CHAPTER 6

Phylogenetic Prediction

INTRODUCTION

A PHYLOGENETIC ANALYSIS OF A FAMILY of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution. The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree. The branching relationships on the inner part of the tree then reflect the degree to which different sequences are related. Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them. The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.

Phylogenetic analysis of nucleic acid and protein sequences is presently and will continue to be an important area of sequence analysis. In addition to analyzing changes that have occurred in the evolution of different organisms, the evolution of a family of sequences may be studied. On the basis of the analysis, sequences that are the most closely related can be identified by their occupying neighboring branches on a tree. When a gene family is found in an organism or group of organisms, phylogenetic relationships among the genes can help to predict which ones might have an equivalent function. These functional predictions can then be tested by genetic experiments. Phylogenetic analysis may also be used to follow the changes occurring in a rapidly changing species, such as a virus. Analysis of the types of changes within a population can reveal, for example, whether or not a particular gene is under selection (McDonald and Kreitman 1991; Comeron and Kreitman 1998; Nielsen and Yang 1998), an important source of information in applications like epidemiology.

Procedures for phylogenetic analysis are strongly linked to those for sequence alignment discussed in Chapters 3 and 4, and similar difficulties are encountered. Just as two very similar sequences can be easily aligned even by eye, a group of sequences that are very similar but with a small level of variation throughout can easily be organized into a tree. Conversely, as sequences become more and more different through evolutionary change, they can be much more difficult to align. A phylogenetic analysis of very different sequences is also difficult to do because there are so many possible evolutionary paths that could have been followed to produce the observed sequence variation. Because of the complexity of this problem, considerable expertise is required for difficult situations.

Phylogenetic analysis programs are widely available at little or no cost. A comprehensive list will not be given here since one has been published previously (Swofford et al. 1996). The main ones in use are PHYLIP (phylogenetic inference package) (Felsenstein 1989 1996) available from Dr. J. Felsenstein at http://evolution.genetics.washington.edu/phylip.html and PAUP (phylogenetic analysis using parsimony) available from Sinauer Associates, Sunderland, Massachusetts, http://www.lms.si.edu/PAUP/. Current versions of these programs provide the three main methods for phylogenetic analysis—parsimony, distance, and maximum likelihood methods (described below)—and also include many types of evolutionary models for sequence variation. Examples using these programs are given later in the chapter. Each program requires a particular type of input sequence format that is described below and in Chapter 2. Another program, MacClade, is useful for detailed analysis of the predictions made by PHYLIP, PAUP, and other phylogenetic programs and is also available from Sinauer (also see http://phylogeny.arizona.edu/macclade/macclade.html). MacClade, as the name suggests, runs on a Macintosh computer. PHYLIP and PAUP run on practically any machine, but the user interface for PAUP has been most developed for use on the Macintosh computer.

There are also several Web sites that provide information on phylogenetic relationships among organisms (Table 6.1). There are several excellent descriptions of phylogenetic

Table 6.1. *Phylogenetic relationships among organisms*

Site name	Address	Description	Reference
Entrez	http://www3.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html	taxonomically related structures or group of organisms	see Web page
RDP (Ribosomal database project)	http://www.cme.msu.edu/RDP/	ribosomal RNA-derived trees	Maidak et al. (1999)
Tree of life	http://phylogeny.arizona.edu/tree/phylogeny.html	information about phylogeny and biodiversity	Maddison and Maddison (1992)

analysis in which the methods are covered in considerable depth (Li and Graur 1991; Miyamoto and Cracraft 1991; Felsenstein 1996; Li and Gu 1996; Saitou 1996; Swofford et al. 1996; Li 1997).

RELATIONSHIP OF PHYLOGENETIC ANALYSIS TO SEQUENCE ALIGNMENT

When the sequences of two nucleic acid or protein molecules found in two different organisms are similar, they are likely to have been derived from a common ancestor sequence. Chapter 3 discusses sequence alignment methods used to determine sequence similarity. Chapter 4 discusses multiple sequence alignment methods that need to be applied to a set of related sequences before a phylogenetic analysis can be performed. Chapter 7 describes methods for searching through a database of sequences to locate sequences that are similar to a query sequence. A sequence alignment reveals which positions in the sequences were conserved and which diverged from a common ancestor sequence, as illustrated in Figure 6.1. When one is quite certain that two sequences share an evolutionary relationship, the sequences are referred to as being homologous.

The commonest method of multiple sequence alignment (the progressive alignment method, p. 152) first aligns the most closely related pair of sequences and then sequentially adds more distantly related sequences or sets of sequences to this initial alignment (see flowchart, p. 144). The alignment so obtained is influenced by the most alike sequences in the group and thus may not represent a reliable history of the evolutionary changes that have occurred. Other methods of multiple sequence alignment attempt to circumvent the influence of alike sequences (see Chapter 4, p. 157). Once a multiple sequence alignment has been obtained, each column is assumed to correspond to an individual site that has

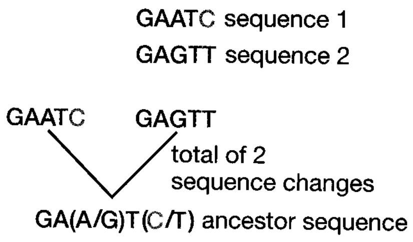


Figure 6.1. Origin of similar sequences. Sequences 1 and 2 are each assumed to be derived from a common ancestor sequence. Some of the ancestor sequence can be inferred from conserved positions in the two sequences. For positions that vary, there are two possible choices at these sites in the ancestor.

been evolving according to the observed sequence variation in the column. Most methods of phylogenetic analysis assume that each position in the protein or nucleic acid sequence changes independently of the others (analysis of RNA sequence evolution is an exception: see Chapter 5).

As indicated above, the analysis of sequences that are strongly similar along their entire lengths is quite straightforward. However, to align most sequences requires the positioning of gaps in the alignment. Gaps represent an insertion or deletion of one or more sequence characters during evolution. Proteins that align well are likely to have the same three-dimensional structure. In general, sequences that lie in the core structure of such proteins are not subject to insertions or deletions because any amino acid substitutions must fit into the packed hydrophobic environment of the core. Gaps should therefore be rare in regions of multiple sequence alignments that represent these core sequences. In contrast, more variation, including insertions and deletions, may be found in the loop regions on the outside of the three-dimensional structure because these regions do not influence the core structure as much. Loop regions interact with the environment of small molecules, membranes, and other proteins (see Chapter 9).

Gaps in alignments can be thought of as representing mutational changes in sequences, including insertions, deletions, or rearrangements of genetic material. The expectation that a gap of virtually any length can occur as a single event introduces the problem of judging how many individual changes have occurred and in what order. Gaps are treated in various ways by phylogenetic programs, but no clear-cut model as to how they should be treated has been devised. Many methods ignore gaps or focus on regions in an alignment that do not have any gaps. Nevertheless, gaps can be useful as phylogenetic markers in some situations.

Another approach for handling gaps is to avoid analysis of individual sites in the sequence alignment and instead to use sequence similarity scores as a basis for phylogenetic analysis. Rather than trying to decide what has happened at each sequence position in an alignment, a similarity score based on a scoring matrix with penalties for gaps is often used. As discussed below, these scores may be converted to distance scores that are suitable for phylogenetic analysis (Feng and Doolittle 1996) by distance methods (p. 254).

GENOME COMPLEXITY AND PHYLOGENETIC ANALYSIS

When performing a phylogenetic analysis, it is important to keep in mind that the genomes of most organisms have a complex origin. Some parts of the genome are passed on by vertical descent through the normal reproductive cycle. Other parts may have arisen by horizontal transfer of genetic material between species through a virus, DNA transformation, symbiosis, or some other horizontal transfer mechanism. Accordingly, when a particular gene is being subjected to phylogenetic analysis, the evolutionary history of that gene may not coincide with the evolutionary history of another.

One of the most significant uses of phylogenetic analysis of sequences is to make predictions concerning the tree of life. For this purpose, a gene should be selected that is universally present in all organisms and easily recognizable by the conservation of sequence in many species. At the same time, there should be enough sequence variation to determine which groups of organisms share the same phylogenetic origin. Ideally, the gene should also not be under selection, meaning that as variation occurs in populations of organisms, certain sequences are not favored with a loss of the more primitive variation.

Two molecules of this type that carry a great deal of evolutionary history in inter-species sequence variations are the small rRNA subunit and mitochondrial sequences. A large number of rRNA sequences from a variety of organisms were aligned and the secondary structure was deduced following methods discussed in Chapter 5. Phylogenetic predictions were then made using the distance method described below (Woese 1987). On the basis of rRNA sequence signatures, or regions within the molecule that are conserved in one group of organisms but different in another (Fig. 6.2), Woese (1987) predicted that early life diverged into three main kingdoms—Archaea, Bacteria, and Eukarya—a view that has been challenged (Mayr 1998). Evidence for the presence of additional organisms in these groups has since been found by PCR amplification of environmental samples of RNA (Barns et al. 1996). A more detailed analysis was used to find relationships among individual species within each group. The types of relationships found among the prokaryotic organisms are illustrated in Figure 6.3. The use of mitochondrial sequences for analysis of primate evolution is given below in the description of the parsimony method of phylogenetic analysis.

Although these studies of rRNA sequences suggest a quite clear-cut model for the evolution of life, phylogenetic analysis of other genes and gene families has revealed that the situation is probably more complex and that a more appropriate model might be the one shown in Figure 6.4. There are now many examples of horizontal or lateral transfer of genes between species (see Fig. 3.3, p. 55) that introduce new genes and sequences into an organism (Brown and Doolittle 1997; Doolittle 1999). These types of transfers are inferred from the finding that the phylogenetic histories of different genes in an organism, such as genes for metabolic functions, are not the same or that codon use in different genes varies (see Chapter 10). Another type of phylogenetic analysis is based on the number of genes shared between genomes and produces a tree that is similar to the rRNA tree (Snel et al. 1999).

To track the evolutionary history of genes, more attention has also been paid to the methodology of phylogenetic analysis and to the inherent errors in many of the assumptions (Doolittle 1999). Problems associated with variations between rates of change in different sites and of analyzing more distantly related sequences are discussed below. Moreover, there is evidence that genomes undergo extensive rearrangements, placing sequences of different evolutionary origin next to each other and even causing rearrangements within protein-encoding genes (Henikoff et al. 1997).

The different regions of independent evolutionary origin in a sequence therefore need to be identified. As discussed in Chapter 9, proteins are modular with functional domains, sometimes repeated within a protein and sometimes shared within a protein family. These regions are identified by their sharing of significant sequence similarity. The remainder of the aligned regions in the group may have variable levels of similarity. In nucleic acid sequences, a given sequence pattern may provide a binding site for a regulatory molecule, leading to promoter function, RNA splicing, or some other function. It may be difficult to decide the extent of these patterns for phylogenetic analysis; however, statistical approaches discussed in Chapter 4 may be used.

Another feature of genome evolution that should be considered in phylogenetic analysis is the occurrence of gene duplication events that create tandem copies of a gene. These two copies may then evolve along separate pathways leading to different functions. However, these copies maintain a certain level of similarity and undergo concerted evolution, a process of acquiring mutations in a coordinated way, probably through gene conversion or recombination events. Speciation events following gene duplications will give rise to two independent sets of genes and sequences, one set for each gene copy. As discussed in Chapter 3 and illustrated in Figure 3.3, two genes in the same lineage can have different relationships. In the example shown in Figure 3.3, genes a1 and a2 have been derived from gene a. The pair is then segregated by speciation such that there is one a1 a2 pair in one species evolving along one path and a second a1 a2 pair in a second species evolving along

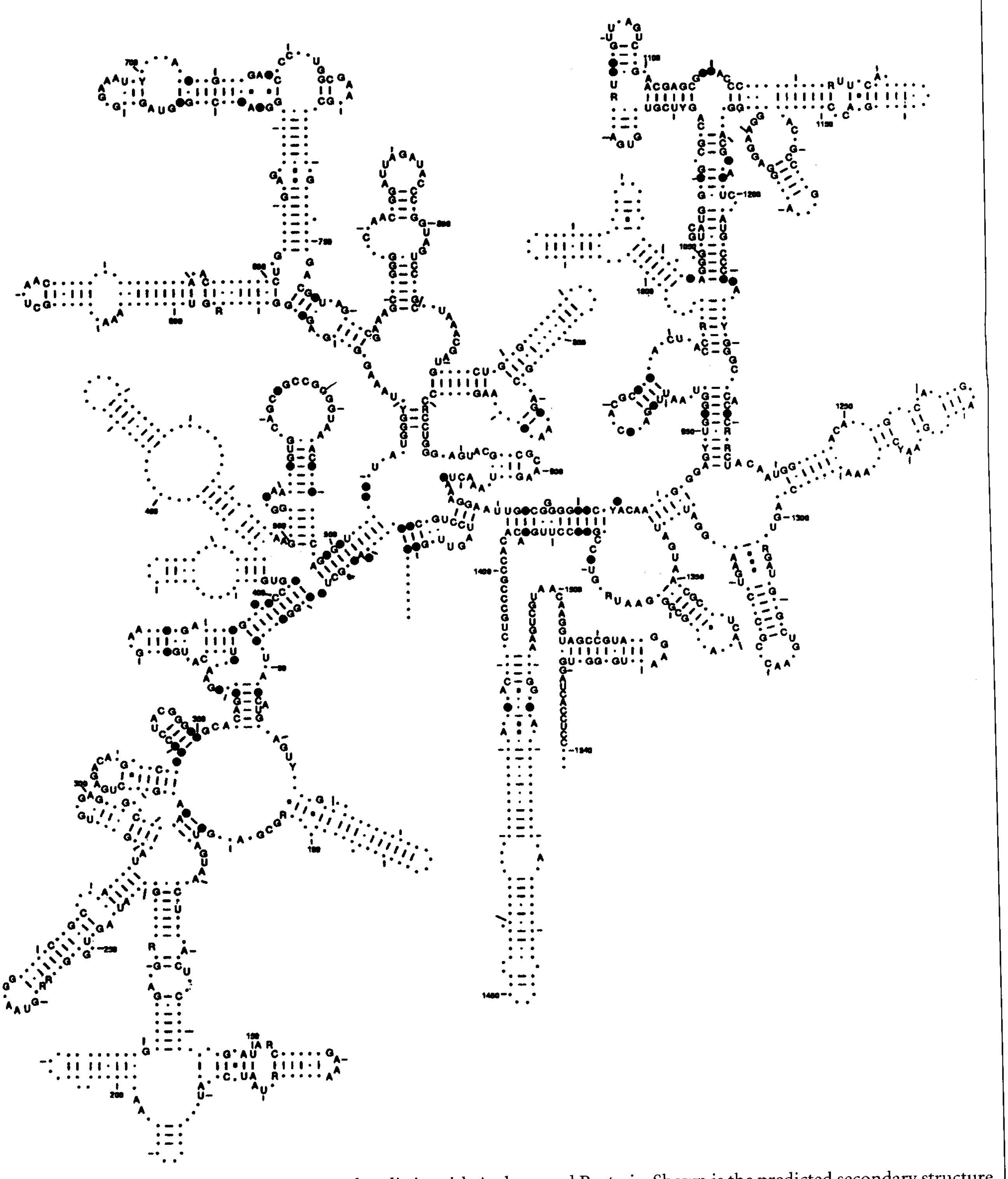


Figure 6.2. The signature positions in rRNA that distinguish Archaea and Bacteria. Shown is the predicted secondary structure for *E. coli* 16S ribosomal RNA with the most highly conserved sequence positions marked by the sequence character and the positions that distinguish Archaea and Bacteria shown by a black dot. Other marker positions in the sequence were used to define the third group, the Eukarya. (Reprinted, with permission, from Woese 1987 [copyright American Society for Microbiology].)

