in step 5.

$$d_{CV} = [d_{CU} + (r'_C - r'_U)]/2 = [0.2 + (0.75 - 0.65)]/2 = 0.15$$
$$d_{UV} = [d_{CU} + (r'_U - r'_C)]/2 = [0.2 + (0.65 - 0.75)]/2 = 0.05$$



10. Because D is the last branch to be decomposed from the star tree, there is no need to convert to r and r′ because r′ is infinitely large when $n - 2 = 0$. Its branch length is calculated as one half of the sum of D to node V and D to C, subtracted from respective branch lengths.

$$d_D = [(d_{DU} - d_{UV}) + (d_{DC} - d_{CV})]/2 = [(0.45 - 0.05) + (0.55 - 0.15)]/2 = 0.4$$



11. When the overall branch lengths are compiled into a distance matrix, which is used to compare with the original distance matrix, it is clear that the estimated distances completely match the actual evolutionary distances, indicating that this treeing method is able to satisfy the constraint of the experimental observation in this case.

	U	B
C	-0.5	
D	-0.375	-0.325

of neighbors to join, leading to multiple trees. Ignoring these multiple options may yield a suboptimal tree. To overcome the limitations, a generalized NJ method has been developed, in which multiple NJ trees with different initial taxon groupings are generated. A best tree is then selected from a pool of regular NJ trees that best fit the actual evolutionary distances. This more extensive tree search means that this approach has a better chance of finding the correct tree.

Optimality-Based Methods

The clustering-based methods produce a single tree as output. However, there is no criterion in judging how this tree is compared to other alternative trees. In contrast, optimality-based methods have a well-defined algorithm to compare all possible tree topologies and select a tree that best fits the actual evolutionary distance matrix. Based on the differences in optimality criteria, there are two types of algorithms, Fitch–Margoliash and minimum evolution, that are described next. The exhaustive search for an optimal tree necessitates a slow computation, which is a clear drawback especially when the dataset is large.

Fitch–Margoliash

The Fitch–Margoliash (FM) method selects a best tree among all possible trees based on minimal deviation between the distances calculated in the overall branches in the tree and the distances in the original dataset. It starts by randomly clustering two taxa in a node and creating three equations to describe the distances, and then solving the three algebraic equations for unknown branch lengths. The clustering of the two taxa helps to create a newly reduced matrix. This process is iterated until a tree is completely resolved. The method searches for all tree topologies and selects the one that has the lowest squared deviation of actual distances and calculated tree branch lengths. The optimality criterion is expressed in the following formula:

$$E = \sum_{i=1}^{T-1} \sum_{j=j+1}^{T} \frac{(d_{ij} - p_{ij})^2}{{d_{ij}}^2} \qquad \text{(Eq. 11.5)}$$

where E is the error of the estimated tree fitting the original data, T is the number of taxa, d_{ij} is the pairwise distance between ith and jth taxa in the original dataset, and p_{ij} is the corresponding tree branch length.

Minimum Evolution

Minimum evolution (ME) constructs a tree with a similar procedure, but uses a different optimality criterion that finds a tree among all possible trees with a minimum overall branch length. The optimality criterion relies on the formula:

$$S = \Sigma b_i \qquad \text{(Eq. 11.6)}$$

where b_i is the ith branch length. Searching for the minimum total branch length is an indirect approach to achieving the best fit of the branch lengths with the original

dataset. Analysis has shown that minimum evolution in fact slightly outperforms the least square-based FM method.

Pros and Cons

The most frequently used distance methods are clustering based. The major advantage is that they are computationally fast and are therefore capable of handling datasets that are deemed to be too large for any other phylogenetic method. The methods, however, are not guaranteed to find the best tree. Exhaustive tree-searching algorithms such as FM and ME have better accuracies overall. However, they can be computationally prohibitive to use when the number of taxa is large (e.g., >12), because the overall number of tree topologies becomes too large to handle. A compromise between the two types of algorithm is a hybrid approach such as the generalized NJ, with a performance similar to that of ME but computationally much faster.

The overall advantage of all distance-based methods is the ability to make use of a large number of substitution models to correct distances. The drawback is that the actual sequence information is lost when all the sequence variation is reduced to a single value. Hence, ancestral sequences at internal nodes cannot be inferred.

CHARACTER-BASED METHODS

Character-based methods (also called *discrete methods*) are based directly on the sequence characters rather than on pairwise distances. They count mutational events accumulated on the sequences and may therefore avoid the loss of information when characters are converted to distances. This preservation of character information means that evolutionary dynamics of each character can be studied. Ancestral sequences can also be inferred. The two most popular character-based approaches are the maximum parsimony (MP) and maximum likelihood (ML) methods.

Maximum Parsimony

The parsimony method chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths. It is based on a principle related to a medieval philosophy called *Occam's razor*. The theory was formulated by William of Occam in the thirteenth century and states that the simplest explanation is probably the correct one. This is because the simplest explanation requires the fewest assumptions and the fewest leaps of logic. In dealing with problems that may have an infinite number of possible solutions, choosing the simplest model may help to "shave off" those variables that are not really necessary to explain the phenomenon. By doing this, model development may become easier, and there may be less chance of introducing inconsistencies, ambiguities, and redundancies, hence, the name Occam's razor.

For phylogenetic analysis, parsimony seems a good assumption. By this principle, a tree with the least number of substitutions is probably the best to explain the differences among the taxa under study. This view is justified by the fact that evolutionary

taxa \ sites	1	2	3	4	5	6	7	8
I	A	A	T	T	A	G	C	T
II	G	G	T	C	G	T	A	G
III	A	A	T	G	C	G	C	T
IV	A	G	T	A	A	G	C	A
V	A	C	T	T	C	G	C	G
VI	A	C	A	T	G	G	C	A

Figure 11.1: Example of identification of informative sites that are used in parsimony analysis. Sites 2, 5, and 8 (*grey boxes*) are informative sites. Other sites are noninformative sites, which are either constant or having characters occurring only once.

changes are relatively rare within a reasonably short time frame. This implies that a tree with minimal changes is likely to be a good estimate of the true tree. By minimizing the changes, the method minimizes the phylogenetic noise owing to homoplasy and independent evolution. The MP approach is in principle similar to the ME approach albeit the latter is distance based instead of character based.

How Does MP Tree Building Work?

Parsimony tree building works by searching for all possible tree topologies and reconstructing ancestral sequences that require the minimum number of changes to evolve to the current sequences. To save computing time, only a small number of sites that have the richest phylogenetic information are used in tree determination. These sites are the so-called informative sites, which are defined as sites that have at least two different kinds of characters, each occurring at least twice (Fig. 11.1). Informative sites are the ones that can often be explained by a unique tree topology. Other sites are *noninformative*, which are constant sites or sites that have changes occurring only once. *Constant sites* have the same state in all taxa and are obviously useless in evaluating the various topologies. The sites that have changes occurring only once are not very useful either for constructing parsimony trees because they can be explained by multiple tree topologies. The noninformative sites are thus discarded in parsimony tree construction.

Once the informative sites are identified and the noninformative sites discarded, the minimum number of substitutions at each informative site is computed for a given tree topology. The total number of changes at all informative sites are summed up for each possible tree topology. The tree that has the smallest number of changes is chosen as the best tree.

