Chapter 11

Phylogenetic Inference

David L. Swofford, Gary J. Olsen,
Peter J. Waddell, and David M. Hillis

INTRODUCTION

Inferring phylogenetic relationships from molecular data requires the selection of an appropriate method from the many techniques that have been described. Unfortunately, phylogenetic analysis is frequently treated as a black box into which data are fed and out of which "The Tree" springs. Our goal in this chapter is to provide more than a cursory description of the available analytical methods; rather, we hope to develop a conceptual framework for understanding the theoretical and practical distinctions among alternative methodologies. Phylogenetic analysis of molecular data is in the midst of a remarkable transformation. The most striking theme in this shift is an increased emphasis on the use of methods that are based on models of evolutionary change. Moreover, users of methods that do not require explicit models are now much more likely to incorporate modifications based on reasonable assumptions about the evolutionary process than when the first edition of *Molecular Systematics* appeared only six years ago. We view this trend as a positive one and have reorganized our chapter accordingly.

Regrettably, we cannot accomplish all of the above objectives and at the same time provide an exhaustive review of the voluminous literature on phylogenetic reconstruction; however, Felsenstein (1982, 1988a, 1993) and Hillis et al. (1993a) have presented general reviews of methods for inferring phylogenies. Instead, we will focus on methods that are currently in widespread use or that are likely to be used in the foreseeable future. We will also avoid the temptation to cite every relevant paper, limiting our citations to papers that are either of fundamental importance to the development of a method or that provide the clearest explanations of that method.

As any reader even moderately familiar with the current state of affairs in phylogenetics already knows, debates among proponents of rival methodologies are often intense and sometimes unnecessarily acrimonious. Consequently, we will offer recommendations where we deem them appropriate, but will deliberately avoid taking strong positions on or making controversial assertions about issues where there is room for legitimate disagreement. Instead, we hope to provide sufficient background so that readers will be able to make informed decisions regarding the techniques most appropriate for their own data. Our treatment in this chapter will be limited to the inference of the phylogenetic history of the genes under study. For a variety of reasons, these "gene trees" may fail to reflect the relationships of the organisms from which the genes were sampled. A discussion of these and related issues is presented in Chapter 1; we will not address them further here.

Algorithms versus Optimality Criteria

Inferring a phylogeny is an estimation procedure; we are making a "best estimate" of an evolutionary history based on the incomplete information contained in the data. In the context of molecular systematics, we generally do not have direct information about the past—we only have access to contemporary species and molecules. Because we can postulate evolutionary scenarios by which *any* chosen phylogeny could have produced the observed data, we must have some basis for selecting one or more preferred trees from among the set of possible phylogenies. Phylogenetic inference methods seek to accomplish this goal in one of two ways: (1) by defining a specific sequence of steps (an **algorithm**) that leads to the determination of a tree; or (2) by defining a criterion for comparing alternative phylogenies to one another and deciding which is better (or that they are equally good).

Purely algorithmic methods combine tree inference and the definition of the preferred tree into a single statement. These methods include all forms of pair-group cluster analysis (e.g., UPGMA) and some other distance methods such as neighbor joining (discussed later in this chapter). The methods tend to be computationally fast because they proceed directly toward the final solution without requiring evaluation of large numbers of competing trees.

The second class of methods has two logical steps. The first step is to define an **optimality criterion** (formally described by an **objective function**) for evaluating a given tree—i.e., a score is assigned and subsequently used for comparing one tree to another. The second is to use specific algorithms for computing the value of the objective function and for finding the trees that have the best values according to this criterion (a maximum or minimum value, as appropriate). Thus, the evolutionary assumptions made in the first step are decoupled from the computer science of the second step. The price of this logical clarity is that the methods tend to be much slower than those of the first class, a consequence of having to search for the tree(s) with the best score. For data sets containing more than about 8 to 20 taxa, the search for the best tree is usually not exact (because of the large number of possible solutions), and thus we must add caveats regarding the thoroughness of the search for the optimal tree. These issues are covered in detail below.

It is important to distinguish between the uses of algorithms in the two approaches. In a purely algorithmic method, the algorithm *defines* the tree selection criterion and takes on fundamental importance. In a criterion-based method, however, the algorithms are merely tools used in evaluating the objective function and searching for trees that optimize it.* Because criterion-based methods can assign scores to every tree examined, phylogenies can be ranked in order of preference according to the chosen criterion. This is an enormous advantage over purely algorithmic methods. If a criterion-based method finds that there are thousands or millions of trees that explain the data about equally well, the user of the method

*Actually, the same algorithm may be used in both approaches, albeit for very different goals. For instance, an algorithm used to specify a final tree in a purely algorithmic method may be used to find an initial tree for a criterion-based method (e.g., as a starting point for branch-swapping rearrangement algorithms).

will not be misled into believing that any particular tree is well-specified. On the other hand, when a purely algorithmic method determines a single tree, the user will have no immediate knowledge about the strength of support for that tree. Some workers (e.g., Hedges et al., 1992b) have argued that algorithmic methods can be rescued by using statistical methods such as nonparametric bootstrapping (see the section "Reliability of Inferred Trees," later in this chapter) to assess the confidence in a tree found using an algorithmic method. This position fails to address the criticism that algorithmic methods generally do not address the operational evolutionary assumptions. As an extreme example, consider an algorithm that chooses trees independently of the data, for example by labeling the tips of a maximally asymmetric tree in alphabetic order of the species names. Repeated analyses using different re-samplings of the data will always generate the same tree, leading to the obviously absurd conclusion that the tree is extraordinarily reliable.

Use of Models and Assumptions in Phylogenetics

Although we will deal extensively with specific models of the evolutionary change of molecules, a preliminary discussion of the relevance of models in general is in order at the outset. Phylogenetic inferences are premised on the inheritance of ancestral characteristics, and on the existence of an evolutionary history defined by changes in these characteristics. The stable inheritance of characteristics is mediated by the genome. Differences due to epigenetic or environmental factors do not provide useful phylogenetic information and must be specifically avoided; all characteristics of interest are genetically mediated. Therefore, the data for phylogenetic inference reflect, more or less directly, genomic information. From this reductionistic perspective, a complete evolutionary history is synonymous with an event-by-event accounting of fixed mutations in every genomic lineage of interest. This view of the problem provides a common framework, albeit a purely conceptual one, for analyzing and comparing types of molecular data and analysis techniques.

If a phylogenetic inference method could be based upon a complete knowledge of the evolutionary process, it would be free of systematic error (i.e., if enough data were obtained, the method would consistently obtain the true phylogeny). Even in the absence of such complete knowledge, hypothetical models of the evolutionary process could be used to derive (or otherwise justify) tree inference methods that would be free of systematic error, *if the assumed model were correct*. A variety of inference techniques have been formulated on the basis of explicit evolutionary assumptions. These model-based methods are not necessarily invalidated when one or more of their assumptions is violated—a model does not have to be perfect in order to be useful. That is, although the assumptions may be sufficient to ensure the validity of a technique, under special circumstances they might not all be necessary, and the method may be robust to violation of its assumptions. Furthermore, model assumptions need not be accepted in a vacuum; data can and should be allowed to reject the model if the model is inadequate.

Although almost all methods accept the appropriateness of a tree-like model of evolution (a strong assumption in itself), many commonly used methods of phylogenetic inference are not explicitly based on a set of evolutionary assumptions. However, the lack of stated assumptions does not mean that a method is assumption-free; the assumptions are simply implicit rather than explicit. For example, the widely used method of maximum parsimony does not depend on a precise model, but believing its results does require one to believe that plausible evolutionary scenarios that could cause it to fail have not taken place. It is often argued that it is circular to model character change for the purpose of estimating a phylogeny because we cannot begin to understand the processes of character change without first knowing the tree. We prefer, instead, to think of the problem as one of "reciprocal illumination" (Hennig, 1966): having some idea of the phylogeny is relevant to the development of good models, but ever-improving models can also lead to better phylogenetic inferences. Thus, both classes of methods are useful and important. We see it as unfortunate that some workers, in their zeal to avoid circularity, limit themselves to "model-free" methods that may be more likely to violate their (implicit) assumptions than the methods they reject, for which the assumptions are more explicit.

One assumption implicit in this general view concerns the uniqueness of the genomic lineage. The potential confusion due to lateral gene transfer has received much recent attention. When transfer is common among the lineages of interest, a population genetic analysis (Chapter 10) is most appropriate. Our presentation is appropriate for cases in which interspecies differences are large compared to intraspecific variation.

Definitions of Terms

Most of the analytical techniques that we will discuss result in the inference of an **unrooted tree** or unrooted phylogeny—a phylogeny in which the earliest point in time (the location of the common ancestor) is not identified. (We generally use *tree* and *phylogeny* interchangeably.) Also, biologists often refer to an unrooted tree as a network; however, this usage conflicts with the definition applied to that term by mathematicians and should be avoided (the section "Split Decomposition" uses *network* in the correct sense). When we find it necessary to distinguish between rooted and unrooted phylogenies or trees, we will do so explicitly.

The components of a phylogenetic tree go by a variety of names. The contemporary taxa correspond to **terminal nodes** or tips, also called leaves or external nodes. The branch points within a tree are called **internal nodes**. Nodes are called vertices or points by some authors. The branches connecting (incident to) pairs of nodes are also called edges, links, or segments. We will use the terms **peripheral branches** to refer to branches that end at a tip and **interior branches** (or, in the case of a tree with four terminal nodes, **central branch**) to refer to branches that are not incident to a tip.

If just three branches connect to an internal node, then the node represents a **bifurcation**, or **dichotomy**. If there are more than three branches connected to an internal node, then the node represents a **multifurcation**, or **polytomy**. A tree in which all internal nodes represent bifurcations is said to be binary, fully resolved, or strictly bifurcating. A tree that contains a single internal node is called a **star tree**.

An unrooted, fully resolved tree has T terminal nodes (corresponding to the taxa) and $T - 2$ internal nodes. The tree has $2T - 3$ branches, of which $T - 3$ are interior and T are peripheral. The total number of distinct unrooted, strictly bifurcating, trees for T taxa is

$$B(T) = \prod_{i=3}^{T} (2i - 5) \qquad \textbf{(1)}$$

(Felsenstein, 1978b). Adding a root adds one more internal node and one more interior branch. Since the root can be placed along any of the $2T - 3$ branches, the number of possible rooted trees is increased by a factor of $2T - 3$.

TYPES OF DATA

All of the experimental data gathered by the techniques in this volume fall into one of two broad categories: discrete **characters**, and similarities or distances. A discrete character provides data about an individual species or sequence. Character data are often transformed into similarity or distance values representing quantitative comparisons of two species or sequences; each such measure describes a pairwise relationship. Of the methods discussed in this book, only DNA–DNA hybridization data are collected directly in the form of pairwise distance comparisons. Appropriate distance measures and transformations for DNA–DNA hybridization data are discussed in Chapter 6. Our discussion here focuses on character data.

Discrete character data are those for which a data matrix **X** assigns a **character state** x_{ij} to each taxon i for each character j. Although systematists sometimes disagree about the terminological distinction between *character* and *character state*, we prefer to think of characters as independent variables whose possible values are collections of mutually exclusive character states.

The assumption of independence among characters is common to most character-based methods of analysis. When we can not assume independence, we are forced to take covariances among characters into account, and the computational methods become considerably more complicated. Furthermore, the assumption of independence enables us to treat each position separately in certain time-consuming stages of computational algorithms, thereby allowing problems to be subdivided into a number of much simpler subproblems. (For example, numbers of substitutions can be minimized separately position-by-position and then summed over positions in a parsimony algorithm, or probabilities can be multiplied over positions in a maximum likelihood approach.)

A second assumption required of character data is that the characters be homologous. As articulated in Chapter 1, the concept of homology is complicated by the variety of meanings that have been applied to the term. In general, by *homology* we mean that a character must be defined in such a way that all of the states observed over taxa for that particular character must have been derived, perhaps with modification, from a corresponding state observed in the common ancestor of those taxa. When we are interested in relationships among species rather than among genes, we further restrict this definition to include only orthologous, as opposed to paralogous or xenologous, genes.

In general, character data are either qualitative, in which case the possible states are two or more discrete values; or quantitative, in which case the characters vary continuously and are measured on an interval scale. Qualitative characters may be further subdivided into binary (two possible states) and multistate (three or more possible states). Binary characters typically represent the presence or absence of some item, such as the recognition sequence for a restriction endonuclease at a certain map location (restriction site) or a particular allele at an isozyme locus.

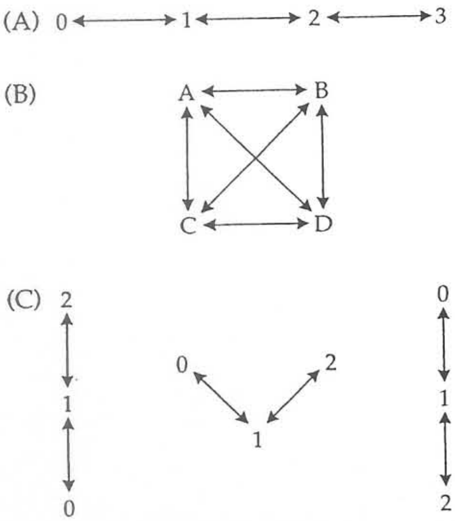


Figure 1 Ordered and unordered characters. (A) Ordered multistate character (transformation between any two states that are not directly connected implies passage through one or more intermediate states). (B) Unordered multistate character (any state can transform directly into any other state). (C) Ordered multistate characters in which the polarity is indicated (the ordering relation is the same in all three cases but the ancestral state differs).

Multistate characters may be **ordered** or **unordered**, depending on whether an ordering relationship is imposed upon the possible states (Figure 1). For example, nucleotide sequence data are generally treated as unordered multistate characters, since there is no *a priori* reason to assume, for instance, that state C is intermediate between states A and G. In the context of phylogenetic analysis, we say that any state is allowed to transform directly into any other state. If, on the other hand, we are willing to make assumptions involving the relationships among the states of a character, we can rank the character states into an ordered series (i.e., a **linearly ordered** character) or a branching diagram (**partially ordered** character or **character-state tree.**) Multistate ordered characters are not commonly encountered in molecular data sets, but they are sometimes used in the analysis of allozyme data.

The concepts of *character order* and *character polarity* should not be confused. The former defines the allowed character-state transformations, whereas the latter refers to the *direction* of character evolution. Estimation of character polarity generally involves an assessment of the observed character state most likely to represent the ancestral condition (i.e., the state found in the most recent common ancestor of the taxa under study). An excellent discussion of character ordering and polarity (in a non-molecular context) can be found in Mabee (1989). We will return to the subject of character polarity in the discussion of parsimony methods.

Quantitative characters are less commonly used as character data in molecular systematics. The prominent exception occurs when polymorphic characters such as allelic isozymes or mtDNA haplotypes are coded as frequencies.

Sequence Data

In principle, the use of sequence data as characters for phylogenetic analysis is straightforward. Given a set of sequences, the characters are represented by corresponding positions (offsets) in the sequences, and the character states are the nucleotide or amino acid residues observed at those positions. For example, if nucleotide A is observed to occur at position 139 in a sequence, "position 139" is the character and "A" is the state assigned to that character. To simplify our exposition, we will usually confine our descriptions to nucleotide sequences unless the distinction is important.

Unfortunately, this simplicity is deceiving. In addition to requiring the use of homologous molecules (see Chapter 1), phylogenetic analysis of sequence data requires **positional homology**. That is, the nucleotides observed at a given position in the taxa under study should all trace their ancestry to a single position that occurred in a common ancestor of those taxa. Except for highly conserved sequences, insertion and deletion events must nearly always be postulated in order to make believable the assumption that nucleotides at corresponding positions in the various sequences are in fact homologs. An **alignment** of the sequences is obtained by inserting **gaps**, which correspond to insertions or deletions, into one or more of the sequences in order to place positions inferred to be homologous into the same column of the data matrix. Alignment is often the most difficult and least understood component of a phylogenetic analysis from sequence data. Methods for alignment are discussed in Chapter 9.

Restriction Endonuclease Data

Restriction endonuclease analysis provides character data in one of two forms, both of which lead to a set of binary characters for each taxon. Ideally, the characters are map locations and character states are presences or absences of the recognition sequences for particular endonucleases at those locations (restriction-site data). However, because the construction of restriction maps is time-consuming (see Chapter 8), some workers simply treat the presence or absence of restriction fragments of a given length as character states (restriction-fragment data).

We do not recommend the use of restriction-fragment data for input to phylogenetic analysis, primarily because these data violate the assumption of independence among characters. If a new site evolves between two preexisting sites, one (longer) fragment disappears and two new (shorter) ones appear. Thus, even though two species may share two of the three restriction sites, they have no fragments in common—a potentially serious source of error. Some authors (e.g., B. Bremer, 1991) have recognized this difficulty and argue that it can be overcome by looking at "enough" fragment data so that each occurrence of this kind of error will be swamped by other data. We are unconvinced by this argument, however, because there is no guarantee that if something is done inappropriately enough times, all will work out in the end (and the amount of systematic error introduced by this shortcut will increase substantially with increasing divergence among the taxa in the analysis). A second and related problem with fragment data is that insertions or deletions are difficult to handle. For example, the insertion of a length of DNA long enough to alter the mobility of the fragment (but not containing a restriction site) requires the worker to assert that a species lacks a fragment found in one or more other species, even though the restriction sites responsible for the fragment are at homologous points on the map (see Chapter 8).

Even when sites are mapped, restriction endonuclease data are problematic for phylogenetic analysis due to the asymmetry in the probabilities of gaining and losing sites. If a particular sequence of six base pairs is only one substitution away from equalling the recognition sequence of a particular endonuclease (a "one-off" site), then given that a substitution occurs within the six-base sequence, only one of the 18 possible substitutions of one base for another will convert the sequence to a restriction site. On the other hand, if the six-base sequence is already a restriction site, then a substitution at any of the six positions will cause the site to be lost. Thus, losing an existing restriction site is much more likely than gaining a site at a particular location. (For more complete discussions, see Templeton, 1983a, 1983b and DeBry and Slade, 1985.) Note that this argument applies only to *particular* sites in the genome; it does not imply a net loss of restriction sites during evolution. Because of these gain–loss asymmetries, special handling may be required for restriction-site data.

Isozyme Data

Allozyme (allelic isozyme) data represent the only type of isozyme data routinely used in phylogenetic analysis (but see Buth, 1984a, and Chapter 4 for a discussion of other data types). These data are usually presented as a three-dimensional array that specifies the frequency of each allele at each locus in each population or taxon.* Two controversial issues confront the researcher attempting to estimate phylogenies from allozyme data. The first concerns whether or not to transform the data to genetic similarities or distances. Probably due more to inertia than anything else, the predominant mode of analysis throughout the 1970s and into the 1980s was to compute a matrix of pairwise similarities or distances between taxa that served as the input to cluster analysis or additive-tree methods. The stereotypical way in which these data were treated tended to retard the development of approaches that made direct use of the character information.

With the development of character-based methods, however, came a second controversy, this one involving the importance of allele frequency information. Some authors (e.g., Mickevich and Johnson, 1976) argued that the presence or absence of an allele was of more fundamental evolutionary importance than was its frequency (which was subject to modification by drift and/or selection), and that frequency information should therefore be discarded. These authors preferred to recast the data into presence/absence form. However, other authors (e.g., Swofford and Berlocher, 1987) have argued that there is no reason to ignore frequency information in analyzing allozyme data.

The earliest attempts to use allozyme characters directly in a phylogenetic analysis generally treated the allele as the character and either its presence/absence (e.g., Mickevich and Johnson, 1976) or its frequency (e.g., Buth, 1979b; Simon, 1979) as the character state. This procedure, however, is open to the same criticism leveled at the use of restriction fragment data: the assumption of independence of characters is violated. Specifically, since the frequencies of the alleles at a locus in a given taxon are constrained to sum to one, if the frequency of one allele increases, the frequency of at least one other allele must decrease. This property leads to problems, for example, when allele-as-character data are subjected to maximum parsimony analysis, where ancestors are often inferred to contain no alleles at all (presence/absence coding) or frequencies that do not sum to one (frequency coding) for some loci.

Because of these difficulties, Buth (1984a) and others have advocated an approach that recognizes the locus as the character and the allelic composition at the locus in each taxon (i.e., allele or combination of alleles present) as the character state. For example, if some taxa are fixed either for allele *a* or for allele *b*, whereas others are polymorphic for both alleles, then three states would be recognized: "only *a*," "only *b*," and "*a* plus *b*."

*It is customary to refer to loci as *putative* or *presumptive* and to use the term *electromorphs* rather than *alleles* because of the indirect nature of the data and the usual absence of crossing experiments to confirm the mode of inheritance. For our purposes here, the simpler terms suffice.

The resulting discrete character states ("particulate data") are either left unordered or ordered into some logical progression (see Buth, 1984a, for details) for subsequent analysis.

Despite its intuitive appeal, several factors limit the utility of the particulate data, locus-as-character approach. When many different alleles occur in various combinations across taxa, the number of unique combinations may approach or even equal the number of taxa. Such characters will contain little or no information if the character states are left unordered. Ordering the character states helps somewhat, but the ordering criteria often seem subjective and arbitrary.

Buth (1984a) distinguished qualitative coding, in which observed combinations of alleles are used regardless of frequencies, and quantitative coding, in which estimated allele frequencies are used to assess "whether the states expressed by two taxa are statistically identical." Obviously, qualitative coding is extremely susceptible to sampling error. Consider the example in the above paragraph. Taxa that were in reality polymorphic for alleles *a* and *b* would often be incorrectly scored as "fixed" if one allele were rare, unless sample sizes were large. (Swofford and Berlocher, 1987, give a table showing the probability of failing to detect low-frequency alleles in samples of various sizes; see also Chapter 2). Even if allele frequencies could somehow be determined without error, it would be unreasonable to argue that allele frequencies are so irrelevant that the distinction between allele frequency arrays of, say, [0.01, 0.99] and [0.99, 0.01] is unimportant.

Quantitative coding presumably makes use of contingency-table analysis to test whether two or more samples could have come from a single homogeneous population. In most cases involving interspecific comparisons, however, we know beforehand or from the analysis of other loci that such is not the case, even if the difference between the allele frequency arrays of two taxa at a particular locus is not deemed significant. Furthermore, the power of these tests to detect heterogeneity is weak unless sample sizes are large. Therefore, failure to reject the null hypothesis of homogeneity should not usually be taken as evidence that the taxa are "statistically identical." Because of these considerations, methods that require re-coding of allele frequency arrays into discrete states should be used only when levels of polymorphism are low, with problematic loci excluded from the data set.

J.S. Rogers (1984, 1986) and Swofford and Berlocher (1987) have developed methods of analysis that use the observed allele frequencies directly in character-based analyses rather than requiring their recoding as discrete states (see the section on "Parsimony on Allozyme Data"). Felsenstein's (1981b) maximum likelihood method for continuous characters evolving under a Brownian motion process can also be applied to gene frequency data (after an appropriate transformation).