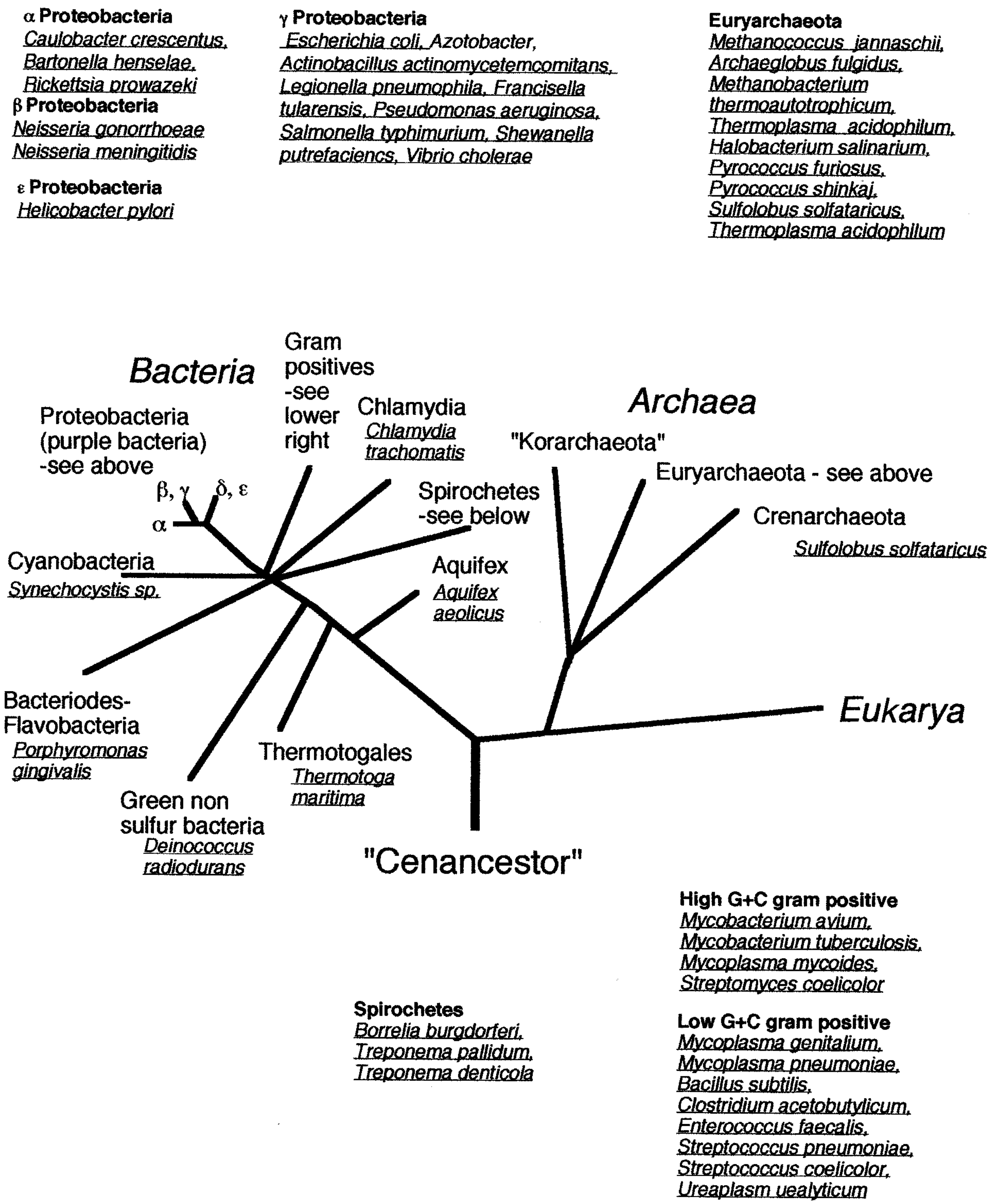


Figure 6.3. Rooted tree of life showing principal relationships among prokaryotic domains Bacteria and Archaea (Woese 1987; Barns et al. 1996; Brown and Doolittle 1997). Branch lengths are approximate only. Species that have been sequenced or are being sequenced are shown. A comprehensive database of sequenced microbial genomes is maintained at http://www.tigr.org/.

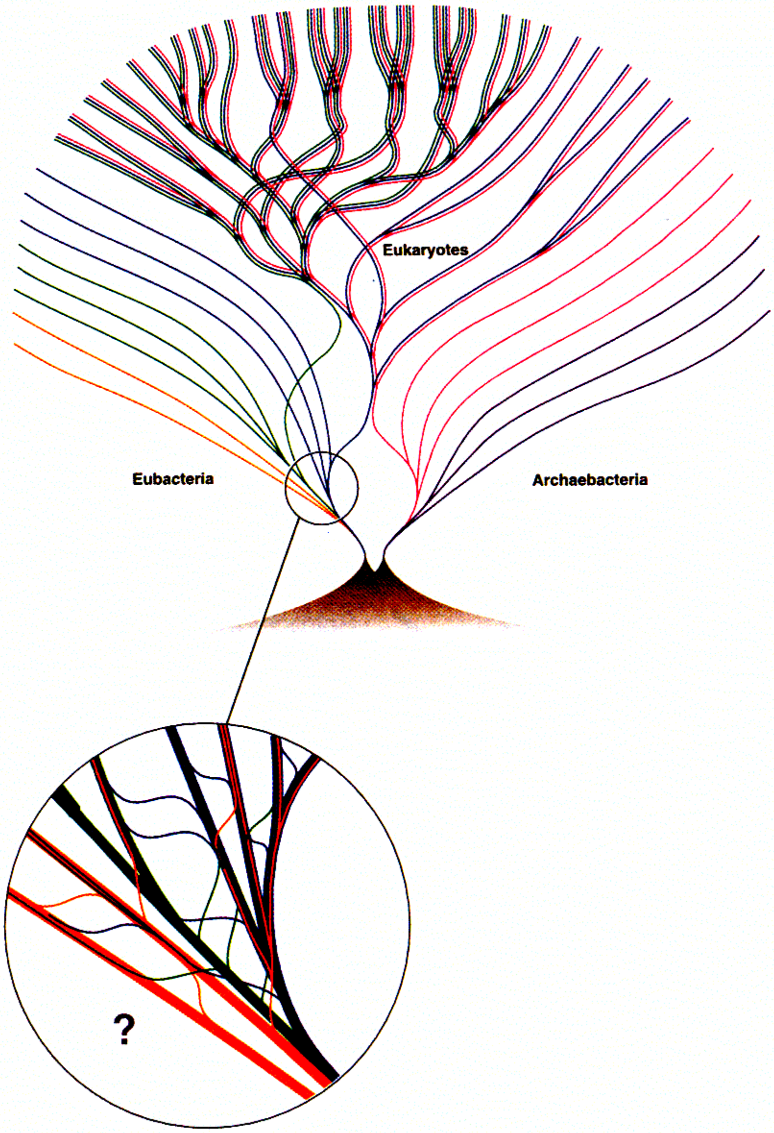


Figure 6.4. The reticulated or net-like form of the tree of life. Analysis of rRNA sequences originally suggested three main branches in the tree of life, Archaea, Bacteria, and Eukarya. Subsequent phylogenetic analysis of genes for some metabolic enzymes is not congruent with the rRNA tree. Hence, for these metabolic genes, the tree has a reticulated form due to horizontal transfer of these genes between species. (Reprinted, with permission, from Martin 1999 [copyright Wiley-Liss, Inc.].)

a second path, reproductively and genetically isolated from each other. The a1 genes in the different species are orthologous to each other, as are the a2 genes, but the a1 and a2 genes are paralogous because they arose from a gene duplication event. These relationships can be determined by a careful analysis of genomes and sequence relationships (Tatusov et al. 1997) that is discussed further in Chapter 10.

THE CONCEPT OF EVOLUTIONARY TREES

An evolutionary tree is a two-dimensional graph showing evolutionary relationships among organisms, or in the case of sequences, in certain genes from separate organisms.

The separate sequences are referred to as taxa (singular taxon), defined as phylogenetically distinct units on the tree. The tree is composed of outer branches (or leaves) representing the taxa and nodes and branches representing relationships among the taxa, illustrated as sequences A–D in Figure 6.5. Thus, sequences A and B are derived from a common ancestor sequence represented by the node below them, and C and D are similarly related. The A/B and C/D common ancestors also share a common ancestor represented by a node at the lowest level of the tree. It is important to recognize that each node in the tree represents a splitting of the evolutionary path of the gene into two different species that are isolated reproductively. Beyond that point, any further evolutionary changes in each new branch are independent of those in the other new branch. The length of each branch to the next node represents the number of sequence changes that occurred prior to the next level of separation. Note that, in this example, the branch length between the A/B node and A is approximately equal to that between the A/B node and B, indicating the species are evolving at the same rate.

The amount of evolutionary time that has transpired since the separation of A and B is usually not known. What is estimated by phylogenetic analysis is the amount of sequence change between the A/B node and A and also between the A/B node and B. Hence, judging by the branch lengths from this node to A and B, the same number of sequence changes has occurred. However, it is also likely that for some biological or environmental reason unique to each species, one taxon may have undergone more mutations since diverging from the ancestor than the other. In this case, different branch lengths would be shown on the tree. Some types of phylogenetic analyses assume that the rates of evolution in the tree branches are the same, whereas others assume that they vary, as discussed below. The assumption of a uniform rate of mutation in the tree branches is known as the molecular clock hypothesis and is usually most suitable for closely related species (Li and Graur 1991; Li 1997). Tests for this hypothesis have been devised as described below. Even if there is a common rate of evolutionary change, statistical variations from one branch to another can influence the analysis. The number of substitutions in each branch is generally assumed to vary according to the Poisson distribution (see Chapter 3, p. 103, for an explanation of the Poisson distribution), and the rate of change is assumed to be equal across all sequence positions (Swofford et al. 1996).

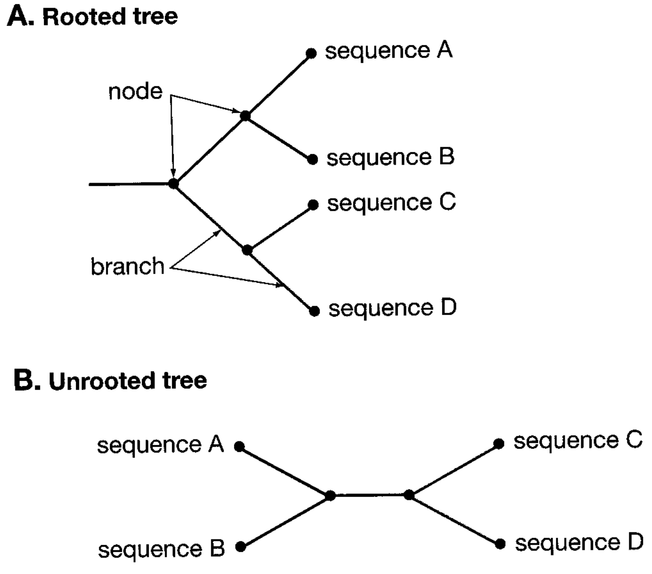


Figure 6.5. Structure of evolutionary trees.

The tree shown is only one of many, each predicting a different evolutionary relationship among the sequences or taxa. The number of possible rooted trees increases very rapidly with the number of sequences or taxa, as shown in Table 6.2. A root has been placed at this position indicating that in this evolutionary model of the sequences this basal node is the common ancestor of all of the other sequences. A unique path leads from the root node to any other node, and the direction of the path indicates the passage of evolutionary time. The root is defined by including a taxon that we are reasonably sure branched off earlier than the other taxa under study but should be related to the remaining taxa. It is also possible to predict a root, assuming that the molecular clock hypothesis holds.

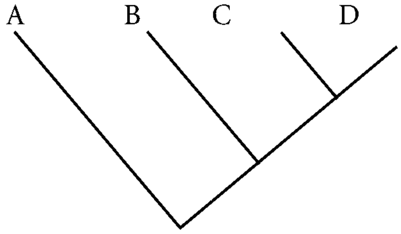
The sum of all the branch lengths in a tree is referred to as the tree length. The tree is also a bifurcating or binary tree, in that only two branches emanate from each node. This situation is what one would expect during evolution—only one splitting away of a new species at a time. Trees can have more than one branch emanating from a node if the events separating taxa are so close that they cannot be resolved, or to simplify the tree.

An alternative representation of the relationships among sequences A–D in Figure 6.5A is shown in Figure 6.5B. The difference between the tree in A and that in B is that the tree in B is unrooted. The unrooted tree also shows the evolutionary relationships among sequences A–D, but it does not reveal the location of the oldest ancestry. B could be converted into A by placing another node and adjoining root to the black line. A root could also be placed anywhere else in the tree. Hence, there are a great many more possibilities for rooted than for unrooted trees for a given number of taxa or sequences, as shown in Table 6.2.

Three methods—maximum parsimony, distance, and maximum likelihood—are generally used to find the evolutionary tree or trees that best account for the observed variation in a group of sequences. Each of these methods uses a different type of analysis as described below. The flowchart on page 247 descibes the types of considerations that need to be made in choosing a method. These methods may find that more than one tree meets the criterion chosen for being the most likely tree. The branching patterns in these trees may be compared to find which branches are shared and therefore are more strongly supported. PAUP provides methods for finding consensus trees, and such trees are also calculated by the CONSENSE program in the PHYLIP package. Trees are stored as a tree file that shows the relationships in nested-parenthesis notation, i.e., a file with the line (A,(B,(C,D))) represents the tree shown below in Table 6.2. Sometimes, branch lengths are

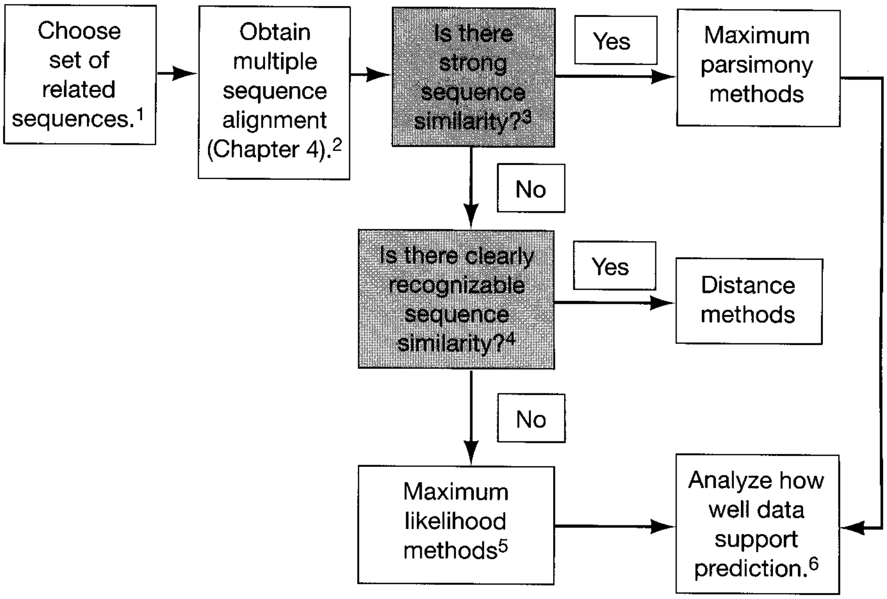
Table 6.2. *Number of possible evolutionary trees to consider as a function of number of sequences*

Taxa or sequence no.	No. of rooted trees	No. of unrooted trees
3	3	1
4	15	3
5	105	15
—	—	—
7	10,395	954



also included next to the names, e.g., A:0.05. From this information, a tree-drawing program may be used to produce a tree representation of the data.

METHODS


Choose set of related sequences.[1]
Obtain multiple sequence alignment (Chapter 4).[2]
Is there strong sequence similarity?[3]
Yes
Maximum parsimony methods
No
Is there clearly recognizable sequence similarity?[4]
Yes
Distance methods
No
Maximum likelihood methods[5]
Analyze how well data support prediction.[6]

1. The sequences chosen can be either DNA or protein sequence: Different programs and program options are used for each type. RNA sequences are analyzed by covariation methods and by analyzing changes in secondary structure, as outlined in Chapter 5. The selected sequences should align with each other along their entire lengths, or else each should have a common set of patterns or domains that provides a strong indication of evolutionary relatedness.
2. The alignment of the sequence pairs should not have a large number of gaps that are obviously necessary to align identical or related characters (see Chapter 3 flowchart, p. 58). A phylogenetic analysis should only be performed on parts of sequences that can be reasonably aligned. In general, phylogenetic methods analyze conserved regions that are represented in all the sequences. The more similar the sequences are to each other, the better. The simplest evolutionary models assume that the variation in each column of the multiple sequence alignment represents single-step changes and that no reversals (A → T → A) have occurred. As the observed variation increases, more multiple-step changes (A → T → G) and reversions are likely to be present. Corrections may be applied for such variation, thereby increasing the observed amount of change to a more reasonable value. These corrections assume a uniform rate of change at all sequence positions over time. Gaps in the multiple sequence alignment are usually not scored because there is no suitable model for the evolutionary mechanisms that produce them.
3. This question is designed to select sequences suitable for maximum parsimony analysis. Other methods may also be used with these same sequences. For parsimony analysis, the best results are obtained when the amount of variation among all pairs of sequences is similar (no very different sequences are present) and when the amount of variation is small. Some columns in the multiple sequence alignment will have the same residue in all sequences; other columns will include both conserved and nonconserved residues. There should be a clear-cut majority of certain residues in some columns of the alignment but also some variation. These more common residues are taken to represent an earlier group of sequences from which others were derived. If there is too much variation, there will be too many possible ancestral relationships. Because the maximum parsimony method has to attempt to fit all possible trees to the data, the method is not suitable for more than 11 or 12 sequences because there are too many trees to test. More than one tree may be found to be equally parsimonious. A consensus tree representing the conserved features of the different trees may then be produced.