The key to counting a minimum number of substitutions for a particular site is to determine the ancestral character states at internal nodes. Because these ancestral character states are not known directly, multiple possible solutions may exist. In this case, the parsimony principle applies to choose the character states that result in a minimum number of substitutions. The inference of an ancestral sequence is made by first going from the leaves to internal nodes and to the common root to determine all possible ancestral character states and then going back from the common root to the leaves to assign ancestral sequences that require the minimum number of substitutions. An example of predicting ancestral sequences at internal nodes is given

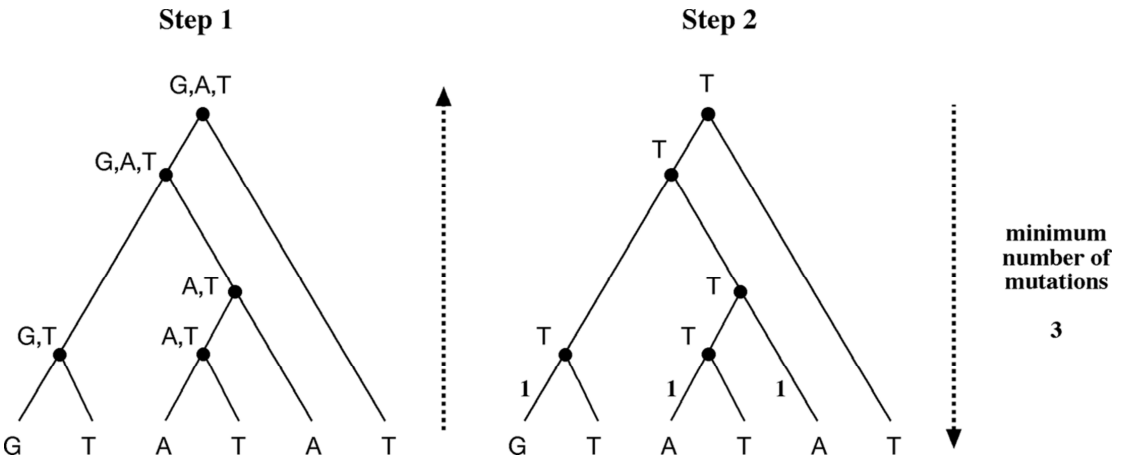


Figure 11.2: Using parsimony to infer ancestral characters at internal nodes involves a two-step procedure. The first step involves going from the leaves to the root and counting all possible ancestral characters at the internal nodes. The second step goes from the root to the leaves and assigns ancestral characters that involve minimum number of mutations. In this example, the total number of mutations is three if T is at the root, whereas other possible character states increase that number.

in Figure 11.2. It needs to be emphasized that, in reality, the ancestral node sequence cannot always be determined unambiguously. Sometimes, there may be more than one character that gives a total minimum number for a given tree topology. It is also possible that there may be two or more topologies that have the same minimum number of total substitutions. In that case, equally parsimonious trees are produced. A consensus tree has to be built that represents all the parsimonious trees (see Chapter 10).

Weighted Parsimony

The parsimony method discussed is unweighted because it treats all mutations as equivalent. This may be an oversimplification; mutations of some sites are known to occur less frequently than others, for example, transversions versus transitions, functionally important sites versus neutral sites. Therefore, a weighting scheme that takes into account the different kinds of mutations helps to select tree topologies more accurately. The MP method that incorporates a weighting scheme is called *weighted parsimony*. In the example shown in Figure 11.3, different branch lengths are obtained using weighted parsimony compared with using unweighted parsimony. In some cases, the weighting scheme may result in different tree topologies.

Tree-Searching Methods

As mentioned, the parsimony method examines all possible tree topologies to find the maximally parsimonious tree. This is an exhaustive search method. It starts by building a three taxa unrooted tree, for which only one topology is available. The choice of the first three taxa can be random. The next step is to add a fourth taxon to the existing branches, producing three possible topologies. The remaining taxa are progressively added to form all possible tree topologies (Fig. 11.4). Obviously, this brute-force approach only works if there are relatively few sequences. The exponential increase in possible tree topologies with the number of taxa means that this exhaustive

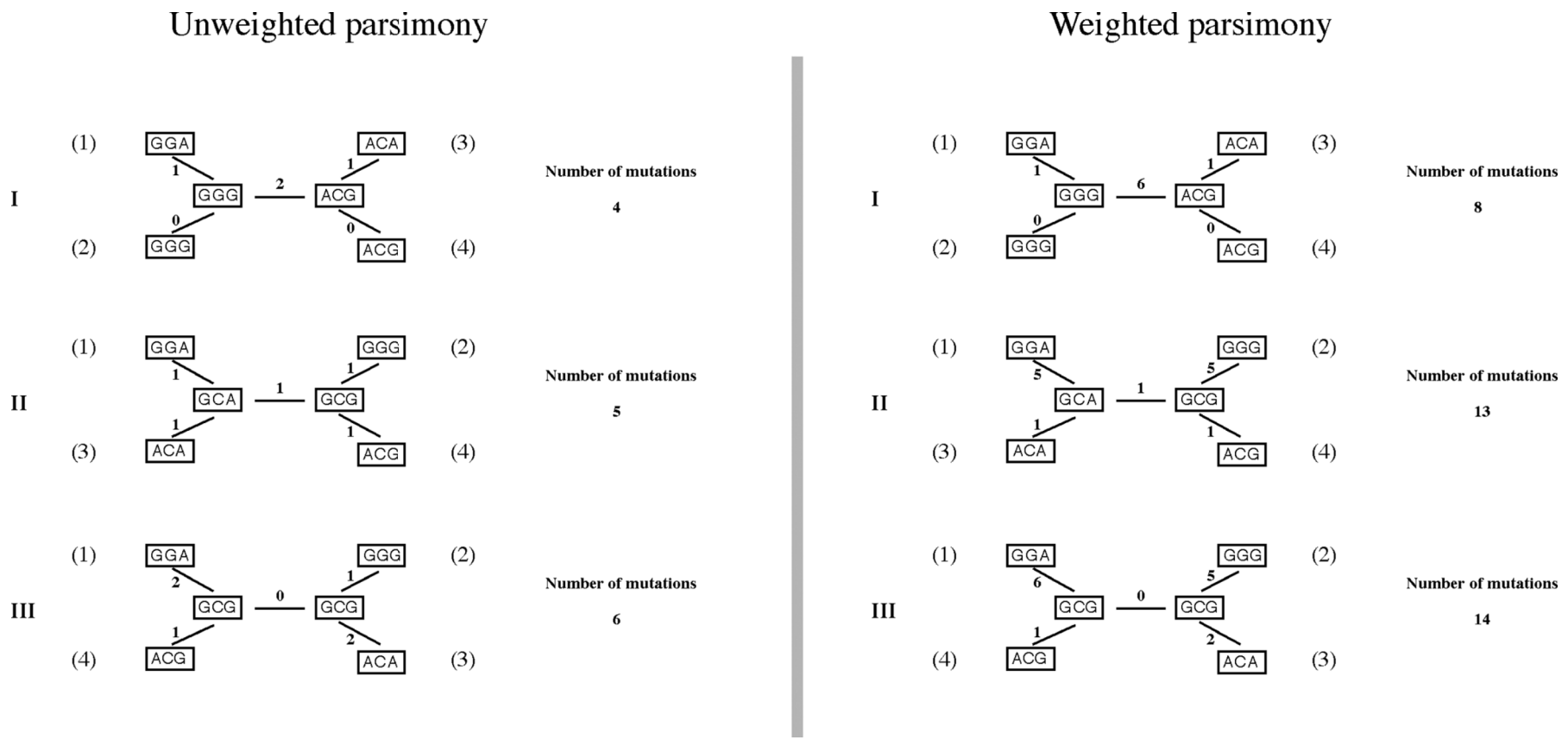


Figure 11.3: Comparison of unweighted and weighted parsimony. In the latter, transitions are weighted as 1 and transversions are weighted as 5.