Gene Order Data

Phylogenetic inference based on the structural arrangement of genes, particularly in organellar genomes, provides a useful alternative to the more traditional comparison of the sequences of one or more genes (or indirect measures thereof). Although we will not discuss the use of gene-order data in detail, there is growing evidence that such data will provide important information on relationships, particularly when trying to resolve ancient divergences. Sankoff et al. (1992) used gene-order comparisons to estimate a phylogeny for 16 taxa, including fungi and other eukaryotes, and obtained a tree highly compatible with our current understanding of metazoan and fungal relationships. More recently, Boore et al. (1995) have used gene-order data to address longstanding questions regarding arthropod relationships. They were able to draw strong conclusions about relationships that previously had been highly ambiguous. Boore et al. (1995), Downie and Palmer (1992b), and others have argued that gene rearrangements are potentially more informative because they occur less frequently (and hence are less subject to parallelism and convergence) than sequence data, and because the large number of possible character states makes it unlikely that the same gene order will evolve independently in different lineages. Thus, while gene-order characters typically are insufficient to obtain a fully resolved tree, one can generally have high confidence in the groups that are supported.

Phylogenetic analysis of gene-order data is in its infancy (although the problems are similar to those encountered in the analysis of chromosomal inversions and other rearrangements). A serious complication is that the characters can no longer be assumed to evolve independently, because it is the relationships of the genes to each other that define the characters. Sankoff et al. (1992) have developed and implemented a method for minimizing the number of evolutionary events (inversions, transpositions, insertions, and deletions) required to convert one circular genome into another. This quantity then serves as the basis for a distance metric. Others (e.g., Boore et al., 1995) have performed parsimony analysis on special codings of the gene order data, despite the nonindependence of the data. It is likely that methods of analysis for gene-rearrangement comparisons will be an active area of research for the next few years.

OPTIMALITY CRITERIA I: PARSIMONY METHODS

Of the existing numerical approaches to inferring phylogenies directly from character data, methods based on the principle of **maximum parsimony** have been the most widely used by far. Most biologists are familiar with the usual notion of parsimony in science, which essentially maintains that simpler hypotheses are preferable to more complicated ones and that *ad hoc* hypotheses should be avoided whenever possible. Methods for estimating trees under the criterion of parsimony equate "simplicity" with the explanation of attributes shared among taxa as due to their inheritance from a common ancestor (e.g., Sober, 1989). When character conflicts occur, however, *ad hoc* hypotheses cannot be avoided if the observed character distributions are to be explained, and assumptions of **homoplasy** (convergence, parallelism, or reversal) must be invoked.

In general, parsimony methods for inferring phylogenies operate by selecting trees that minimize the total **tree length**: the number of evolutionary steps (transformations from one character state to another) required to explain a given set of data. For example, the steps might be base substitutions for nucleotide sequence data, or gain and loss events for restriction-site data. Obviously, a tree that minimizes the total number of steps also minimizes the number of extra steps (homoplasies) needed to explain the data.

In more mathematical terminology, we can define the general maximum parsimony problem as the following. From the set of all possible trees, find all trees τ such that

$$L(\tau) = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j \cdot \mathrm{diff}(x_{k'j}, x_{k''j}) \qquad \textbf{(2)}$$

is minimal, where $L(\tau)$ is the length of tree τ, B is the number of branches, N is the number of characters, k' and k'' are the two nodes incident to each branch k, $x_{k'j}$, and $x_{k''j}$ represent either elements of the input data matrix or optimal character-state assignments made to internal nodes, and diff(y,z) is a function specifying the cost of a transformation from state y to state z along any branch. The coefficient w_j assigns a weight to each character; it is often set to 1, but this need not be the case. Note also that diff(y,z) need not equal diff(z,y), although for methods that yield unrooted trees, diff(y,z) = diff(z,y). As discussed below, the definition of *optimal character-state assignments* may include restrictions on the nature of permissible character-state changes.

Any discussion of parsimony methods must distinguish between the optimality criterion (minimal tree length under a specified set of restrictions on permissible character-state changes) and the actual algorithm used to search for optimal trees. Early descriptions of parsimony methods (e.g., Farris, 1970) were presented in a way that tended to obscure the boundaries between criteria and algorithms. Biologists attempting to understand a method should not become so mired in algorithmic details that they lose track of the underlying biological principles and assumptions (Felsenstein, 1982). Algorithms tend to have short life spans, because better ones are constantly being invented. For example, Farris's (1970) *algorithm* for estimating minimum-length trees under the Wagner parsimony criterion is not, to our knowledge, used in any modern, widely used parsimony computer program (e.g., Farris's Hennig86, Felsenstein's PHYLIP-MIX, or Swofford's PAUP), but his *criterion* forms the basis for all of them. For these reasons, the conceptual framework in which we will discuss parsimony (and other) criteria assumes that the problem of finding optimal trees is not at issue. We assume, for the moment, that every possible tree can be evaluated, optimizing each one according to the chosen criterion and ranking them according to that criterion. We will take up the matter of searching for optimal trees in a subsequent section.

A common misconception regarding the use of parsimony methods is that they require *a priori* determination of character polarities (see above). In morphologically based studies, character polarity is often inferred using the method of **outgroup comparison**, and the resulting "polarized" characters form the basis of the analysis. Furthermore, since a "hypothetical ancestor" is implied by the polarity assignments, the output of an analysis of polarized characters is a rooted tree. Whereas specification of polarities provides a sufficient basis for obtaining rooted (rather than unrooted) trees, it is by no means prerequisite to the use of parsimony methods. This circumstance is fortunate, since the estimation of character polarity is both more difficult and less meaningful for most kinds of molecular data. All that is required to obtain rooted trees from parsimony analysis is to include in the data set one or more assumed outgroup taxa. The location at which the outgroup joins the unrooted tree implies a root with respect to the **ingroup** taxa. We emphasize, however, that the assignment of taxa to the outgroup constitutes an assumption that the remaining taxa (the ingroup) are **monophyletic** (an assumption that hopefully is justified by evidence extrinsic to the data at hand). If this assumption is wrong, the tree will be rooted incorrectly.

Parsimony analysis actually comprises a group of related methods, united by the goal of minimizing some evolutionarily significant quantity but differing in their underlying evolutionary assumptions. We will now address each of these methods in turn. The methods are presented in a logical progression rather than in chronological order of their introduction into the literature. In describing the procedures used to compute the minimum length required by a tree under a particular optimality criterion, we will consider a single character (position) in isolation from the rest. Because of the assumption of independence among characters, we can compute the overall tree length by summing, over all characters, the lengths required by each individual character. For the simplest procedures (Fitch and Wagner parsimony), we provide pencil-and-paper algorithms for computing tree lengths and determining optimal character-state assignments. Again, we are concerned only with calculating the length of a single tree, which is taken as a given; this tree may not be a most-parsimonious arrangement for our example character (or even over all characters); it is simply a tree that we wish to evaluate.

Fitch and Wagner Parsimony

These are the simplest parsimony methods, imposing no (Fitch) or minimal (Wagner) constraints on permissible character-state changes. The Wagner method, formalized by Kluge and Farris (1969) and Farris (1970), assumes that characters are measured on an interval scale; thus it is appropriate for binary, ordered multistate, and continuous characters. Fitch (1971b) generalized the method to allow unordered multistate characters (e.g., nucleotide and protein sequences). Wagner parsimony assumes that any transformation from one character state to another also implies a transformation through any intervening states, as defined by the ordering relationship. Fitch parsimony allows any state to transform directly to any other state. Both methods permit free reversibility; that is, change of character-states in either direction is assumed to be equally probable, and character states may transform from one state to another and back again. A consequence of reversibility is that the tree may be rooted at any point with no change in the tree length.

To determine the minimum length required by a given character j under either the Wagner or Fitch criteria, only a single pass over the tree is required, proceeding from the tips toward the arbitrary root. Computer scientists call this pass a *postorder traversal*. Although the computation can be performed in other ways, we recommend rooting the tree at one of the terminal taxa, denoted r, as shown in Figure 2. The algorithm for computing the length of a strictly bifurcating tree under the Wagner parsimony criterion then proceeds as follows (see Swofford and Maddison, 1987, for a more rigorous presentation).

1. To each terminal node i (including the one at the root), assign a **state set** S_i containing the character state assigned to the corresponding taxon in the input data matrix ($= x_{ij}$). Initialize the tree length to zero.
2. Visit an internal node k for which a state set S_k has not been defined but for which the state sets of k's two immediate descendants has been defined. Let i and j represent k's two immediate descendants. Assign to k a state set S_k according to the following rules:

   2a. If the intersection of the state sets assigned to nodes i and j is non-empty ($S_i \cap S_j \neq \varnothing$), let k's state set equal this intersection (i.e., $S_k = S_i \cap S_j$). The intersection can be represented as a closed interval $[a_k,b_k]$.

   2b. Otherwise ($S_i \cap S_j = \varnothing$), let k's state set equal the smallest closed interval $[a_k,b_k]$ containing an element from each of the state sets assigned to i and j. Increase the tree length by $b_k - a_k$.
3. If node k is located at the basal fork of the tree (i.e., the immediate descendant of the terminal node placed at the root), the traversal has been completed; proceed to step 4. Otherwise, return to step 2.
4. If the state assigned to the terminal node at the root of the tree (x_r) is not contained in the state set just assigned to the node at the basal fork of the tree (S_k), increase the tree length by the distance from x_r to S_k. (This distance equals $a_k - x_r$ if $x_r < a_k$ or $x_r - a_k$ if $x_r > b_k$.)

An application of the above algorithm is presented in Figure 2. We wish to compute the length of the unrooted tree of Figure 2A. (Although the more usual situation for molecular data would involve binary rather than multistate characters, we treat the multistate case to demonstrate the generality of the algorithm. Binary characters are simply a special case.) We first re-root the tree at node

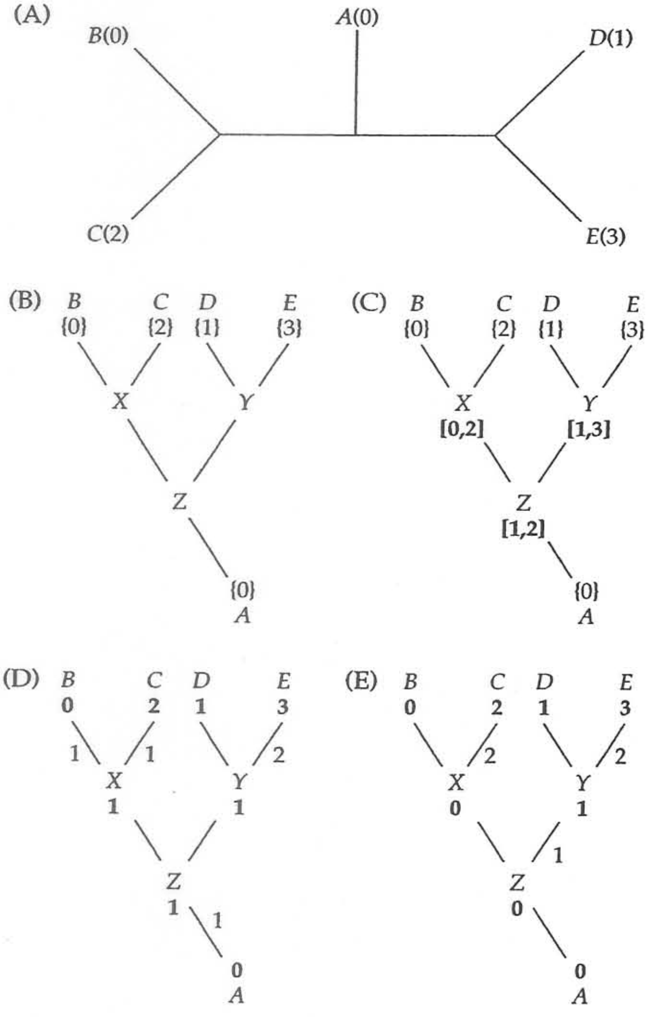


Figure 2 Steps in the algorithm for computing the length of an ordered character under Wagner parsimony. (A) The unrooted tree and character states. (B) Tree obtained by rooting at terminal node A and initial state sets assigned to teminal nodes. (C) State sets computed for interior nodes (bold). (D) Reconstruction obtained according to the algorithm described in the text. (E) An alternative, equally parsimonious reconstruction.

A (although we could have chosen any node), yielding the rooted tree shown in Figure 2B. Also shown in Figure 2B are the state sets assigned to the terminal nodes according to step 1 of the algorithm. Visiting internal node X in the first invocation of step 2, we observe that $S_B \cap S_C = \{0\} \cap \{2\} = \varnothing$, and hence assign the interval [0,2] to S_X, adding $2 - 0 = 2$ to the tree length. Similarly, we let $S_Y = [1,3]$ in the second invocation of step 2, and add $3 - 1 = 2$ to the length, which is now 4. In the third and final invocation of step 2, we observe that the intersection $S_X \cap S_Y = [0,2] \cap [1,3]$ is not empty, and therefore assign the interval [1,2] to S_Z. The situation as we arrive at step 4 is shown in Figure 2C. Since $x_r = 0$ is not an element of $S_Z = [1,3]$, we add an additional $1 - 0 = 1$ to the length. Thus, evolution of this character requires a minimum of five steps on our given tree.

The procedure outlined above is sufficient to obtain the minimal length required by any character on a given tree. However, it does not actually assign optimal character states to the hypothetical ancestors (internal nodes) of the tree to yield a **most-parsimonious reconstruction (MPR)**. To obtain such a reconstruction we can make a second pass over the tree, this time proceeding from the root toward the tips (a preorder traversal):

5. Visit an internal node k for which an optimal state assignment x_k has not yet been made but for which such an assignment has been made to k's immediate ancestor, denoted m. (Note that the first time this step is invoked, k corresponds to the node at the basal fork of the tree and $m = r$, the terminal taxon at the root of the tree.)

6. Assign to k the state from the state set computed in the first-pass, S_k (= $[a_k,b_k]$), that is closest to x_m. Specifically, if x_m is contained in S_k, we let $x_k = x_m$. Otherwise, we let $x_k = a_k$ if $x_m < a_k$ or $x_k = b_k$ if $x_m > b_k$.

7. If all internal nodes have been visited, stop. Otherwise return to step 1.

Applying steps 5–7 to the example of Figure 2, we first assign state 1 (the closest state in [1,2] to 0) to node Z. We then assign state 1 (the closest state in [0,2] to 1) to node X; likewise we assign state 1 (the closest state in [1,3] to 1) to node Y. The resulting reconstruction is shown in Figure 2D, and confirms the value of 5 as the minimum length for this character.

It is important to remember that this method finds only a single MPR, although others may exist. For instance, the reconstruction in Figure 2E also requires 5 steps. Swofford and Maddison (1987) described an exact algorithm for obtaining all MPRs for discrete character data under the Wagner parsimony criterion.

Simple modifications of the above algorithm provide for the treatment of multistate unordered characters (e.g., nucleotide sequence positions) under the Fitch (1971b) parsimony criterion. In the initial pass (computation of state sets and tree lengths), modify steps 2 and 4 as follows:

> 2a′. If the intersection of the state sets assigned to nodes i and j is non-empty ($S_i \cap S_j \neq \varnothing$), let k's state set equal this intersection (i.e., $S_k = S_i \cap S_j$).
>
> 2b′. Otherwise ($S_i \cap S_j = \varnothing$), let k's state set equal the union of the state sets assigned to nodes i and j ($S_i \cup S_j$), and increase the tree length by 1.

4′. If the state assigned to the terminal node at the root of the tree (x_r) is not contained in the state set just assigned to the node at the basal fork of the tree (S_k), increase the tree length by 1.

In order to obtain an MPR, modify step 6 above as follows:

6′. If x_m is contained in the state set assigned to k in the first-pass (S_k), assign this state to k as well. Otherwise, arbitrarily assign any state from S_k to k.

An example of the application of the above algorithm is shown in Figure 3. We are interested in computing the length required by a single character on the unrooted tree of Figure 3A. As before, we re-root the tree arbitrarily at node A, yielding the tree shown in Figure 3B. The state sets assigned to the terminal nodes are indicated on the figure. Visiting node X in the first invocation of step 2′, we see that $\{A\} \cap \{C\} = \varnothing$, and hence assign the union {A,C} as the state set S_X and set the

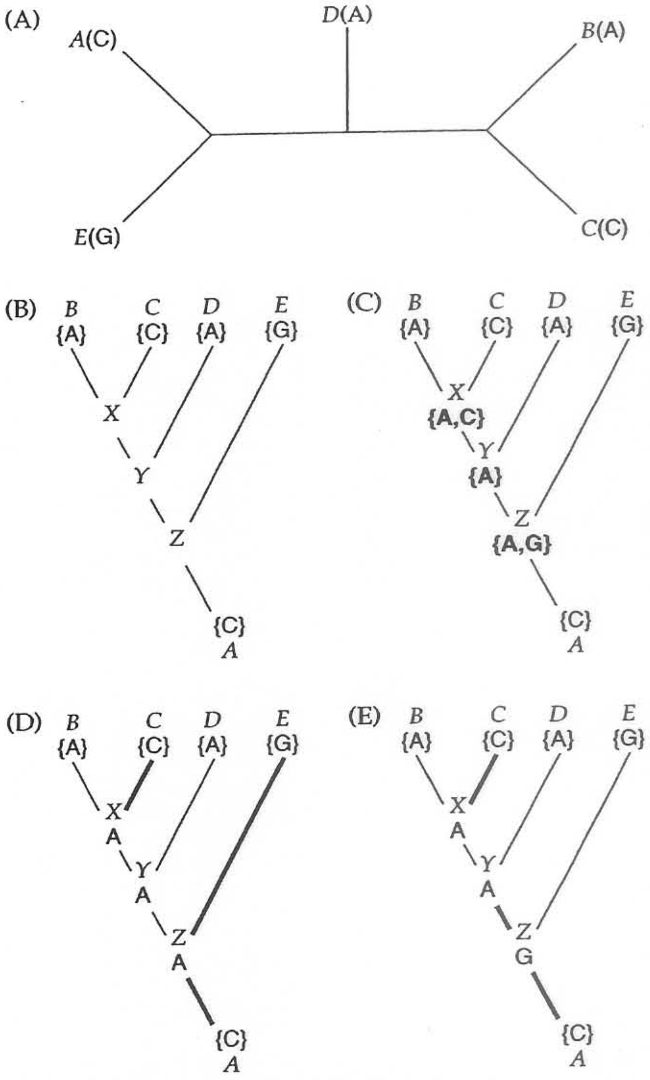


Figure 3 Steps in the algorithm for computing the length of an unordered character under Fitch parsimony. (A) The unrooted tree and character states. (B) Tree obtained by rooting at terminal node A and initial state sets assigned to teminal nodes. (C) State sets computed for interior nodes (bold). (D) Reconstruction obtained according to the algorithm described in the text. Branches on which character-state change occur are indicated in bold. (E) An alternative, equally parsimonious reconstruction.

tree length for this character to 1. Moving to node Y, we assign $\{A,C\} \cap \{A\} = \{A\}$ to S_Y. Finally, since $\{A\} \cap \{G\} = \varnothing$, we assign the state set {A,G} to node Z, again adding 1 to the tree length. Thus, at the beginning of step 4′, the state sets are as indicated in Figure 3C. Since x_r = C is not an element of S_Z = {A,G}, we add an additional step to the length, so that a total of 3 steps (nucleotide substitutions) are required on this tree.

If we wish to obtain one of the MPRs, we observe that the state C taken by the terminal taxon at the root of the tree is not contained in the set {A,G} assigned to the node at the first fork, and we may arbitrarily choose to assign state A to this node. We then assign state A to node Y as well (since the state set was a singleton no decision need be made). Finally, since state A is contained in node X's state set {A,C}, we assign it to the node, yielding the reconstruction shown in Figure 3D.

As was the case for the ordered character example, more than one MPR exists. For example, if we had chosen to assign state G rather than state A to node Z, we would have obtained the reconstruction shown in Figure 3E. Still another MPR exists, however, in which state C is assigned to all three internal nodes. That C was a possible state for node Z was not readily apparent from the state set {A,G} originally assigned to that node. In fact, a second pass over the tree is necessary in order to obtain all of the possible state assignments to each interior node. Fitch (1971b) described one such method and gave an algorithm for enumerating all of the possible MPRs.

Although all the algorithms described above are restricted to strictly bifurcating trees, they can easily be modified to handle multifurcations (polytomies). W.P. Maddison (1989) reviewed algorithms for obtaining MPRs on polytomous trees under a variety of evolutionary models, including the introduction of some novel approaches.

Other Parsimony Variants

Dollo Parsimony

The Wagner and Fitch parsimony criteria are appropriate under the assumption that probabilities of character change are symmetrical (i.e., the probability of a transformation from state 0 to state 1 in some small unit of evolutionary time is equivalent to that of a change from state 1 to state 0). As discussed above, this assumption is proba-

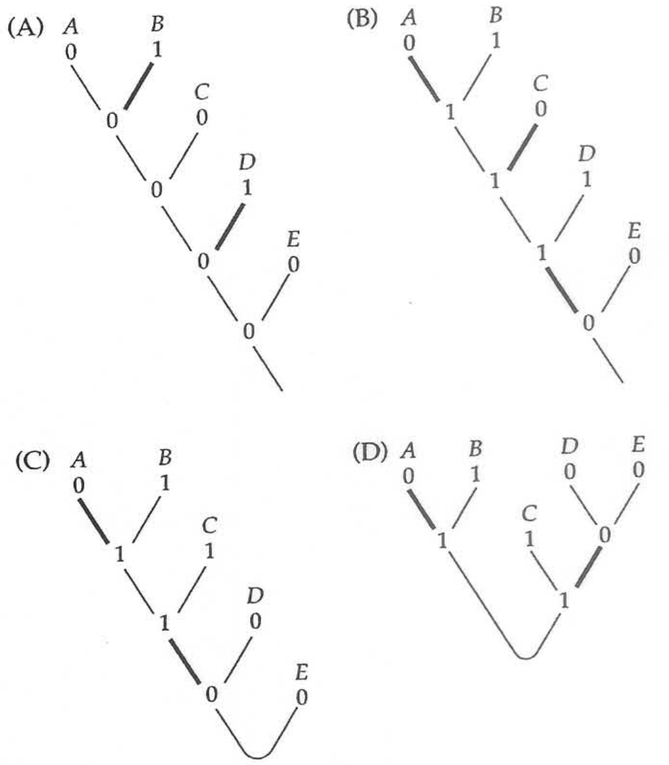


Figure 4 Character-state reconstructions demonstrating Dollo parsimony criterion. Branches on which character state changes occur are indicated in bold. (A) Most parsimonious reconstruction if multiple originations of state 1 are allowed. (B) Most parsimonious reconstruction under Dollo parsimony, in which only a single origination of state 1 is permitted. (C,D) Reconstructions obtained under unrooted Dollo model. Either rooting of the tree implies a minimum of two character state changes and only a single origination of state 1.

bly unreasonable for restriction-site characters, since the loss of an existing restriction site is more probable than a parallel gain of the same site at any particular location.