4. The purpose of this question is to select sequences for phylogenetic analysis by distance methods. Distance methods are able to predict an evolutionary tree when variation among the sequences is present (some sequences are more alike than others) and when the amount of variation is intermediate. The number of changed positions in an alignment between two sequences divided by the total number of matched positions is the distance between the sequences. As distances increase, corrections are necessary for deviations from single-step changes between sequences (see note 3). Of course, as distances increase, the uncertainty of alignments also increases (see Chapter 4), and a reassessment of the suitability of the multiple sequence alignment method may be necessary. Sequences with this type of variation may also be suitable for phylogenetic analysis by maximum likelihood methods. Distance methods may be used with a large number of sequences. The program CLUSTALW produces a distance-based tree at the same time as a multiple sequence alignment (Higgins et al. 1996).
5. Maximum likelihood methods may be used for any set of related sequences, but they are particularly useful when the sequences are more variable. These methods are computationally intense, and computational complexity increases with the number of sequences since the probability of every possible tree must be calculated as described in the text. An advantage of these methods is that they provide evolutionary models to account for the variation in the sequences.
6. The data in the multiple sequence alignment columns is resampled to test how well the branches on the evolutionary tree are supported (boot-strapping).

MAXIMUM PARSIMONY METHOD

This method predicts the evolutionary tree (or trees) that minimizes the number of steps required to generate the observed variation in the sequences. For this reason, the method is also sometimes referred to as the minimum evolution method. A multiple sequence alignment is required to predict which sequence positions are likely to correspond. These positions will appear in vertical columns in the multiple sequence alignment. For each aligned position, phylogenetic trees that require the smallest number of evolutionary changes to produce the observed sequence changes are identified. This analysis is continued for every position in the sequence alignment. Finally, those trees that produce the smallest number of changes overall for all sequence positions are identified. This method is used for sequences that are quite similar and for small numbers of sequences, for which it is best suited. The algorithm followed is not particularly complicated, but it is guaranteed to find the best tree, because all possible trees relating a group of sequences are examined. For this reason, the method is quite time-consuming and is not useful for data that include a large number of sequences or sequences with a large amount of variation. One or more unrooted trees are predicted and other assumptions must be made to root the predicted tree.

PAUP offers a number of options and parameter settings for a parsimony analysis in the Macintosh environment. The main programs for maximum parsimony analysis in the PHYLIP package (Felsenstein 1996) are listed below.

For analysis of nucleic acid sequences, programs are:

1. DNAPARS, which treats gaps as a fifth nucleotide state.
2. DNAPENNY, which performs parsimonious phylogenies by branch-and-bound search that can analyze more sequences (up to 11 or 12).
3. DNACOMP, which performs phylogenetic analysis using the compatibility criterion. Rather than searching for overall parsimony at all sites in the multiple sequence alignment, this method finds the tree that supports the largest number of sites. This method is recommended when the rate of evolution varies among sites.
4. DNAMOVE, which performs parsimony and compatibility analysis interactively.

For analysis of protein sequences, the program is:

1. PROTPARS, which counts the minimum number of mutations to change a codon for the first amino acid into a codon for the second amino acid, but only scores those mutations in the mutational path that actually change the amino acid. Silent mutations that do not change the amino acid are not scored on the grounds that they have little evolutionary significance.

The maximum parsimony analysis is illustrated in the following example of four sequences shown in Table 6.3 and Figure 6.6 (adapted from Li and Graur 1991). An example of a parsimony analysis of mitochondrial sequences using PAUP and MacClade is then given. Note that in a multiple sequence alignment, only certain sequence variations at a given site are useful for a parsimony analysis. In the analysis, all of the possible unrooted trees (three trees for four sequences) are considered. The sequence variations at each site in the alignment are placed at the tips of the trees, and the tree that requires the smallest number of changes to produce this variation is determined. This analysis is repeated for each informative site, and the tree (or trees) that supports the smallest number of changes overall is found. The length of the tree, defined as the sum of the number of steps in each branch of the tree, will be a minimum.

Example: Maximum Parsimony Analysis of Sequences

Table 6.3 shows an example of phylogenetic analysis by maximum parsimony. This method finds the tree that changes any sequence into all of the others by the least number of steps.

Rules for analysis by maximum parsimony in this example are:

1. There are four taxa giving three possible unrooted trees.
2. Some sites are informative, i.e., they favor one tree over another (site 5 is informative but sites 1, 6, and 8 are not).
3. To be informative, a site must have the same sequence character in at least two taxa (sites 1, 2, 3, 4, 6, and 8 are not informative; sites 5, 7, and 9 are informative).
4. Only the informative sites need to be analyzed.

The three possible trees are shown in Figure 6.6. The optimal tree is obtained by adding the number of changes at each informative site for each tree, and picking the tree requiring the least number of changes. A scoring matrix may be used instead of scoring a change as 1. Tree 1 is the correct one and the tree length will be 4 (one change at each of positions 5 and 7 and two changes at position 9).

In the above example, because there were only four sequences to consider, it was necessary to consider only three possible unrooted trees. For a larger number of sequences, the number of trees becomes so large that it may not be feasible to examine all possible trees. The example of 12 sequences below took only a few seconds on a Macintosh G3. The exhaustive and branch-and-bound options of the program PAUP will analyze all possible trees, and if the number is too large, the program can keep running for a very long time.

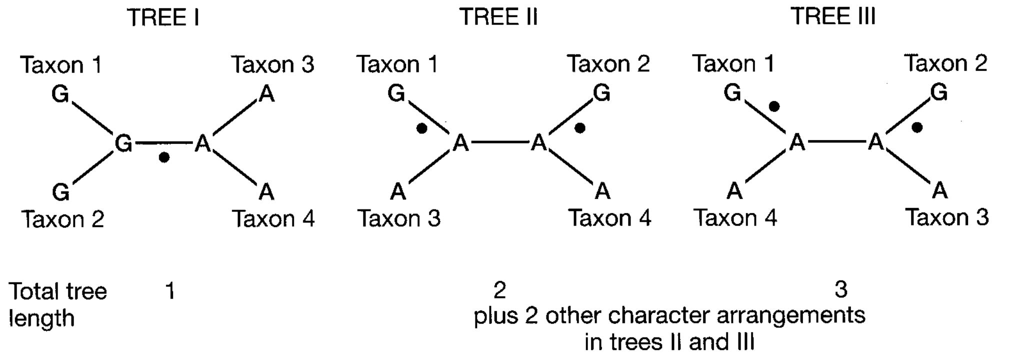
Branch-and-bound is a method that stops analyzing a particular branching pattern in trees when it is not possible to obtain a more parsimonious solution than has been already found.

For large numbers of sequences, PAUP provides a program option called "heuristic," which searches among all possible trees and keeps representative trees that best fit the data. The presence of common branch patterns in these trees reveals some of the broader features of the phylogenetic relationships among the sequences.

Table 6.3. *Example of phylogenetic analysis to find the correct unrooted tree from four aligned sequences by the maximum parsimony method*

Taxa	Sequence position (sites) and character								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Adapted from Li and Graur 1991.



• is a substitution

Figure 6.6. Example of phylogenetic analysis using the maximum parsimony method. (Redrawn, with permission, from Li and Graur 1991 [copyright Sinauer Associates].)

Analysis of Mitochondrial Sequences by PAUP

To search for this tree, which best fits all the sequence data, the trees that best fit each vertical column of sequence characters in Figure 6.7A were first determined. In some columns, the data are not informative, as in the case of all nucleotides being the same. For a nucleotide position to be informative, at least two different nucleotides must be present in at least two of the sequences. A tree that provides the least number of evolutionary steps to satisfy the data in all columns, the most parsimonious tree, is then found.

Parsimony can give misleading information when rates of sequence change vary in the different branches of a tree that are represented by the sequence data. These variations produce a range of branch lengths, long ones representing more extended periods of time and short ones representing shorter times. For example, the real tree shown below in Figure 6.8A includes two long branches in which G has turned to A independently, probably with a number of intermediate changes that are not observed in the sequence data. Because in a parsimony analysis rates of change along all branches of the tree are assumed to be equal, the tree predicted by parsimony and shown in Figure 6.8B will not be correct.

Although other columns in the sequence alignment that show less variation may provide the correct tree, the columns representing greater variation dominate the analysis

(Swofford et al. 1996). Such long branches may be broken down if additional taxa are present that are more closely related to taxa 1 and 4, thereby providing branches that intersect the long branches and give a better resolution of the changes.

Another method for identifying such long branches is called Lake's method of invariants or evolutionary parsimony, available in PAUP. In this method, four of the sequences are chosen at a time, and only transversions in the aligned positions are scored as changes on the grounds that transversions are the most significant base changes during evolution. Transversions of any base to each possible derivative, e.g., A → C or T, are assumed to change at the same rate to create a balanced distribution, and the changes in each column of the alignment (each sequence position) are assumed to occur independently of each other. Suppose that there are two long branches as in the case discussed immediately above. The correct tree is shown in Figure 6.9A, and one of the sites has changed multiply but ends up as the same base A by chance. Traditional parsimony will identify this tree incorrectly, as indicated above. If these long branches do indeed exist, then other sites should give the type of transversion events shown in Figure 6.9B. The greater the number of B-type sites, the less one can depend on the A-type sites revealed in A. The evolutionary parsimony method subtracts the number of type B from the number of type A. If, on the one hand, long branches are not present in the quartet of sequences, there will be very few type B, and type A will be taken as evidence for the correct tree. On the other hand, if many examples of type B are present, the A type will carry little weight. These calculations are performed for all three possible unrooted trees and all possible types of transversions for the four sequences, and the tree receiving the most support is chosen. These methods and other more sophisticated methods for correcting uneven branch lengths are discussed in detail in Swofford et al. (1996). The PHYLIP program DNAINVAR computes Lake's and other phylogenetic invariants for nucleic acid sequences. PAUP also includes an option for Lake's invariant.

Compared to the above methods, maximum likelihood and distance methods provide more reliable predictions when corrections are made for multiple substitutions. Distance methods such as neighbor joining discussed below have been shown generally to be better predictors than both standard and evolutionary parsimony methods when branch lengths are varying (Jin and Nei 1990; Swofford et al. 1996).

There are options in PAUP and MacClade for selecting among the most parsimonious trees. With MacClade it is possible to view the changes in sequence characters in each branch of the tree to arrive at the current base in each sequence or taxon, as shown below. As these characters are traced from positions lower in the tree to upper positions, some nodes in the tree may be assigned an unambiguous character (shown in color, Fig. 6.10). For other nodes, the assignment may be ambiguous because the node is leading to two different characters above (thin black line). It is possible to arrange these ambiguities optionally in two ways: one is to delay them going as far up the tree away from the root as possible (the Deltran option; not shown in figure); a second is to introduce them as soon as possible and as close to the root as possible (the Acctran option; not shown in figure). The effect of using Deltran is to force parallel changes in the upper branches of the tree, that of Acctran is to force reversals in the upper branches. Using these options is not recommended unless such variations are expected, as in analysis of more divergent sequences (Maddison and Maddison 1992).

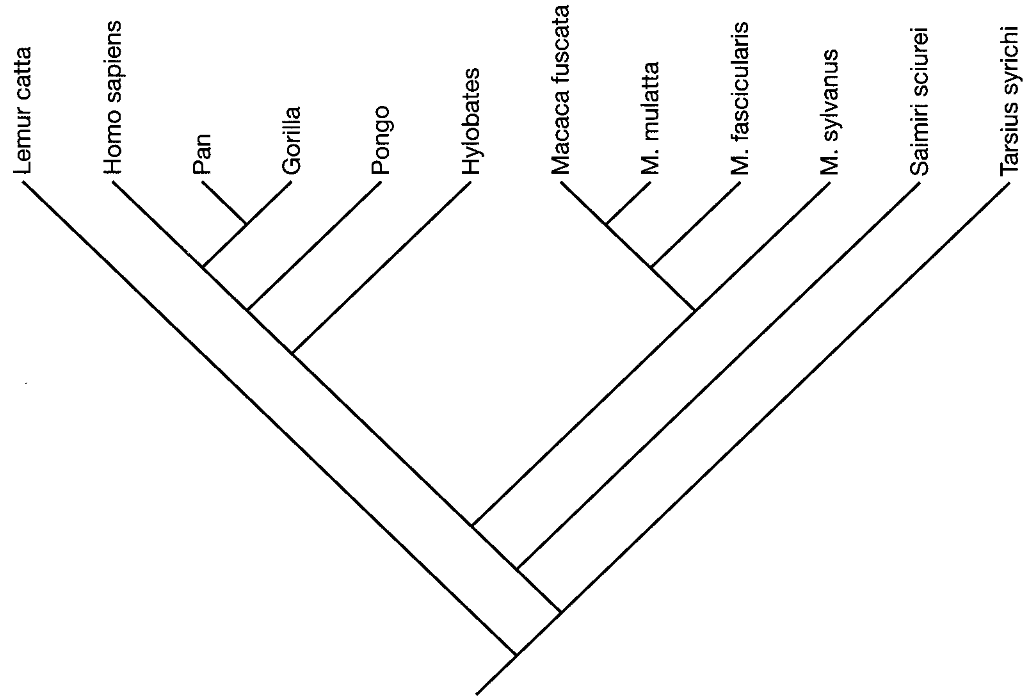
Homoplasy refers to the occurrence of the same sequence change in more than one branch of the tree. If all the sequence character changes support the same tree, there is no homoplasy. In reality, homoplasy is usually found for some characters for any tree. MacClade allows changing of the tree to avoid homoplasy at a sequence position, but the new tree length will often increase, thus making the tree a less parsimonious choice than the

This sequence format is the NEXUS format, which allows additional information about the sequences, species relationship, and a scoring system for base substitution referred to as a cost or step matrix.

A. Mitochondrial sequences.

```
#NEXUS

begin taxa;
      dimensions ntax=12;
end;

begin characters;
      dimensions nchar=898;
      format missing=? gap=- matchchar=. interleave datatype=dna;
      options gapmode=missing;
      matrix

Lemur_catta       AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCATATTATT
Homo_sapiens      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTATT
Pan               AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATT
Gorilla           AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATT
Pongo             AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTT
Hylobates         AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATT
Macaca_fuscata    AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCATATATTT
M._mulatta        AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCATATATTT
M._fascicularis   AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCATGTATTT
M._sylvanus       AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCATATACTT
Saimiri_sciureus AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTATGCTATT
Tarsius_syrichta AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCCTATTATT

Lemur_catta       CTGTCTAGCCAACTCTAACTACGAACGAATCCATAGCCGTACAATACTACTAGCACGAGG
Homo_sapiens      CTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATCCTCTCTCAAGG
Pan               CTGCCTAGCAAACTCAAATTATGAACGCACCCACAGTCGCATCATAATTCTCTCCCAAGG
Gorilla           CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGG
Pongo             CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATCCTCTCTCAAGG
Hylobates         CTGCCTTGCAAACTCAAACTACGAACGAACTCACAGCCGCATCATAATCCTATCTCGAGG
Macaca_fuscata    CTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTACCATACTACTGTCCCGAGG
M._mulatta        CTGCCTAGCCAATTCAAACTATGAACGCACTCACAACCGTACCATACTACTGTCCCGGGG
M._fascicularis   CTGCTTGGCCAATTCAAACTATGAGCGCACTCATAACCGTACCATACTACTATCCCGAGG
M._sylvanus       CTGCTTGGCCAACTCAAACTACGAACGCACCCACAGCCGCATCATACTACTATCCCGAGG
Saimiri_sciureus CTGCCTAGCAAACTCAAATTACGAACGAATTCACAGCCGAACAATAACATTTACTCGAGG
Tarsius_syrichta TTGCCTAGCAAATACAAACTACGAACGAGTCCACAGTCGAACAATAGCACTAGCCCGTGG
.
.
.
end;
```

B. Phylogenetic tree



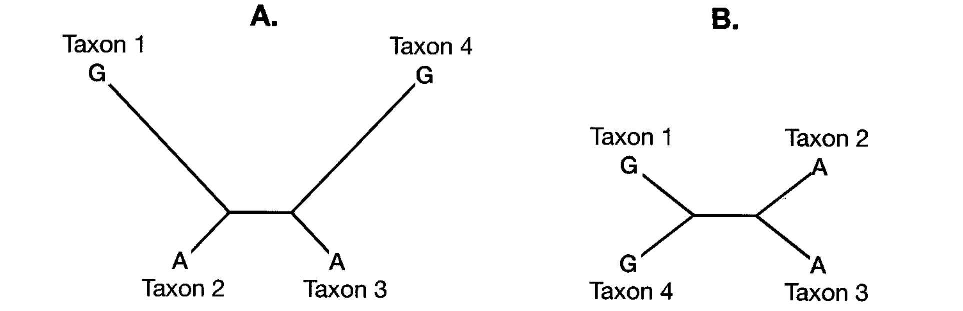


Figure 6.8. Type of sequence variation that leads to an incorrect prediction by the maximum parsimony method.