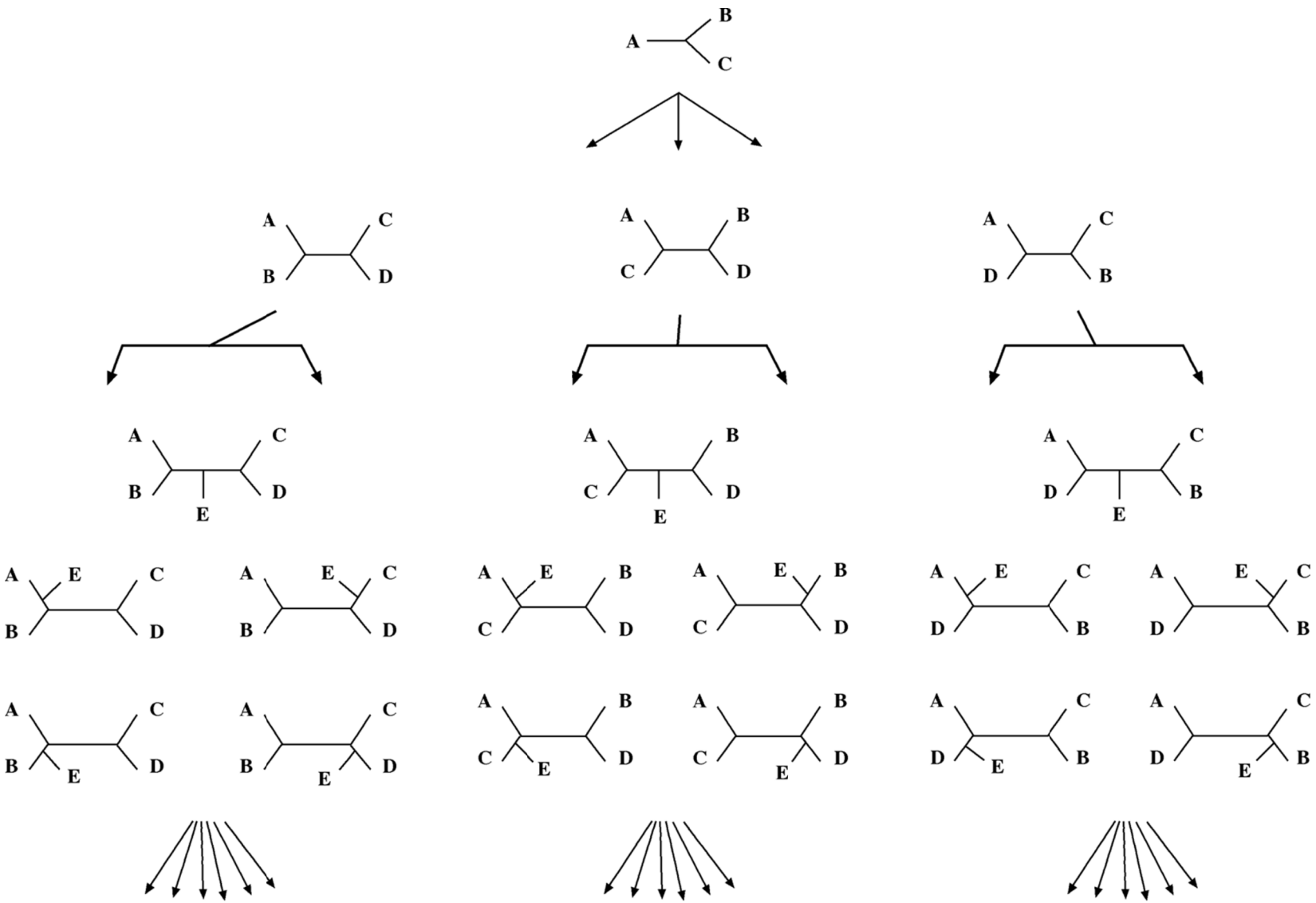


Figure 11.4: Schematic of exhaustive tree construction in the MP procedure. The tree starts with three taxa with one topology. One taxon is then added at a time in an progressive manner, during which the total branch lengths of all possible topologies are calculated.

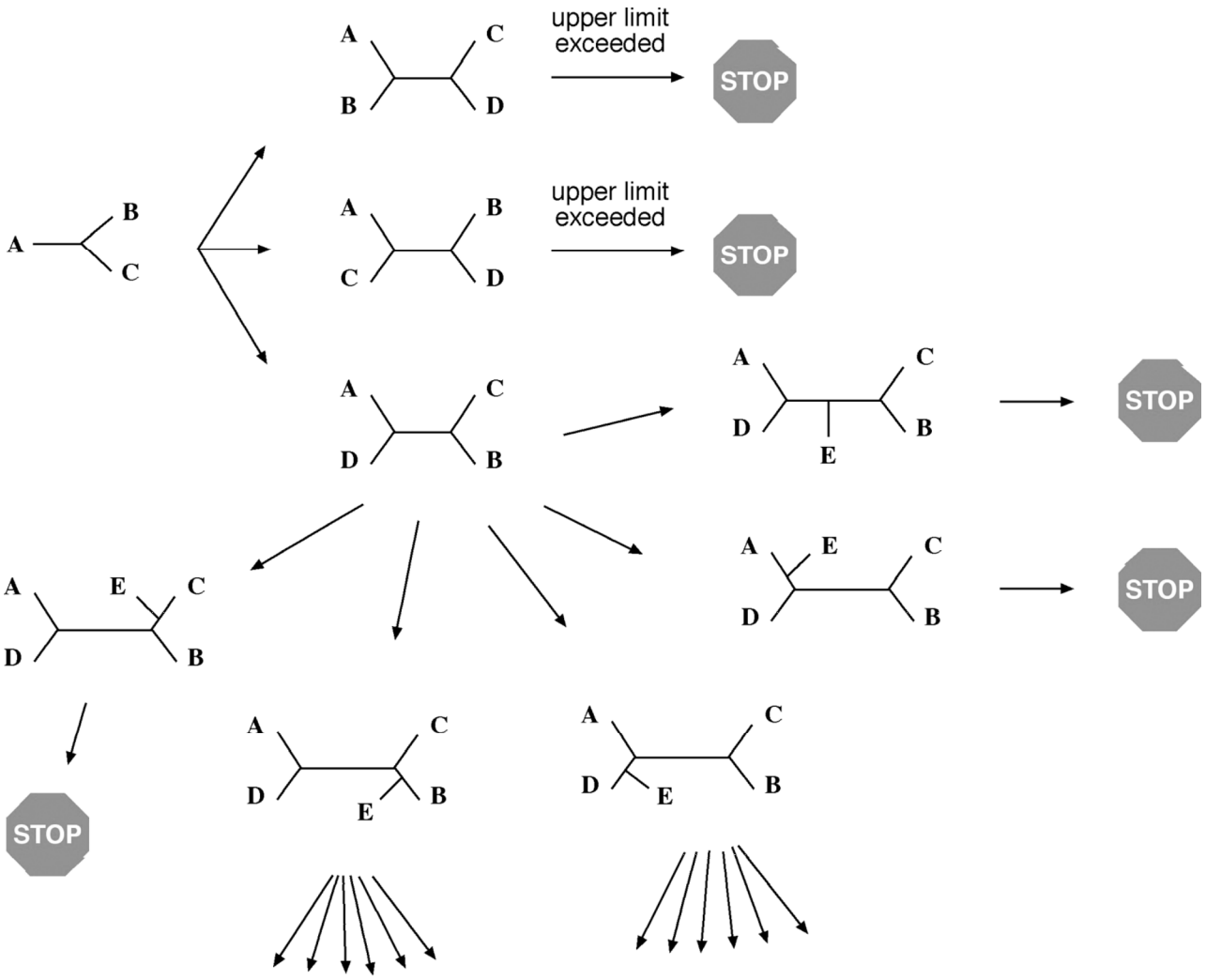


Figure 11.5: Schematic illustration of the branch-and-bound algorithm. Tree building starts with a stepwise addition of taxa of all possible topologies. Whenever the total branch length for a given topology exceeds the upper bound, the tree search in that direction stops, thereby reducing the total computing time.

method is computationally too demanding to use when the number of taxa is more than ten. When this is the case, some simplified steps have to be introduced to reduce the complexity of the search.

One such simplified method is called branch-and-bound, which uses a shortcut to find an MP tree by establishing an upper limit (or upper bound) for the number of allowed sequence variations. It starts by building a distance tree for all taxa involved using either NJ or UPGMA and then computing the minimum number of substitutions for this tree. The resulting number defines the upper bound to which any other trees are compared. The rationale is that a maximally parsimonious tree must be equal to or shorter than the distance-based tree.