Because of this asymmetry, DeBry and Slade (1985) and others have suggested that the Dollo parsimony model (Farris, 1977) is more appropriate for restriction-site data. The Dollo parsimony criterion can be applied to binary or linearly ordered multistate characters for which we can reasonably hypothesize an ancestral condition (polarity). As for Wagner and Fitch parsimony, the preferred tree is the one requiring the fewest steps, but the character-state reconstruction (and hence the tree length assigned) must be consistent with the constraint that every derived character state be *uniquely* derived. If a hypothetical ancestor (a hypothetical taxon to which the assumed ancestral states for each character have been assigned) is included in the analysis, this definition corresponds to the traditional Dollo model (Farris, 1977): each character state is allowed to originate only once on the tree, and any required homoplasy takes the form of reversals to a more ancestral condition (i.e., parallel or convergent gains of the derived condition are not allowed). In the context of restriction-site data, each site may be gained once, with as many parallel losses of the site being assumed as are necessary to explain the data. For example, for the tree and character states shown in Figure 4 and with state 0 (site absent) assumed to be ancestral, the reconstruction of Figure 4A, requiring only two steps, is not acceptable under the Dollo model because two gains are indicated. Consequently, three steps would be required under the Dollo criterion (Figure 4B): a single gain followed by two losses.

Use of the Dollo parsimony criterion does not require inclusion of a hypothetical ancestor; it can be applied to unrooted trees as well. Stated another way, although the Dollo criterion requires specification of character polarity in a universal sense, it does not require us to know the state occurring in the most recent ancestor of the ingroup taxa. Specifically, the unrooted Dollo model forces us to assign character states to the interior nodes of the tree such that if a path is traced from any terminal taxon to any other, a backward change (from a more derived state to a more ancestral state) is never followed by a forward change (from a more ancestral state to a more derived state). Under this definition, the position of the root affects neither the assignment of character states to interior nodes nor the length of the tree. For example, both of the trees shown in Figures 4C and 4D, which differ only in the placement of the root, require two steps under the unrooted Dollo model (assuming that state 1 is the derived state). Neither tree requires more than a single origination of state 1. (Note that in the tree of Figure 4D, the derived state 1 is assumed to be ancestral with respect to the group ABCD, but derived relative to some more inclusive group.)

The unrooted Dollo approach is particularly convenient for restriction-site characters since it does not require the construction of a hypothetical ancestor, only the inclusion of one or more outgroup taxa. If a site is present in some of the ingroup taxa and in one or more of the outgroup taxa as well, then the most recent common ancestor of the ingroup is assumed to have had the site. The analysis will then seek to minimize the number of losses of the site over the full tree (ingroup and outgroup). If, on the other hand, the site is found only in some of the ingroup taxa but not in the outgroup, then the site is assumed to be ancestrally absent with respect to the ingroup, and a single gain will be postulated at an optimal location within the ingroup. Remember that the specification of "site absent" as the ancestral condition does not imply that the site was absent in the most recent common ancestor of the ingroup taxa, only that the site was absent in some, perhaps quite distant, ancestor.

The drawback to use of Dollo parsimony for restriction-site characters is demonstrated in Figure 5. If, despite its unlikelihood, a particular restriction site does originate independently in two lineages (Figure 5A), then the actual number of evolutionary changes can be drastically overestimated (Figure 5B) due to the strict enforcement of the requirement for unique originations. This pathological behavior may occur more often than the reader might suspect. Suppose one particular position within the restriction site were less constrained than the others, and further suppose that transition substitutions at this position were much more likely to occur than transversions. Then it is easy to imagine that the nucleotide at this position would, on an evolutionary time scale, toggle between the two purines (or pyrimidines). The site would then "blink" on and off, depending on which base was present at any particular point on a lineage. If we permitted only a single origination of the site, the number of losses we would be forced to postulate could become large.

One way to avoid this problem is to adopt a "relaxed" Dollo criterion. For example, we might prefer one gain and two losses to two independent gains, but we might prefer two independent gains to one gain and ten losses. The generalized parsimony method, discussed later, provides a mechanism for implementing a relaxed Dollo model.

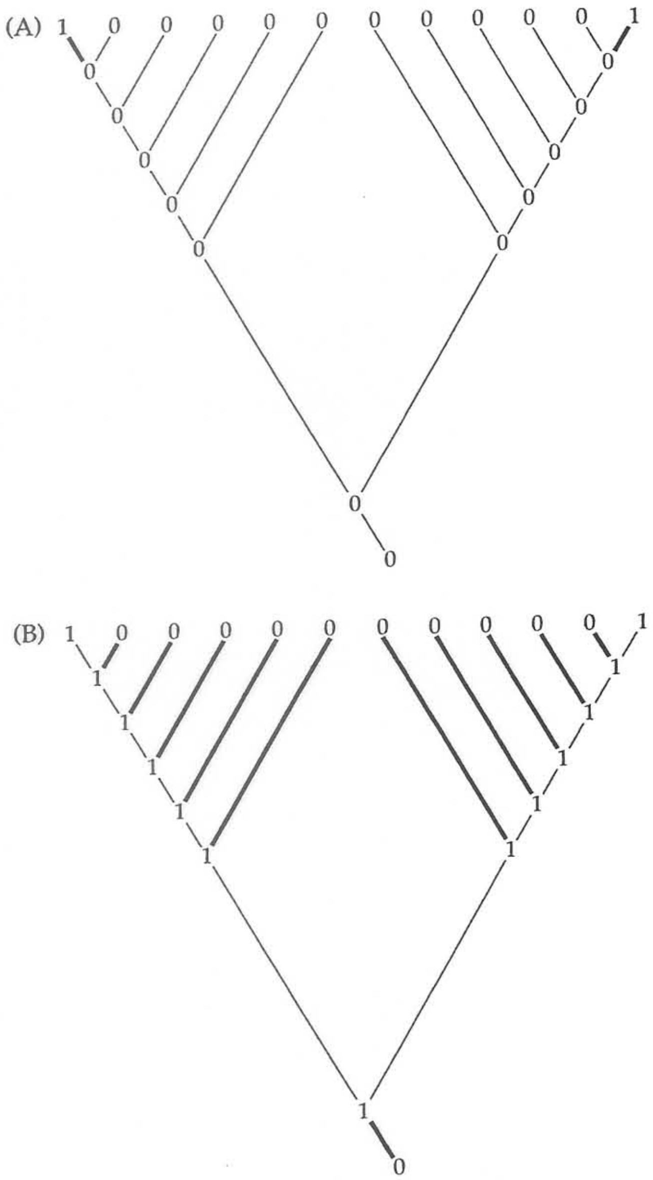


Figure 5 Demonstration of problems affecting Dollo parsimony if multiple originations of the derived state actually occur. (A) "True" tree has two steps due to independent derivations of state 1. (B) Reconstruction obtained under Dollo parsimony requires 11 steps (one derivation of state 1 and ten reversals to the ancestral state 0).

Camin-Sokal Parsimony

The method of Camin and Sokal (1965) was actually the first discrete-character parsimony approach to be described. It makes the strongest assumption of any of the methods discussed so far, namely, that evolution is irreversible.* We mention it here only for the sake of completeness, since it is highly unlikely that the assumption of irreversibility could be justified for any type of molecular data.

Transversion Parsimony

A common observation (e.g., W.M. Brown et al., 1982) is that transition substitutions occur more frequently than transversions in a given gene. For some molecules, it might even be argued that transitions occur so frequently that they quickly degenerate into noise and should therefore be ignored altogether. A simple method for ignoring transitions is to re-code the four nucleotides as either R (purine; A or G) or Y (pyrimidine; C or T). Standard Wagner parsimony may then be applied to the resulting binary-coded matrix.

A disadvantage to the complete rejection of information on transitions is that, while transitions may become saturated over long evolutionary distances, they may nonetheless be highly informative with respect to relationships among closely related taxa. One way around the dilemma is to assign greater weight to transversions than transitions, without going so far as to give transitions zero weight, as does transversion parsimony. Generalized parsimony can also be used for this purpose, as outlined below.

Note that some authors (e.g., Lake, 1987a) use the term *transversion parsimony* in a different sense than we describe here.

Generalized Parsimony

All of the above parsimony variants can be subsumed into a generalized method that assigns a cost for the transformation of each character state to the other possible states (Sankoff, 1975; Sankoff and Rousseau, 1975; see Sankoff and Cedergren, 1983, for a somewhat more consumer-oriented treatment and note that *generalized parsimony* is our term, not theirs). The costs can be represented as an m-by-m matrix S, where S_{ij} represents the increase in tree length (weight) associated with a transformation from state i to state j, and m is the total number of possible states. Three such weighting matrices, corresponding to the Wagner, Fitch, and Dollo parsimony criteria, are shown in Figure 6A–C. An exact, dynamic programming algorithm can be used to determine the minimum length required on a given tree for any particular choice of costs and to obtain one or all of the MPRs that yield this length (Sankoff and Cedergren, 1983); because of the complexity of this algorithm, we will not attempt to describe it here (but see Swofford and Maddison, 1992, for an introductory presentation).

Unfortunately, the generalized parsimony approach is much more computationally expensive than the algorithms described above for certain special cases (although a new procedure described by Wheeler and Nixon, 1995, may provide a faster approximation). Its advantage lies in its generality. For instance, S is not required to be symmetric. Relaxation of this requirement provides a means of implementing a relaxed Dollo criterion: by making the cost of a forward transformation greater than that of a backward transformation, we can prefer single-gain, multiple-loss scenarios until the number of losses becomes great enough that we are willing to allow independent gains. For example, the step matrix shown in Figure 6D would prefer one gain and *two* losses over two gains, but would prefer two gains over one gain and *four* losses. Generalized parsimony can also be used to attach greater importance to transversions than to transitions by assigning costs such that changes between two purines or between two pyrimidines receive lower weight than changes from a purine to a pyrimidine or vice versa (e.g., Figure 6E).

Perhaps the most troublesome aspect of generalized parsimony is determining how to choose the costs for different kinds of transformations.

*Some readers, familiar with "Dollo's Law of Irreversibility," may be confused at this point. The Dollo parsimony model does not assume complete irreversibility, only that a derived character state cannot be lost and then regained. The Camin-Sokal model does not permit a derived character state to return to the ancestral condition.

(A)

	a	b	c	d
a	–	1	2	3
b	1	–	1	2
c	2	1	–	1
d	3	2	1	–

(B)

	a	b	c	d
a	–	1	1	1
b	1	–	1	1
c	1	1	–	1
d	1	1	1	–

(C)

	a	b	c	d
a	–	*M*	2*M*	3*M*
b	1	–	*M*	2*M*
c	2	1	–	*M*
d	3	2	1	–

(D)

	0	1
0	–	3
1	1	–

(E)

	A	C	G	T
A	–	5	1	5
C	5	–	5	1
G	1	5	–	5
T	5	1	5	–

Figure 6 Cost matrices for generalized parsimony. (A) Cost matrix equivalent to Wagner parsimony (ordered characters). (B) Cost matrix equivalent to Fitch parsimony (unordered characters). (C) Cost matrix equivalent to Dollo parsimony. *M* is an arbitrarily large number, guaranteeing that only one transformation to each derived state will be permitted. (D) Cost matrix that assigns greater weight to gains (0 → 1 changes) than to losses (1 → 0 changes). (E) Cost matrix that assigns greater weight to transversions than to transitions.

One approach is to assign weights consistent with the researcher's assumptions about the relative frequency of different kinds of events. As a matter of general principle, we disagree with those who argue that *a priori* weighting of different kinds of changes introduces an unacceptable level of subjectivity into the analysis; an assumption of equal weights is itself a strong assumption. If, for example, we examined an alignment and observed that of 200 variable positions (columns), 80 contained only A and G, 80 contained only C and T, and only 40 contained a mixture of purines and pyrimidines, the conclusion that transitions occur much more frequently than transversions would not be controversial. In this case, a transversion:transition weighting of 1:1 would certainly represent a stronger assumption than a 2:1 weighting. Even if we have no idea how much more frequently transitions occur than transversions, a transversion:transition weight such as a 1.1:1 weighting may be desirable. Suppose that under equal weighting one tree required 5 homoplastic transversions and 3 homoplastic transitions, while another tree required 1 homoplastic transversion and 7 homoplastic transitions. Whether the "optimal" transversion:transition weighting is 2:1, 3:1, or 20:1, the tree requiring only 1 "extra" transversion would be preferable and would be chosen as superior under the 1.1:1 weighting scheme. Similar arguments can also be advanced for the use of gain:loss weights other than 1:1 for restriction sites.

An alternative to assuming a particular set of costs based on extrinsic criteria is to estimate the appropriate weights from the data themselves. Williams and Fitch (1989) discussed methods for choosing initial weights and for refining them by iterative improvement. Unfortunately, these methods may be sensitive to the starting point, a frequent drawback to successive approximation methods. Iterative approximation of optimal weights remains an area of active research, and further developments may be expected in the near future (for more on this subject see the section on "Reliability of Inferred Trees").

The methods developed by Sankoff and his colleagues were also designed to construct optimal alignments on a given tree by incorporating insertion/deletion weights (with insertions of gaps as appropriate) in addition to the substitution weights. This strategy is very appealing in that it effectively merges the problems of alignment and tree selection into a single problem; insertions and deletions are treated as events localized to particular branches on the tree in order to maximize the overall parsimony. The alternative method, construction of a multiple alignment prior to the phylogenetic analysis, is vastly inferior, since the topology of the tree cannot be ignored when deciding where to place gaps.

Unfortunately, rigorous application of Sankoff's method is computationally difficult for more than three sequences and one interior node. Sankoff et al. (1976) described an iterative procedure that rigorously aligns within local regions of a tree (three sequences adjacent to a single interior node), sacrificing the guarantee of global optimality but providing greater tractability. Nanney et al. (1989) described and programmed a more approximate, but much faster, procedure that operates by assuming that lengths of insertions and deletions are sufficiently small to allow alignment within a local "window" rather than obtaining a global alignment for any triplet of sequences. Hein (1990a,b) and Wheeler and Gladstein (1994) have developed useful programs for simultaneous alignment and tree optimization (see Chapter 9 for details).

Parsimony on Protein Sequences

Because this book does not specifically deal with amino acid sequencing, our discussion of parsimony methods for treating these sequences will be brief. Three general procedures have been used. The first, and simplest, is to minimize the number of amino acid replacements by using Fitch parsimony as described above (i.e., each position in the aligned sequences is a multistate unordered character, of which the possible states are the 20 possible amino acid residues). This approach, apparently used first by Eck and Dayhoff (1966), ignores the genetic code by failing to consider the minimal number of nucleotide substitutions required for the replacement of one amino acid by another (i.e., some replacements require a single nucleotide substitution, whereas others require two or even three substitutions).

Goodman, Moore, and their colleagues developed a more sophisticated approach (reviewed by Goodman, 1981) that seeks trees requiring the fewest number of nucleotide substitutions at the mRNA level. They produced an algorithm that generalizes the Fitch parsimony approach to codons, taking into account the degeneracy of the genetic code and guaranteeing that one obtains the minimum number of nucleotide substitutions required by any given tree. (A highly readable presentation of the algorithm, including a worked example, appears in G.W. Moore, 1976; see also Goodman et al., 1979). A more recent modification to their algorithm, by J. Czelusniak, permits the mixture of amino acid and nucleotide sequences (when available) in the same analysis (Goodman, 1981). Despite its elegance, the Moore–Goodman–Czelusniak algorithm may be overkill in the sense that it pays too much attention to silent substitutions (e.g., substitutions at third positions that do not change the corresponding amino acid). If silent substitutions occur so frequently that information from third positions quickly reaches saturation, then these positions would contribute mainly noise (or worse, systematic error) and should therefore be ignored. Weighting methods presumably could be used to minimize the contribution of third positions without ignoring them entirely. To our knowledge, however, such methods have not been used.

A third approach, intermediate between the first two, has been implemented by Felsenstein (1993) in his PROTPARS program from the PHYLIP package but has yet be formally described in the literature. Unlike the Eck–Dayhoff approach, it does consider the genetic code, but it also deviates from the Moore–Goodman–Czelusniak method by ignoring silent substitutions. Although ignoring silent substitutions sounds like extra work, the required bookkeeping is in fact simplified considerably because the program does not need to consider all the potential mRNA codons responsible for a particular amino acid residue or all of the potential synonymous codon assignments to the interior nodes. For example, PROTPARS would assign one step to a change from lysine to arginine (e.g., AAA → AGA), but two steps to a change from lysine to proline (e.g., AAA → CAA (glutamine) → CCA). Changes such as phenylalanine to glutamine require three nucleotide substitutions (e.g., AAA → GAA (leucine) → GAT (leucine) → GTT) but are counted as only two steps, since the middle substitution is silent.

One could take Felsenstein's argument a step further. Because of the biochemical properties of the various amino acids, there is often little selection against changes between amino acids having similar properties (e.g., between aspartic and glutamic acids). If changes between similar residues occur very frequently, perhaps we should ignore them as well (or at least give them less weight). The generalized parsimony method can be used to implement this strategy (Marsh et al., 1994), with the weights derived from the matrices presented by Dayhoff (1978) or Henikoff and Henikoff (1992).

(A)

	Allele frequencies					Presence/Absence				
Taxon	a	b	c	d	e	a	b	c	d	e
A	0.8	0	0	0	0.2	1	0	0	0	1
B	1.0	0	0	0	0	1	0	0	0	0
C	0	0.5	0	0.5	0	0	1	0	1	0
D	0	0	1.0	0	0	0	0	1	0	0
E	0	0	1.0	0	0	0	0	1	0	0

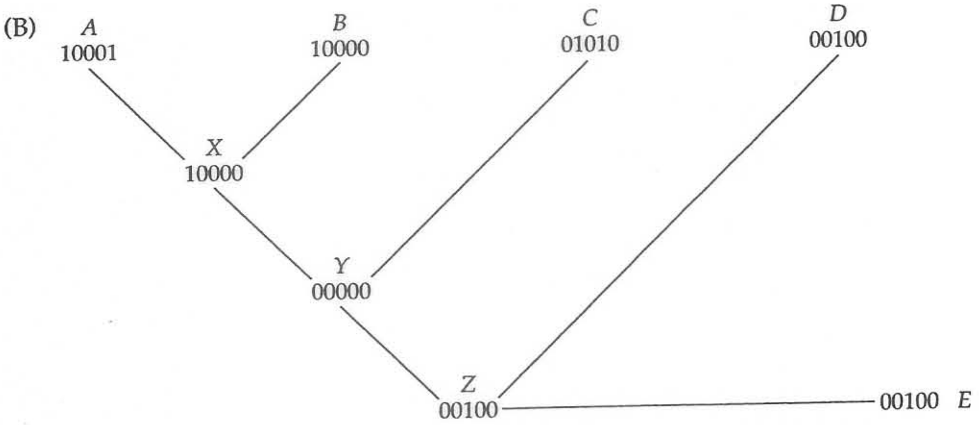


Figure 7 Demonstration of one problem with "independent alleles" coding. (A) Allele frequencies and data matrix resulting from presence/absence coding. (B) The most parsimonious reconstruction for the tree indicated assigns no alleles to ancestor *Y*.

Parsimony on Allozyme Data

The problems with treating allele frequencies or presence/absence as characters in a phylogenetic analysis were discussed above (see "Isozyme Data" in the section "Types of Data"). To clarify these issues within the context of parsimony analysis, consider the example in Figure 7. If the alleles indicated in Figure 7A are scored as present or absent and then treated as independent characters, the most parsimonious reconstruction under Wagner parsimony assigns no alleles to ancestor Y (Figure 7B), an outcome that most biologists would find unacceptable. A similar example could have been constructed using allele frequencies rather than presence/absence, in which the most parsimonious reconstruction assigned ancestral frequencies that summed to a value less than 1.

J.S. Rogers (1984, 1986) and Swofford and Berlocher (1987) developed methods for minimizing the total amount of frequency change on a given tree, subject to the constraint that the array of allele frequencies (for a particular locus) assigned to each interior node of the tree must exist in "allele frequency space" (a hyperplane in which the sum of the frequencies for all alleles is 1). These methods differ only in the choice of distances used for measuring branch lengths. Rogers' original method (1984) was derived for his earlier (1972) distance measure; he later extended it to a variety of other (mostly Euclidean) distance measures. His procedure uses the optimization technique of "hyperboloid approximation," which requires that the distance measure be representable as a differentiable function. Swofford and Berlocher (1987) argued for the superiority of the Manhattan metric and were forced to solve the problem via linear programming.

Methods that use allele frequencies rather than presence/absence are often criticized on the grounds that the allele frequencies are too easily modified by random drift and/or selection, and therefore do not provide reliable information for phylogenetic analysis (e.g., Mickevich and Johnson, 1976). In some cases, allele frequencies are known to vary temporally over the span of a few years, and this observation also has been used to question their relevance to phylogeny (Crother, 1990). We would argue, however, that even if the information contained in allele frequencies is somewhat unreliable, the frequencies at least provide a way to weight the presence or absence of particular alleles. For example, if an allele were detected sporadically in the taxa being analyzed, but never at frequencies higher than 0.04, we would be hesitant to attach much importance to the shared presence of that allele in some of the taxa; it could easily be present in other taxa at similar frequencies, but missed due to sampling error. On the other hand, an allele that is either fixed or nearly fixed whenever it occurs is probably more indicative of relationship. It should be emphasized that adopting a cutoff frequency (typically 0.05) does not solve the problem unless a researcher is willing to assert that an allele known to occur in a sample at an estimated frequency of, say, 0.04 is "not present."

Although the Rogers and the Swofford–Berlocher methods are conceptually simple, the computer algorithms used to implement them are quite complex; the interested reader should refer to the original papers for details. These methods are also much slower than other parsimony methods. However, Berlocher and Swofford (1996; see Swofford, 1996) have developed a fast approximation using generalized parsimony on single-locus Manhattan distance matrices (for a given tree, this algorithm obtains an exact solution to Swofford and Berlocher's 1987 MANOB criterion).

OPTIMALITY CRITERIA II: METHODS BASED ON MODELS OF EVOLUTIONARY CHANGE

The Utility of Models

Although the parsimony methods described above are based on specific optimality criteria, they do not require explicit models of evolutionary change. Considerable disagreement exists as to whether the "model-free" nature of parsimony is an advantage or a disadvantage. Regardless of where one stands on this issue, however, one thing is clear: parsimony *does* make assumptions, and violation of these assumptions can lead to problems. The difficulty lies in stating precisely what the assumptions are. At a minimum, acceptance of an optimal tree under the parsimony criterion requires one to assume that conditions that can cause parsimony to estimate an incorrect tree are unlikely to have occurred.* The ability of an estimation method to converge to a true value (in this case the correct tree) as more data are accumulated is known as **consistency**. Felsenstein (1978a) showed that parsimony methods can make inconsistent estimates of the true phylogeny under one simple evolutionary model.

Parsimony and Inconsistency

Examination of the conditions under which parsimony† is an inconsistent estimator will be helpful in understanding the usefulness of explicit evolutionary models. We will first present a non-technical examination of the problem; in a later section ("Model-Based Corrections for Character Data: Hadamard Conjugation") we will look at the issue more rigorously. Suppose that the true phylogeny for a group of four taxa is as shown in Figure 8A, where the lengths of the branches indicate the relative expected amount of evolutionary change along each branch under some model of evolution (e.g., the model of Jukes and Cantor, 1969). For whatever reason, the rate of evolution has been accelerated in the peripheral branches leading to taxa 1 and 4. Each nucleotide position will have some ancestral nucleotide (e.g., A in Figure 8A). Suppose that the short branches are so short that there are essentially no changes along the lineages leading to taxa 2 and 3. One of four possible classes of nucleotide patterns will then

*An alternative position is that parsimony is required as a method of scientific inquiry regardless of any considerations about whether it is more or less likely to recover the true phylogeny than other methods. Some proponents of this view hold that since the truth is essentially unknowable, we should abandon the search for it and simply choose the most parsimonious solution for its own sake. We do not subscribe to this position. Although the true phylogeny may be "unknowable," it can nonetheless be estimated, and we view phylogenetic methods as means to that end rather than an end in themselves.