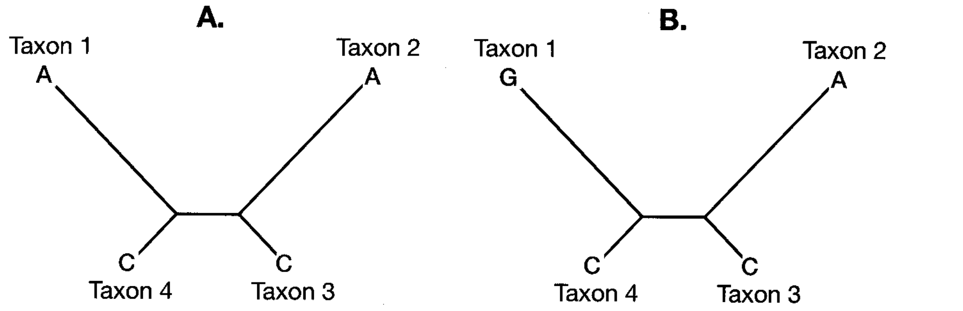


Figure 6.9. Type of sequence variation that, if detected, can reduce incorrect predictions by the maximum parsimony method.

original. Another parameter used is the consistency index (CI), which is the minimum possible tree length divided by the actual tree length. The more homoplasy, the greater the actual tree length, and the smaller the value of CI.

Parsimony methods can use information on the number of changes required or steps to change one residue into another. For example, the number of mutations required to change one amino acid into another in one branch of a tree can be taken into account. The parsimony method then attempts to minimize the number of such steps. This number of steps for interchanging characters can be incorporated into a matrix, called a step or cost matrix for programs such as PAUP and MacClade to use.

A program designated PROTPARS for protein squences in the PHYLIP package scores only those mutations that produce amino acid changes (Felsenstein 1996). This program uses an algorithm similar to one described by Sankoff (1975) for determining the mini-

Figure 6.7. Analysis of mitochondrial sequences using the maximum parsimony method provided by the PAUP program. (*A*) Portion of a multiple sequence alignment of the mitochondrial sequences provided in the PAUP distribution package. PAUP will import sequences in other multiple sequence alignment format and convert them into the NEXUS format. The program READSEQ will reformat multiple sequence alignments into the NEXUS format. This format includes information about type of sequence, coding information, codon positions, differential weights for transitions and transversions, treatment of gaps, and preferred groupings (see Chapter 2). Only a portion of the NEXUS file is shown. In this analysis, branch-and-bound and otherwise default options were used. Gaps are treated as missing information. The number of sequences is indicated as ntaxa, number of alignment columns as nchar, and the interleave command allows the data to be entered in readable blocks of sequence 60 characters long. (*B*) One of the two predicted trees. The tree file of PAUP was edited in MacClade and output as a graphics file.

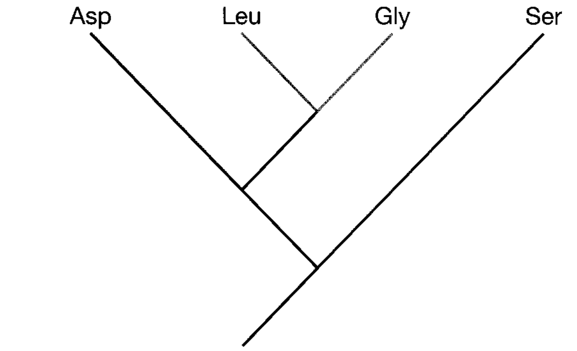


Figure 6.10. Tracing of sequence characters in an evolutionary tree by MacClade.

mum number of mutations in a tree for changing one sequence into another. Similar types of analyses for proteins are also available in PAUP and MacClade. The PAUP program uses a 3+1 option in the stepmatrices option, which is a short cut for analyzing trees that represent the most possible ancestors of an amino acid (PAUP vers 3.1 manual, pp. 124–126).

DISTANCE METHODS

The distance method employs the number of changes between each pair in a group of sequences to produce a phylogenetic tree of the group. The sequence pairs that have the smallest number of sequence changes between them are termed "neighbors." On a tree, these sequences share a node or common ancestor position and are each joined to that node by a branch. The goal of distance methods is to identify a tree that positions the neighbors correctly and that also has branch lengths which reproduce the original data as closely as possible. Finding the closest neighbors among a group of sequences by the distance method is often the first step in producing a multiple sequence alignment, as discussed in Chapter 4.

The distance method was pioneered by Feng and Doolittle, and a collection of programs by these authors will produce both an alignment and tree of a set of protein sequences (Feng and Doolittle 1996). The program CLUSTALW, discussed in Chapter 4, uses the neighbor-joining distance method as a guide to multiple sequence alignments. PAUP version 4 has options for performing a phylogenetic analysis by distance methods. Programs of the PHYLIP package that perform a distance analysis include the following programs, which automatically read in a sequence in the PHYLIP infile format (see Chapter 2) and automatically produce a file called outfile with a distance table.

1. DNADIST computes distances among input nucleic acid sequences. There are choices given for various models of evolution as described below and a choice for the expected ratio of transitions to transversions.
2. PROTDIST computes a distance measure for protein sequences, based on the Dayhoff PAM model (see p. 78) or other models of evolutionary change in proteins (Felsenstein 1996).

Once distance matrices have been produced, they may be used as input to the following distance analysis programs in PHYLIP. The PHYLIP programs all automatically read an input file called infile and produce an output file called outfile. Hence, file names have to be edited when using these programs. In this example, the distance outfile must be edited to include only the distance table and the number of taxa, and then the file is saved under the sequence name infile.

Distance analysis programs in PHYLIP:

1. FITCH estimates a phylogenetic tree assuming additivity of branch lengths using the Fitch-Margoliash method described below and does not assume a molecular clock (allows rates of evolution along branches to vary).
2. KITSCH estimates a phylogenetic tree using the Fitch-Margoliash method but under the assumption of a molecular clock.
3. NEIGHBOR estimates phylogenies using the neighbor-joining or unweighted pair group method with arithmetic mean (UPGMA) described below. The neighbor-joining method does not assume a molecular clock and produces an unrooted tree. The UPGMA method assumes a molecular clock and produces a rooted tree.

Recall that in aligning sequences, we normally calculate a similarity score, defined as the sum of the number of identities and number of conservative substitutions in the alignment of the two sequences, with gaps being ignored. An identity score between the sequences showing just the identities may also be found from the alignment. For phylogenetic analysis, the distance score between two sequences is used. This score between two sequences is the number of mismatched positions in the alignment or the number of sequence positions that must be changed to generate the other sequence. Gaps may be ignored in these calculations or treated like substitutions. When a scoring or substitution matrix is used, the calculation is slightly more complicated, but the principle is the same. These methods are described below.

The success of distance methods depends on the degree to which the distances among a set of sequences can be made additive on a predicted evolutionary tree. Suppose there are four sequences, A–D, as shown below in Figure 6.11A, and that they were derived from evolutionary changes reflected by the tree in Figure 6.11D. The number of changes along the branches of the tree corresponds to distances between the sequences shown in Figure 6.11, B and C. In this tree, each change only occurs once, and there are no examples of the same change occurring twice (homoplasy). Although this pattern of change is idealized and most groups of sequences would have examples of the same change occurring more than once, as well as reversions, this example illustrates the additivity principle for four sequences. The principle is that for four sequences predicted by this tree, $d_{AB} + d_{CD} \leq d_{AC} + d_{BD} = d_{AD} + d_{BC}$. In this example the additivity is $3 + 3 \leq 7 + 7 = 8 + 6$. For any other tree, there would be examples of parallel changes and reversions. The additivity condition can be relaxed such that $d_{AB} + d_{CD} \leq d_{AC} + d_{BD}$ and $d_{AB} + d_{CD} \leq d_{AD} + d_{BC}$ will still hold even for sequences in which the changes in the sequence are not fully additive. For each set of four sequences, the tree for which the above additivity condition among the distances best holds provides information as to which sequences are neighbors. This method may be used to evaluate trees and find the minimum evolution tree for four sequences and for any additional number of sequences by extending the analysis to additional groups of four sequences (Sattath and Tversky 1977; Fitch 1981; for references, see Swofford et al. 1996). In order to calculate branch lengths, distance methods assume additivity in the distances between sequences. However, real sequence data may not fit these idealized conditions. As a result, a small positive, zero, or even a negative value may be calculated for a branch length. This result may be due to errors in the sequences or sequence alignment, statistical variation, or simply a reflection of two or more sequences diverging at approximately the same time from a common ancestor.

A. Sequences

sequence A	ACGCGTTGG GCG ATG GCAAC
sequence B	ACGCGTTGG GCG ACG GTAAT
sequence C	ACGCATTGA ATG ATG ATAAT
sequence D	ACACATTGA GTG ATA ATAAT

B. Distances between sequences, the number of steps required to change one sequence into the other.

n_{AB}	3
n_{AC}	7
n_{AD}	8
n_{BC}	6
n_{BD}	7
n_{CD}	3

C. Distance table

	A	B	C	D
A	–	3	7	8
B	–	–	6	7
C	–	–	–	3
D	–	–	–	–

D. The assumed phylogenetic tree for the sequences A-D showing branch lengths. The sum of the branch lengths between any two sequences on the trees has the same value as the distance between the sequences.

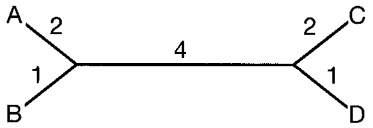


Figure 6.11. Set of idealized sequences for which the branch lengths of an assumed tree are additive.

An even more demanding condition, rarely found in real distance data, is that the distances are ultrametric, meaning that for three taxa, $d_{AC} \leq \max(d_{AB}, d_{BC})$. If the data meet this condition, the distances between two taxa and their common ancestor are equal (Swofford et al. 1996). If the distances follow this relationship, the rates of evolution in the tree branches are approximately the same, thereby meeting the expectations of the molecular clock hypothesis. If these conditions are not met, an analysis based on the assumption of a molecular clock may give misleading results. One method of finding the best tree under such conditions is to transform the sequences after identifying one or more sequences that are least like the rest, called an outgroup (Li and Graur 1991). Some distance methods are based on this assumption and others are not. The overall objective of the distance methods described below is to find this tree by the identification of consecutive sets of neighbors starting with the most alike sequence pair.

Fitch and Margoliash Method and Related Methods

The Fitch and Margoliash (1987) method uses a distance table illustrated in Figure 6.11C. The sequences are combined in threes to define the branches of the predicted tree and to calculate the branch lengths of the tree. The branch lengths are assumed to be additive, as described above. This method of averaging distances is most accurate for trees with short branches. The presence of long branches tends to decrease the reliability of the predictions (Swofford et al. 1996). The following first example describes the use of the algorithm for three sequences, and the second example expands the analysis to more than three sequences.

Example 1: Use of Fitch Margoliash Algorithm for Three Sequences

Steps in algorithm for three sequences:

1. Draw an unrooted tree with three branches emanating from a common node and label ends of branches as shown in Figure 6.12. Given the closer distance between A and B, the branch lengths between these sequences are expected to be shorter, as indicated.
2. Calculate lengths of tree branches algebraically:

   Distances among sequences A, B, and C are shown in the following table.

	A	B	C
A	—	22	39
B	—	—	41
C	—	—	—

The branch lengths may be calculated algebraically using the branch labels *a–c* in Figure 6.12:

distance from A to B = $a + b = 22$ (1)
distance from A to C = $a + c = 39$ (2)
distance from B to C = $b + c = 41$ (3)

subtract (3) from (2), $a - b = -2$ (4)
add (1) and (4), $2a = 20$, $a = 10$
from (1) and (2), $b = 12$, $c = 29$

Note that this calculation finds that the branch lengths of A and B from their common ancestor are not the same. Hence, A and B are diverging at different rates of evolution by this calculation and model. For the rates to be the same, these distances would be the same and equal to the distance from A to B divided by 2 = 22/2 = 11.

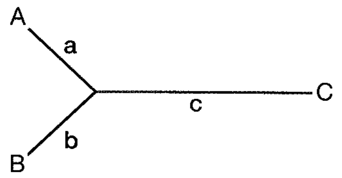


Figure 6.12. Tree showing relationship among three sequences A, B, and C.

Example 2: Use of Fitch-Margoliash Algorithm for Five Sequences

	A	B	C	D	E
A	—	22	39	39	41
B	—	—	41	41	43
C	—	—	—	18	20
D	—	—	—	—	10
E	—	—	—	—	—

These distance data are derived from the unrooted tree shown in Figure 6.13. The Fitch-Margoliash method may be extended from three sequences as shown in example 1 by following the steps shown in the box below, Steps in Fitch-Margoliash algorithm for more than three sequences. The method will find the correct tree and provide the branch lengths *a*–*g*, as illustrated below.

1. The most closely related sequences given in the distance table are D and E. A new table is made with the remaining sequences combined.
2. The average distances from D to A, B and C and from B to A, B and C are calculated.

	D	E	ave. ABC
D	—	10	32.7
E	—	—	34.7
average ABC	—	—	—

3. The average distances from D to ABC and from E to ABC can also be found by averaging the sum of the appropriate branch lengths *a*–*g*.

   Distance between D and E = $d + e$
   Average distance between D and ABC = $d + m$, $m = g + [(c + 2f + a + b)/2]$
   Average distance between E and ABC = $e + m$

   By subtracting the third from the second equation and adding the result to the first equation, $d = 4$ and $e = 6$.
4. D and E are now treated as a single composite sequence (DE), and a new distance table is made. The distance from A to (DE) is the average of the distance of A to D and of A to E. The other distances to (DE) are calculated accordingly.

	A	B	C	(DE)
A	—	22	39	40
B	—	—	41	42
C	—	—	—	19
(DE)	—	—	—	—

5. The next most closely related sequences are identified, in this case C with the (DE) composite group. The new table is:

	DE	C	Ave. AB
DE	—	19	41
C	—	—	40
Ave. AB	—	—	—

By algebraic manipulations similar to those described above, $c = 9$ and the composite distance of $g + [(e + f)/2] = 10$.

6. Given the above composite distance and the previously calculated values of *e* and *f*, then $g = 10 - + [(e + f)/2] = 5$.

The next round of tree-building is that A and B are the next matching pair, giving $a = 10$ and $b = 12$, and a composite distance of $29.7 = [3f + c + 2g + d + e]/3$ giving $f = 29.7 - [(9 + 10 + 10)/3] = 20$. These values are precisely those given in the original tree.

7. Although by design we have generated the correct tree, normally the next step is to repeat the process starting with another sequence pair, such as A and B. We will leave this step as a student exercise to show that the correct tree will again be predicted.

The procedure generally followed is to join all combinations of sequences in pairs to find a tree that best predicts the data in the distance table. The percent change from the actual to the predicted distance is determined for each sequence pair. These values are squared and summed over all possible pairs. This sum divided by the number of pairs = $n(n-1)/2$ less one (the number of degrees of freedom) provides the square of the percent standard deviation of the result.

Steps Followed by Fitch-Margoliash Algorithm for Phylogenetic Analysis of More Than Three Sequences

Steps in algorithm for more than three sequences:

1. Find the most closely related pair of sequences, for example, A and B.
2. Treat the rest of the sequences as a single composite sequence. Calculate the average distance from A to all of the other sequences, and B to all of the other sequences.
3. Use these values to calculate the distances *a* and *b* as in the above example with three sequences.