The branch-and-bound method starts building trees in a similar way as in the exhaustive method. The difference is that the previously established upper bound limits the tree growth. Whenever the overall tree length at every single stage exceeds the upper bound, the topology search toward a particular direction aborts (Fig. 11.5).

By doing so, it dramatically reduces the number of trees considered hence the computing time while at the same time guaranteeing to find the most parsimonious tree.

When the number of taxa exceeds twenty, even the branch-and-bound method becomes computationally unfeasible. A more heuristic search method must be used. As a reminder, a computer heuristic procedure is an approximation strategy to find an empirical solution for a complicated problem (see Chapter 4). This strategy generates quick answers, but not necessarily the best answer. In a heuristic tree search, only a small subset of all possible trees is examined. This method starts by carrying out a quick initial approximation, which is to build an NJ tree and subsequently modifying it slightly into a different topology to see whether that leads to a shorter tree.

The modification includes cutting a branch or subtree and regrafting it to other parts of the tree (Fig. 11.6). The total branch length for the new tree is recomputed. If the tree is found to be shorter through rearrangement, it is used as a starting point for another round of rearrangement. The iteration continues until no shorter trees are found. This method is very fast, but does not guarantee to find the most parsimonious tree. The commonly used branch-swapping algorithms are nearest neighbor interchange, tree bisection and reconnection, and subtree pruning and regrafting.

The pitfall with branch swapping is that the tree rearrangement tends to focus on a local area and stalls when a local branch length minimum is reached. To avoid getting stuck in a local minimum, a "global search" option is implemented in certain programs. This allows the removal of every possible subtree and its reattachment in every possible way, to increase the chance of finding the most parsimonious tree. This approach significantly increases the computing time and thus compromises the trade-off between obtaining an optimal tree and obtaining a tree within a realistic time.

Pros and Cons

The main advantage of MP is that it is intuitive – its assumptions are easily understood. In addition, the character-based method is able to provide evolutionary information about the sequence characters, such as information regarding homoplasy and ancestral states. It tends to produce more accurate trees than the distance-based methods when sequence divergence is low because this is the circumstance when the parsimony assumption of rarity in evolutionary changes holds true. However, when sequence divergence is high, or the amount of homoplasies is large, tree estimation by MP can be less effective, because the original parsimony assumption no longer holds. Estimation of branch lengths may also be erroneous because MP does not employ substitution models to correct for multiple substitutions. This drawback can become prominent when dealing with divergent sequences. In addition, MP only considers informative sites, and ignores other sites. Consequently, certain phylogenetic signals may be lost. MP is also slow compared to the distance methods, and more important, is very sensitive to the "long-branch attraction" (LBA) artifacts.

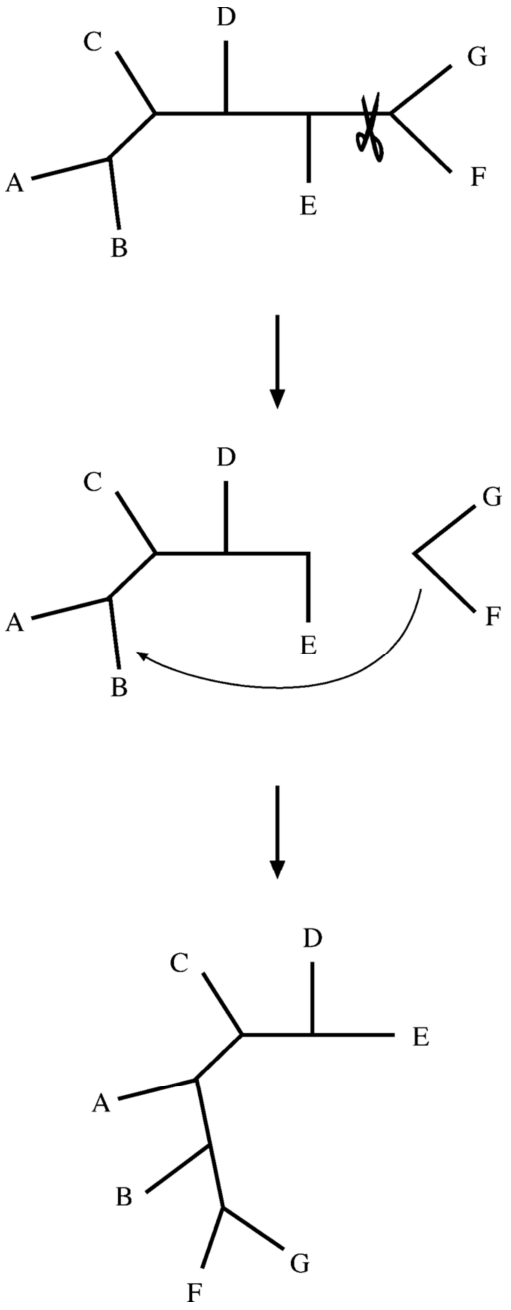


Figure 11.6: Schematic representation of a typical branch swapping process in which a branch is cut and moved to another part of the tree, generating a new topology.

Long-Branch Attraction

LBA is a particular problem associated with parsimony methods. It refers to a phylogenetic artifact in which rapidly evolving taxa with long branches are placed together in a tree, regardless of their true positions in a tree (Fig. 11.7). This is partly due to the assumption in parsimony that all lineages evolve at the same rate and that all mutations (transitions versus transversions) contribute equally to branch lengths. It may also be partly owing to multiple substitutions at individual sites and among-site rate heterogeneity for which MP is not capable of correcting.

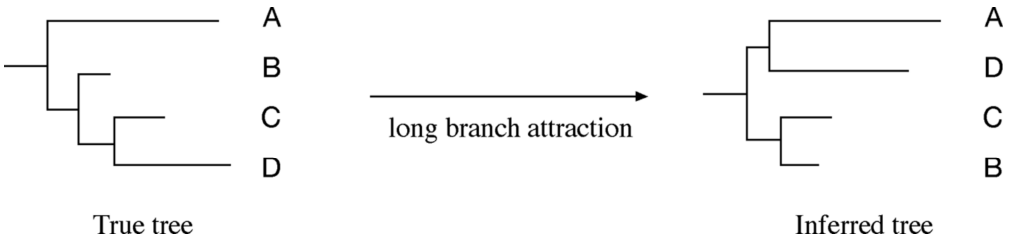


Figure 11.7: The LBA artifact showing taxa A and D are artifactually clustered during phylogenetic construction.

There are several possible solutions to the LBA artifact. For homoplasies that cause LBA, distance and likelihood (discussed below) methods that employ substitution models and rate heterogeneity models should be able to alleviate the problem. In addition, weighted parsimony should be more advantageous than unweighted parsimony in countering the transitional bias when transitions occur more often than transversions. Increasing the taxon sampling size may also help because introduction of intermediate taxa breaks up the long branches. A dataset with concatenated multiple genes also has less chance of LBA because the combined gene analysis may dampen the effect of a single gene having a high rate of evolution.

Maximum Likelihood Method

Another character-based approach is ML, which uses probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data. It finds a tree that most likely reflects the actual evolutionary process. ML is an exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites. By employing a particular substitution model that has probability values of residue substitutions, ML calculates the total likelihood of ancestral sequences evolving to internal nodes and eventually to existing sequences. It sometimes also incorporates parameters that account for rate variations across sites.

How Does the Maximum Likelihood Method Work?