†Parsimony in the traditional sense, i.e., "uncorrected parsimony"; see the end of this section.

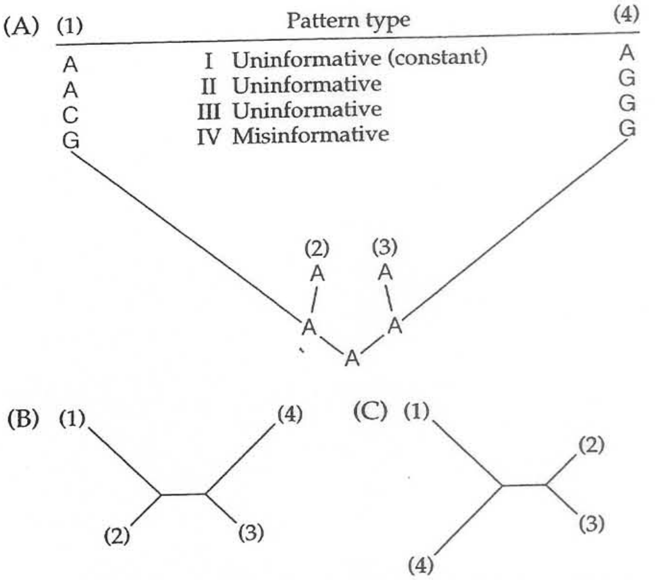


Figure 8 Demonstration of the potential inconsistency of parsimony methods. (A) Hypothetical four-taxon tree containing two long peripheral branches, with all other branches being very short. (B) Unrooted equivalent for the tree shown in (A). (C) Incorrect tree selected by maximum parsimony. See text.

result from changes along the lineages leading to taxa 1 and 4. In patterns of type I, taxa 1 and 4 both retain the ancestral nucleotide, so the position is constant and therefore uninformative under the parsimony criterion. (Note that the observation of pattern I does not imply that no substitutions have occurred—only that the nucleotides observed in the terminal taxa are identical at the tips of the tree, regardless of the number of changes that have actually occurred.) Patterns of types II (a change along only one of the two long branches) and III (a change to a different nucleotide along each long branch) are likewise uninformative, as one can explain each pattern with a single change (pattern II) or two changes (pattern III) along the peripheral branches of any of the three possible unrooted trees. Only patterns of type IV are informative under parsimony. Unfortunately, this pattern supports an incorrect tree (Figure 8C); it is actually misinformative about evolutionary relationships. Patterns that support the true tree (Figure 8B) will occur only rarely, because they require an unlikely change along the central branch (or, even less likely, two parallel changes along the long and short branches on the same side of the tree).

Felsenstein (1978a) called the behavior of parsimony in situations like that shown in Figure 8 "positively misleading" because as the number of characters (e.g., sequence length) increases, we become more and more certain to infer an incorrect tree. Stated another way, when we are in the **Felsenstein zone** (of inconsistency), the only hope of getting the correct tree is by sampling few enough characters that we may be lucky enough to obtain more of the character patterns favoring the true tree than of the more probable character patterns favoring the wrong tree. As we will see below, methods that attempt to account for unobserved as well as observed substitutions (maximum likelihood and distance methods using "multiple-hit" corrections) will not be inconsistent under this model and will estimate the correct tree as long as enough data are available to overcome sampling error.* Although in this case the inconsistency is due to strongly unequal rates of change along different branches, Hendy and Penny (1989) demonstrated other scenarios that lead to inconsistency even with equal rates of change throughout the tree (i.e., a molecular clock) and suggested the term "long-branch attraction" for this general phenomenon.

Steel et al. (1993a) have emphasized that parsimony is a criterion for choosing an optimal tree for a data set, whether the data are the original data or some transformation of those data. They show that, for conditions such as those shown above, parsimony can still make a consistent estimate of the phylogeny if the data are first corrected for unobserved substitutions using a Hadamard conjugation (see below). The correc-

*Felsenstein's results have often been criticized (e.g., Farris, 1983, 1986b) because they are based on unrealistic and restrictive models of the evolutionary process. This criticism is unjustified, however, as the point could equally well be made with much more general and believable models, but requiring more complex mathematics. Farris's (1983, 1986b) point that a maximum likelihood method will guarantee consistency only if evolution proceeds according to the assumed model is of course true, a point to which we will return later.

tion formally involves transformation of the original data matrix to a new data matrix containing 2^{T-1} characters (in the case of two states), each with an associated weight. Weighted parsimony analysis of this new data set will not be inconsistent as long as the other assumptions of the model are satisfied (e.g., equal rates of change at each site). In extreme cases, this new data set may contain highly weighted character patterns that were completely absent from the original data set, so the method is quite different from the conventional usage of the term *parsimony*. Consequently, we will use *parsimony* to mean *uncorrected parsimony* unless otherwise indicated.

Differences in Perspective between Parsimony and Likelihood

Even under conditions where parsimony is consistent, alternative methods that incorporate models of evolutionary change can make more effective use of the data, as demonstrated in the example of Figure 9. The tips of the tree in Figure 9A are labeled by the nucleotides observed at one sequence position. (Although the tree is shown as a rooted tree, it is formally unrooted, with the path between ancestor 1 and the outgroup treated as a single branch.) As a preliminary step to our formal introduction to maximum likelihood, it will be instructive to examine (qualitatively) the perspectives of parsimony and maximum likelihood with respect to the identity of the corresponding nucleotide in ancestor 1 of this tree. Using the methods of Fitch parsimony described above, we find that the most parsimonious state assignment for ancestor 2 is an A (an obvious choice, as all of ancestor 2's descendants possess A as well). Thus, ancestor 1 has given rise to a lineage with an A and a lineage with a C. It is also related to a lineage (leading to the outgroup) with a G. Assignment of any one of these three nucleotides to ancestor 1 would be equally parsimonious, with each reconstruction explaining all of the tip nucleotides at this position with exactly two changes. (If ancestor 1 had a T at this position, three character-state changes would be required.) Consequently, a new sequence with a C could be inserted equally parsimoniously (with respect to this position) into branches α, β, or γ of the tree (Figure 9B–D). More generally, because adding a sequence with an A to any branch of the tree would require no additional steps and adding a sequence with a T would require a single additional step for every branch, this sequence position would be uninformative regarding the placement of a sequence with an A or a T. This position would predispose a lineage containing only a C or a G to originate from branches α, β, or γ, because connecting such a lineage anywhere in the subtree descending from ancestor 2 would entail an extra nucleotide substitution.

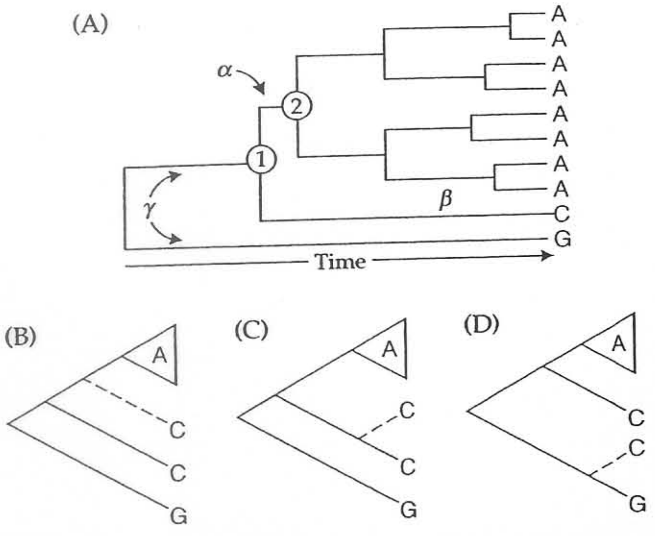


Figure 9 Example used to show difference in perspective between parsimony and likelihood methods. (A) Hypothetical tree with branch lengths drawn proportionally to expected number of substitutions, labeled by base observed at a particular site. (B,C,D) Insertion of a new taxon containing a C at the site of interest to branches α, β and γ, respectively, of the tree shown in (A).

Now consider the maximum likelihood perspective. In maximum likelihood estimation, we choose the hypothesis that maximizes the probability of observing the data we have obtained (i.e., the tip sequences). To calculate this probability, we need a model of evolutionary change. For now, suppose that the rate of substitution from any nucleotide to any other nucleotide is the same for all nucleotide pairs, and that the expected number of such substitutions along any one branch is a function of this substitution rate and the length of the branch in evolutionary time. (This is an oversimplified version of the Jukes-Cantor model of nucleotide sequence change, discussed in more detail below.) For the moment, also assume that the substitution rate is the same throughout the tree (we will see later that this assumption is not necessary; it merely allows us to think of branch lengths as amounts of evolutionary time). The observation that all eight descendants of ancestor 2 have nucleotide A is most consistent with change being rare, so postulated histories with fewer changes are more plausible than histories with more changes. Thus, from a maximum likelihood perspective, ancestor 2 would have an A in those histories (ancestral state reconstructions) having the highest probability of giving rise to the observed nucleotides. Although histories in which ancestor 2 had a C, G, or T would also contribute to the overall probability of the specified tree having generated the observed data, if all branches in the subtree were very short, histories with an A at ancestor 2 would contribute the vast majority of the total probability. This is as close as maximum likelihood gets to saying "ancestor 2 had an A."

We now move to ancestor 1. The branches connected to this ancestor lead to ancestor 2 (probably an A) and to sequences known to possess a C and a G (the outgroup), respectively. Ignoring the G for the moment, let us consider whether ancestor 1 is more likely to have possessed an A or a C, given the topology of the tree and the nucleotides found in the tip sequences. If ancestor 2 indeed possessed an A as expected, at least one change must have occurred along the path between ancestor 2 and the tip having a C (i.e., branches α and β). Because branch lengths represent the expected number of character-state changes along a branch, when a branch is short, there is a relatively low probability of a single change occurring along that branch, and an almost negligible probability of more than one change. Thus, given that a character change (probably) occurred somewhere along branches α or β, it is far more likely to have occurred along the long branch β than the short branch α. Thus, ancestor 1 is much more likely to have possessed an A than a C. Remember, however, that the estimate of A at ancestor 1 is a probabilistic statement. When the same configuration of tip states arises at different sites, the nucleotide found in the actual ancestor will usually be an A, but it would sometimes be a C—and occasionally it would even be a G or a T.

Returning our attention to the full tree, we know that at least two changes must have occurred, and since change is rare in this example, histories with three or more changes are less likely than those with only two changes. But on which two of the three branches (α, β, or γ) are the changes most likely to have occurred? Because branch α is so short, it is much more likely that the two changes have occurred on branches β and γ than on any pair of branches involving branch α. Therefore, histories with an A in ancestor 1 are more likely than others of having generated the observed data, and due to the greater length of branch γ, histories with a C in ancestor 1 are more likely than are those with a G. It seems extremely unlikely under our model that ancestor 1 would have possessed a T. Thus, we obtain a clear ordering of preferences for all residues. An important practical consequence is that, unlike parsimony, this sequence position would be informative with respect to the placement of a new sequence containing a C at the site, biasing the decision toward connecting this new sequence to branch β (Figure 9C).

It is important to remember that the only reason for the appropriate predisposition toward tree 9C is the short length of branch α and the low overall rate of change. In this case, an improbable substitution along branch α is avoided by placing the change along the branch leading to the tips with nucleotide C in Figure 9C. For either of the trees of Figure 9B and 9D, avoiding a substitution along the original branch α would require parallel A → C changes along the lineages terminating at taxa possessing nucleotide C. These parallelisms would be improbable events if the rate of change is low, but they become more probable as rates increase. Thus, as branch α becomes longer and the rates of change grow faster, the preference for tree 9C will decrease.

In summary, whereas parsimony ignores information on branch lengths when evaluating a tree, maximum likelihood considers that changes are more likely along long branches than short ones, and estimation of branch lengths is an important component of the method. This difference explains the consistency of maximum likelihood under many situations in which parsimony is inconsistent. In the example of Figure 8, maximum likelihood will not be fooled by the "misinformative" pattern IV, because this pattern is very likely to occur even on the true tree. Distance methods that adequately account for unobserved substitutions will also succeed in this case, although they tend to be less efficient, requiring more data to achieve the same level of accuracy (e.g., see Hillis et al., 1994b; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995a,b).

Maximum Likelihood Methods

Maximum likelihood methods of phylogenetic inference evaluate a hypothesis about evolutionary history in terms of the probability that a proposed model of the evolutionary process and the hypothesized history would give rise to the observed data. It is conjectured that a history with a higher probability of giving rise to the current state of affairs is a preferable hypothesis to one with a lower probability of reaching the observed state. Maximum likelihood estimation was first used in phylogenetic inference by Cavalli-Sforza and Edwards (1967). However, because they did not use sequence data, this work remained relatively obscure. Felsenstein (1981a, 1993) brought the maximum likelihood framework to nucleotide-based phylogenetic inference. Later, maximum likelihood was applied to amino acid sequence data as well (Kishino et al., 1990; Adachi and Hasegawa, 1992).

In addition to its consistency properties, maximum likelihood is useful because it often yields estimates that have lower variance than other methods (i.e., it is frequently the estimation method least affected by sampling error). It also tends to be robust to many violations of the assumptions used in its models. Part of its power in this respect is that many models of sequence evolution that assume identical distributions across sites can safely assume that the actual substitution processes taking place at different sites have much in common, even if they are not exactly identical. Consequently, the major components determining the evolution of sequences can be described by just a few parameters. The overall result of both improved compensation for superimposed changes and of sampling variance is that even with very short sequences, maximum likelihood tree inference tends to outperform alternative methods (e.g. parsimony or additive distances) when evaluated under many models of sequence evolution (see, e.g., Hasegawa and Fujiwara, 1993; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995a).

Several areas of biological research, notably genetic mapping and clinical testing, routinely use maximum likelihood methods for testing hypotheses. However, the perceived and actual complexities of obtaining maximum likelihood solutions to problems that involve numerous alternative hypotheses has inhibited the more general use of these techniques. The following discussion attempts to outline the elements of a maximum likelihood formation of phylogenetic inference. For additional perspective, Goldman (1990) provides a very accessible introduction to these concepts.

Objective

Phylogenetic analysis seeks to infer the history (or set of histories) that are most consistent with a set of observed data. In the present case, the data are observed nucleotide (or protein) sequences; the unknowns are the branching order and branch lengths of the tree. To apply a maximum likelihood approach, a concrete model of the evolutionary process that accounts for the conversion of one sequence into another must be specified. This model may be fully defined; alternatively, it may contain many parameters that are to be estimated from the data. A maximum likelihood approach to phylogenetic inference evaluates the probability that the chosen evolutionary model will have generated the observed sequences (the probability of the data under the model); phylogenies are then inferred by finding those trees that yield the highest likelihoods.

The basic principles involved in calculating the likelihood of a tree are introduced in Figure 10. Figure 10A shows a set of aligned nucleotide sequences for four taxa. Suppose we want to evaluate the likelihood of the unrooted tree shown in Figure 10B; that is, what is the probability that this tree could have generated the data of Figure 10A under our chosen model? Because most of the models currently used are **time-re-**

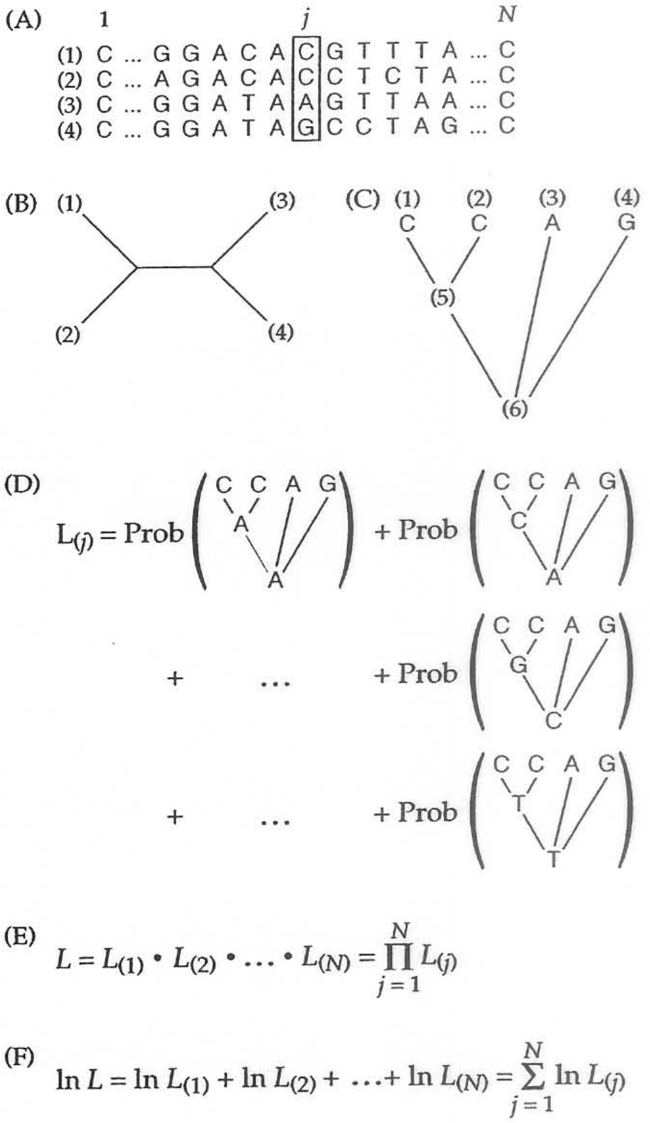

(A)
1
j
N
(1) C ... G G A C A C G T T T A ... C
(2) C ... A G A C A C C T C T A ... C
(3) C ... G G A T A A G T T A A ... C
(4) C ... G G A T A G C C T A G ... C
(B)
(1)
(2)
(3)
(4)
(C)
(1) C
(2) C
(3) A
(4) G
(5)
(6)
(D)
$L_{(j)}$ = Prob
+ Prob
+ ...
C C A G
A
C
G
T
(E)
$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$
(F)
$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$

Figure 10 Overview of the calculation of the likelihood of a tree. (A) Hypothetical sequence alignment. (B) An unrooted tree for the four taxa whose sequences appear in (A). (C) Tree after rooting at an arbitrary internal node. (D) The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution. (E) The likelihood of the full tree is the product of the likelihoods at each site. (F) The likelihood is usually evaluated by summing the log of the likelihoods at each site, and reported as the log likelihood of the full tree.

versible, the likelihood of the tree is generally independent of the location of the root. It is convenient to root the tree at an arbitrary internal node (e.g., Figure 10C).

Under the assumption that nucleotide sites evolve independently, we can calculate the likelihood for each site separately, and combine the likelihoods into a total value at the end. To calculate the likelihood for some site *j*, we must consider all of the possible scenarios by which the tip sequences could have evolved. Obviously, some of these scenarios are much more plausible than others, but every scenario has at least some probability of generating any pattern of observed nucleotides. More specifically, for any given site, the node at the root of Figure 10C might have possessed an A, a C, a T, or a G. For each of these possibilities, the other internal node might also have possessed any of the four nucleotides. Thus, there are 4 × 4 = 16 possibilities to consider. Since any one of these scenarios could have led to the nucleotide configuration at the tips of the tree, we must calculate the probability of each and sum them to obtain the total probability for each site *j*. This calculation is illustrated schematically in Figure 10D. Because we assume a **Markov model** (see below), we assume that changes along different branches are independent. Thus, the probability of any single scenario is equal to the product of the probabilities of the changes required by that scenario. For instance, the probability of the scenario represented by the first term of Figure 10D is equal to the prior probability that the nucleotide at node 6 is an A (typically 1/4, or the average frequency of A in the original sequences, depending on the specifics of the model) times the probability of retaining an A along the branch leading from node 6 to node 5, times the probability of an A → C change along the peripheral branch leading to tip 1, and so on.

Having calculated the likelihoods at each site, the joint probability that the tree and model confer upon all sites is computed as the product of the individual-site likelihoods (Figure 10E). Because the probability of any single observation is an extremely small number (much too small to represent using standard floating-point representations on a computer), we almost always evaluate the log of the likelihood instead, so the probabilities are accumulated as the sum of the logs of the single-site likelihoods (Figure 10F).

Models of Sequence Evolution

The critical element missing from the above overview is how the probabilities of the various changes are calculated. These probabilities depend on several assumptions about the process of nucleotide substitution, which define a substitution model. We will restrict our attention here to Markov models, in which the probability of a change from state i to state j at a given site does not depend on the history of the site prior to its possession of state i. For example, if a sequence position has base A at some time t_0, the probability that it will have base T at a later time t_1 depends only on the fact that it has base A at t_0; knowing that it had state C at some time prior to t_0 would be irrelevant to the probability. We will also assume that the substitution probabilities do not change in different parts of the tree (i.e., that the evolutionary mechanisms responsible for sequence change constitute a **homogeneous Markov process**). The use of Markov processes to model nucleotide substitution has been discussed by Felsenstein (1981a), Lanave et al. (1984), Tavaré (1986), Barry and Hartigan (1987a,b), Kishino and Hasegawa (1990), Rodríguez et al. (1990), and Zharkikh (1994), among others.

The mathematical expression of a substitution model is a table of rates (substitutions per site per unit evolutionary distance) at which each nucleotide is replaced by each alternative nucleotide. For DNA sequences, these rates can be expressed as a 4 × 4 instantaneous rate matrix, **Q**, in which each element Q_{ij} represents the rate of change from base i to base j during some infinitesimal time period dt. The most general form of this matrix is

$$\mathbf{Q} = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix} \tag{3}$$

where the rows (and columns) correspond to the bases A, C, G, and T, respectively. The factor μ represents the mean instantaneous substitution rate. This mean rate is modified by the relative rate parameters $a, b, c, \ldots, l$, which correspond to each possible transformation from one base to a different base. The product of a relative rate parameter and the mean instantaneous substitution rate constitutes a *rate parameter*. The remaining parameters, π_A, π_C, π_G, and π_T, are *frequency parameters* that correspond to the frequencies of the bases A, C, G, and T, respectively (Z. Yang, 1994a). We assume that these frequencies remain constant over time (i.e., they are always at equilibrium), and that the rate of change *to* each base is proportional to the equilibrium frequency but independent of the identity of the starting base. The diagonal elements of **Q** are always chosen so that the elements in the corresponding row sum to zero. It is sometimes convenient to decompose **Q** into two matrices **R** and **Π**, where

$$\mathbf{R}=\begin{pmatrix} - & \mu a & \mu b & \mu c \\ \mu g & - & \mu d & \mu e \\ \mu h & \mu j & - & \mu f \\ \mu i & \mu k & \mu l & - \end{pmatrix}$$

and

$$\mathbf{\Pi}=\begin{pmatrix} \pi_{\mathrm{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathrm{C}} & 0 & 0 \\ 0 & 0 & \pi_{\mathrm{G}} & 0 \\ 0 & 0 & 0 & \pi_{\mathrm{T}} \end{pmatrix}$$

The off-diagonal elements of **Q** are then equal to the off-diagonal elements of the matrix product **RΠ**, and the diagonal elements of **Q** are once again set to the negative of the sum of the off-diagonal elements for the corresponding row. Analogous matrices can be defined for protein sequence data, except that there are 20 states rather than 4.