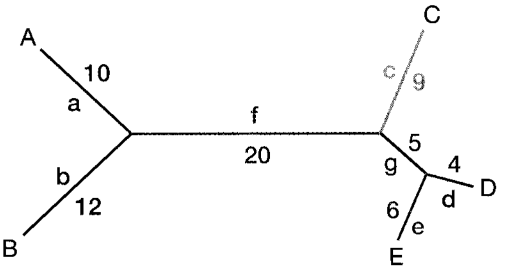


Figure 6.13. Tree showing relationships among sequences A–E.

4. Now treat A and B as a single composite sequence AB, calculate the average distances between AB and each of the other sequences, and make a new distance table from these values.
5. Identify the next pair of most closely related sequences and proceed as in step 1 to calculate the next set of branch lengths.
6. When necessary, subtract extended branch lengths to calculate lengths of intermediate branches.
7. Repeat the entire procedure starting with all possible pairs of sequences A and B, A and C, A and D, etc.
8. Calculate the predicted distances between each pair of sequences for each tree to find the tree that best fits the original data.

The Neighbor-joining Method and Related Neighbor Methods

The neighbor-joining method (Saitou and Nei 1987) is very much like the Fitch-Margoliash method except that the choice as to which sequences to pair is determined by a different algorithm. The neighbor-joining method is especially suitable when the rate of evolution of the separate lineages under consideration varies. When the branch lengths of trees of known topology are allowed to vary in a manner that simulates varying levels of evolutionary change, the neighbor-joining method and the Sattath and Taversky method, described below, are the most reliable in predicting the correct tree (Saitou and Nei 1987). Pearson et al. (1999) have enhanced the neighbor-joining method so that a set of trees that fit the data, rather than just a single tree, may be determined. The general neighbor joining (GNJ) is available from ftp.virginia.edu/pub/fasta/GNJ.

Neighbor-joining chooses the sequences that should be joined to give the best least-squares estimates of the branch lengths that most closely reflect the actual distances between the sequences. It is not necessary to compare all possible trees to find the least-squares fit as in the Fitch-Margoliash method. The method pairs sequences based on the effect of the pairing on the sum of the branch lengths of the tree. To start, the distances between the sequences are used to calculate the sum of the branch lengths for a tree that has no preferred pairing of sequences. The star-like appearance of such a tree and the calculation of the length of the tree using the data in Example 2 above are shown in Figure 6.14.

The next step in the neighbor-joining algorithm is to decompose or modify the star-like tree in Figure 6.14 by combining pairs of sequences. When this step is performed for sequences D and E in Example 2, the new tree shown in Figure 6.15 will be produced. The tree has A and B paired from a common node that is joined by a new branch *j* to a second node to which C, D, and E are joined. The sum of the branch lengths of this new tree is calculated as shown in Figure 6.15.

In the neighbor-joining algorithm, each possible sequence pair is chosen and the sum of the branch lengths of the corresponding tree is calculated. For example, using the data of Example 2, $S_{AB} = 67.7$, $S_{BC} = 81$, $S_{CD} = 76$, and $S_{DE} = 70$, plus six other possible combinations. Of these, S_{AB} has the lowest value. Hence, A and B are chosen as neighbors on the grounds that they reduce the total branch length to the largest extent. Once the choice of neighbors has been made, the branch lengths *a* and *b* and the average distance from AB to CDE may be calculated by the FM method, as described in the last section. *a* is calculated by $a = [d_{AB}+(d_{AC}+d_{AD}+d_{AE})/3-(d_{BC}+d_{BD}+d_{DE})/3]/2 = (22+39.7-41.70)/2=10$, and *b* is calculated by $b = [d_{AB}+(d_{BC}+d_{BD}+d_{BE})/3-(d_{AC}+d_{AD}+d_{AE})/3]/2 = (22+41.7-39.7)/2=12$.

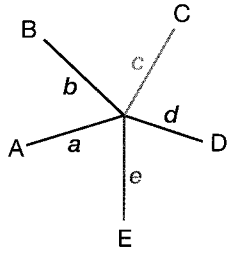


Figure 6.14. Tree for five sequences with no pairing of sequences. In the neighbor-joining method, the sum of the branch lengths $S_0 = a + b + c + d + e$ is calculated. The known distances from (1) A to B, $D_{AB} = a + b$; (2) A to C = $D_{AC} = a + c$; (3) B to C = $D_{BC} = b + c$; and finally (4) D to E, $D_{DE} = d + e$ for a total of 4 + 3 + 2 + 1 =10 combinations. In summing the 10 distances = 22 + 39 + . . . + 10 = 314, each branch a, b, c, etc., is counted four times. Hence, the sum of branch lengths is 314/4 = 78.5. In general, for N sequences, $S_0 = \Sigma\ D_{ij} / (N - 1)$, where D_{ij} represents the distances between sequences i and j, $i < j$.

The next step of the neighbor-joining algorithm is like that of the Fitch-Margoliash method: a new distance table with A and B forming a single composite sequence is produced. The neighbor-joining algorithm is then used to find the next sequence pair and Fitch-Margoliash is then used to find the next branch lengths. The cycle is repeated until the correctly branched tree and the branch distances on that tree have been identified.

The neighbors relation method (Sattath and Tversky 1977; Li and Graur 1991) also is a reliable predictor of trees when the rate of evolution varies. In this method, the sequences are divided into all possible groups of four. The sum of the pair-wise distances for the three possible neighbor groupings (AB/CD, AC/BD, AD/BC) for each group are then compared to find which grouping of the three gives the lowest sum of pairs. This procedure is repeated for all possible groups of four. The pair that appears most often in the lowest sum of pairs is selected as neighbors. An example of this method is shown in Table 6.4. The pair is then treated as a composite grouping and the entire process is repeated to find the next closest neighbor until all of the sequences have been included.

The Unweighted Pair Group Method with Arithmetic Mean

The above distance methods provide a good estimate of an evolutionary tree and are not influenced by variations in the rates of change along the branches of the tree. The UPGMA

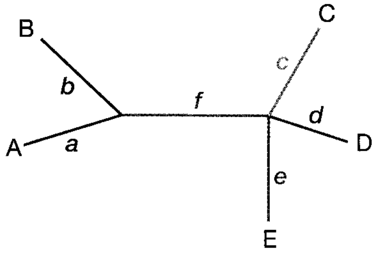


Figure 6.15. Tree for five sequences with pairing of A and B. The sum of the branch lengths $S_{ab} = a + b + c + d + e + f$ is calculated algebraically from the original distance data. The sum is given by $S_{ab} = [(d_{AC} + d_{AD} + d_{CE} + d_{BC} + d_{BD} + d_{BE})/6) + d_{AB}/2 + [(d_{CD} + d_{CE} + d_{DE})/3] = 244/6 + 22/2 + 48/3 = 67.7$. In general, the formula for N sequences when m and n are paired is $S_{mn} = [(\Sigma\ d_{im} + d_{in})/2(N - 2)] + d_{mn}/2 + \Sigma\ d_{ij}/N - 2$ where i and j represent all sequences except m and n, and $i < j$.

Table 6.4. *The Sattath and Tversky (1977) method for finding repeated neighbors*

Chosen set of 4	Sum of distances	Pairs chosen
ABCD	$n_{AB} + n_{CD} = 22 + 18 = 40$	AB, CD
	$n_{AC} + n_{BD} = 39 + 41 = 80$	
	$n_{AD} + n_{BC} = 39 + 41 = 80$	
	$n_{AB} + n_{CE} = 22 + 20 = 42$	
	$n_{AC} + n_{BE} = 39 + 43 = 82$	
ABCE	$n_{AE} + n_{BC} = 39 + 41 = 82$	AB, CE
	$n_{AB} + n_{DE} = 22 + 10 = 32$	
ABDE	$n_{AD} + n_{BE} = 39 + 43 = 82$	AB, DE
	$n_{AE} + n_{BD} = 41 + 41 = 82$	
	$n_{AC} + n_{DE} = 39 + 10 = 49$	
ACDE	$n_{AD} + n_{CE} = 39 + 20 = 59$	AC, DE
	$n_{AE} + n_{CD} = 41 + 18 = 59$	
	$n_{BC} + n_{DE} = 41 + 10 = 51$	
BCDE	$n_{BD} + n_{CE} = 41 + 20 = 61$	BC, DE
	$n_{BE} + n_{CD} = 43 + 18 = 61$	

Totals from Column 3 giving the number of times a pair gives the lowest score: AB (3), DE (3), CD (1), CE (1), and BC (1). AB and DE are therefore closest neighbors.

The five sequences used in the above example (see Fig. 6.13) are divided into the five possible groups of four. The sums of distances for each set of sequence pairs for the three possible groupings are then determined and the closest pairs in each grouping are determined. The closest neighbors overall are those that appear as neighbors most often. In this example, AB and DE appear most often as neighbors. These sequences are then chosen as neighbors to calculate the branch lengths on the phylogenetic tree by the method of Fitch and Margoliash.

method is a simple method for tree construction that assumes the rate of change along the branches of the tree is a constant and the distances are approximately ultrametric (see above). There are also a number of variations of this method for pairing or clustering sequences. The UPGMA method starts by calculating branch lengths between the most closely related sequences, then averages the distance between this pair or sequence cluster and the next sequence or sequence cluster, and continues until all the sequences are included in the tree. Finally, the method predicts a position for the root of the tree.

Using Example 2 from the above analysis:

Example: UPGMA Analysis

1. Sequences D and E are the most closely related. The branch distances *d* and *e* to the node below them are calculated as $d = e = n_{de}/2 = 5$ based on the assumption of an equal rate of change in each branch of the tree. The tree is often drawn in a form (Fig. 6.16a) where only the horizontal lines indicate branch lengths, but the branches are intended to be joined to a common node as in Figure 6.16B.

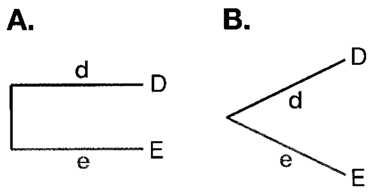


Figure 6.16. Branch lengths of most closely related sequences by UPGMA method.

2. Treating D and E as a composite sequence pair, find the next most related pair. The calculations will be similar to the FM method above and the distance between DE and C, $n_{DE,C} = 19$, will be the shortest one. Because we are assuming an equal rate of change in each branch of the tree, there will be two equal length branches, one including D and E and passing to a common node for C and DE, and a second from the common node to C. Some simple arithmetic gives $c = 19/2 = 9.5$ and $g = 9.5 - 5 = 4.5$ (Fig. 6.17).

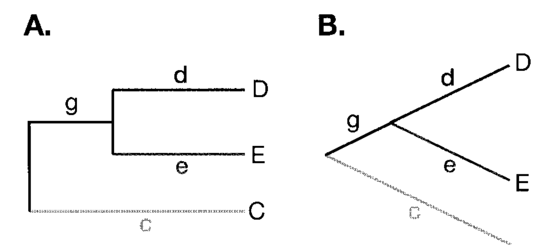


Figure 6.17. Inclusion of third sequence for calculation of branch lengths by UPGMA method.

3. With CDE now being treated as a composite trio of sequences, the next closest pair is A and B, giving an estimate of the distance between them and a common node in the tree of $a = b = n_{AB}/2 = 11$ (Fig. 6.18).

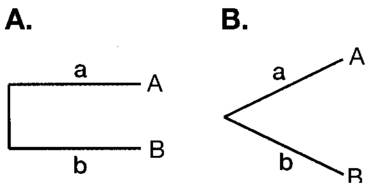


Figure 6.18. Inclusion of fourth and fifth sequences in UPGMA tree.

4. The final calculation is to take the average distance between the two composite sets of sequences CDE and AB. The average of n_{AC}, n_{AD}, n_{AE}, n_{BC}, n_{BD}, and $n_{BE} = 39 + 39 + 41 + 41 + 41 + 43 = 40.7$. One half of this distance $40.7/2 = 20.35$ is included in the part of the tree that goes from the root to CDE, and the other half goes from the root to AB. Note also that the presence of the root breaks the branch between AB and CDE, previously denoted f in this example, into two components $f1$ and $f2$. Hence, $f2 + g + d = 20.35$, $f2 = 20.35 - 4.5 - 5 = 10.85$, and $f1 + a = 20.35$, $f1 = 20.35 - 11 = 9.35$ (Fig. 6.19).

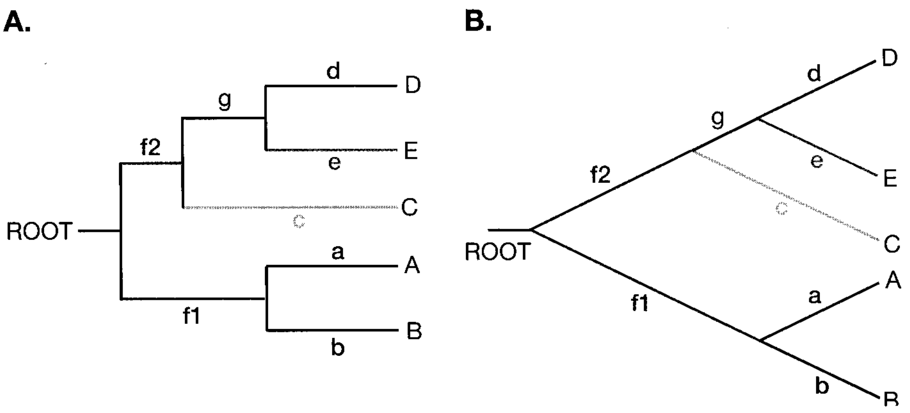


Figure 6.19. Final UPGMA rooted tree for five sequences.

The UPGMA method can lead to an erroneous tree if the rates of mutation in the branches of the tree are not uniform (Li and Graur 1991; Li 1997).

Choosing an Outgroup

If we have independently obtained information that certain sequences are more distantly related, a procedure may be followed which ensures that those sequences are added last to the tree and are closest to the root. This modification can improve the prediction of trees by the above methods by forcing the addition of the outgroup at a later stage in the procedure. One or more sequences of this type are referred to as an outgroup. Suppose, for example, that sequences A and B are from species that are known to have separated from the others at an early evolutionary time based on the fossil record. A and B may then be treated as an outgroup. Choosing one or more outgroups with the distance method can also assist with localization of the tree root (Swofford et al. 1996). The root will be placed between the outgroup and the node that connects the rest of the sequences. It is important that the sequence of the outgroup be closely related to the rest of the sequences, but also that there are significantly more differences between the outgroup and the other sequences than there are among the other sequences themselves. Choosing too distant a sequence as the outgroup may lead to incorrect tree predictions due to the more random nature of the differences between the distant outgroup and the other sequences (Li and Graur 1991; Li 1997). Multiple sequence changes at each site are more possible, and there has been more time for complex genetic rearrangements. For the same reason, using sequences that are too different in the distance method of phylogenetic prediction can lead to errors (Swofford et al. 1996). As the number of differences increases, the history of sequence changes at each site becomes more and more complex, and therefore much more difficult to predict. In choosing an outgroup, one is assuming that the evolutionary history of the gene under study is the same as that provided by the external information. If this assumption is incorrect, such as if horizontal gene transfer has occurred, an incorrect analysis could result.

Converting Sequence Similarity to Distance Scores

For determining phylogenetic relationships among a group of sequences, it is necessary to know the distances between the sequences. The majority of the available sequence alignments determine degree of similarity between sequences rather than distances. For simple scoring systems, similarity is a measure of the number of sequence positions that match in an alignment, whereas distance is the number of positions that are different and that must be changed to convert one sequence into the other. This difference reflects the number of changes that occurred since the sequences diverged from a common ancestor.

As outlined in Chapter 3, similarity methods provide an alignment score, and the significance of this score can be quite reliably calculated based on the probability that a score between unrelated sequences could achieve that score. What is needed is a way to convert such a score to a distance equivalent so that the appropriate phylogenetic analysis can be performed. A simple method, described and used above, is to count the number of different sequence pairs in an alignment. Another method is to convert the similarity score between two sequences to a normalized measure of similarity that varies from 0 for no similarity to 1 for full similarity. The distance can then be readily calculated.

Feng and Doolittle (1996) describe a method for calculating such a normalized score between a pair of aligned sequences. They calculate the similarity score between two sequences S_{real} for a given scoring matrix and gap penalty using a Needleman-Wunsch alignment algorithm (see Chapter 3). They then shuffle both sequences many times, align pairs of shuffled sequences using the same scoring system, and obtain a background average score S_{rand} for unrelated sequences. Finally, each sequence is aligned with itself to give a maximum score that could be obtained in an alignment of two identical sequences with the scoring system used, and the average of these two scores, S_{ident}, is calculated. The normalized similarity score S between the proteins is then given by

$$S = (S_{\mathrm{real}} - S_{\mathrm{rand}})/(S_{\mathrm{ident}} - S_{\mathrm{rand}}) \qquad (1)$$

A different method for calculating $\mathrm{S}_{\mathrm{rand}}$ from the scoring matrix, amino acid composition, and number of gaps in a multiple sequence alignment is also given (Feng and Doolittle 1996).