ML works by calculating the probability of a given evolutionary path for a particular extant sequence. The probability values are determined by a substitution model (either for nucleotides or amino acids). For example, for DNA sequences using the Jukes–Cantor model, the probability (P) that a nucleotide remains the same after time t is:

$$P(t) = 1/4 + 3/4e^{-\alpha t} \qquad \text{(Eq. 11.7)}$$

where α is the nucleotide substitution rate in the Jukes–Cantor model, which is either empirically assigned or estimated from the raw datasets. In Figure 11.8, the elapsed time t from X to A can be assigned as 1 and from Z to A as 2. For a nucleotide to change into a different residue after time t, the probability value is determined by the

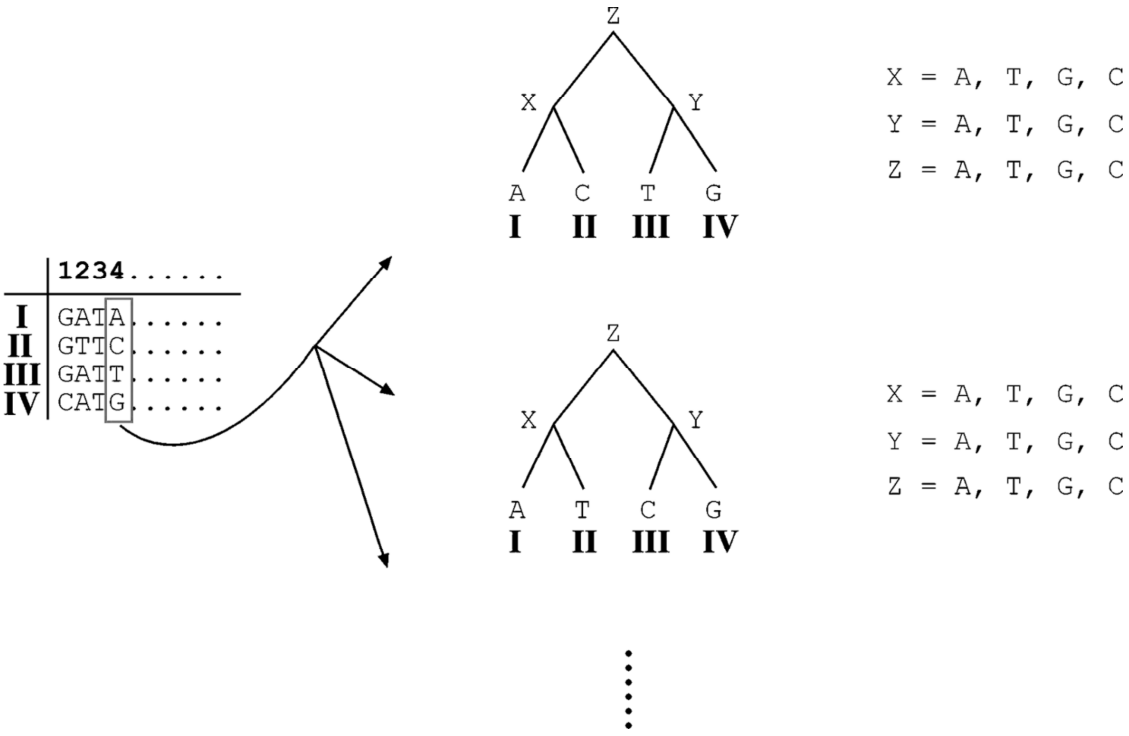


Figure 11.8: Schematic representation of the ML approach to build phylogenetic trees for four taxa, I, II, III, and IV. The ancestral character states at the internal nodes and root node are assigned X, Y, and Z, respectively. The example only shows some of the topologies derived from one of the sites in the original alignment. The method actually uses all the sites in probability calculation for all possible trees with all combinations of possible ancestral sequences at internal nodes according to a predefined substitution model.

following formula:

$$P(t) = 1/4 - 1/4e^{-\alpha t} \qquad \text{(Eq. 11.8)}$$

For other substitution models, the formulas are much more complex and are not described here. For a particular site, the probability of a tree path is the product of the probability from the root to all the tips, including every intermediate branches in the tree topology. Because multiplication often results in very small values, it is computationally more convenient to express all probability values as natural log likelihood (lnL) values, which also converts multiplication into summation. Because ancestral characters at internal nodes are normally unknown, all possible scenarios of ancestral states (X, Y, and Z in Fig. 11.8) have to be computed.

After logarithmic conversion, the likelihood score for the topology is the sum of log likelihood of every single branch of the tree. After computing for all possible tree paths with different combinations of ancestral sequences, the tree path having the highest likelihood score is the final topology at the site. Because all characters are assumed to have evolved independently, the log likelihood scores are calculated for each site independently. The overall log likelihood score for a given tree path for the entire sequence is the sum of log likelihood of all individual sites. The same procedure has to be repeated for all other possible tree topologies. The tree having the highest

likelihood score among all others is chosen as the best tree, which is the ML tree. This process is exhaustive in nature and therefore very time consuming.

$$\begin{aligned} L_{(4)} &= \Pr(Z \to X) * \Pr(Z \to Y) * \Pr(X \to A) * \Pr(X \to C) * \Pr(Y \to T) * \Pr(Y \to G) \\ \ln L_{(4)} &= \ln \Pr(Z \to X) + \ln \Pr(Z \to Y) + \ln \Pr(X \to A) + \ln \Pr(X \to C) \\ &\quad + \ln \Pr(Y \to T) + \ln \Pr(Y \to G) \end{aligned}$$

Pros and Cons

ML is based on well-founded statistics instead of a medieval philosophy. It is thus considered mathematically more rigorous than MP. In fact, it is the most rigorous among all approaches. ML uses the full sequence information, not just the informative sites and therefore may be more robust. ML employs substitution models and is not sensitive to LBA. Some of these strengths, however, can also be the weakness of ML depending on the context. For example, accuracy depends on the substitution model used. Choosing an unrealistic substitution model may lead to an incorrect tree. Because of the exhaustive nature of the ML method, when the number of taxa increases to a modest size, it becomes impossible to use. To overcome the problem, several heuristic or alternative approaches have been proposed. These alternative methods include quartet puzzling, genetic algorithms (GAs), and Bayesian inference, which are introduced in the following sections.

Quartet Puzzling

The most commonly used heuristic ML method is called quartet puzzling, which uses a divide-and-conquer approach. In this approach, the total number of taxa are divided into many subsets of four taxa known as *quartets*. An optimal ML tree is constructed from each of these quartets. This is a relatively easy process as there are only three possible unrooted topologies for a four-taxon tree. All the quartet trees are subsequently combined into a larger tree involving all taxa (Fig. 11.9). This process is like joining pieces in a jigsaw puzzle, hence the name. The problem in drawing a consensus is that the branching patterns in quartets with shared taxa may not agree. In this case, a majority rule is used to determine the positions of branches to be inserted to create the consensus tree.

The reason that quartet puzzling is computationally faster than exhaustive ML is because there are fewer tree topologies to search. To take four-taxon subsets out of n sequences, there are total C_n^4 combinations. Each subset has only three possible trees, and so the total number of trees that need to be computed are $3 \times C_n^4$. For instance, for twenty taxa, there are $3 \times C_{20}^4 = \frac{3 \times 20!}{(20-4)! \times 4!} = 14{,}535$ tree topologies to search, compared with 2×10^{20} trees if using the exhaustive search strategy. Thus, the method significantly reduces the computing time. The caveat of using the puzzling approach is that it does not necessarily return a tree with ML, but instead produces a consensus tree that is supported by the results of most quartets. Although the heuristic method is not as robust as regular ML, it has become a popular choice with many researchers because of its computational feasibility with large datasets.