Almost all of the DNA substitution models proposed to date are special cases of matrix (3). It is usually assumed that the overall rate of change from base *i* to base *j* in a given length of time is the same as the rate of change from base *j* to base *i*. Such models are said to be time-reversible. This corresponds to the rate parameter restrictions $g = a$, $h = b$, $i = c$, $j = d$, $k = e$, and $l = f$. One byproduct of time reversibility is that the likelihood of a tree generally does not depend on how the tree is rooted. Consequently, as for most of the parsimony methods discussed above, maximum likelihood estimation is usually limited to the inference of unrooted trees, and other assumptions must be invoked to convert an unrooted tree into a rooted one. Although it is possible to relax the time-reversibility assumption, this relaxation introduces additional computational complications, including the need to consider rooted trees. Thus, we will only consider symmetric **R** matrices of the form

$$\mathbf{R}=\begin{pmatrix} - & \mu a & \mu b & \mu c \\ \mu a & - & \mu d & \mu e \\ \mu b & \mu d & - & \mu f \\ \mu c & \mu e & \mu f & - \end{pmatrix}$$

The most general time-reversible model (GTR) is then represented by

$$\mathbf{Q}=\begin{pmatrix} -\mu(a\pi_{\mathrm{C}}+b\pi_{\mathrm{G}}+c\pi_{\mathrm{T}}) & \mu a\pi_{\mathrm{C}} & \mu b\pi_{\mathrm{G}} & \mu c\pi_{\mathrm{T}} \\ \mu a\pi_{\mathrm{A}} & -\mu(a\pi_{\mathrm{A}}+d\pi_{\mathrm{G}}+e\pi_{\mathrm{T}}) & \mu d\pi_{\mathrm{G}} & \mu e\pi_{\mathrm{T}} \\ \mu b\pi_{\mathrm{A}} & \mu d\pi_{\mathrm{C}} & -\mu(b\pi_{\mathrm{A}}+d\pi_{\mathrm{C}}+f\pi_{\mathrm{T}}) & \mu f\pi_{\mathrm{T}} \\ \mu c\pi_{\mathrm{A}} & \mu e\pi_{\mathrm{C}} & \mu f\pi_{\mathrm{G}} & -\mu(c\pi_{\mathrm{A}}+e\pi_{\mathrm{C}}+f\pi_{\mathrm{G}}) \end{pmatrix} \quad \textbf{(4)}$$

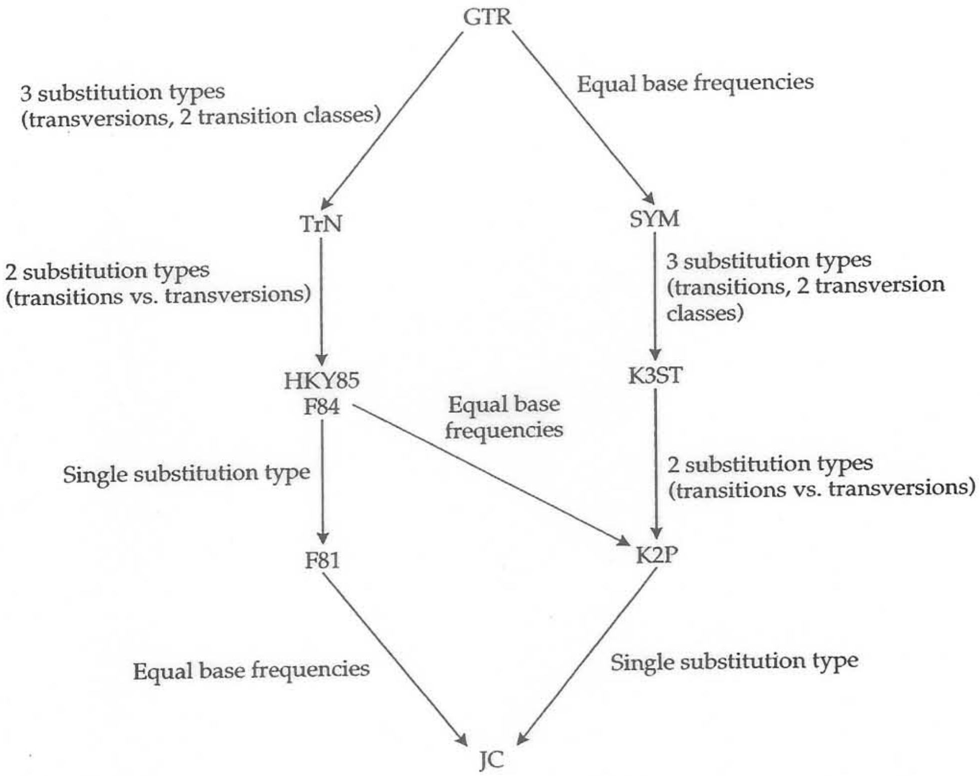

GTR
3 substitution types (transversions, 2 transition classes)
Equal base frequencies
TrN
SYM
2 substitution types (transitions vs. transversions)
3 substitution types (transitions, 2 transversion classes)
HKY85
F84
K3ST
Equal base frequencies
Single substitution type
2 substitution types (transitions vs. transversions)
F81
K2P
Equal base frequencies
Single substitution type
JC

Figure 11 Relationship between special cases of the general time-reversible family of substitution models. Arrow labels indicate restrictions that convert a more general model to a more specific one. Model abbreviations: F81, model of Felsenstein, 1981a (equivalent to the "equal input" model of Tajima and Nei, 1982); F84, model used in versions 2.6 and later of PHYLIP (Felsenstein, 1993; Kishino and Hasegawa, 1989); GTR, General time-reversible (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990); HKY85, Hasegawa-Kishino-Yano model (Hasegawa et al., 1985b); JC, Jukes and Cantor (1969) model; K2P, Kimura (1980) two-parameter model; K3ST, Kimura (1981) three-substitution-type model; SYM, model described by Zharkikh (1994); TrN, Tamura and Nei (1993) model.

(Lanave et al., 1984; Tavaré, 1986; Barry and Hartigan, 1987b; Rodríguez et al., 1990). Most of the remaining models commonly used either for maximum likelihood tree inference or estimation of pairwise evolutionary distances can be obtained by restricting the parameters in matrix (4), as shown in Figure 11. For instance, if the substitution types are divided into transversions, transitions between purines, and transitions between pyrimidines, we obtain the model of Tamura and Nei (1993; TrN) by requiring that $a = c = d = f$. Similarly, we can obtain Kimura's (1981) three-substitution-type (K3ST) model by requiring that all bases occur in equal frequency ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$) and dividing the substitution types into transitions ($b = e$), A ↔ T or C ↔ G transversions ($c = d$), and A ↔ C or G ↔ T transversions ($a = f$). Zharkikh (1994) described a model (SYM) that is equivalent to GTR except that it assumes equal base frequencies. Any other restriction of the relative rates from the general time-reversible model (e.g., $a = c$, $e = f$) is possible; all such models are also time-reversible.

Further restrictions on the parameters in matrix (4) lead to more familiar models. If we assume that the equilibrium frequencies of all bases are the same

($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$) and that all substitutions occur at the same rate ($a = b = c = d = e = f = 1$, the model reduces to that of Jukes and Cantor (JC) (1969):

$$\mathbf{Q} = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

The base frequency and substitution rate are typically combined into a single parameter $\alpha = \mu/4$, leading to the simpler form:

$$\mathbf{Q} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

Kimura's (1980) two-parameter model (K2P) takes into account the common observation that transitions and transversions occur at different rates, but still assumes equal base frequencies. Thus we set $a = c = d = f = 1$ and $b = e = \kappa$ and obtain

$$\mathbf{Q} = \begin{pmatrix} -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\ \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) \end{pmatrix}$$

Letting the transition rate $\alpha = \mu\kappa/4$ and the transversion rate $\beta = \mu/4$, the above can be rewritten as

$$\mathbf{Q} = \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix}$$

Note that $\kappa = \alpha/\beta$ represents the transition bias. When $\kappa = 1$, there is no preference for transitions and the model reduces to the JC model. However, because there are twice as many kinds of transversions as transitions, the expected transition:transversion ratio is 1:2. Similarly, if $\kappa = 4$, we would then expect twice as many transitions as transversions.

The K2P model can easily be generalized to allow unequal equilibrium base frequencies (Hasegawa et al., 1985b). The instantaneous rate matrix for this model (HKY85) is then given by

$$\mathbf{Q}=\begin{pmatrix} -\mu(\kappa\pi_G+\pi_Y) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T+\pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A+\pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C+\pi_R) \end{pmatrix} \qquad (5)$$

where $\alpha = \mu$, $\beta = \mu\kappa$, $\pi_R = \pi_A + \pi_G$, and $\pi_Y = \pi_C + \pi_T$. This corresponds to the GTR model with the constraints $a = c = d = f = 1$ and $b = e = \kappa$. The JC model can likewise be generalized to allow for unequal base frequencies (Felsenstein, 1981a; the F81 model) by setting $\kappa = 1$ in matrix (5) or, equivalently, requiring that $a = b = c = d = e = f = 1$ in matrix (4):

$$\mathbf{Q}=\begin{pmatrix} -\mu(\pi_Y+\pi_G) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_R+\pi_T) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_Y+\pi_A) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_R+\pi_C) \end{pmatrix} \qquad (6)$$

This model was also described as the "equal input" model by Tajima and Nei (1982).

Felsenstein (1984) used a different method to accommodate unequal base frequencies in a two-parameter model (the F84 model, formally described in Kishino and Hasegawa, 1989). The F84 model divides the substitution process into two components: a *general* substitution rate capable of producing all types of substitutions, and a *within-group* substitution rate that produces only transitions. The instantaneous rate matrix for the F84 model can be obtained from matrix (4) by setting $a = c = d = f = 1$, $b = (1 + K/\pi_R)$, and $e = (1 + K/\pi_Y)$:

$$\mathbf{Q}=\begin{pmatrix} - & \mu\pi_C & \mu\pi_G(1+K/\pi_R) & \mu\pi_T \\ \mu\pi_A & - & \mu\pi_G & \mu\pi_T(1+K/\pi_Y) \\ \mu\pi_A(1+K/\pi_R) & \mu\pi_C & - & \mu\pi_T \\ \mu\pi_A & \mu\pi_C(1+K/\pi_Y) & \mu\pi_G & - \end{pmatrix}$$

where K is the parameter determining the transition:transversion ratio, $\pi_R = \pi_A + \pi_G$, $\pi_Y = \pi_C + \pi_T$, and the diagonal elements are set to the negative of the sum of the off-diagonal elements in the corresponding row. The elements of the above matrix corresponding to transitions each have two components, because transitions can occur due to either the general substitution rate or the within-group rate. When $K = 0$, this model collapses to the F81 model. As K increases above zero, transitions occur more and more frequently relative to transversions.

Calculating Change Probabilities

The instantaneous rate matrix **Q** specifies the rates of change between pairs of nucleotides per instant of time *dt*, but in order to calculate likelihoods we need the probabilities of change from any state to any other along a branch of length *t*. The substitution probability matrix* is calculated as

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

(e.g., Cox and Miller, 1977; Hasegawa et al., 1985b; Z. Yang, 1994a). The exponential can be evaluated by decomposing the instantaneous rate matrix **Q** into its eigenvalues and eigenvectors (we omit the details of how this is done, but see Lewis et al., 1996, for an introductory explanation of the techniques used). For several models, simple expressions exist for the eigenvalues, allowing direct analytic calculation of the elements of the substitution probability matrix. For example, in the K2P model of DNA substitution, there are only three probabilities to consider: the probability of a transversion-type substitution; the probability a transition-type substitution; and the probability of no substitution. These probabilities are:

$$\text{K2P}: \quad P_{ij}(t) \;=\; \begin{cases} \dfrac{1}{4}+\dfrac{1}{4}e^{-\mu t}+\dfrac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & (i=j) \\ \dfrac{1}{4}+\dfrac{1}{4}e^{-\mu t}-\dfrac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & (i\neq j,\ \text{transition}) \\ \dfrac{1}{4}-\dfrac{1}{4}e^{-\mu t} & (i\neq j,\ \text{transversion}) \end{cases}$$

The full substitution probability matrix is then given by:

$$\mathbf{P}(t)=\begin{pmatrix} \frac{1}{4}+\frac{1}{4}e^{-\mu t}+\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{1}{4}e^{-\mu t}-\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} \\ \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{1}{4}e^{-\mu t}+\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{1}{4}e^{-\mu t}-\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} \\ \frac{1}{4}+\frac{1}{4}e^{-\mu t}-\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{1}{4}e^{-\mu t}+\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} \\ \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{1}{4}e^{-\mu t}-\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{1}{4}e^{-\mu t}+\frac{1}{2}e^{-\mu t\left(\frac{\kappa+1}{2}\right)} \end{pmatrix}$$

Substitution probabilities for some other DNA models are as follows (see Lewis et al., 1996):

*We refer to this matrix as the substitution probability matrix rather than the more traditional transition probability matrix to avoid confusion with "transition" in the sense of a change between two purines or between two pyrimidines.

$$\text{JC}: \quad P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-\mu t} & (i = j) \\ \frac{1}{4} - \frac{1}{4} e^{-\mu t} & (i \neq j) \end{cases}$$

$$\text{F81}: \quad P_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j) e^{-\mu t} & (i = j) \\ \pi_j (1 - e^{-\mu t}) & (i \neq j) \end{cases} \tag{7}$$

$$\text{HKY85, F84}: \quad P_{ij}(t) = \begin{cases} \pi_j + \pi_j \left(\frac{1}{\Pi_j} - 1 \right) e^{-\mu t} + \left(\frac{\Pi_j - \pi_j}{\Pi_j} \right) e^{-\mu t A} & (i = j) \\ \pi_j + \pi_j \left(\frac{1}{\Pi_j} - 1 \right) e^{-\mu t} - \left(\frac{\pi_j}{\Pi_j} \right) e^{-\mu t A} & (i \neq j, \text{ transition}) \\ \pi_j (1 - e^{-\mu t}) & (i \neq j, \text{ transversion}) \end{cases}$$

where $A = 1 + \Pi_j (\kappa - 1)$ for the HKY85 model and $A = K + 1$ for the F84 model, with $\Pi_j = \pi_A + \pi_G$ if base j is a purine (A or G) and $\Pi_j = \pi_C + \pi_T$ if base j is a pyrimidine (C or T). Substitution probabilities for the remaining models can be calculated by numerical evaluation of the eigenvalues and eigenvectors of **Q** using standard algorithms (Z. Yang, 1994a; Lewis et al., 1996).

CHANGE PROBABILITIES FOR PROTEIN SEQUENCE DATA The techniques described above for DNA sequence data can be applied to protein sequences as well; the difficulty lies in specifying an appropriate model of amino acid replacement. The simplest model is a Poisson model, analogous to the JC model for DNA sequences but extended to 20 states (e.g., Kishino et al., 1990), which assumes that all changes between amino acids occur at the same rate and that the equilibrium frequencies of all amino acids are equal. The change probabilities for this model are given by:

$$\text{Poisson}: \quad P_{ij}(t) = \begin{cases} \frac{1}{20} + \frac{19}{20} e^{-\mu t} & (i = j) \\ \frac{1}{20} - \frac{1}{20} e^{-\mu t} & (i \neq j) \end{cases}$$

The assumption of equal amino acid frequencies is clearly unreasonable for protein sequence data. If substitution rates are still assumed to be equal, an analog to the Felsenstein (1981a) model would have the same basic form as the instantaneous rate matrix of (6), but with 20 states instead of 4. This model has been called the proportional model by Hasegawa and Fujiwara (1993). The corresponding change probabilities are the same as (7):

$$\text{Proportional:}\quad P_{ij}(t) = \begin{cases} \pi_j + \left(1-\pi_j\right)e^{-\mu t} & (i = j) \\ \pi_j\left(1-e^{-\mu t}\right) & (i \neq j) \end{cases}$$

where π_j now represents amino acid frequencies rather than base frequencies. Although this model is preferable to the Poisson model, it still assumes that the relative frequencies of the amino acids are constant across sites. This assumption is clearly violated as well (e.g., hydrophobic amino acids predominate in some regions of a protein, while hydrophilic amino acids predominate in others).

A large body of empirical evidence demonstrates that an amino acid is more likely to be replaced by a physicochemically similar amino acid than would be predicted by an equal-change-probability model (Dayhoff et al., 1978). Kishino et al. (1990) were able to derive a maximum likelihood method analogous to the general time-reversible model for DNA sequences by using an instantaneous rate matrix derived from Dayhoff et al.'s (1978) empirical substitution matrix. This model has been implemented as the Dayhoff model in the PROTML program of the MOLPHY package (Adachi and Hasegawa, 1992). More recently, a model (JTT) based on the updated empirical substitution matrix of D.T. Jones et al. (1992) has been added to PROTML; preliminary evidence indicates that this modification provides a better model for the evolution of diverse proteins than the Dayhoff model (Cao et al., 1994).

Protein-coding DNA sequences can be analyzed using either the original DNA sequences or the translated proteins (with some complications). Some information is lost in the translation to protein sequences. On the other hand, an obvious limitation to use of the original DNA sequences is that the assumption of equal rates of change for all sites is violated due to the degeneracy of the genetic code; a greater proportion of synonymous changes allows third positions to evolve at a much more rapid rate than first and second positions. This problem is easily corrected by allowing relative rates to be specified on a site-specific basis (see below). However, selection at the amino acid or codon level will cause the assumption of independence among sites to be violated as well. Consequently, maximum likelihood analyses of protein-coding DNA sequences probably should be conducted at the protein level unless the sequences are not very divergent (see Reeves, 1992, for a discussion of these and related issues). An alternative is to use a model of codon evolution with 61 states (Muse and Gaut, 1994; Goldman and Yang, 1994), retaining the full information content of the DNA sequences. Unfortunately, codon-based models are still in their infancy and are much more computationally intensive than 4-state (or even 20-state) models.

THE RELATIONSHIP BETWEEN SUBSTITUTION RATE AND TIME For all of these models, the probability of a change from state i to state j depends on the interaction of the duration of time t and the substitution rate μ only through their product μt (Felsenstein, 1981a). Thus, a branch could be "long" either because it represents a long period of evolutionary time or because the rate of substitution has been high. In general, it is impossible to tease these two components apart unless one is willing to assume a perfect molecular clock. Consequently, the mean substitution rate μ is usually set to 1 and the relative rate parameters $a, b, ..., f$ are scaled so that the average rate of substitution at equilibrium is 1 (e.g., Z. Yang, 1994a). The length of a branch then represents the expected number of substitutions per site along that branch, with no implication as to the actual amount of evolutionary time it represents.

These models allow the expected number of substitutions to be different for each branch of the tree. As noted above, one consequence of this freedom is that the likelihood of a tree can be calculated independently of the location of the root. If one is willing to assume that the substitution rate is approximately homogeneous across lineages, then the likelihood can be estimated under a molecular clock model by estimating branching times rather than the lengths of each branch (Bishop and Friday, 1985; Felsenstein, 1993). (Note that this model then requires evaluation of rooted rather than unrooted trees.) Because the clock model requires estimation of only about half as many parameters as the unconstrained model $[(T-1)/(2T-3)]$, it will be more efficient (in the sense of requiring less data to achieve the same level of accuracy) if the clock assumptions are valid. Felsenstein (1993) outlined a likelihood ratio test of the molecular clock that compares the likelihoods of the more constrained clock model to the unconstrained-branch-length model.

CHOOSING AN APPROPRIATE MODEL In a phylogenetic analysis, model selection and evaluation are interrelated. There are two main criteria for evaluating a phylogenetic model: how well it fits the data at hand, and how well it fits with other reliable data (sometimes called *congruence* in the case of comparing trees). In selecting a model based on fit of data at hand, there are tradeoffs to consider. We can always improve the apparent fit of a model by adding additional parameters, but estimating these additional parameters also leads to higher sampling variances. Measures of fit are useful in deciding whether it is worth adding an extra parameter (see A.J. Miller, 1990).

The general approach is to choose an overall goodness-of-fit statistic and then search for a model that maximizes this statistic without adding unnecessary parameters that do little more than explain random fluctuations in the data. If we can assume that sites in the sequence evolve independently, then the data represent a multinomial sample, so goodness-of-fit statistics such as a χ^2 or the log likelihood ratio test (e.g., *G* of Sokal and Rohlf, 1981) can be used to measure the fit of the observed data to the predictions of the model (see Navidi et al., 1991 for a general discussion, and Ritland and Clegg, 1987 for examples). In phylogenetics it is more common to use the likelihood ratio statistic, which (unlike the χ^2 statistic) does not require the expected probability of all distinct nucleotide patterns to be calculated. As with a contingency table analysis, we expect that with a large amount of data, the G statistic will behave like a χ^2-distributed random variable, assuming the model is correct. (Likelihood-ratio tests of model fit are further described in the section on "Reliability of Inferred Trees.") A related measure, the Akaike information criterion (Akaike, 1974), can also be used to choose the most appropriate model (e.g., Kishino and Hasegawa, 1990), although in practice this measure is similar to a variety of other model selection criteria (see A.J. Miller, 1990). It is also important to avoid overconfidence when one model fits the data much better than another if the overall fit is not good, since both models could be quite inadequate.

Calculating the Likelihood of a Tree

To calculate the likelihood of a full tree, it is necessary to consider the likelihoods of the occurrence of each state at each node in the tree as a function of the tree topology and branch lengths. As with other methods that define the optimal tree in terms of an optimality criterion (e.g., least-squares and parsimony), we will assume that the tree is given, and that the present task is to determine how good it is. The method for evaluating the likelihood of a given tree proceeds from a hypothetical root node at any convenient location in the tree, and combines the likelihoods of each of its daughter trees (i.e., descendant lineages). (For time-reversible models, the choice of root location will not change the likelihood of the tree.) If A is an ancestor that gave rise to sequences B and C, then the conditional likelihood of state i at sequence position j in A is

$$L\left(x_{Aj}=i\right)=\left[\sum_k P_{ik}\left(v_{AB}\right)L\left(x_{Bj}=k\right)\right] \times \left[\sum_l P_{il}\left(v_{AC}\right)L\left(x_{Cj}=l\right)\right] \qquad (8)$$

where v_{xy} is the length of the branch joining sequence x to sequence y. We say "conditional" likelihood because this value actually represents the likelihood of the subtree descending from node A given that $x_A = i$. In words, the conditional likelihood that A has state i is the product of the likelihoods that the i could have given rise to the outcomes in B and C. The first term on the right-hand side is the probability of state i changing to state k in the interval v_{AB}, $P_{ik}(v_{AB})$, times the likelihood that sequence B has state k at the corresponding position, summed over all possible values of k. If B is a known sequence, then the likelihood that position j has state k is 1 if k is equal to the observed state in the sequence, or zero otherwise. On the other hand, if B is an ancestor, then the likelihoods of it having state k are derived recursively, by inserting another copy of the right-hand side of (8) into the equation. The second term in equation (8) is analogous to the first, but refers to the lineage leading to C. Calculating the likelihood of the entire evolutionary tree at sequence position j requires multiplying the conditional likelihood of each possible state at the root node, $L(x_{Aj} = i)$, by its prior probability, π_i, and summing over all ancestral states i. Usually the root node will be made coincident with one of the other nodes in the tree, eliminating one branch and one summation, as shown in Figure 10C. The product of the position-specific likelihoods is the overall likelihood of the tree. Again, this is usually expressed as a sum of the log-likelihoods for each position.