If, instead, a local alignment based on the Smith-Waterman algorithm is obtained (see Chapter 3), then the statistics of local similarity scores can be used. If λ and K have been calculated for a given scoring matrix and gap penalty combination, the standardized score of an alignment of score S_{rand} is given by

$$S' = \lambda S_{\mathrm{rand}} - \log Kmn \qquad (2)$$

where m and n are the sequence lengths. Recall that S' gives approximate probability of a higher score by $e^{-S'}$ (see Chapter 3, p. 109). A conservative value of 5 for S' corresponds to a probability of 7×10^{-3}. A value of S_{rand} is then given by

$$S_{\mathrm{rand}}(p = 0.007) = 1 / \lambda\,(5 + \log Kmn) \qquad (3)$$

An expected value for S_{ident}, $S_{\mathrm{ident(calc)}}$, is provided by the scoring matrix as the score for a match of identical amino acids (the scores along the diagonal of the log odds form of the amino acid substitution matrix) averaged over amino acid composition for the matrix. If s_{ii} is the score for a match and p_i is the proportion of each amino acid, the predicted score for an alignment of sequences of length m and n, $S_{\mathrm{ident(calc)}}$, where n is the length of the shorter sequence, is given by

$$S_{\mathrm{ident(calc)}} = n \sum_{i=1}^{20} p_i s_{ii} \qquad (4)$$

where $\Sigma\, p_i = 1$. For the PAM250 matrix, the average expected score for a matched pair of identical amino acids is 4.95. Subtracting S_{rand} from this value is not appropriate because the score is not a local alignment score but a global one that grows proportional to sequence length. With the above changes, Equation 1 becomes

$$S = (S_{\mathrm{real}} - S_{\mathrm{rand}(p\,=\,0.007)}\,) / S_{\mathrm{ident(calc)}} \qquad (5)$$

Once the similarity score S has been obtained, it is tempting to calculate the distance between the sequences as $1 - S$. Recall that a simple model of amino acid substitutions is a constant probability of change per site per unit of evolutionary time. Accordingly, some of the observed substitutions in a sequence alignment represent a single amino acid change between the two sequences, but others represent two or more sequential changes. The model predicts that the expected number of 0, 1, 2, . . . substitutions is expected to follow the Poisson distribution, where D is the average number of substitutions. The calculated probability of zero changes is e^{-D}. The probability of one or more changes, which corresponds to S, is then given by $1 - e^{-D}$ such that

$$S = 1 - e^{-D} \tag{6}$$

Taking logarithms of both sides and rearranging then gives

$$D = -\log(S) \tag{7}$$

which is used to calculate D.

Example: Distance Calculation

Two sequences of length 250 have an alignment score of 700, using the PAM250 scoring matrix and gap penalties of −12, −2, which are small enough to give a long but local alignment score, then λ = 0.145 and K = 0.012 (Altschul and Gish 1996). Then $S_{\text{rand}(p = 0.007)} = 1 / 0.145\ (5 + \log 0.012 \times 250 \times 250) = 80$ and $S_{\text{ident(calc)}} = 4.95 \times 250 = 1238$. Then, $S = (700 - 80) / 1238 = 0.50$, and $D = -\log 0.50 = \log 2 = 0.69$.

There are some additional points to make about the above procedure for calculating genetic distance from similarity scores:

1. Use of scoring matrices that are based on an evolutionary model are much preferred to matrices that are based on some other criterion. The Dayhoff PAM matrices meet this criterion but are based on a small data set. A more recent set of PAM matrices (Jones et al. 1992) discussed in Chapter 3 are based on a much larger data set and are based on the same evolutionary model as the Dayhoff matrices.
2. A scoring matrix that models the amino acid substitutions expected for a particular distance should be used. The PAM250 matrix models a separation giving only a remaining level 20% similarity. In the above example, the alignment should be rescored using the log odds PAM80 matrices, which model the expected substitution proteins that are 50% similar, and a better alignment score may be obtained. Suitable gap penalties will have to be found by trial and error, and statistical parameters will be calculated as described above. One must also be sure that the scoring system chosen provides a local alignment by demonstrating a logarithmic dependence of the growth of the alignment score on sequence length.
3. Note that Equation 7 provides an estimate of distance based on the observed similarity. The relationship only holds for sequences that are 50% or more similar. Beyond that point, so many multiple substitutions are possible that the distance essentially becomes 1.
4. When Feng and Doolittle perform distance calculations, they use multiple sequence alignments to assess the changes that occur in a family of related proteins. This method is a large improvement over aligning sequence pairs because

the presumed evolutionary changes can be seen in perspective of a whole related family of proteins. However, using multiple sequence alignment presents a brand new set of challenges that are discussed in Chapter 4.

The following sections describe two entirely different approaches for determining the evolutionary distance between related sequences.

Correction of Distances between Nucleic Acid Sequences for Multiple Changes and Reversions

In the above examples, the assumption is made that each observed sequence change represents a single mutational event. This assumption may be reasonable for sequences that are very much alike, but as the number of observed changes increases, the chance that two or more changes actually occurred at the same site and that the same site changed in both sequences increases. Some of the types of changes that may have occurred are illustrated in Figure 6.20. Note that of all the possible changes, only certain classes shown cause sequence variations.

In the PAM model of evolutionary change described in Chapter 3, such multiple evolutionary changes and reversions are taken into account for a fixed period of evolutionary time called 1 PAM, where 1 PAM roughly equals 10 million years (my). Such tables provide a way to score a sequence alignment by taking into account all possible changes that may have occurred. The PAM table is chosen that provides the highest log odds score between two sequences, and the PAM value of this table then provides a measure of the evolutionary distance between the sequences.

There are several models of evolutionary change of increasing complexity for correcting for the likelihood of multiple mutations and reversions in nucleic acid sequences. These models use a normalized distance measurement that is the average degree of change per length of aligned sequence. For example, in the 20-amino-acid-long sequence alignment given above, there are three changes between sequences A and B. Hence, $d_{AB} = n_{AB} / N = 3/20 = 0.15$.

The simplest model, called the Jukes-Cantor model, is that there is the same probability of change at each sequence position, and that once a mutation has occurred, that position is also just as likely to change again. The model also assumes that each base will eventually have the same frequency in DNA sequences (0.25) once equilibrium has been reached. It may be shown (Li and Graur 1991; Li 1997) that the average number of substitutions per site K_{AB} between two sequences A and B by this model is given by

$$K_{AB} = -3/4 \log_e [1 - 4/3\, d_{AB}] \qquad (8)$$

Thus, K_{AB} in the above example is $K_{AB} = -3/4 \log_e [1 - (4/3 \times 0.15)] = 0.17$, which is slightly greater than the observed number of changes (0.15) to compensate for some mutations that may have reverted. For more different sequences, such as A and D ($d_{AD} = 8/20 = 0.4$), the number of substitutions will be relatively higher than the observed number of changes. $K_{AD} = -3/4 \log_e [1 - (4/3 \times 0.4) = 0.57]$. Hence, the difference between the estimated and observed substitution rates will increase as the number of observed substitutions increases.

The Jukes-Cantor model has been modified to take into account unequal base frequencies (Swofford et al. 1996), which may be calculated from the multiple sequence alignment of the sequences.

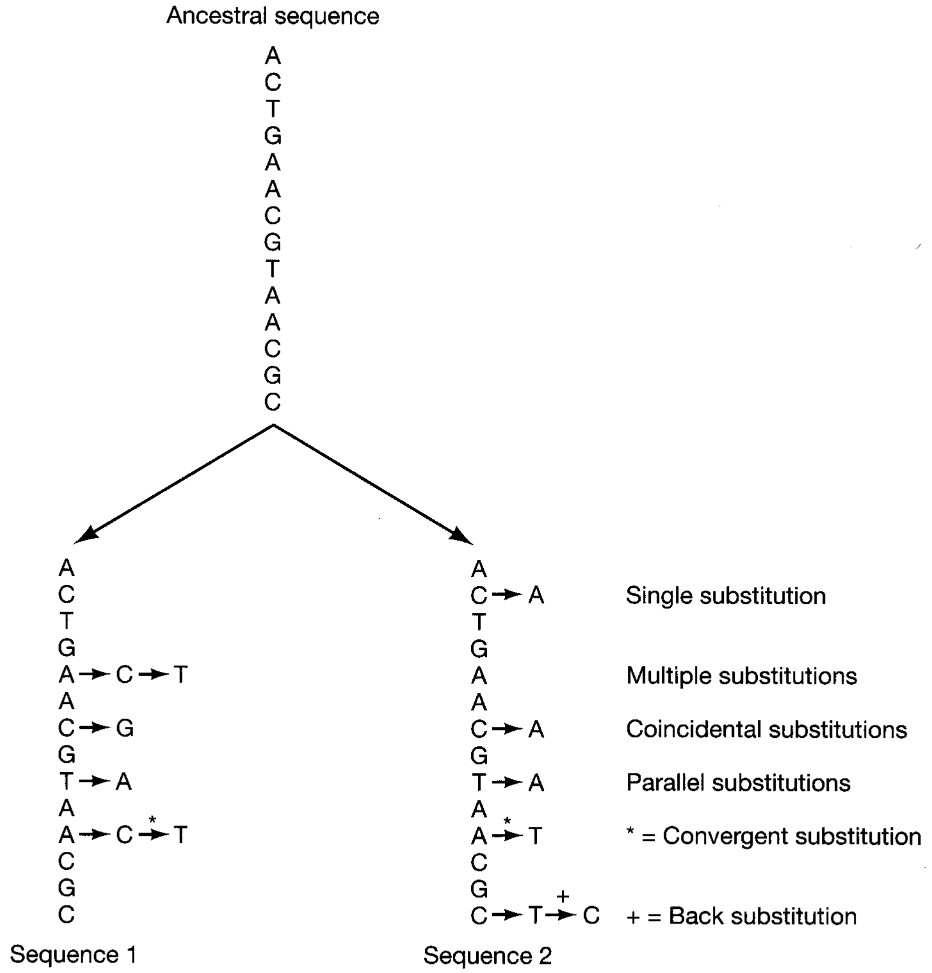


Figure 6.20. Types of mutational changes in nucleic acid sequences that have diverged during evolution. Note that the observed sequence changes between these homologous sequences represent only a fraction of the actual number of sequence variations that may have occurred during evolution and that multiple changes may have occurred at many sites. (Redrawn, with permission, from Li and Graur 1991 [copyright Sinauer Associates].)

$$K_{AB} = -B \log_e [1 - d_{AB}/B] \qquad (9)$$

where B is given by $B = 1 - (f_A{}^2 + f_G{}^2 + f_C{}^2 + f_T{}^2)$ and f_A is the frequency of A in the set of sequences, etc.

A slightly more complex model of change, the so-called Kimura two-parameter model (Kimura 1980), assumes that transition mutations should occur more often than transversions. However, there are four ways of obtaining a transition mutation A ↔ G and C ↔ T, but eight ways of making transversions, A ↔ C, A ↔ T, G ↔ T, and G ↔ C. Thus, in general, transversions can more readily be produced by multiple changes than transitions, and the frequency of each should be adjusted separately. This model also assumes that the eventual frequency of each base in the two sequences will be 1/4. In this case, it is necessary to calculate the proportion of transition and transversion mutations between two sequences. If the frequencies of transitions and transversions between two sequences A and B are $d_{\text{ABtransition}}$ and $d_{\text{ABtransversion}}$, respectively, if $a = 1 / (1-2d_{\text{ABtransition}} - d_{\text{ABtransversion}})$

and $b = 1 / (1-2d_{\text{ABtransversion}})$, and if the basic mutation rate to transitions and transversions is the same, the number of substitutions per site K_{AB} (Li and Graur 1991) is given by

$$K_{AB} = 1/2 \log_e (a) + 1/4 \log_e (b) \qquad (10)$$

For example, suppose that between two 20-nucleotide-long aligned sequences there are six transitions and two transversions, then $a = 1 / (1 - 2 \times 0.3 - 0.1) = 3.33$, $b = 1 / (1 - 2 \times 0.1) = 1.25$, and $K_{AB} = 1/2 \log_e (3.33) + 1/4 \log_e (1.25) = 0.66$. For comparison, by the Jukes-Cantor model, $K_{AB} = -3/4 \log_e [1 - 4/3 \times 8/20] = 0.57$. The larger predicted distance between A and B in the Kimura two-parameter model is due to the greater number of sequence changes in this model that could have given rise to the two observed transversion mutations.

The Jukes-Cantor and Kimura two-parameter models can be modified to take into account variations in the rates of mutation at different sites in the sequence alignment (see Swofford et al. 1996, p. 436), and there is also a Kimura three-parameter model that distinguishes between A ↔ T / G ↔ C transversions with A ↔ C / G ↔ T transversions. These various models are used in the distance methods for phylogenetic construction described above.

For distance calculations between sequences, these base-change models provide ways to improve estimates of the average mutation rate between sequences. They have less effect on phylogenetic predictions of closely related sequences and of the tree branch lengths, but a stronger effect on the more distantly related sequences.

Comparison of Protein Sequences and Protein-encoding Genes

One of the commonest types of phylogenetic comparisons made by biologists is to perform a multiple sequence alignment of a set of proteins using the BLOSUM50 or BLOSUM62 scoring matrix and then to design a phylogenetic tree using the neighbor-joining method. The fraction of sequence positions in an alignment that match provides a similarity score. Ambiguous matches and gaps may also be included in the scoring system for similarity. The distance, 1 minus the similarity score, is calculated and used to produce a tree. CLUSTALW and other programs described in Chapter 4 provide both an alignment and a tree.

Using amino acid variations for phylogenetic predictions offers several advantages. Amino acids confer structure and function to proteins. The order of variations in the tree may therefore provide information concerning the influence of the amino acids on function and of mutations associated with conservation of function and others with changes in function. The difficulty of using the above methods with protein sequences is that, in many cases, no evolutionary model of protein sequence variation is being used. Some amino acid substitutions are much more rare than others and should therefore reflect a longer evolutionary interval. Therefore, treating the substitutions equally may not provide the best phylogenetic prediction.

Another method for circumventing this problem is to use PAM scoring tables. Recall that as evolutionary distance between proteins increases, the expected pattern of amino acid changes varies. Rarer substitutions come into play, and the rate of increase of other changes with increasing time slows down. The Dayhoff PAM amino acid scoring matrices were designed to predict the expected substitutions for proteins separated by different evolutionary distances. The PAM score of the matrix that provides the best alignment score between two sequences reflects the evolutionary separation of the proteins, a distance of 1

PAM being approximately 10 my. Some phylogenetic programs use these original Dayhoff PAM tables. Another updated set of protein PAM tables based on changes in 40-fold more proteins (the PAM250 equivalent is called PET91) is also available (Jones et al. 1992). Some phylogenetic prediction methods use these PAM tables.

The PAM tables have been criticized for failure to take the mutational origin of amino acid changes into account. Although useful for analyzing amino acid variation, they do not allow for the multiple mutations required for some amino acid changes (see Chapter 3, p. 83). Amino acid variation arises through mutation and natural selection acting on DNA sequences. Some amino acid changes require several mutations in codons and should therefore be more rare than amino acid mutations, which require only one mutation in a codon.

Another method for comparing protein sequences is to assess the number of nucleic acid changes that are likely to generate the amino acid differences. In the original Fitch-Margoliash method, when only amino acid sequences were available, the distance between an amino acid pair was chosen to be the minimum number of base changes that would be required to change from a codon for the first amino acid into a codon for the second.

With the availability of the cDNA sequences that encode proteins, cDNA sequences may be compared instead of the amino acid sequences of the encoded proteins. Distance methods may be applied directly to the DNA sequence after the number of different positions in the sequences has been determined. If the protein sequences are very similar, most of the changes that will be observed are silent changes that do not change the amino acid and should provide an accurate representation of the phylogenetic history without the complications of evolutionary selection. However, as the amount of variation increases, the number of silent changes will increase and multiple mutations at some of these sites will occur, whereas at other sites, other more rare types of changes will appear. It is very difficult to make accurate predictions when faced with such variation in the rate of change at different sites. One method around this difficulty is to analyze changes in only the first and second base positions in each codon, ignoring the third position, which is the source of most silent mutations (Swofford et al. 1996). A comparison of nucleic acid sequences that encode proteins for mutations that either (1) change the amino acid or (2) do not change the amino acid may be made. Once these types of changes have been distinguished, phylogenetic predictions based on only one of them may be made.