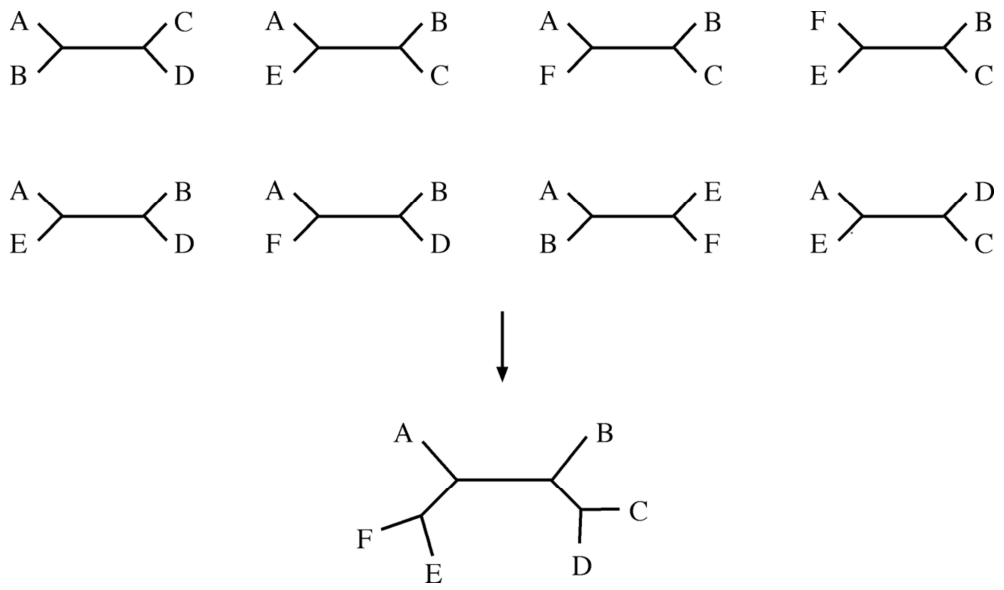


Figure 11.9: Schematic illustration of quartet puzzling in deriving a consensus tree by combining multiple quartet trees.

NJML

NJML is a hybrid algorithm combining aspects of NJ and ML. It constructs an initial tree using the NJ method with bootstrapping (which will be described). The branches with low bootstrap support are collapsed to produce multifurcating branches. The polytomy is resolved using the ML method. Although the performance of this method is not yet as good as the complete ML method, it is at least ten times faster.

Genetic Algorithm

A recent addition to fast ML search methods is the GA, a computational optimization strategy that uses biological terminology as a metaphor because the method involves "crossing" mathematical routines to generate new "offspring" routines. The algorithm works by selecting an optimal result through a mix-and-match process using a number of existing random solutions. A "fitness" measure is used to monitor the optimization process. By keeping record of the fitness scores, the process simulates the natural selection and genetic crossing processes. For instance, a subroutine that has the best score (best fit process) is selected in the first round and is used as a starting point for the next round of the optimization cycle. Again using biological metaphors, this is to generate more "offspring," which are mathematical trials with modifications from the previous ones. Different computational routines (or "chromosomes") are also allowed to combine (or "crossover") to produce a new solution. The iteration continues until an optimal solution is found.

When applying GA to phylogenetic inference, the method strongly resembles the pruning and regrafting routines used in the branch-swapping process. In GA-based tree searching, the fitness measure is the log likelihood scores. The tree search begins with a population of random trees with an arbitrary branch lengths. The tree with a highest log likelihood score is allowed to leave more "offspring" with "mutations"

on the tree topology. The mutational process is essentially branch rearrangement. Mutated new trees are scored. Those that are scored higher than the parent tree are allowed to mutate more to produce even higher scored offspring, if possible. This process is repeated until no higher scored trees can be found. The advantage of this algorithm is its speed; a near optimal tree can often be obtained within a limited number of iterations.

Bayesian Analysis

Another recent development of a speedy ML method is the use of the Bayesian analysis method. The essence of Bayesian analysis is to make inference on something unobserved based on existing observations. It makes use of an important concept of known as *posterior probability*, which is defined as the probability that is revised from prior expectations, after learning something new about the data. In mathematical terms, Bayesian analysis is to calculate posterior probability of two joint events by using the prior probability and conditional probability values using the following simplified formula:

$$\text{Posterior probability} = \frac{\text{Prior probability} * \text{Conditional likelihood}}{\text{Total probability}} \qquad \text{(Eq. 11.9)}$$

Without going into much mathematical detail, it is important to know that the Bayesian method can be used to infer phylogenetic trees with maximum posterior probability. In Bayesian tree selection, the prior probability is the probability for all possible topologies before analysis. The probability for each of these topologies is equal before tree building. The conditional probability is the substitution frequency of characters observed from the sequence alignment. These two pieces of information are used as a condition by the Bayesian algorithm to search for the most probable trees that best satisfy the observations.

The tree search incorporates an iterative random sampling strategy based on the Markov chain Monte Carlo (MCMC) procedure. MCMC is designed as a "hill-climbing" procedure, seeking higher and higher likelihood scores while searching for tree topologies, although occasionally it goes downhill because of the random nature of the search. Over time, high-scoring trees are sampled more often than low-scoring trees. When MCMC reaches high scored regions, a set of near optimal trees are selected to construct a consensus tree.

In the end, the Bayesian method can achieve the same or even better performance than the complete ML method, but is much faster than regular ML and is able to handle very large datasets. The reason that the Bayesian analysis may achieve better performance than ML is that the ML method searches one single best tree, whereas the Bayesian method searches a set of best trees. The advantage of the Bayesian method can be explained by the matter of probability. Because the true tree is not known, an optimal ML tree may have, say, 90% probability of representing the reality. However, the Bayesian method produces hundreds or thousands of optimal or near-optimal

trees with 88% to 90% probability to represent the reality. Thus, the latter approach has a better chance overall to guess the true tree correctly.

PHYLOGENETIC TREE EVALUATION

After phylogenetic tree construction, the next step is to statistically evaluate the reliability of the inferred phylogeny. There are two questions that need to be addressed. One is how reliable the tree or a portion of the tree is; and the second is whether this tree is significantly better than another tree. To answer the first question, we need to use analytical resampling strategies such as bootstrapping and jackknifing, which repeatedly resample data from the original dataset. For the second question, conventional statistical tests are needed.

What Is Bootstrapping?

Bootstrapping is a statistical technique that tests the sampling errors of a phylogenetic tree. It does so by repeatedly sampling trees through slightly perturbed datasets. By doing so, the robustness of the original tree can be assessed. The rationale for bootstrapping is that a newly constructed tree is possibly biased owing to incorrect alignment or chance fluctuations of distance measurements. To determine the robustness or reproducibility of the current tree, trees are repeatedly constructed with slightly perturbed alignments that have some random fluctuations introduced. A truly robust phylogenetic relationship should have enough characters to support the relationship even if the dataset is perturbed in such a way. Otherwise, the noise introduced in the resampling process is sufficient to generate different trees, indicating that the original topology may be derived from weak phylogenetic signals. Thus, this type of analysis gives an idea of the statistical confidence of the tree topology.

Parametric and Nonparametric Bootstrapping

Bootstrap resampling relies on perturbation of original sequence datasets. There are two perturbation strategies. One way to produce perturbations is through random replacement of sites. This is referred to as *nonparametric bootstrapping*. Alternatively, new datasets can be generated based on a particular sequence distribution, which is *parametric bootstrapping*. Both types of bootstrapping can be applied to the distance, parsimony, and likelihood tree construction methods.

In nonparametric bootstrapping, a new multiple sequence alignment of the same length is generated with random duplication of some of the sites (i.e., the columns in an alignment) at the expense of some other sites. In other words, certain sites are randomly replaced by other existing sites. Consequently, certain sites may appear multiple times, and other sites may not appear at all in the new alignment (Fig. 11.10). This process is repeated 100 to 1,000 times to create 100 to 1,000 new alignments that are used to reconstruct phylogenetic trees using the same method as the originally inferred tree. The new datasets with altered the nucleotide or amino acid composition