Figure 12 illustrates a tree of five sequences. The corresponding likelihood for a position j is:

$$\begin{aligned} L_{(j)} &= \sum_m \pi_m \\ &\times \left[\sum_k P_{m,k}\left(v_{FG}\right)P_{k,x_{Aj}}\left(v_{AF}\right)P_{k,x_{Bj}}\left(v_{BF}\right)\right] \\ &\times \left[\sum_l P_{m,l}\left(v_{GH}\right)P_{l,x_{Dj}}\left(v_{DH}\right)P_{l,x_{Ej}}\left(v_{EH}\right)\right] \\ &\times\ P_{m,x_{Cj}}\left(v_{CG}\right) \end{aligned} \qquad (9)$$

where A, B, C, D, and E are the original sequences; F, G, and H are the labels of the internal nodes; and the hypothetical root has been placed at node G. The overall likelihood would be the product over positions. The four factors of the outer summation are: (1) the prior probability of a state with identity m; (2) the conditional likelihood of state m at node G giving rise to state k at node F, and k giving rise to x_{Aj} at node A, and k giving rise to x_{Bj} at node B; (3) the conditional likelihood of m giving rise to l at node H, and l giving rise to x_{Dj} at node D, and l giving rise to x_{Ej} at node E; and (4) the conditional likelihood of m giving rise to x_{Cj} at node C. This basic pattern can be expanded to trees of any size.

In the above description, we implicitly assumed that the branch lengths were known, but of course these are in general unknown and must be estimated as part of the process of computing a likelihood. The methods for finding the branch lengths that maximize the value of the likelihood function are beyond the scope of this chapter, but

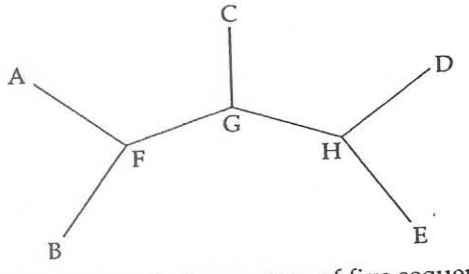


Figure 12 An evolutionary tree of five sequences. The known sequences are at the terminal nodes and are labeled A, B, C, D, and E. The nodes F, G, and H represent ancestral sequences. The likelihood of this tree for a particular site is calculated using equation (9).

typically involve an iterative approach in which each branch is optimized separately by Newton's method (e.g., Kishino et al., 1990; G.J. Olsen et al., 1992; Tillier, 1994; Lewis et al., 1996). This method is guaranteed to find globally optimal branch lengths for a given tree topology only if there is at most one maximum on the likelihood surface. Although Fukami and Tateno (1989) claimed to have proved this to be the case, Steel (1994b) presented a simple counterexample demonstrating that multiple optimality peaks could occur and found the error in Fukami and Tateno's proof. Steel's example was artificial, but preliminary results (J.S. Rogers and D.L. Swofford, unpublished data) indicate that the problem can occur with real data sets as well. So far, local optima seem to occur only on trees that provide extremely poor explanations of the data (e.g., random trees).

It is important to emphasize that the method for calculating likelihoods described in this section does not require calculation of the probabilities of each possible reconstruction of ancestral states as was shown in the conceptual example of Figure 10. The two methods are in fact equivalent, but if we were indeed required to consider all possible reconstructions, the problem would become essentially intractable, as there are 4^{T-2} possible reconstructions for DNA sequence data and 20^{T-2} possible reconstructions for protein sequence data. For example, a data set of 20 taxa and DNA sequences of length 2000 would require calculation of the probabilities of 1.4×10^{14} reconstructions for a given topology and set of branch lengths, and adjustment of even one branch length would require recalculation of all of them. It is extremely fortuitous that the probability summations can be rearranged into forms like equation (9) (corresponding to the "pruning" algorithm of Felsenstein, 1981a).

Evaluation of the likelihood of a tree and counting the number of changes of a tree under the general parsimony criterion are similar in several respects. The cost of a given change under parsimony is analogous to the likelihood of the given change from the substitution matrix, $\mathbf{P}(t)$. In parsimony, the cost of placing a given state at an internal node is the sum of the costs of deriving both of the daughter trees from that state, whereas the likelihood of an ancestral state is the product of the likelihoods of the state giving rise to the daughter trees. In parsimony, the total cost of the tree is the sum of the costs at each position, whereas the net log-likelihood of a tree is the sum of the log-likelihoods of the evolution at each sequence position. Essential differences between the general parsimony approach and the maximum likelihood approach include: the cost of a change in parsimony is not a function of branch length, unlike maximum likelihood; and maximum parsimony looks only at the single, lowest cost solution, whereas maximum likelihood looks at the combined likelihood for all solutions (ancestral states) consistent with the tree and branch lengths (see the discussion of integrated likelihood in Goldman, 1990). Felsenstein has used the relationship between likelihood and parsimony to gain several insights into the parsimony criterion, including the discovery of the potential for inconsistency due to unequal rates (Felsenstein, 1978a) and the inference of a character-weighting rationale (Felsenstein, 1981c).

Accommodating Rate Heterogeneity across Sites

The maximum likelihood models described above all assume that every site evolves at the same rate. Violation of this assumption can have devastating consequences. For instance, Gaut and Lewis (1995) showed that maximum likelihood inference under the assumption of rate homogeneity can become inconsistent when the true evolutionary process exhibits site-to-site rate variation, even when all other aspects of the process are modeled perfectly. If there is strong variation in rates across sites, sites that are resistant to change (e.g., due to strong selective constraints) can hide the actual amount of change that has occurred at more rapidly evolving sites. This causes maximum likelihood to underestimate the number of multiple changes; the longer the branch the greater the underestimation. Thus, maximum likelihood can become "positively misleading" (Felsenstein, 1978a) for exactly the same reasons as parsimony (Figure 8): highly divergent sequences will appear to be more closely related than they actually are (see Lockhart et al., 1995a, for a probable example of this problem with real data).

Rate heterogeneity can be incorporated into likelihood analyses by including an additional relative rate component, r, into the substitution probability expressions. In the JC model, for example, we let

$$P_{ij}(t,r) = \begin{cases} \frac{1}{4}+\frac{3}{4}e^{-\mu rt} & (i=j) \\ \frac{1}{4}-\frac{1}{4}e^{-\mu rt} & (i \neq j) \end{cases}$$

If the relative rates r are scaled so that the mean substitution rate remains 1, branch lengths will still reflect the number of substitutions per site. In the simplest case, we simply assign a rate r_j to each site j. Typically, the basis for this assignment would be some *a priori* classification of sites into functional categories and assignment of relative rates to the categories. Categorizations might be first, second, and third positions of a protein-coding gene, or paired versus unpaired sites for a ribosomal RNA gene. It is also possible to assign sites to rate categories based on the observed pattern of residue change. Van de Peer et al. (1993) proposed a way to do this by observing the frequency with which sequence pairs differ at each site as a function of the distance between the sequence pair. G.J. Olsen has written a program (DNArates; see Appendix) that performs a maximum likelihood estimate of the rate at each site for a given phylogenetic tree.

Several stochastic models that explicitly incorporate site-to-site rate variation are available. In these models, each site has a certain probability of evolving at any rate contained in some probability distribution, which may either be discrete or continuous. For a discrete rate distribution, the full likelihood for a given site is obtained by summing over rate categories the likelihoods of the site given each rate, weighted by the probability that the site is drawn from each category (Felsenstein, 1981a). Site likelihoods are calculated analogously for a continuous rate distribution except that the likelihoods must be integrated over the entire distribution.

The simplest model based on a discrete distribution is an invariable-sites model that assumes some fraction of the sites is incapable of accepting substitutions (perhaps due to strong functional constraint), but that the remaining sites all vary at the same rate (Hasegawa et al., 1985b; Churchill et al., 1992; Reeves, 1992; Sidow et al., 1992). In this case, when $r = 0$, $P_{ii}(t,r) = 1$ and $P_{ij}(t,r) = 0$ for all $i \neq j$. The proportion of invariable sites must either be estimated separately (see below) or treated as a parameter that is optimized for each tree. There is no reason in principle to restrict the rate of one of the categories to 0 (no change), or to limit the number of categories to 2, but estimation of the proportion of sites within each category and the relative rates among categories becomes much more complicated otherwise.

The most commonly used continuous distribution for modeling rate heterogeneity is the gamma (Γ) distribution (e.g., Z. Yang, 1993; Steel et al., 1993c). The Γ distribution has two parameters, a shape parameter α and a scale parameter β. By setting β to $1/\alpha$, a distribution with a mean rate of 1 is obtained, and a wide variety of rate distributions can be obtained by varying α (Figure 13).

The shape parameter α is equal to the inverse of the squared coefficient of variation of the substitution rate, so that as α increases, the distribution converges to an equal-rates model. Obtaining likelihoods by integrating over the Γ distribution (or any other continuous distribution) is usually computationally intensive (Z. Yang, 1993; see the section on Hadamard conjugation for a fast method under some models). Z. Yang (1994b) evaluated an alternative procedure in which the Γ distribution is divided into several rate categories by finding boundaries in the distribution such that each category has equal probability. The mean (or median) of each category is then used to represent all of the rates within that category. Z. Yang (1994b) found that this "discrete gamma" model can provide a good approximation with as few as four rate categories. The advantage of using a discrete model is that it requires only a tiny fraction of the computer time needed for the continuous Γ model. The discrete Γ distribution, like the continuous case, only adds one extra parameter to the model (the shape parameter), no matter how may rate categories are considered.

In some situations, mixtures of rate hetero-

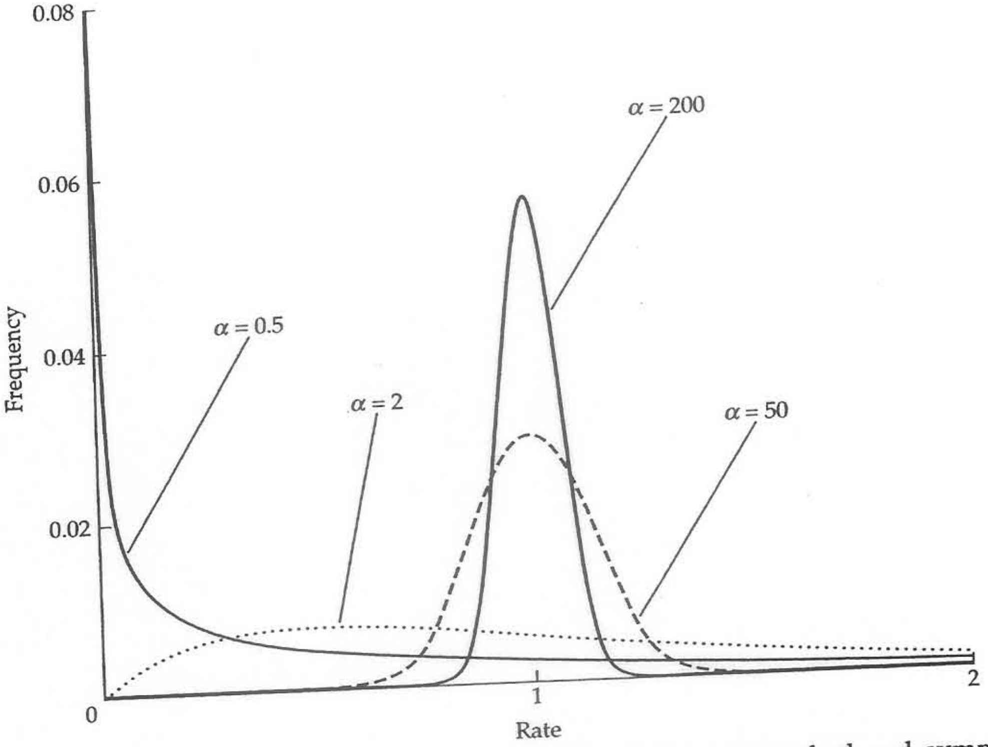


Figure 13 The gamma distribution for four different values of the shape parameter (α). When α is small, most of the sites evolve very slowly, but a few sites have moderate-to-high rates. As α increases, the distribution becomes more peaked and symmetrical about a mean rate of 1.0. When α is infinity, all sites have relative rate 1.0, so that an equal-rates model can be obtained as a special case of the gamma model.

geneity models may be appropriate. For example, Gu et al. (1995) and Waddell and Penny (1996a) have proposed an "invariant + gamma" model, in which some fraction of the sites, θ, are invariable, with the remaining rates distributed according to a Γ distribution with shape parameter α.

Estimating Model Parameters

The models described above contain a variety of parameters that must be estimated from the data or supplied on the basis of extrinsic evidence. These parameters include: the tree topology; the branch-length estimates (which are specific to each topology); the relative rate parameters of the substitution models ($a, b, ..., f$) in matrix (4) or related parameters such as κ and K; the base-frequency parameters (π_A, π_C, π_G, and π_T), and the parameters used in modeling rate heterogeneity (gamma shape parameter, proportion of invariable sites, etc.). Ideally, we would search for globally optimal values of these parameters in the n-dimensional parameter space. That is, we would consider every possible tree and optimize (jointly) all parameters of the model for each tree, choosing the resulting tree(s) of highest likelihood. For a given tree, one could perform a multidimensional optimization using Newton's method (e.g., A.W.F. Edwards, 1972). Unfortunately, this approach is difficult to implement becaause it requires knowledge of the first and second partial derivatives (and second cross-derivatives) of the likelihood function with respect to each of the parameters. Even when these derivatives are available, their computation can be quite slow.

In the section "Calculating the Likelihood of a Tree," we described a procedure that finds branch lengths that are at least locally optimal, given the values of any other parameters in the model. For any model more complex than the

JC/Poisson models, the values of additional parameters should be simultaneously optimized. When the model contains only one additional parameter (e.g., κ in the K2P model or the shape parameter in the JC+Γ model), it is relatively easy to plot the likelihood function evaluated at various values of the parameter of interest and thereby find a value that approximately maximizes the likelihood (e.g., Felsenstein, 1993). Obviously, this procedure can be quite tedious.

A method that has worked well for one of us (DLS) is the use of derivative-free methods for function minimization developed by Brent (1973) for a single variable and M.J.D. Powell (1964; as modified by Brent, 1973) for two or more variables. The procedure implemented in PAUP* (Swofford, 1996) is to use the Brent–Powell methods to find optimal values for all parameters other than branch lengths. When these algorithms need to evaluate the likelihood function, optimal branch lengths (conditional on the current values of the other parameters) are obtained using Newton's method as described above. Thus, optimal values of all parameters are obtained when the algorithm converges. (As for all heuristic methods, however, there is no guarantee that the resulting solution is globally optimal.) For small data sets (4–8 taxa), this strategy can be used for every tree evaluated due to the small size of the trees and the modest number of topologies tested. However, optimization of all model parameters on every tree tested dramatically slows the search using larger data sets. Z. Yang and coworkers (Yang, 1994a,b,c; Yang et al., 1994) have suggested that parameter estimates are fairly stable across tree topologies as long as the trees are not "too wrong" (Yang, 1995). Estimates of the shape parameter for the Γ model of site-to-site rate variation appear to be somewhat more sensitive to the tree topology than substitution-rate parameters (Yang, 1995; Sullivan et al., 1995b), although these conclusions are largely based on comparison of trees that probably fall into the "too wrong" category (e.g., random trees or star trees).

As long as parameter estimates are not wildly unstable across tree topologies, a potentially useful method would be to estimate the model parameters on some reasonably good tree for the data (e.g., a parsimony tree, or a maximum likelihood tree inferred under the model of Jukes and Cantor, 1969) and then "fix" the resulting estimates in a search for better trees under the desired model. A successive approximations approach might work very well in this case. That is, if a tree of higher likelihood is found, the parameters could be re-optimized on this new tree and fixed for yet another search, alternating between estimation and tree-searching until the same tree is found in successive iterations. Although this strategy seems quite promising, its effectiveness needs to be confirmed in empirical studies. Note that one of the limitations ascribed to the use of successive approximations in parsimony character weighting is not relevant in this case, because the likelihood function provides an objective function that is comparable across parameter values and trees.

An alternative to the methods presented above is to estimate the model parameters using methods other than likelihood. For example, the Γ shape parameter can be approximated by fitting a negative binomial distribution to a frequency distribution of the number of changes required at each site under the parsimony criterion (e.g., Uzzell and Corbin, 1971; Kocher and Wilson, 1991; Wakeley, 1993; Sullivan et al., 1995a). A similar approach can be used to estimate the proportion of invariable sites using the Poisson distribution (Fitch and Markowitz, 1970; Markowitz, 1970). Sidow et al. (1992) described another interesting method for estimating the proportion of invariable sites based on a mark–recapture model (Seber, 1982). These estimates require different assumptions than maximum likelihood tree models and can be calculated quickly, so they may be useful as a first approximation for selecting a model, obtaining starting parameter values for maximum likelihood estimation, or examining the effect of tree topology on parameter estimates (e.g., Sullivan et al., 1995a).

Maximum Likelihood Methods for Other Data Types

Maximum likelihood methods also can be applied to other data types, such as gene frequencies (Felsenstein, 1981b) or restriction sites (Felsenstein, 1992b). The basic approach is the same as that described above for sequence data: one formulates a model of evolutionary change and calculates the probability that the observed data (in this case, restriction site presences/absences or arrays of gene frequencies) would have been generated by a particular tree topology under the model. The mechanics of estimating branch lengths and other model parameters are essentially equivalent; the differences lie in the form of the models and how change probabilities are calculated.

Pairwise Distance Methods

A critical point made in the comparison of parsimony and likelihood methods above was that parsimony methods seek solutions that minimize the amount of evolutionary change required to explain the data, whereas likelihood methods attempt to estimate the actual amount of change according to an evolutionary model. This distinction is relevant because as mutations are fixed in the genome, there is an ever-increasing chance of **superimposed changes** occurring at a single sequence position: changes at a particular site along a lineage of the phylogeny may mask earlier changes at that site, and parallel or convergent changes may occur at the same site in different lineages. Thus, estimates of the amount of evolutionary change implied by parsimony will be underestimates of the true amount of change, unless the actual rate of change is extremely small.

An alternative to the use of likelihood for minimizing the impact of the underestimation problem is the use of corrected distances that account for superimposed changes by estimating the number of unseen events using the same sorts of models employed in maximum likelihood analysis. The corrected distances are then estimates of the true **evolutionary distance**, which reflects the actual mean number of changes per site that have occurred between a pair of sequences since their divergence from a common ancestor. Thus, following Cavalli-Sforza and Edwards (1967), we view distance methods as less desirable approximations to a full maximum likelihood approach. In recent simulation studies, maximum likelihood methods have consistently outperformed distance methods in choosing the correct tree (e.g., Kuhner and Felsenstein, 1994; Z. Yang, 1994c; Huelsenbeck, 1995a). Although some other studies have reported better performance of some distance methods (Saitou, 1988; Saitou and Imanishi, 1989; Tateno et al., 1994), these results have subsequently been shown to be based on inadequate computer programs and/or inappropriate comparisons (Hasegawa et al., 1991; Z. Yang, 1994c; Huelsenbeck, 1995b).

For some sources of data, including immunology and nucleic acid hybridization, there is no alternative to the use of distance methods. For other types of data, including macromolecular sequence, restriction site, and allozyme data, distances can provide a way to take advantage of models of evolutionary change when likelihood methods are either unavailable or intractable. Until recently, computers have been too slow and algorithms too inefficient to exploit fully the advantages of maximum likelihood techniques, and distance methods played a more important role. Even with the availability of faster maximum likelihood computer programs (see Appendix), distance methods remain useful, particularly for the analysis of large data sets, where their increased speed allows more thorough testing of alternative tree topologies.

The negative side of reducing character data to pairwise distances is that information is lost in the transformation. For instance, Penny (1982) has shown examples in which several different sets of sequences yield the same distance matrix, but given only the distances it is impossible to go back to the original sequences. Although this loss of information probably explains the better performance of character-based maximum likelihood inference, it clearly is not devastating. In fact, many sequence data sets yield identical conclusions with character-based and distance-based analyses (e.g., G.J. Olsen, 1987). Another drawback to distance analysis is that it does not lend itself to the combination of different kinds of data into the same analysis, as is possible for character-based analyses (e.g., Miyamoto, 1985). Finally, only through character-based analysis can a researcher identify particularly informative characters (or regions) in order to limit subsequent studies to those characters that are most useful (e.g., the detection of so-called "signature" events; Woese et al., 1980).