A final type of correction that may be made to phylogenetic predictions is for the increase in multiple substitutions as the evolutionary distance between protein expected sequences increases. Although use of the PAM matrices provides this type of correction, another way is to adapt the Jukes-Cantor model for nucleic acid sequences to protein sequences. The correction to the distance is given by Equation 9, where $B = 19/20$ for the assumption of equal amino acid representation and $B = 1 - \Sigma f_{\mathrm{aa}i}$ for unequal representation of the amino acids, where $f_{\mathrm{aa}i}$ is the frequency of amino acid i, and the sum is taken over all 20 amino acids. The second representation is, of course, much preferred, since amino acid frequencies in proteins vary.

Another correction that may be applied to protein distances is due to Kimura (1983). This correction is based on the Dayhoff PAM model of amino acid substitution. If K is the corrected distance and D the observed distance (number of exact matches between two sequences divided by total number of matched residues in alignment), then

$$K = -\ln(1 - D - 0.2\,D^2) \qquad (11)$$

This formula may be used up to values of $D = 0.75$. Above this value, tables based on the Dayhoff PAM model at these distances are used. This correction is applied by

CLUSTALW, a commonly used program for multiple sequence alignment and phylogenetic analysis (Higgins et al. 1996).

Comparison of Open Reading Frames by Distance Methods

When nucleic acid sequences that encode proteins first became available, the appearance of synonymous substitutions that do not change the amino acid (silent changes) and nonsynonymous substitutions (replacement changes) that do change the amino acid was analyzed. Separate analyses of these two kinds of substitutions can help remove site-to-site variation in more closely related sequences and background noise of silent mutations in more distantly related sequences (Swofford et al. 1996).

One method of estimating the rates of synonymous and nonsynonymous mutations (Li et al. 1985; Li and Graur 1991; Li 1997) employs the following steps:

1. The fraction of substitutions at each codon position that can give rise to synonymous substitutions and the fraction that can give rise to nonsynonymous substitutions are counted. The first two positions of most codons count as two nonsynonymous sites because the amino acid will change regardless of the substitution. Similarly, many third-codon substitutions are synonymous. Other sites contribute synonymous and nonsynonymous substitutions. The total number of each of these two possible substitutions is determined for each sequence, and the average of these two values for the two sequences is then calculated. N_{syn} is the average number of synonymous sites and N_{nonsyn} is the average number of nonsynonymous sites in the two sequences.
2. Each pair of codons in the alignment is then compared to classify nucleotide differences into synonymous and nonsynonymous types. A single base difference can readily be designated as synonymous or nonsynonymous. When the codons differ by more than one substitution, all of the possible pathways of sequence change must be considered, and the number of synonymous and nonsynonymous changes in each pathway is identified. The average of each type of change in the two pathways is then calculated. Weights derived from the frequency of these pathways for known codon pairs may be used to derive this average, or else the pathways may be weighted equally. These calculations give the number of synonymous differences M_{syn} and the number of nonsynonymous differences M_{nonsyn} between the sequences.
3. The fraction of synonymous differences per synonymous site ($f_{\text{syn}} = N_{\text{syn}} / M_{\text{syn}}$) and the fraction of nonsynonymous differences per nonsynonymous site ($f_{\text{nonsyn}} = N_{\text{nonsyn}} / N_{\text{nonsyn}}$) are calculated. These fractions may then be corrected for the effect of multiple changes at the same site by the Jukes-Cantor formula (Eq. 8) or by some alternative method.

An alternative method for estimating synonymous and nonsynonymous substitutions (Li et al. 1985; Li and Graur 1991; Li 1993, 1997) is to classify each nucleotide position in the coding sequences as nondegenerate, twofold degenerate, or fourfold degenerate. The Genetics Computer Group program DIVERGE uses this method. A site is nondegenerate if all possible changes at this site are nonsynonymous, twofold degenerate if one of the three possible changes is synonymous, and fourfold degenerate if all possible changes are synonymous. For simplification, the third position of isoleucine codons (ATA, ATC, and ATT in the universal code) is treated as a twofold degenerate site even though in reality it is threefold degenerate. The number of each type of site in each of the two sequences is calculated and the average values for the two sequences are calculated. Each pair of codons in the sequence alignment is then compared to classify nucleotide differences as to type of site (nondegenerate, twofold degenerate, or fourfold degenerate) and as to whether the change is a transition or a transversion.

Calculation of Nonsynonymous and Synonymous Changes

To calculate these values, note that by definition all substitutions at nondegenerate sites are nonsynonymous, and all substitutions at fourfold degenerate sites are synonymous. At twofold degenerate sites, transitions nearly always produce synonymous changes, and transversions nearly always produce nonsynonymous changes. Hence, counting transitions and transversions at these sites provides a nearly exact count of the number of synonymous and nonsynonymous substitutions, respectively. One exception to this scoring scheme in the universal genetic code is that one type of transversion in the first position of the arginine codons produces a synonymous change, whereas the other transversion and the transition produce a synonymous change. Another exception is in the last position of the three isoleucine codons. When the codons differ by more than one substitution, a method similar to that described above is used to evaluate each possible pathway for changing one codon into the other, and the average of each type of change in the pathways is then calculated.

The scored codon differences are then used to calculate the proportions of each type of site that are transitions or transversions. The proportion of synonymous substitutions per synonymous site and the corresponding proportion for transversions may then be calculated. The two-parameter model of Kimura may be used to correct for multiple mutations and for differences between rates of transitions and transversions before these calculations are performed.

Example of Distance Analysis: Using the PHYLIP Programs DNADIST and FITCH (Fitch-Margoliash Distance Method)

A set of aligned DNA sequences was converted to the PHYLIP format and placed in a text file called infile in the same folder/directory as the programs (Fig. 6.21A). READSEQ may be used to produce a file with this format from a multiple sequence alignment. Note the required spacing of the sequences including spaces for a sequence name at the start of each row of sequence, and note that line 1 includes two numbers giving the number of sequences and the length of the alignment. Note also the presence of ambiguous sequence characters that are recognized appropriately by the program. Longer sequence alignments may be continued in additional blocks without the identifying names.

DNADIST was invoked, the program automatically read the infile, and after setting various options on a menu, an outfile was produced (Fig. 6.21B). This file was edited to remove all but the distance matrix shown. Note the required number on line 1 giving the number of taxa or sequences. Each distance is given twice as a mirror image about the upper-right to lower-left diagonal.

The predicted unrooted tree is given in the outfile and the treefile by the FITCH program. The average percent standard deviation of the predicted intersequence distance was 14, and 990 trees were analyzed to produce this result. The treefile was used as input to the program DRAWTREE, and shown in Figure 6.21C.

A. Sequences in Phylip format

```
  20    60
MACHIERH     AACNGGCCTT CTACTAGCCA TACACTACAC CGCAGACACC ACCCTAGCCT TTTCATCTGT
CIRCUS       AACTGGCCTN CTACTAGCAA CACACTATTC CGCAGACACT ACCCTGGCTT TCTCATCCGT
LOPHICTI     AACTGGCCTC CTACTGGCCA TGCACTACAC CGCAGACACA TCACTAGCCT TCTCGTCCGT
AQUILA       AACCGGCCTC CTATTAGCCA TACACTACAC GGCAGACACC ACCCTAGCCT TCTCATCCGT
ACCIPITE     AACCGGCCTC CTCCTAGCAA TACACTACAC CGAAGACACC ACCCTAGCCT TTTCATCAGT
BUTASTUS     AACCGGCCTC CTCCTAGCAA TACACTACAC CGCAGACACC ACCCTAGCCT TTTCATCAGT
HAERAETU     AACCGGCCTC CTACTAGCCA TGCACTACAC CGCAGACACC ACCCTAGCCT TCTCGTCCGT
```

B. DNA distances.

```
  20
MACHIERH    0.0000  0.1739  0.1705  0.0899  0.0899  0.0711  0.0899  0.1496 0.1292  0.1705  0.10
0.1292  0.1496
CIRCUS      0.1739  0.0000  0.2373  0.1921  0.2144  0.1921  0.1921  0.1292 0.1496  0.1496  0.21
0.2144  0.2853
LOPHICTI    0.1705  0.2373  0.0000  0.1674  0.2326  0.2102  0.0883  0.1885 0.1674  0.2557  0.18
0.1674  0.1468
AQUILA      0.0899  0.1921  0.1674  0.0000  0.1268  0.1073  0.0698  0.1268 0.1468  0.1885  0.08
0.0698  0.1674
ACCIPITE    0.0899  0.2144  0.2326  0.1268  0.0000  0.0169  0.1268  0.1468 0.1268  0.1674  0.14
0.1885  0.2326
BUTASTUS    0.0711  0.1921  0.2102  0.1073  0.0169  0.0000  0.1073  0.1268 0.1073  0.1468  0.12
0.1674  0.2102
HAERAETU    0.0899  0.1921  0.0883  0.0698  0.1268  0.1073  0.0000  0.1268 0.1073  0.1674  0.08
0.1268  0.1468
ELANUS      0.1496  0.2853  0.1468  0.1674  0.2326  0.2102  0.1468  0.2102 0.2326  0.2795  0.21
0.1268  0.0000
```

C. Fitch tree

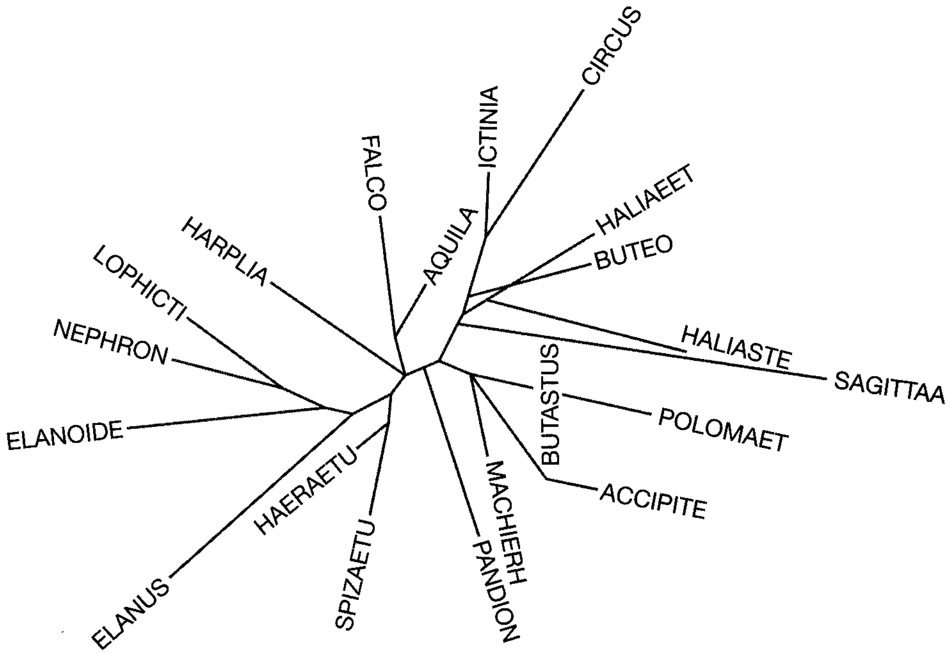


Figure 6.21. Tree predicted by FITCH (Fitch-Margoliash distance method) for the DNA sequences given in the example above.

THE MAXIMUM LIKELIHOOD APPROACH

This method uses probability calculations to find a tree that best accounts for the variation in a set of sequences. The method is similar to the maximum parsimony method in that the analysis is performed on each column of a multiple sequence alignment. All possible trees are considered. Hence, the method is only feasible for a small number of sequences. For each tree, the number of sequence changes or mutations that may have occurred to give the sequence variation is considered. Because the rate of appearance of new mutations is very small, the more mutations needed to fit a tree to the data, the less likely that tree (Felsenstein 1981). The maximum likelihood method resembles the maximum parsimony method in that trees with the least number of changes will be the most likely. However, the maximum likelihood method presents an additional opportunity to evaluate trees with variations in mutation rates in different lineages, and to use explicit evolutionary models such as the Jukes-Cantor and Kimura models described in the above section with allowances for variations in base composition. Thus, the method can be used to explore relationships among more diverse sequences, conditions that are not well handled by maximum parsimony methods. The main disadvantage of maximum likelihood methods is that they are computationally intense. However, with faster computers, the maximum likelihood method is seeing wider use and is being used for more complex models of evolution (Schadt et al. 1998). Maximum likelihood has also been used for an analysis of mutations in overlapping reading frames in viruses (Hein and Støvlbæk 1996). PAUP version 4 can be used to perform a maximum likelihood analysis on DNA sequences. The method has also been applied for changes from one amino acid to another in protein sequences.

PHYLIP includes two programs for this maximum likelihood analysis:

1. DNAML estimates phylogenies from nucleotide sequences by the maximum likelihood method, allowing for variable frequencies of the four nucleotides, for unequal rates of transitions and transversions, and for different rates of change in different categories of sites, as specified by the program.
2. DNAMLK estimates phylogenies from nucleotide sequences by the maximum likelihood method in the same manner as DNAML, but assumes a molecular clock.

One starts with an evolutionary model of sequence change that provides estimates of rates of substitution of one base for another (transitions and transversions) in a set of nucleic acid sequences, as illustrated in Table 6.5. The rates of all possible substitutions are chosen so that the base composition remains the same. The set of sequences is then aligned, and the substitutions in each column are examined for their fit to a set of trees that describe possible phylogenetic relationships among the sequences. Each tree has a certain likelihood based on the series of mutations that are required to give the sequence data. The probability of each tree is simply the product of the mutation rates in each branch of the tree, which itself is the product of the rate of substitution in each branch times the branch length. There are multiple sets of possible base changes within each tree to consider. For each column in the aligned sequences, the probability of each set of changes is found and the probabilities are then added to produce a combined probability that a given tree will produce that column in the alignment. A simple example of this approach is shown in Figure 6.22. Once all positions in the sequence alignment have been examined, the likelihoods given by each column in the alignment for each tree are multiplied to give the likelihood of the tree. Because these likelihoods are very small numbers, their logarithms are usually added to give the logarithm likelihood of each tree. The most likely tree given the data is then identified.

Table 6.5. *General model of sequence evolution*

Base	A	C	G	T
A	$-u(a\pi_C+b\pi_G+c\pi_T)$	$ua\pi_C$	$ub\pi_G$	$uc\pi_T$
C	$ug\pi_A$	$-u(g\pi_A+d\pi_G+e\pi_T)$	$ud\pi_G$	$ue\pi_T$
G	$uh\pi_A$	$uj\pi_G$	$-u(h\pi_A+j\pi_G+f\pi_T)$	$uf\pi_T$
T	$ui\pi_A$	$uk\pi_G$	$ul\pi_T$	$-u(i\pi_A+k\pi_G+l\pi_T)$

The table gives rates for any substitution in a nucleic acid sequence or for no substitution at all (the diagonal values). Base frequencies are given by π_A, π_C, π_G, and π_T, the mutation rate by u, and the frequency of change of any base to any other by *a, b, c..,l.* Rates of substitutions in one direction, i.e., A→G, are generally considered to be the same as that in the reverse direction so that $a = g$, $b = h$, etc. In the JC model these frequencies are all equal, and in the Kimura two-parameter there are only two frequencies, one for transitions (α) and the other for transversions (β), and the frequency for transitions is twice that for transversions. PAUP allows these numbers to be varied. This model assumes that changes in a sequence position constitute a Markov process, with each subsequent change depending only on the current base. Furthermore, the model assumes that each base position has the same probability of change in any branch of the tree (Swofford et al. 1996).

SEQUENCE ALIGNMENT BASED ON AN EVOLUTIONARY MODEL

Thorne et al. (1991, 1992) have introduced a method of sequence alignment based on a model (Bishop and Thompson 1986) that predicts the manner in which DNA sequences change during evolution. Although this method has limitations and is only considered by these authors to be preliminary, it will be outlined here because of its relationship to the maximum likelihood method for phylogenetic analysis. The basis of this method is to devise a scheme for introducing substitutions, insertions, and gaps into sequences and to provide a probability that each of these changes occurs over certain periods of evolutionary time. Given each of these predicted changes, the method examines all the possible combinations of mutations to change one sequence into another. One of these combinations will be the most likely one over time. Once this combination has been determined, a sequence alignment and the distance between the sequences will be known. This method is different from the Smith-Waterman local alignment algorithm in identifying the most probable (maximum likelihood probability alignment) based on an evolutionary model of change in sequences, as opposed to a score based on observed substitutions in related proteins and a gap scoring system. The underlying mutational theory is, however, like those used to produce the PAM matrices for predicting changes in DNA and protein sequences.