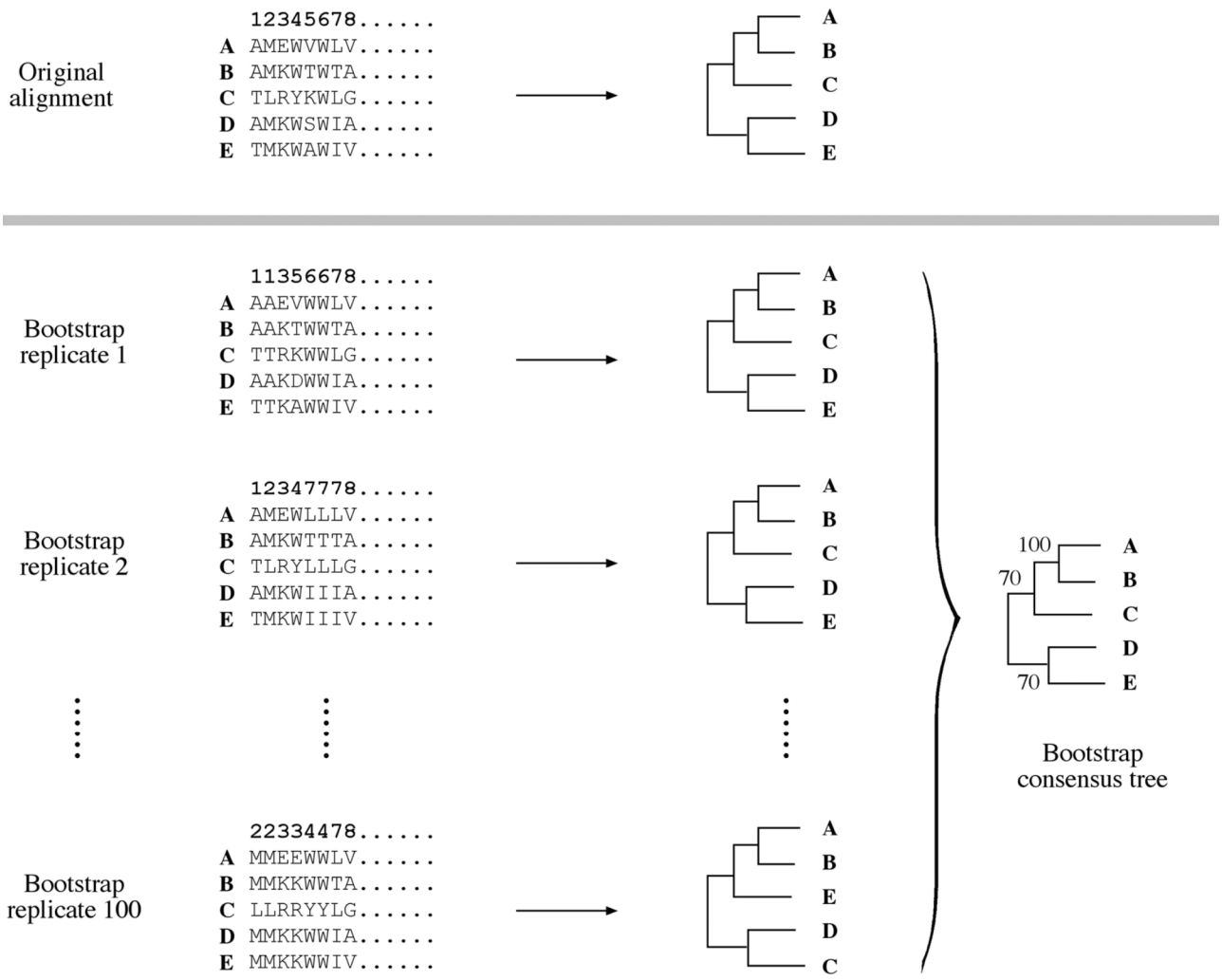


Figure 11.10: Schematic representation of a bootstrap analysis showing the original alignment and modified replicates in which certain sites are randomly replaced with other existing sites. The resulting altered replicates are used to building trees for statistical analysis at each node.

and rate heterogeneity may result in certain parts of the tree having a different topology from the original inferred tree.

All the bootstrapped trees are summarized into a consensus tree based on a majority rule. The most supported branching patterns shown at each node are labeled with bootstrap values, which are the percentage of appearance of a particular clade. Thus, the bootstrap test provides a measure for evaluating the confidence levels of the tree topology. Analysis has shown that a bootstrap value of 70% approximately corresponds to 95% statistical confidence, although the issue is still a subject of debate.

Instead of randomly duplicating sites to generate new datasets, parametric bootstrapping uses altered datasets with random sequences confined within a particular sequence distribution according to a given substitution model. For instance, for a nucleotide dataset, according to the Juke–Cantor model, all four nucleotides are identically distributed, whereas the Kimura model provides a different distribution (see Fig. 10.8). The parametric bootstrapping method may help avoid the problem of certain sites being repeated too many times as in nonparametric bootstrapping resulting

in skewed sequence distribution. If a correct nucleotide/amino acid distribution model is used, parametric bootstrapping generates more reasonable replicates than random replicates. Thus, this procedure is considered more robust than nonparametric bootstrapping.

Caveats

Strictly speaking, bootstrapping does not assess the accuracy of a tree, but only indicates consistency and stability of individual clades of the tree. This means that, because of systematic errors, wrong trees can still be obtained with high bootstrap values. Therefore, bootstrap results should be interpreted with caution. Unusually high GC content in the original dataset, unusually accelerated evolutionary rates and unrealistic evolutionary models are the potential causes for generating biased trees, as well as biased bootstrap estimates, which come after the tree generation.

In addition, from a statistical point of view, a large number of bootstrap resampling steps are needed to achieve meaningful results. It is generally recommended that a phylogenetic tree should be bootstrapped 500 to 1,000 times. However, this presents a practical dilemma. In many instances, it may take hours or days to construct one ML or MP tree. So the multiplication of computing time makes bootstrapping virtually impossible to use with limited computing resources.

Jackknifing

In addition to bootstrapping, another often used resampling technique is jackknifing. In jackknifing, one half of the sites in a dataset are randomly deleted, creating datasets half as long as the original. Each new dataset is subjected to phylogenetic tree construction using the same method as the original. The advantage of jackknifing is that sites are not duplicated relative to the original dataset and that computing time is much shortened because of shorter sequences. One criticism of this approach is that the size of datasets has been changed into one half and that the datasets are no longer considered replicates. Thus, the results may not be comparable with that from bootstrapping.

Bayesian Simulation

In terms of statistical evaluation, the Bayesian method is probably the most efficient; it does not require bootstrapping because the MCMC procedure itself involves thousands or millions of steps of resampling. As a result of Bayesian tree construction, posterior probabilities are assigned at each node of a best Bayesian tree as statistical support. Because of fast computational speed of MCMC tree searching, the Bayesian method offers a practical advantage over regular ML and makes the statistical evaluation of ML trees more feasible. Unlike bootstrap values, Bayesian probabilities are normally higher because most trees are sampled near a small number of optimal trees. Therefore, they have a different statistical meaning from bootstrap.

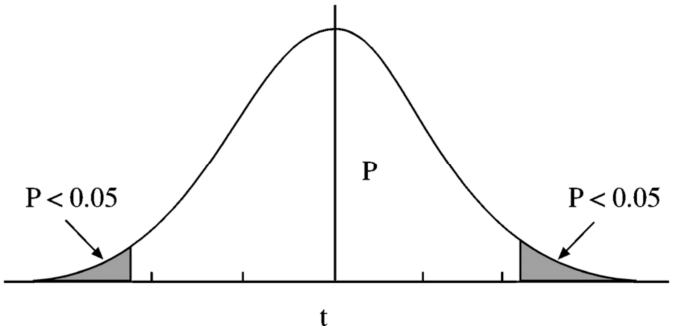


Figure 11.11: A *t*-distribution curve showing highlighted areas with margins of statistical significance.

Kishino–Hasegawa Test

In phylogenetic analysis, it is also important to test whether two competing tree topologies can be distinguished and whether one tree is significantly better than the other. The task is different from bootstrapping in that it tests the statistical significance of the entire phylogeny, not just portions of it. For that purpose, several statistical tests have been developed specifically for each of the three types of tree reconstruction methods, distance, parsimony, and likelihood. A test devised specifically for MP trees is called the Kishino–Hasegawa (KH) test.