Additive Distances

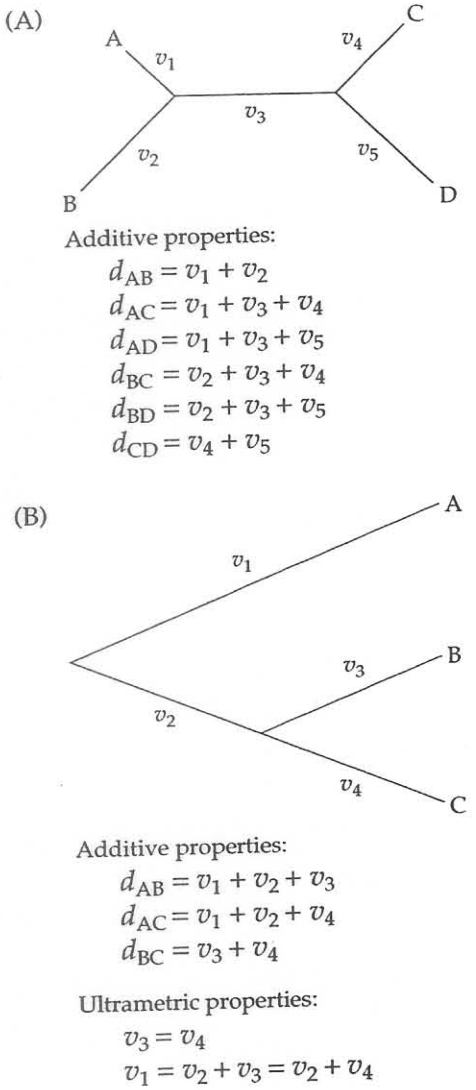
If we could determine exactly the true evolutionary distance implied by a given amount of observed sequence difference between each pair of taxa under study, these distances would have the very useful property of **tree additivity** (Figure 14): the evolutionary distance between each pair of taxa would be equal to the sum of the lengths of each branch lying on the path between the members of each pair. (The branch lengths also represent evolutionary distances between pairs of sequences, but at least one member of the pair is a hypothetical ancestral taxon.) Additive distances satisfy the *four-point metric condition* (Buneman, 1971): for any four taxa A, B, C and D,

$$\begin{gathered} d_{\mathrm{AB}} + d_{\mathrm{CD}} \leq \\ \max\left(d_{\mathrm{AC}} + d_{\mathrm{BD}},\ d_{\mathrm{AD}} + d_{\mathrm{BC}}\right) \end{gathered} \tag{10}$$

where d_{ij} is the distance between taxa i and j, and "max" is the maximum value function. Conceptually, this simply means that of the three sums of distances $d_{ij} + d_{kl}$ where $i \neq j \neq k \neq l$, one of these must be as small or smaller than the other two, and these other two must be equal. For example, in Figure 14A:

$$\begin{aligned} d_{\mathrm{AB}} + d_{\mathrm{CD}} &= v_1 + v_2 + v_4 + v_5 \\ d_{\mathrm{AC}} + d_{\mathrm{BD}} &= \left(v_1 + v_3 + v_4\right) + \left(v_2 + v_3 + v_5\right) = \\ &\quad v_1 + v_2 + v_4 + v_5 + 2v_3 \\ d_{\mathrm{AD}} + d_{\mathrm{BC}} &= \left(v_1 + v_3 + v_5\right) + \left(v_2 + v_3 + v_4\right) = \\ &\quad v_1 + v_2 + v_4 + v_5 + 2v_3 \end{aligned}$$

Tree-additive distances can be fitted to an unrooted tree such that all pairwise distances are equal to the sum of the lengths of the branches along the path connecting the corresponding taxa (Figure 14A). Unfortunately, due to the finite amount of available data, stochastic (random) errors will cause deviation of the estimated evolutionary distances from perfect tree additivity even



Additive properties:

$d_{\mathrm{AB}} = v_1 + v_2$
$d_{\mathrm{AC}} = v_1 + v_3 + v_4$
$d_{\mathrm{AD}} = v_1 + v_3 + v_5$
$d_{\mathrm{BC}} = v_2 + v_3 + v_4$
$d_{\mathrm{BD}} = v_2 + v_3 + v_5$
$d_{\mathrm{CD}} = v_4 + v_5$

Additive properties:

$d_{\mathrm{AB}} = v_1 + v_2 + v_3$
$d_{\mathrm{AC}} = v_1 + v_2 + v_4$
$d_{\mathrm{BC}} = v_3 + v_4$

Ultrametric properties:

$v_3 = v_4$
$v_1 = v_2 + v_3 = v_2 + v_4$

Figure 14 Additive and ultrametric trees. (A) An additive tree relating four taxa: A, B, C, and D. It also lists the relationships between the six taxon-to-taxon distances (d_{AB} through d_{CD}) and the five branch lengths (v_1 through v_5). Additive distances and trees do not make any assumption about the rooting; hence the relationships are displayed in an unrooted format. All sets of pairwise distances that satisfy the four-point condition (see text) can be represented as a unique additive tree. (B) An ultrametric tree relating three taxa: A, B, and C. In addition to having additive properties (all taxon-to-taxon distances are the total of the branch lengths joining them), every common ancestor is equidistant from all its descendants. Thus, the most recent common ancestor of B and C is v_3 from B and v_4 from C, therefore $v_3 = v_4$. Likewise, the common ancestor of A and B is v_1 from A and $v_2 + v_3$ from B, therefore $v_1 = v_2 + v_3$.

when evolution proceeds exactly according to the model used for distance correction. Many methods have been described that derive a tree and an associated set of branch lengths that comes closest (in some sense) to being additive for a matrix of pairwise distances. These methods typically, but not always, attempt to optimize an objective function that quantifies the degree of "distortion" between the path length and observed distances. The original descriptions of these methods often confound the choice of an optimality criterion with the algorithms used to select an optimal tree, but we will separate these two components, deferring the latter to the "Searching for Optimal Trees" section.

Additive-Tree Methods

A complete record of all genetic events would constitute a set of perfectly additive distances. We will treat the experimentally derived distances, which estimate the (unknown) number of genetic events that have actually occurred from the number of differences actually observed between each pair of taxa, as approximations of this ideal. To emphasize the uncertainty in the values, we will call them **distance estimates**. We can now address the problem of choosing a tree from the following conceptual perspective: We have uncertain data that we want to fit to a particular mathematical model (an additive tree) and find the optimal value for the adjustable parameters (the branching pattern and the branch lengths).

FITCH–MARGOLIASH AND RELATED METHODS

Several methods depend on a definition of the disagreement between a tree and the data based on the following family of objective functions:

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} w_{ij} \mid d_{ij} - p_{ij} \mid^{\alpha} \qquad \textbf{(11)}$$

where E defines the error of fitting the distance estimates to the tree, T is the number of taxa, w_{ij} is the weight applied to the separation of taxa i and j, d_{ij} is the pairwise distance estimate, p_{ij} is the length of path connecting i and j in the given tree, the vertical bars represent the absolute value, and $\alpha = 1$ or 2. A value of α and a weighting scheme must be chosen.

Setting α to 2 represents a weighted least-squares criterion; the weighted squared deviation of the path-length distances from the distance estimates will be minimized. If $\alpha = 1$, then the weighted absolute differences will be minimized. If the errors in the distance estimates are distributed uniformly across the data, then the least-squares criterion is preferred. If some estimates are apt to be particularly bad, there are two considerations. First, if the identities of the least certain estimates are known, this knowledge can be accommodated in the least-squares method by assigning particularly low weights to these uncertain values. If, however, it is not known *a priori* which estimates are apt to be erroneous, then using the minimum absolute deviations will reduce the overall perturbation caused by spurious data values. This last condition might pertain to direct experimental determinations of the distance data, a situation in which unrecognized experimental artifacts could substantially flaw some values.

The four most commonly used weighting schemes are:

$$w_{ij} = 1 \qquad \textbf{(12a)}$$

$$w_{ij} = 1 / d_{ij} \qquad \textbf{(12b)}$$

$$w_{ij} = 1 / d_{ij}^2 \qquad \textbf{(12c)}$$

$$w_{ij} = 1 / \sigma_{ij}^2 \qquad \textbf{(12d)}$$

where σ_{ij}^2 is the expected variance of measurements of d_{ij}. The first three equations amount to implicit assumptions about the uncertainty of the measurements: equation (12a) (Cavalli-Sforza and Edwards, 1967) assumes that all distance estimates are subject to the same magnitude of error; equation (12c) (Fitch and Margoliash, 1967) assumes that the estimates are uncertain by the same percentage; and equation (12b) could be viewed as a compromise that assumes the uncertainties are proportional to the square roots of the values (Felsenstein, 1993). Note that missing data can correctly be handled by setting the corresponding weight to zero; that is, if d_{ij} is unknown, setting $w_{ij} = 0$ will cause this observation to be ignored (although most currently available software does not allow specification of individual pairwise weights).

If there is a rational method for estimating σ_{ij}^2, then use of equation (12d) is preferable. Theoretical variance formulas are available for most of the model-based distances described below (although space limitations preclude their inclusion here, they are available in the original references). These theoretical variances can be used for DNA and protein sequence data, restriction site data, and gene frequency data. An important property of these formulas is that they explicitly state the dependence of uncertainty on the amount of data; e.g., for sequence-based distances, the variance is inversely proportional to the sequence length N. A problem, however, is that if two sequences are identical, the estimated uncertainty will be zero, which causes equation (12d) to be undefined and would be a questionable conclusion in any case. A practical treatment is to assume that the minimum measurable dissimilarity is one-half of a substitution, yielding (approximately) $1/(2N^2)$, as a minimum value to be imposed on the estimated variance.

For other kinds of data, including indirect methods such as DNA hybridization or immunological distances, random errors can be estimated by comparing replicate experiments or using reciprocal comparisons (where appropriate; see Chapter 6). These concepts are discussed in the corresponding experimental chapters.

For an unrooted tree of T taxa, there are $2T - 3$ independent branches that define the p_{ij} values, and there are $T(T - 1)/2$ distinct pairwise distances. To represent mathematically the relationships between the branch lengths, v_k, and the path lengths between pairs of taxa, we need an appropriate representation of the tree topology. Let $\mathbf{A}$ be a matrix of $T(T - 1)/2$ rows and $2T - 3$ columns such that the element $A_{(ij)k}$ is equal to 1 if the branch k is part of the path connecting taxon i to taxon j, otherwise $A_{(ij)k}$ is equal to 0. With this definition it follows that

$$p_{ij} = \sum_{k=1}^{2T-3} A_{(ij)k} v_k$$

Thus, a system of equations such as that of Figure 14A can be represented in matrix notation as

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{pmatrix} = \begin{pmatrix} p_{\mathrm{AB}} \\ p_{\mathrm{AC}} \\ p_{\mathrm{AD}} \\ p_{\mathrm{BC}} \\ p_{\mathrm{BD}} \\ p_{\mathrm{CD}} \end{pmatrix} \qquad \textbf{(13)}$$

$$\mathbf{A} \qquad \mathbf{v} = \mathbf{p}$$

If the distances were additive, then $p_{ij} = d_{ij}$ for all (i, j) pairs, and we could solve (13) directly. In general, however, due to the imperfect additivity of the distances, we must use (13) to eliminate p_{ij} from (11) and seek a solution to the v_k's that minimizes E. This minimization can be accomplished using special-purpose linear or quadratic programming algorithms (e.g., Barrodale and Roberts, 1973), by iterative successive refinement techniques ("alternating least-squares;" Felsenstein, 1993), or—when $\alpha = 2$ and $w_{ij} = 1$—by using ordinary linear algebra (e.g., Cavalli-Sforza and Edwards, 1967; Kidd and Sgaramella-Zonta, 1971; G.J. Olsen, 1988) using the equation:

$$\mathbf{v} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{d}\right) \qquad \textbf{(14)}$$

For weighted least-squares criteria like that of Fitch and Margoliash (1967), the linear algebraic solution is

$$\mathbf{v} = \left(\mathbf{A}^T\mathbf{W}\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{W}\mathbf{d}\right) \qquad \textbf{(15)}$$

where **W** is a $T(T-1)/2 \times T(T-1)/2$ matrix with diagonal elements equal to the weights associated with each pairwise comparison and all off-diagonal elements equal to 0.

The methods in the previous paragraph fit the data to a specific tree topology, and thus assume that an appropriate search strategy will be used to find the best topology. In an alternative approach described by De Soete (1983a,b), the values of p_{ij} are initially set to the observed distances (d_{ij}), and then they are gradually adjusted by an optimization regimen that keeps them at a local minimum of equation (11), while improving their fit to inequality (10) for all sets of four taxa. At the end of the process, all sets of p_{ij} satisfy inequality (10)—so they will perfectly fit some additive tree—and they are at a minimum of equation (11).

A problem that sometimes arises with the above methods is that full minimization of equation (11) requires that some of the v_k be negative. A negative branch length does not correspond to any meaningful biological process and should probably be avoided (e.g., Kidd and Sgaramella-Zonta, 1971). Allowing branches to have negative values when E is evaluated is probably inappropriate because some highly suboptimal trees can use negative values to produce a low apparent error. Several methods for dealing with negative branch lengths have been proposed. Some authors (e.g., Cavalli-Sforza and Edwards, 1967; Kidd and Sgaramella-Zonta, 1971) have favored outright rejection of any tree that requires a negative optimal value for any branch. This extreme approach runs the risk of rejecting the correct tree in certain realistic situations. An alternative strategy (Felsenstein, 1993) is to constrain the optimization process so that the negative branch lengths are disallowed; a solution that optimizes E under the constraint that all branch lengths be non-negative is obtained. If (14) or (15) is used to determine least-squares branch lengths, the only alternative is simply to set any negative branch lengths to zero and then calculate E without readjusting the other branches. This method gives exact values of E for trees that have no negative branch lengths and overestimates the value of E otherwise. The amount of overestimation is small as long as there are no large negative branch lengths.

Table 1 summarizes the results of a least-

Table 1

Optimal 5S rRNA tree by weighted least-squares criterion

Sequence pair[a]	Estimated distance[b]	Expected distance[c]	Distance difference[d]	Expected uncertainty[e]	Error contribution[f]
Bsu–Bst	0.1717	0.1655	0.0062	0.0522	0.00133
Bsu–Lvi	0.2147	0.2269	−0.0122	0.0600	0.00415
Bsu–Amo	0.3091	0.2895	0.0196	0.0758	0.00667
Bsu–Mlu	0.2326	0.2414	−0.0088	0.0630	0.00194
Bst–Lvi	0.2991	0.2958	0.0033	0.0743	0.00020
Bst–Amo	0.3399	0.3584	−0.0185	0.0809	0.00521
Bst–Mlu	0.2058	0.2058	0.0000	0.0584	0.00000
Lvi–Amo	0.2795	0.2795	0.0000	0.0708	0.00000
Lvi–Mlu	0.3943	0.3716	0.0227	0.0902	0.00633
Amo–Mlu	0.4289	0.4343	−0.0054	0.0906	0.00031

Data from G.J. Olsen, 1988. The corresponding tree is illustrated in Figure 15A.

[a] Abbreviations are as in Figure 15.

[b] Distance estimate from sequence comparisons, using equations (4) and (5), with $b = 3/4$.

[c] Sum of appropriate branch lengths along the path joining the taxa in the inferred tree.

[d] Difference of the two previous columns.

[e] Square root of the variance estimate from equation (16).

[f] The individual terms of the summation in equation 14, with $\alpha = 2$ and $w_{ij} = \sigma_{ij}^{-2}$.

squares calculation for a tree of five rRNA sequences. The table presents the pairwise distance estimates with their expected uncertainties, the corresponding path lengths through the inferred tree, and the error contributed by each distance to the overall value of E. As expected for a least-squares methodology, the paths through the best fitting tree will sometimes exceed the corresponding distance estimates (e.g., Bsu to Lvi) and sometimes they will be shorter (e.g., Bsu to Bst). It might be noticed that two distances are fitted exactly. Tree branch lengths assigned by most methods will exactly reproduce the distances between sister taxa in a tree (as long as negative numbers are not involved). The inferred tree is shown in Figure 15A.

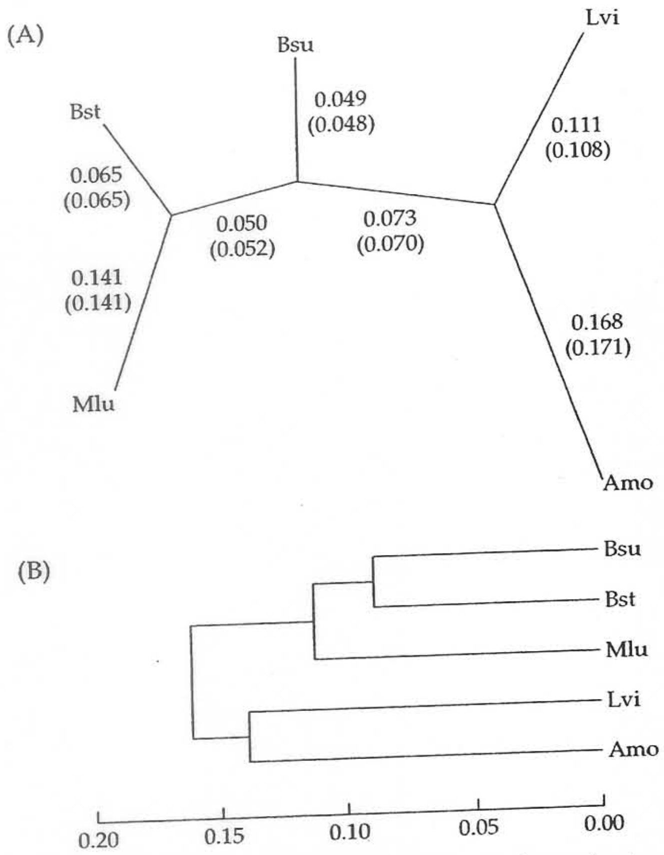


Figure 15 Comparison of 5S rRNA phylogenies inferred by different pairwise distance methods (data from Olsen, 1988). (A) Trees obtained using neighbor joining and weighted least-squares. The upper branch lengths (expected substitutions per sequence position) are from the neighbor-joining analysis in Figure 30, and the parenthetical values are from the weighted least-squares analysis in Table 1. Although the tree is unrooted, the *M. luteus* sequence is considered the outgroup. (B) Tree obtained by cluster analysis (UPGMA) from the analysis in Figure 29. It can be seen that the neighbor-joining and least-squares procedures produced very similar trees, but the cluster analysis tree is very different. Two of the sequences, those of *L. viridescens* and *A. modicum*, are very much more diverged than are the others, an effect to which cluster analysis is particularly sensitive. Abbreviations used to identify the taxa: Bsu, *Bacillus subtilis*, Bst, *Bacillus stearothermophilus*; Lvi, *Lactobacillus viridescens*; Amo, *Acholeplasma modicum*; and Mlu, *Micrococcus luteus*.

The least-squares and minimum-absolute-deviation approaches implicitly assume that each pairwise distance measurement is independent. Because of the common evolutionary history of the molecules in question, this assumption is not generally true. The primary consequence of violating this assumption is purely statistical; trees will be less well resolved than they would be if the samples were in fact independent. However, a second consequence is that any systematic errors in the distance estimates can also be multiply sampled, and thus the pairwise methods are potentially more sensitive to undercompensation for homoplasy in the data (see the section on "Systematic Error" later in this chapter). Felsenstein (1986, 1988a) discussed methods for dealing with interdependencies in pairwise distance data. In practice, none of these methods are used regularly because of their computational complexity and other limitations. Note that neither parsimony nor maximum likelihood suffers from this difficulty.

THE MINIMUM EVOLUTION METHOD Kidd and Sgaramella-Zonta (1971) suggested using the unweighted least-squares criterion (equation 11, with $w_{ij} = 1$ and $\alpha = 2$; Cavalli-Sforza and Edwards, 1967) to fit the branch lengths, but a different criterion to evaluate and compare trees:

$$\text{LS length} = \sum_{k=1}^{2T-3} |v_k| \qquad \textbf{(16)}$$

That is, the optimality criterion is simply the sum of the absolute values of the branch lengths that minimize the sum of squared deviations between observed (estimated) and path-length distances. Subsequent simulations indicated that the LS length criterion consistently outperformed least-squares criteria based on (11) (Kidd and Cavalli-Sforza, 1971). Apparently unaware of this work, Rzhetsky and Nei (1992a) described a method based on essentially the same criterion, calling it the minimum evolution method:

$$S = \sum_{k=1}^{2T-3} v_k \qquad \textbf{(17)}$$

The only difference between the two methods is that Rzhetsky and Nei drop the absolute values in equation (16), which has the seemingly undesirable property of allowing negative branch lengths to improve the apparent goodness-of-fit of the tree. In practice, however, the two methods are little different, because the branch lengths are usually non-negative (or very close to zero if negative) (Swofford, unpublished observations) on trees scoring well according to equation (16). The choice of the name "minimum evolution" is unfortunate, as the same name had been used earlier for a quite different method (Cavalli-Sforza and Edwards, 1967; Thompson, 1973). Because the earlier method was never widely used and the Rzhetsky–Nei method is becoming very popular, it seems best to refer to the methods defined by equations (16) and (17) as the minimum evolution (ME) method.

Rzhetsky and Nei (1992b) have provided a theoretical argument for the superiority of the ME method over the Fitch–Margoliash (FM) and related methods due to a bias in the latter methods when the variance of the estimated distances is high (e.g., due to large differences between short sequences). Although their computer simulations appeared to reinforce this conclusion, the actual reason for the better performance of ME is unclear, as the bias quickly becomes inconsequential as sequence length increases. It seems more plausible that the enhanced ME performance is due to a reduced impact of negative branch lengths in the ME method. Kidd et al. (1974) reported that if trees containing negative branch lengths are automatically rejected, the ME and FM methods give essentially identical results. Felsenstein and Kuhner's (1994) simulations also demonstrated a striking improvement in the performance of the FM method when branch lengths were constrained to be non-negative; in their study the performance of the FM method slightly surpassed an approximate method closely related to ME (the neighbor-joining method; see below), but only if negative branch lengths were disallowed.

Ultrametric Distances

Ultrametric distances are more constrained than tree-additive distances. Mathematically, ultrametric distances are defined by satisfaction of the three-point condition, which requires that for any three taxa A, B, and C,

$$d_{AC} \leq \max(d_{AB}, d_{BC}) \qquad \textbf{(18)}$$

This inequality simply states that two of the three pairwise distances between three taxa are equal and at least as large as the third. Phylogenetically, ultrametric distances will precisely fit a tree so that the distance between any two taxa is equal to the sum of the branches joining them, *and* the tree can be rooted so that all of the taxa are equidistant from the root (Figure 14B). The first half of this description defines an additive tree (and implies that ultrametric distances are additive). The second half of the description corresponds to the concept of a molecular clock that runs at the same rate in all lineages at any given moment. Two potential surprises may emerge, however. First, even with ultrametric data, there is no guarantee that the amount of divergence is *linear* in time. In particular, superimposed sequence changes, which decrease the observed molecular divergence, do not destroy the ultrametric property. Second, obtaining ultrametric data is extremely unlikely; even if the underlying substitution rate is perfectly constant, any finite sample will yield statistical fluctuations in the measured divergences. Consequently, even a universal substitution rate would not give ultrametric data without an infinitely large sample. The closest experimental approximations of infinite samples are genome hybridization measurements (Chapter 6), although measurement errors limit the effective amount of data (Felsenstein, 1987).

If data are nearly ultrametric by equation (18), which is rarely the case, methods that assume a molecular clock can be more efficient (require less data to achieve the same probability of inferring the correct tree). Felsenstein's (1993) KITSCH program uses the same criterion as equation (11) (with $\alpha = 2$), but constrains the lengths of the branches so that the total length from the root of the tree to each terminal taxon is the same. Cluster analysis methods (described below) are also appropriate under the assumption of a molecular clock, and are very fast. Colless (1970) provided a precise definition of how much deviation from ultrametricity can be tolerated without causing the estimation of the tree to become inconsistent. However, there is little practical reason to use cluster analysis because related methods such as neighbor joining are applicable to more general additive distances, require very little additional computation, and are often more efficient in simulation studies under a molecular clock model (Sourdis and Krimbas, 1987; Charleston, 1994) unless rates of substitution are high.

Distance Transformations for Sequence Data

MEASUREMENT OF SEQUENCE DISSIMILARITY By far the most common method of summarizing the relationship between two sequences is by their fractional (or percentage) similarity or dissimilarity. In its simplest form, the sequence dissimilarity is equal to the number of aligned sequence positions containing non-identical residues (bases or amino acids) divided by the number of sequence positions compared (in mathematics this distance is called the Hamming distance). However, we must explicitly address several subtleties and potential ambiguities: alternatives to limiting the comparison to identical residues; terminal length variation of molecules; alignment gaps; and treatment of ambiguities. The following sections assume that the sequence alignment has already been defined (see "Sequence Data" in the section "Types of Data" above, and Chapter 9).