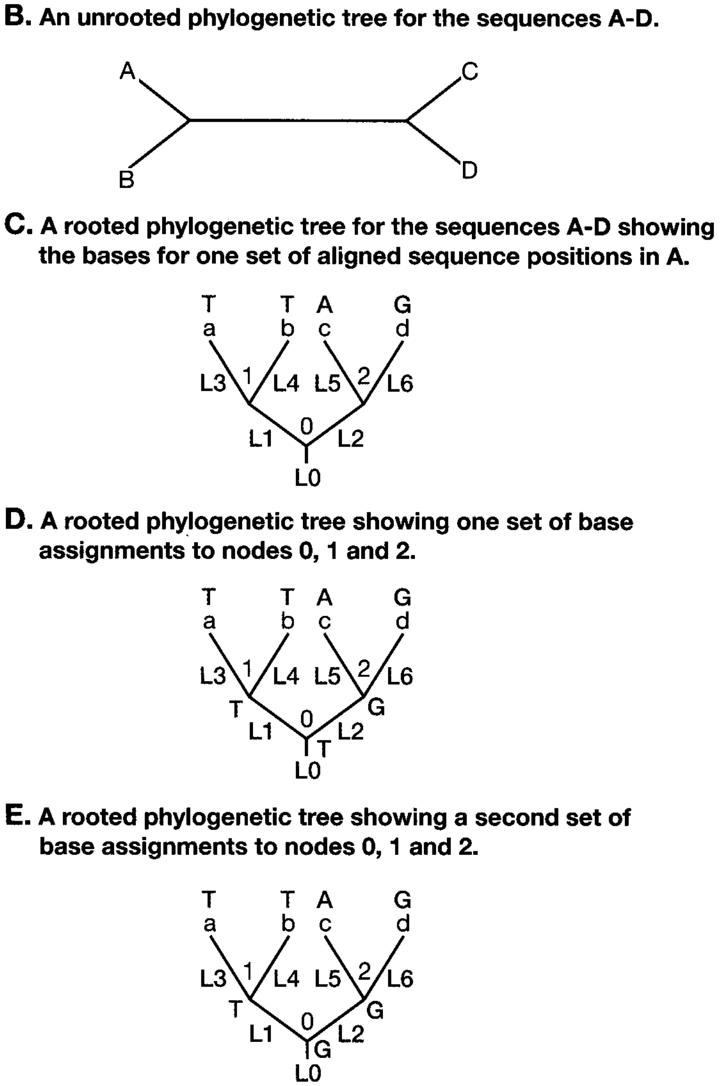
A careful reading of these papers by those interested in evolutionary models of sequence changes is strongly recommended.

Sequences are predicted to change by a Markov process (see Chapter 3 discussion of PAM matrices, p. 78) such that each mutation in the sequence is independent of previous mutations at that site or at other sites. For example, a given nucleotide at any sequence position can mutate into another at the same rate or may not change at all during a period of evolutionary time. This model is very similar to the PAM model of evolutionary change in proteins introduced by Dayhoff and discussed earlier. In the Thorne et al. (1991) model, single insertion–deletion events between any two nucleotides are modeled by a birth–death process that leaves the sequence length roughly the same. Longer insertion–deletion events were modeled in a similar way by considering the sequence to be composed of a set of fragments, and the rate of substitution of these fragments is allowed to vary (Thorne et al. 1992).

A set of transition probabilities for changing from one nucleotide to another or for introducing an insertion or deletion into a sequence is derived mathematically from the evolutionary model. The substitution probabilities are not unlike the substitution probabilities in the protein and DNA PAM matrices. An important difference between the PAM matrices and the transition probabilities is that the insertion/deletion probabilities have been derived from the evolutionary model rather than from the ad hoc gap penalty scoring system (penalty = gap opening penalty + gap extension penalty × length) that is commonly used to produce sequence alignments by dynamic programming. Two algorithms not unlike dynamic programming are then used, one to obtain a sequence alignment and the other to calculate the likelihood that the sequences are related (the likelihood of the

A. Sequences

sequence a	A C G C G T T G G G
sequence b	A C G C G T T G G G
sequence c	A C G C A A T G A A
sequence d	A C A C A G G G A A



B. An unrooted phylogenetic tree for the sequences A-D.

C. A rooted phylogenetic tree for the sequences A-D showing the bases for one set of aligned sequence positions in A.

D. A rooted phylogenetic tree showing one set of base assignments to nodes 0, 1 and 2.

E. A rooted phylogenetic tree showing a second set of base assignments to nodes 0, 1 and 2.

F. L(Tree) = L(Tree1) + L(Tree2) + ... + L(Tree64)

sequences) given the calculated set of parameters. The entries in the scoring matrices are likelihood scores (giving the highest probability of arriving at that position in the scoring matrix by a combination of mutations and gaps) and not a sum of weights for substitutions based on a scoring matrix. To estimate the likelihood of the sequences also requires that the number and types of substitutions, insertions, and deletions be optimized to find the most likely path for changing one sequence into another. This path then provides an indication of the evolutionary distance between the sequences.

Figure 6.22. Maximum likelihood estimation of phylogenetic tree. For the hypothetical sequences shown in *A*, one of three possible unrooted trees is shown in *B*. One column has been set aside for analysis. (*C*) One of five possible rooted derivatives of the unrooted tree is shown. The position of the root is not important since the likelihood of the tree is the same regardless of the root location. This property follows the assumption that the substitutions along each branch are considered to be a Markov chain with reversible steps (Felsenstein 1981). The bases from the marked alignment column are shown on the outer branches of this tree. Also shown are three interior nodes of the tree labeled 0, 1, and 2. The object is to consider every possible base assignment to these three nodes and then to calculate the likelihood of each choice. Since there are four possible bases for each of the three node positions of the tree, there are $4 \times 4 \times 4 = 64$ possible combinations. Also shown on the tree are six likelihood values L1–L5 for the probability of a base change per site along the respective branches of the tree, and a probability L0 for the base at node 0. These probabilities depend on the bases assigned to nodes 0, 1, and 2 and on the resulting types of base substitutions in the particular tree under consideration. The likelihood of a tree with a particular choice of bases at nodes 0, 1, and 2 is given by the product of the probability of the base at node 0 times the product of each of the substitution probabilities, or L(tree) = $\text{L0} \times \text{L1} \times \text{L2} \times \text{L3} \times \text{L4} \times \text{L5} \times \text{L6}$ (Felsenstein 1981). (*D*) A possible tree (tree1) with T assigned to nodes 0 and 1, and G assigned to node 2. L0 will be given by the frequency of T and will have an approximate value of 0.25. L2 will be the probability of a transversion of T to G, and L5 the probability of a transition of G to A in this tree. The remaining likelihoods will have an approximate value of unity with a small adjustment for the possibility that a mutation has occurred and then reverted to the original base so that no substitutions are observed. Assuming that the probabilities of the transition and transversion are 2×10^{-6} and 10^{-6}, respectively, the likelihood of tree1 is approximately $0.25 \times 2 \times 10^{-6} \times 10^{-6} = 5 \times 10^{-13}$. These numbers are usually very small and are therefore handled as logarithms in the computer. (*E*) Another possible arrangement of base assignments in tree2. The likelihood of this tree will take into account the probability of a G to T transversion (L1) and that of a G to A transition (L5). (*F*) The likelihood of the tree in *B* or the tree in *C* is given by the sum of the likelihoods of these two trees. To this sum is added the probability of the other 62 possible arrangements of bases. This calculation is repeated for all other columns in the multiple sequence alignment. The likelihood of the tree given the data in all of the aligned columns, that in the first column, or that in the second, etc., will be the sum of the likelihoods so calculated for each column. Each of the three possible trees for four sequences is then evaluated in this same manner and the one with the highest likelihood score is identified. These calculations can be computationally so intense for a large number of sequences that trees for a fraction of the sequences may first be found. The data for additional sequences will then be sequentially added to refine this initial tree. The procedure may then be repeated with a different starting group of sequences with the hope that the range of trees found will give an indication of the most likely tree (Felsenstein 1981). However, this procedure is not guaranteed to find the optimal tree. Additional calculations are made in the ML method. The probability of each branch in the tree is individually adjusted by a method similar to expectation maximization (see Chapter 3) to maximize the likelihood of the tree while holding the probability of the other branches at a constant value. The rate of evolution of each site or each column in the multiple sequence alignment is also allowed to vary. Otherwise, the method will be biased by sites that do not vary much and the information in variable sites may become lost, a problem shared with the maximum parsimony method. For an average number of mutations *x* over all branches, the number along an individual branch is assumed to vary according to the Poisson distribution $P(n) = e^{-x} x^n / n!$. A continuous variable giving the equivalent probability of observing a given number of changes along a particular branch for various average values of *x* (or a particular mutation rate along that branch) is given by the Γ distribution. These probabilities may then be used in calculations of tree likelihoods (Swofford et al. 1996).

RELIABILITY OF PHYLOGENETIC PREDICTIONS

As discussed earlier in this chapter, phylogenetic analysis of a set of sequences that aligns very well is straightforward because the positions that correspond in the sequences can be readily identified in a multiple sequence alignment of the sequences. The types of changes in the aligned positions or the numbers of changes in the alignments between pairs of sequences then provide a basis for a determination of phylogenetic relationships among the sequences by the above methods of phylogenetic analysis. For sequences that have diverged considerably, a phylogenetic analysis is more challenging. A determination of the sequence changes that have occurred is more difficult because the multiple sequence alignment may not be optimal and because multiple changes may have occurred in the aligned sequence positions. The choice of a suitable multiple sequence alignment method depends on the degree of variation among the sequences, as discussed in Chapter 4. Once a suitable alignment has been found, one may also ask how well the predicted phylogenetic relationships are supported by the data in the multiple sequence alignment.

In the bootstrap method, the data are resampled by randomly choosing vertical columns from the aligned sequences to produce, in effect, a new sequence alignment of the same length. Each column of data may be used more than once and some columns may not be used at all in the new alignment. Trees are then predicted from many of these alignments of resampled sequences (Felsenstein 1988). For branches in the predicted tree topology to be significant, the resampled data sets should frequently (for example, >70%) predict the same branches. Bootstrap analysis is supported by most of the commonly used phylogenetic inference software packages and is commonly used to test tree branch reliability. Another method of testing the reliability of one part of the tree is to collapse two branches into a common node (Maddison and Maddison 1992). The tree length is again evaluated and compared to the original length, and any increase is the decay value. The greater the decay value, the more significant the original branches. In addition to these methods, there are some additional recommendations that increase confidence in a phylogenetic prediction.

One further recommendation is to use at least two of the above methods (maximum parsimony, distance, or maximum likelihood) for the analysis. If two of these methods provide the same prediction, confidence in the prediction is much higher. Another recommendation is to pay careful attention to the evolutionary assumptions and models that are used for both sequence alignment and tree construction (Li and Graur 1991; Swofford et al. 1996; Li 1997).

COMPLICATIONS FROM PHYLOGENETIC ANALYSIS

The above methods provide a further level of sequence analysis by predicting possible evolutionary relationships among a group of related sequences. The methods predict a tree that shows possible ancestral relationships among the sequences. A phylogenetic analysis can be performed on proteins or nucleic acid sequences using any one of the three methods described above, each of which utilizes a different type of algorithm. The reliability of the prediction can also be evaluated.

The traditional use of phylogenetic analysis is to discover evolutionary relationships among species. In such cases, a suitable gene or DNA sequence that shows just enough, but not too much, variation among a group of organisms is selected for phylogenetic analysis. For example, analysis of mitochondrial sequences is used to discover evolutionary relationships among mammals. Two more recent uses of phylogenetic analysis are to analyze gene families and to trace the evolutionary history of specific genes. For example, database similarity searches discussed in Chapter 7 may identify several proteins in a plant genome that are similar to a yeast query protein. From a phylogenetic analysis of the protein family, the plant gene most closely related to the yeast gene and therefore most likely to have the same function can be determined. The prediction can then be evaluated in the laboratory. Tracking the evolutionary history of individual genes in a group of species can reveal which genes have remained in a genome for a long time and which genes have been horizontally transferred between species. Thus, phylogenetic analysis can also contribute to an understanding of genome evolution, as further explored in Chapter 10.

REFERENCES

Altschul S.F. and Gish G. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Barns S.M., Delwiche C.F., Palmer J.D., and Pace N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci.* **93:** 9188–9193.

Bishop M.J. and Thompson E.A. 1986. Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190:** 159–165.

Brown J.R. and Doolittle W.F. 1997. *Archaea* and the procaryotic-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61:** 456–502.

Comeron J.M. and Kreitman M. 1998. The correlation between synonymous and nonsynonymous substitutions in *Drosophila:* Mutation, selection or relaxed constraints? *Genetics* **150:** 767–775.

Doolittle W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284:** 2124–2128.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

———. 1988. Phylogenies from molecular sequences: Inferences and reliability. *Annu. Rev. Genet.* **22:** 521–565.

———. 1989. PHYLIP: Phylogeny inference package (version 3.2). *Cladistics* **5:** 164–166.

———. 1996. Inferring phylogeny from protein sequences by parsimony, distance and likelihood methods. *Methods Enzymol.* **266:** 368–382.

Feng D.F. and Doolittle R.F. 1996. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. *Methods Enzymol.* **266:** 368–382.

Fitch W.M. 1981. A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* **18:** 30–37.

Fitch W.M. and Margoliash E. 1987. Construction of phylogenetic trees. *Science* **155:** 279–284.

Hein J. and Støvlbæk J. 1996. Combined DNA and protein alignment. *Methods Enzymol.* **266:** 402–418.

Henikoff S., Greene E.A., Pietrokovski S., Bork P., Attwood T.K., and Hood L. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278:** 609–614.

Higgins D.G., Thompson J.D., and Gibson T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.

Jin L. and Nei M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* **7:** 82–102.

Jones D.T., Taylor W.R., and Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8:** 275–282.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111–120.

———. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, United Kingdom.

Li W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.

———. 1997. *Molecular evolution.* Sinauer Associates, Sunderland, Massachusetts.

Li W.-H. and Graur D. 1991. *Fundamentals of molecular evolution*, pp. 106–111. Sinauer Associates, Sunderland, Massachusetts.

Li W.-H. and Gu X. 1996. Estimating evolutionary distances between DNA sequences. *Methods Enzymol.* **266:** 449–459.

Li W.-H., Wu C.I., and Luo C.C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2:** 150–174.

Maddison W.P. and Maddison D.R. 1992. MacClade: Analysis of phylogeny and character evolution (version 3). Sinauer Associates, Sunderland, Massachusetts.

Maidak B.L., Cole J.R., Parker C.T., Jr., Garrity G.M., Larsen N., Li B., Lilburn T.G., McCaughey M.J., Olsen G.J., Overbeek R., Pramanik S., Schmidt T.M., Tiedje J.M., and Woese C.R. 1999. A new version of the RDP (ribosomal database project). *Nucleic Acids Res.* **27:** 171–173.

Martin W. 1999. Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. *Bioessays* **21:** 99–104.

Mayr E. 1998. Two empires or three? *Proc. Natl. Acad. Sci.* **95:** 9720–9723.

McDonald J.H. and Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila. Nature* **351:** 652–654.

Miyamoto M.M. and Cracraft J. 1991. *Phylogenetic analysis of DNA sequences.* Oxford University Press, New York.

Nielsen R. and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.

Pearson W.R., Robins G., and Zhang T. 1999. Generalized neighbor-joining: More reliable phylogenetic tree construction. *Mol. Biol. Evol.* **16:** 806–816.

Saitou N. 1996. Reconstruction of gene trees from sequence data. *Methods Enzymol.* **266:** 427–449.

Saitou N. and Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Sankoff D. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **78:** 35–42.

Sattath S. and Tversky A. 1977. Additive similarity trees. *Psychometrika* **42:** 319–345.

Schadt E.E., Sinsheimer J.S., and Lange K. 1998. Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res.* **8:** 222–233.

Snel B., Bork P., and Huynen M.A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.

Swofford D.L., Olsen G.J., Waddell P.J., and Hillis D.M. 1996. Phylogenetic inference. In *Molecular systematics*, 2nd edition (ed. D.M. Hillis et al.), chap. 5, pp. 407–514. Sinauer Associates, Sunderland, Massachusetts.

Tatusov R.L., Koonin E.V., and Lipman D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Thorne J.L., Kishino H., and Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:** 114–134.

———. 1992. Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.* **34:** 3–16.

Woese C.R. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.