The KH test sets out to test the null hypothesis that the two competing tree topologies are not significantly different. A paired Student *t*-test is used to assess whether the null hypothesis can be rejected at a statistically significant level. In this test, the difference of branch lengths at each informative site between the two trees is calculated. The standard deviation of the difference values can then be calculated. This in turn allows derivation of a *t*-value (see Eq. 11.10), which is used for evaluation against the *t*-distribution to see whether the value falls within the significant range (e.g., $P < .05$) to warrant the rejection of the null hypothesis (Fig. 11.11).

$$t = \frac{D_a - D_t}{S_D/\sqrt{n}} \qquad \text{(Eq. 11.10)}$$

$$df = n - 1 \qquad \text{(Eq. 11.11)}$$

where n is the number of informative sites, df is the degree of freedom, t is the test statistical value, D_a is the average site-to-site difference between the two trees, S_D is the standard deviation, and D_t is the total difference of branch lengths of the two trees.

Shimodaira–Hasegawa Test

A frequently used statistical test for ML trees is the Shimodaira–Hasegawa (SH) test (likelihood ratio test). It tests the goodness of fit of two competing trees using the χ^2 test. For this test, log likelihood scores of two competing trees have to be obtained first. The degree of freedom used for the analysis depends on the substitution model used. It relies on the following test formula:

$$d = 2(\ln L_A - \ln L_B) = 2\ln(L_A/L_B) \qquad \text{(Eq. 11.12)}$$

where d is the log likelihood ratio score and $\ln L_A$ and $\ln L_B$ are likelihood scores for tree A and tree B, respectively. The statistical meaning of d can be obtained from calculating the probability value from a χ^2 table.

Once the log ratio of the two scores is obtained, it is used to test against the χ^2 distribution. The resulting probability value (P-value) determines whether the difference between the two trees is significant.

PHYLOGENETIC PROGRAMS

Phylogenetic tree reconstruction is not a trivial task. Although there are numerous phylogenetic programs available, knowing the theoretical background, capabilities, and limitations of each is very important. For a list of hundreds of phylogenetic software programs, see Felsenstein's collection at: http://evolution.genetics.washington.edu/phylip/software.html. Most of these programs are freely available. Some are comprehensive packages; others are more specialized to perform a single task. Most require special efforts to learn how to use them effectively. Because this book is not intended as a computer manual, a brief introduction to several of the most commonly used programs is provided.

PAUP* (Phylogenetic analysis using parsimony and other methods, by David Swofford, http://paup.csit.fsu.edu/) is a commercial phylogenetic package. It is probably one of the most widely used phylogenetic programs available from Sinauer Publishers. It is a Macintosh program (UNIX version available in the GCG package) with a very user-friendly graphical interface. PAUP was originally developed as a parsimony program, but expanded to a comprehensive package that is capable of performing distance, parsimony, and likelihood analyses. The distance options include NJ, ME, FM, and UPGMA. For distance or ML analyses, PAUP has the option for detailed specifications of substitution models, base frequencies, and among site rate heterogeneity (γ-shape parameters, proportion of invariant sites). PAUP is also able to perform nonparametric bootstrapping, jackknifing, KH testing, and SH testing.

Phylip (Phylogenetic inference package; by Joe Felsenstein) is a free multiplatform comprehensive package containing thirty-five subprograms for performing distance, parsimony, and likelihood analysis, as well as bootstrapping for both nucleotide and amino acid sequences. It is command-line based, but relatively easy to use for each single program. The only problem is that to complete an analysis the user is required to move between different subprograms while keeping modifying names of the intermediate output files. The program package is downloadable from http://evolution.genetics.washington.edu/phylip.html. An online version is also available at http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html. A more user-friendly online version is WebPhylip available at http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/.

TREE-PUZZLE is a program performing quartet puzzling. The advantage is that it allows various substitution models for likelihood score estimation and incorporates a discrete γ model for rate heterogeneity among sites (see Chapter 10). Because of the heuristic nature of the program, it allows ML analyses of large datasets. The

resulting puzzle trees are automatically assigned puzzle support values to internal branches. These values are percentages of consistent quartet trees and do not have the same meaning as bootstrap values. TREE-PUZZLE version 5.0 is available for Mac, UNIX, and Windows and can be downloaded from www.tree-puzzle.de/. There is also an online version of the program available at: http://bioweb.pasteur.fr/seqanal/interfaces/Puzzle.html.

PHYML (http://atgc.lirmm.fr/phyml/) is a web-based phylogenetic program using the GA. It first builds an NJ tree and uses it as a starting tree for subsequent iterative refinement through subtree swapping. Branch lengths are simultaneously optimized during this process. The tree searching stops when the total ML score no longer increases. The main advantage of this program is the ability to build trees from very large datasets with hundreds of taxa and to complete tree searching within a relatively short time frame.

MrBayes is a Bayesian phylogenetic inference program. It randomly samples tree topologies using the MCMC procedure and infers the posterior distribution of tree topologies. It has a range of probabilistic models available to search for a set of trees with the highest posterior probability. It is fast and capable of handling large datasets. The program is available in multiplatform versions and can be downloaded from http://morphbank.ebc.uu.se/mrbayes/. A web program that also employs Bayesian inference for phylogenetic analysis is BAMBE (http://bioweb.pasteur.fr/seqanal/interfaces/bambe.html).

SUMMARY

Molecular phylogeny is a fundamental tool for understanding sequence evolution and relationships. The accuracy of the tree-building methods used for phylogenetic analysis depends on the assumption on which each the method is based. Understanding these assumptions is the first step toward efficient use of these methods. The second step is understanding how the methods actually work and what intrinsic limitations these methods have. The third step is choosing suitable phylogenetic method(s) that can give a reasonably correct picture of a phylogenetic tree.

The phylogenetic methods can be divided into distance-based and character-based methods. The distance methods include UPGMA, NJ, Fitch–Margoliash, and minimum evolution. The first two are clustering based, and are fast but not accurate; the latter two are optimality based and are accurate but not fast. Character-based approaches include the MP and ML methods. The principle of parsimony is easy to understand, but has its root in a medieval philosophy. It is slower compared to distance methods. To speed up the computation, branch-and-bound and heuristics tree searching strategies are used. The ML method is the slowest, but is based on a solid statistical foundation. To overcome the bottleneck of computation in ML, faster algorithms such as quartet puzzling, NJML, GA, and Bayesian analysis have been developed to make the method more feasible. To assess the reliability and robustness of

every clade in a phylogenetic tree, bootstrapping and jackknifing are used. The KH and SH tests distinguish the overall topology of two competing trees.

It is important to realize that phylogenetic tree reconstruction is not a trivial matter, but a complicated process that often requires careful thought. Accuracy, reliability, and computational speed are all major factors for consideration when choosing a particular phylogenetic method. It is also important to realize that none of the three phylogenetic reconstruction methods are guaranteed to find the correct tree. All three methods have the potential to produce erroneous trees. To minimize phylogenetic errors, it is recommended that at least two methods be used for any phylogenetic analysis to check the consistency of tree building results obtained. Because the theories behind each of the three methods are fundamentally different, agreement in conclusion by several of these methods provides a particularly strong support for a correct phylogenetic tree. In addition, it is recommended that different rate substitution models, weighting schemes, and resampling strategies with or without exclusion of specific taxa and/or sites be applied. The same analysis should be repeated on multiple genes or proteins as well as the concatenated datasets. If more than one fundamentally different methods provide the same prediction, the confidence in the prediction is higher.

FURTHER READING

Graur, D., and Li, W.-H. 2000. *Fundamentals of Molecular Evolution.* Sunderland, MA: Sinauer Associates.

Hall, B. G. 2001. *Phylogenetic Trees Made Easy. A How-to Manual for Molecular Biologists.* Sunderland, MA: Sinauer Associates.

Huelsenbeck J. P., Ronquist, F., Nielsen R., and Bollback, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*. 294:2310–14.

Nei, M., and Kumar, S. 2000. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Salemi, M., and Vandamme, A.-M. 2003. *The Phylogenetics Handbook – A Practical Approach to DNA and Protein Phylogeny*. Cambridge, UK: Cambridge University Press.

Swofford, D. L., Olsen, G. J., Waddel, P. J., and Hillis, D. M. 1996. "Phylogenetic inference." In *Molecular Systematics*. 2nd ed., edited by D. M. Hillis, C. Moritz, and B. K. Mable Sunderland, MA: Sinauer Associates.