It is frequently of interest to define the similarity of two molecules in terms of a more relaxed criterion than the fraction of identical residues, thereby changing our definition of *sequence dissimilarity* to the number of aligned sequence positions containing "non-synonymous" residues divided by the number of sequence positions compared. For example, "conservative substitutions" are commonly ignored when comparing proteins by pooling the amino acids into six groups: acidic (D, E), aromatic (F, W, Y), basic (H, K, R), cysteine, non-polar (A, G, I, L, P, V), and polar (M, N, Q, S, T). Residues within each group are considered synonymous; residues in different groups are considered non-synonymous.

As discussed above, if the evolution of a gene includes insertions and/or deletions, then gaps must be inserted to adjust for the internal length changes when aligning the contemporary sequences. Although the character state "gap" is sometimes treated as a fifth base or twenty-first amino acid, the processes responsible for base substitution and for insertion and deletion are evolutionarily and mechanistically distinct. Because a proper treatment is not obvious, sequence positions with gaps are usually omitted from analyses in one of two ways (e.g., Kumar et al., 1993; Swofford, 1996). The first (pairwise deletion) omits sites in which one or both sequences have a gap for each affected comparison. This option is appropriate when gaps are short and distributed approximately at random (Kumar et al., 1993). A second treatment (complete deletion) deletes a site from all pairwise comparisons if any of the sequences in the data set have a gap at that site. Although the complete deletion method discards more information, it may be more appropriate when some regions of a sequence (e.g., more rapidly changing regions) are more prone to insertion/deletion events than others, in which case pairwise deletion could introduce a bias. Alignment gaps are usually positioned to maximize the alignment of identical residues in sequences. Thus, additional insertion/deletion events could systematically raise the apparent similarity. Once again we emphasize that regions of the sequence alignment that contain substantial numbers of alignment gaps should be omitted from the analysis; positional homology is too uncertain for reliable estimates to be made from these regions.

Terminal length variation refers to the observation that corresponding molecules from different species (and even within an individual organism) can start and end at different distances from homologous features within the molecules. In addition to insertions and deletions, other genetic and physiological factors (e.g., substitution mutations or alteration of a processing enzyme) could be responsible for these variations. Because of the diversity of mechanisms, omitting the corresponding alignment columns, as in the second treatment above, seems most appropriate.

ACCOUNTING FOR SUPERIMPOSED EVENTS The raw dissimilarity (or similarity) is an appropriate value for summarizing the relationship between sequences. However, it is an inescapable fact that as genes accumulate mutations, there is an ever increasing likelihood that some of the changes will be at the same sequence location. Because pairwise comparisons of sequences are based entirely on the identity or non-identity of residues at corresponding sequence positions, the first substitution at a site will convert identical residues to non-identical residues. Subsequent changes at the same sequence position cannot further decrease the similarity, but they can raise the similarity by converting the compared residues to similar identities (parallelism or reversion). The net effect of this superimposition of substitutions is that dissimilarity does not increase uniformly with the number of events; instead, it increases rapidly at first and more slowly thereafter. Thus, correction of the distance to account for the unobserved substitutions is necessary for the distances to conform to an additive-tree model, unless all sequences are extremely similar. We show some of the more common distance corrections below, but see Kumar et al. (1993) and Swofford (1996) for more complete compilations.

A general framework for describing distance measures under a variety of models uses a divergence matrix $\mathbf{F}_{xy}$ to represent the relative frequencies of each nucleotide (or amino acid) pair in a given pairwise comparison of two sequences X and Y, e.g.:

$$\mathbf{F}_{xy} = \begin{pmatrix} n_{AA}/N & n_{AC}/N & n_{AG}/N & n_{AT}/N \\ n_{CA}/N & n_{CC}/N & n_{CG}/N & n_{CT}/N \\ n_{GA}/N & n_{GC}/N & n_{GG}/N & n_{GT}/N \\ n_{TA}/N & n_{TC}/N & n_{TG}/N & n_{TT}/N \end{pmatrix} \qquad \textbf{(19)}$$

where n_{ij} is the number of times sequence X has state i aligned next to state j in sequence Y, and $N = \Sigma n_{ij}$. Let us represent this matrix as

$$\mathbf{F}_{xy} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{pmatrix}$$

A frequently overlooked issue in pairwise sequence comparison is the treatment of ambiguities (i.e., nucleotide bases or amino acid residues of uncertain identity) in the sequences being compared. For example, counting a purine (R) as synonymous with A and G and non-synonymous with C and T will tend to overestimate the similarity between the affected sequence comparisons. One approach (Swofford, 1996) is to distribute differences between sites with ambiguities based on the frequencies of differences at unambiguous sites. For instance, suppose that a site has an A in sequence X and an R in sequence Y. If for this comparison there are 450 sites that have an A in both sequences, and 50 sites that have an A in one sequence and a G in the other, then the site would contribute 450/500 = 0.9 to the value of $(F_{xy})_{AA}$, and 0.1 to $(F_{xy})_{AG}$. Maximum likelihood distances (see below) can deal with the ambiguities directly (e.g., Felsenstein, 1993) by considering the likelihood of each possible resolution of the ambiguity.

The uncorrected distance, often referred to as the dissimilarity (D) or p-distance (e.g., Kumar et al., 1993), is simply the total number of differences divided by the total number of available sites:

$$p\text{-distance}: \quad \begin{aligned} d_{xy} &= b+c+d+e+g+h+i+j+l+m+n+o \\ &= 1-(a+f+k+p) \end{aligned}$$

A pairwise distance estimate is essentially the branch length in an optimal phylogenetic tree of two taxa. Thus, most of what was said about models for maximum likelihood tree inference (above) also applies here.

The corrected distance for the Jukes–Cantor model, which assumes equal rates of substitution between all pairs of bases, is calculated as

$$\text{JC}: \quad \begin{aligned} D &= 1-(a+f+k+p) \\ d_{xy} &= -\tfrac{3}{4}\ln\left(1-\tfrac{4}{3}D\right) \end{aligned} \qquad \textbf{(20)}$$

Note that the maximum expected dissimilarity is 0.75; if D equals or exceeds this value, the distance becomes undefined because the argument of the logarithm becomes negative. A distance for the model of Felsenstein (1981a), which relaxes the assumption of equal base frequencies, is given by

$$\text{F81}: \quad d_{xy} = -B\ln(1-D/B) \qquad \textbf{(21)}$$

where D is the same as for JC above and $B = 1-\left(\pi_A^2+\pi_C^2+\pi_G^2+\pi_T^2\right)$ (Tajima and Nei, 1982). The base frequencies can either be estimated for each pair of sequences compared or for the full set of sequences; we favor the latter due to lower sampling variance.

Comparison of equations (20) and (21) reveals that (21) can be used to calculate a distance for the JC model if we set $B = 3/4$. In fact, (21) is a very general formula that can also be used for calculating distances from protein sequences. A distance for the Poisson model is obtained if we set $B = 19/20$, and a distance for the Proportional (unequal amino acid frequency) model is obtained by setting

$$B = 1 - \sum_{i=1}^{20} \pi_i^2$$

where the π_i's now represent the frequencies of each amino acid, and D represents the proportion of amino acid differences between the two sequences.

The distance for Kimura's (1980) two-parameter model is calculated from the proportions of transition-type differences (P) and transversion-type differences (Q):

$$\text{K2P:}\quad \begin{aligned} P &= c+h+i+n \\ Q &= b+d+e+g+j+l+m+o \\ d_{xy} &= \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right)+\frac{1}{4}\ln\left(\frac{1}{1-2Q}\right) \end{aligned}$$

Note that the proportion of transitions and transversions is estimated separately for each pair of taxa, in spite of the fact that different pairs of taxa share common lineages on the tree. For many models more complex (and general) than the K2P model, no simple distance formula exists (e.g., Z. Yang, 1994a; Zharkikh, 1994). For example, the HKY model (the unequal base-frequency generalization of the K2P model) does not have a simple distance formula (see Z. Yang, 1994a for an explanation). However, the closely related F84 model does have a simple distance formula (Tateno et al., 1994):

$$\text{F84:}\quad d_{xy} = -2A\ \ln\left(1-\frac{P}{2A}-\frac{(A-B)Q}{2AC}\right)+ \\ 2(A-B-C)\ln\left(1-\frac{Q}{2C}\right)$$

where $\pi_Y = \pi_C + \pi_T$, $\pi_R = \pi_A + \pi_G$, $A = \pi_C\pi_T/\pi_Y + \pi_A\pi_G/\pi_R$, $B = \pi_C\pi_T + \pi_A\pi_G$, and $C = \pi_R\pi_Y$, and P and Q are as defined for the K2P model. The most general model for which a simple distance formula exists (Z. Yang, 1994a) is that of Tamura and Nei (1993) (not shown), which generalizes the HKY model to allow different rates for transitions between purines versus those between pyrimidines.

Lanave et al. (1984) and Rodríguez et al. (1990) have formulated a distance for the most general time-reversible model (GTR). [Although the algorithms described in these two papers are quite different, the methods are actually equivalent (Lewis and Swofford, unpublished; see Swofford, 1996).] The Rodríguez et al. version of this distance is

$$\text{GTR:}\quad d_{xy} = -\,\text{trace}\left[\mathbf{\Pi}\ \ln\left(\mathbf{\Pi}^{-1}\mathbf{F}_{xy}\right)\right]$$

where $\mathbf{\Pi}$ is a diagonal matrix of the average base frequencies in sequences X and Y. (Interpretation of this formula requires some familiarity with matrix algebra. Note in particular that evaluating the log of a matrix requires, among other things, determination of its eigenvalues and eigenvectors.) Lewis and Swofford (unpublished; see Swofford, 1996) have developed an extension of the Lanave et al.–Rodríguez et al. method that allows estimation of distances under any special case of the GTR family of models. When simple formulas such as the ones shown above exist, the Lewis–Swofford method gives identical results, but it allows calculation of distances for many models for which distances were unavailable previously.

ESTIMATING TRANSITION AND TRANSVERSION SUBSTITUTIONS SEPARATELY If transversions occur much less frequently than transitions and the amount of divergence is high, transition differences are likely to approach or reach saturation. When this happens, transitions will contribute little phylogenetic information and will cause inflation of the variance of the evolutionary distance estimates. In such situations, it may be preferable to estimate the phylogeny using transversion data alone, minimizing the impact of the noisy transitions (Goldstein and Pollock, 1994). All of the distance formulas described above can be modified to estimate the number of transitions and transversions per site separately (see Kumar et al., 1993, and Swofford, 1996, for compilations of these methods). Alternatively, one could recode the nucleotide states into R (A or G) and Y (C or T) and apply a two-state distance correction (analogously to the transversion parsimony method). Alternative distances have been proposed for the K2P model that make separate estimates of the number of transition versus transversion substitutions and use a weighted combination of these as the estimate of the evolutionary distance (Schöniger and von Haeseler, 1993; Goldstein and Pollock, 1994; Tajima and Takezaki, 1994). These methods appear to be much more reliable for tree inference than the usual K2P distance (Pollock and Goldstein, 1994).

PROTEIN-CODING DNA SEQUENCES In principle, knowledge of the gene sequence should be more informative than the corresponding protein sequence. In practice, at least two factors call this assertion into question. First, silent substitutions in protein-coding genes are much more frequent than replacement substitutions; thus the third codon positions tend to become randomized quickly and convey very little information about distant phylogenetic relationships. Second, the base composition of the third codon position appears to vary systematically between some species, thereby indicating that it can be subject to at least a moderately strong selective force that is different in different lineages. The presence of directional selection can lead to profound sequence convergences and consequent errors in inferred relationships. With these considerations in mind, three relatively simple strategies can be used to analyze protein-coding sequences, and a host of moderately to extremely complex alternatives exists.

The simplest method of calculating distances between sequences for protein-coding genes is to apply the distance formulas above directly to the gene sequence without special treatment. This method is reasonable, or even preferred, when the total amount of divergence is very small, in which case the resulting trees are based primarily on silent substitutions in the genes. The main drawback is that a systematic undercorrection for superimposed substitutions will result, since the assumption that all positions are equally subject to change will clearly be violated. If the amount of sequence divergence is truly small, then superimposed changes will be rare and the undercorrection will be negligible.

The second approach is to restrict the analysis to the first two nucleotides of each codon. This strategy is appropriate when a substantial sequence divergence is apparent. The rationale is that the third codon position will be largely randomized and hence phylogenetically uninformative. This approach, by definition, also circumvents the problem of the third codon position changing more rapidly than the first two and reduces the degree of violation of the assumption that all sites are changing at the same rate.

The third basic method is to infer the protein sequence from the gene sequence and perform the phylogenetic analysis at the protein level. This approach has two merits: (1) the protein is the most biologically relevant aspect of the gene (taken as a whole); and (2) the sequence can be compared with homologous molecules that were sequenced at the protein level, for which nucleotide sequences are therefore unknown. In addition to the distances for the Poisson and Proportional models described above, PHYLIP (Felsenstein, 1993) provides a distance under the Dayhoff model.

The more complex methods involve estimating the numbers of synonymous (silent) and non-synonymous (replacement) substitutions separately. When the maximum divergence between taxa is low, distances based on synonymous changes may reduce the effect of among-site rate variation, as synonymous substitutions are largely neutral (Kumar et al., 1993). For more distantly related taxa, restriction to non-synonymous changes tends to minimize the impact of noise contributed by a large number of silent changes. Many methods have been proposed for estimating synonymous versus non-synonymous substitutions (W.-H. Li et al., 1985b; Nei and Gojobori, 1986; W.-H. Li, 1993b; and references cited therein). These methods differ in the details of how they deal with multiple substitution pathways when two codons are more than one substitution apart and how they account for different levels of degeneracy (e.g., a site in a sequence is twofold degenerate if one of the three possible changes is synonymous and fourfold degenerate if all possible changes at the site are synonymous).

MAXIMUM LIKELIHOOD DISTANCES The most straightforward (and computationally intensive) method for estimating evolutionary distances is to apply maximum likelihood according to the models described under "Models of Sequence Evolution." As noted above, the "tree" in this case is a single branch, and we estimate the branch length (expected number of substitutions per site) that maximizes the probability of one sequence evolving from the other. (Because of the time-reversibility of the models, it makes no difference which sequence is considered ancestral.) Felsenstein's (1993) DNADIST program obtains maximum likelihood estimates of distance under the JC, K2P (with or without a gamma-correction for among-site rate variation), and F84 models, but the same approach easily could be adapted to accommodate other models. Many (but not all) of the distance formulas presented above are maximum likelihood estimators (e.g, see Zharkikh, 1994). However, direct use of maximum likelihood to calculate the distance has a number of advantages. Most importantly, it allows model parameters, such as the transition:transversion ratio, to be maintained at a consistent value across all pairwise comparisons (e.g., although the standard K2P distance formula is a maximum likelihood estimate when estimating the transition:transversion ratio independently for every pair, the distance must be numerically evaluated using maximum likelihood in order to use a fixed ratio as a means of reducing sampling variance). Maximum likelihood estimation also provides a very clean way of handling missing or ambiguous data, as the probability of observing each of the bases allowed by the ambiguity can be explicitly evaluated.

Although maintenance of substitution-model parameters at a consistent value is an advantage of maximum likelihood distances, it adds the burden of specifying their values. One possible way of estimating these parameters is to perform phylogenetic analyses using a range of parameter values, then choose the parameter settings that maximize the additivity of the distances on the best tree(s) found (e.g., that minimize the value of E in equation 11). Alternatively, parameters may be estimated using maximum likelihood on a few "reasonable" trees obtained using simpler distances. If the parameter estimates are reasonably similar across these trees, it is probably safe to use their mean value (or the value from the tree that had the highest likelihood) as the parameter value for calculating distances as input to a tree search using a distance criterion. This hybrid approach can be an effective compromise between a full search under the maximum likelihood criterion (which may be computationally infeasible) and an arbitrary choice of parameter values using a distance criterion.

TREATMENT OF UNDEFINED VALUES Distance values become undefined if the apparent sequence divergence exceeds the maximum possible (true) distance under the assumed model of evolution. For example, in the JC model, complete randomization of sequences would lead to $D = 0.75$ (i.e., even for two random sequences, one-fourth of the nucleotides are expected to be identical by chance). If the observed dissimilarity equals or exceeds 0.75 due to sampling error or violation of the model, the logarithm in equation (20) cannot be taken. In this situation, it is probably wise not to proceed without taking steps to avoid problems due to excessive saturation. If only one or two sequences are causing the problem, they can be eliminated from the analysis. If the problem is mostly due to high rates of transition-type differences, transversion-only distances (or maximum likelihood distances with a high transition:transversion ratio) can be employed. As a last resort, any undefined distances can be replaced by an arbitrarily large distance value, such as twice the maximum observed distance.

ACCOMMODATING AMONG-SITE RATE VARIATION IN DISTANCE CORRECTIONS Distance corrections that assume equal rates of change across sites will be affected by the same problem that complicates maximum likelihood analysis when among-site rate heterogeneity exists: distances will underestimate the actual number of substitutions (Golding, 1983). Fortunately, this rate heterogeneity can be accommodated without too much difficulty. For maximum likelihood distances, any of the model variations described under "Accommodating Rate Heterogeneity Across Sites" in the "Maximum Likelihood Methods" section can be applied directly. If rates are assumed to follow a gamma distribution, special modifications of the distances described above are available for the JC and K2P models (Jin and Nei, 1990) and TrN models (Tamura and Nei, 1993). Although not noted by these authors, these "gamma" distances can be obtained from the usual distances simply by replacing the function $\ln(x)$ with $\alpha(1 - x^{-1/\alpha})$ in the original formulas, where α is the shape parameter of the gamma distribution (this function is the inverse of the moment generating function for the distribution). In fact, this method also works for most (if not all) of the other time-reversible distances (Waddell and Steel, 1995; Lewis and Swofford, unpublished; see Swofford, 1996). For example, the general time-reversible distance with a distribution of rates across sites can be written as $d_{xy} = -\text{trace}\{\mathbf{\Pi}\ [\mathbf{M}^{-1}(\mathbf{\Pi}^{-1}\mathbf{F}_{xy})]\}$, where $\mathbf{M}^{-1}$ is the same function used for the Jin–Nei and Tamura–Nei distances in the case of the gamma distribution, but can be the inverse of the moment-generating function for other distributions as well (Waddell and Steel, 1995). The value of α must be determined independently using one of the methods outlined above. Choice of an α value based on results from previous studies is also an option (e.g., Kumar et al., 1993), although evidence is accumulating that levels of rate heterogeneity vary widely among different genes, regions of genes, and organisms.

The invariable sites model (see above) can also be applied to distance estimation by removing a certain fraction of the constant sites from the data matrix. The easiest way to accomplish this is to subtract the constant $\phi N/4$ from the diagonal entries of n_{ij} in matrix (19) (and adjusting N accordingly) before calculating the distance, where ϕ is the desired proportion of invariable sites and N is the total number of sites (Waddell, 1995). If base frequencies are unequal, it is preferable to subtract $\pi_k\phi N$ from the kth diagonal element of the divergence matrix, where π_k is the frequency of base k. When base composition is not homogeneous throughout the tree, or in other situations where constant sites have a different composition than the variable sites, the base frequencies used in this correction should be estimated from the constant sites alone.

LOG-DETERMINANT DISTANCES The models described above for maximum likelihood and distance estimation assume that the substitution probability matrices remain constant throughout the tree (i.e., they are **stationary**) and that they have the property of time reversibility (which jointly imply that base frequencies remain at a constant, equilibrium value). The LogDet (Steel, 1994a; Lockhart et al., 1994) or paralinear distance (Lake, 1994) is a transformation that yields additive distances under a much wider set of models. Perhaps most importantly, this transformation is robust to changing base composition (e.g., GC bias) among the taxa being studied—a potential source of systematic error if stationary models are assumed. The LogDet transformation will yield an additive distance (in expectation) under any Markov model of evolution (see above) as long as sites evolve identically and independently and rates of substitution are equal across sites. This general Markov model is described by a rooted tree, where the root can have any base composition (as long as all states have a non-zero frequency). There are no constraints on the parameters in each substitution probability matrix $\mathbf{P}(t)$ (all 12 substitutions are free to occur at different rates), and $\mathbf{P}(t)$ can be different for each branch or at different points along the same branch. Each $\mathbf{P}(t)$ matrix implies its own set of stationary base composition values, so these are also allowed to vary throughout the tree. These assumptions correspond to those of the maximum likelihood model proposed by Barry and Hartigan (1987a).

The basic form of the log-determinant distances is

$$d_{xy} = -\ln\left(\det\ \mathbf{F}_{xy}\right) \tag{22}$$

(Steel, 1994a), where "det" refers to the determinant* of a matrix and $\mathbf{F}_{xy}$ is an $r \times r$ divergence matrix for sequences X and Y (e.g., equation 19)

*The definition of the determinant of a matrix is beyond the scope of this chapter. Introductions to matrix algebra can be found in many statistics texts or any linear algebra text. An excellent introduction for biologists is Bulmer (1994, p. 298 ff.).

with r equal to the number of character states (e.g., $r = 4$ for DNA sequences). For identical sequences, d_{xy} should be set to zero, although in practice, equation (23) is used instead, in which case no explicit treatment is needed for this case. If evolution proceeds according to the model described in the above paragraph, distances calculated using equation (22) will have the property of tree additivity (apart from sampling error), but in general, this expression cannot be used to estimate the number of nucleotide substitutions per site (evolutionary distance). However, for stationary models, the value obtained from (22) can be scaled to a distance that is *proportional to* the evolutionary distance using the formula

LogDet:

$$d_{xy} = \left[-\ln\left(\det \mathbf{F}_{xy}\right) + \tfrac{1}{2}\ln\left(\det \mathbf{\Pi}_x\mathbf{\Pi}_y\right)\right]\Big/ r$$
$$= -\tfrac{1}{r}\ln\left(\frac{\det \mathbf{F}_{xy}}{\sqrt{\det \mathbf{\Pi}_x\mathbf{\Pi}_y}}\right) \qquad (23)$$

where $\mathbf{\Pi}_x$ and $\mathbf{\Pi}_y$ are diagonal matrices of the character-state frequencies in sequences X and Y, respectively (Lockhart et al., 1994). The expected value of this distance will be equal to the mean number of substitutions per site if base frequencies are all equal, in which case

$$\tfrac{1}{2}\ln\left(\det \mathbf{\Pi}_x\mathbf{\Pi}_y\right) = -r\ln r$$

Otherwise, it will overestimate the evolutionary distance by a constant factor that becomes larger as base composition becomes more unequal (Waddell, 1995). Note that equation (23) is equivalent to Lake's (1994) paralinear distance except for the scaling by $1/r$. (Lake did observe, however, that his paralinear distance was approximately equal to r times the mean number of substitutions per site.) For non-stationary models, (23) tends to overestimate the mean number of substitutions, but it can also be an underestimate, depending on the base composition at internal points of the tree. Even under non-stationary models, however, the LogDet distance often provides better estimates of the number of substitutions per site than any of the standard distance transformations, because non-stationary base composition can lead the standard formulas to over- or underestimate the true distance by large amounts (Waddell, 1995). Thus, as a general rule, the branch lengths of a tree estimated using the LogDet distance should be considered just as useful as any other distance when base frequencies are not homogeneous.

A concern with using any distance transformation derived from a very general model is that it will suffer from inflated sampling errors, making it less reliable for tree selection unless sequences are very long. This concern appears to be unjustified, however, as the sampling variance of LogDet distances can approximate that of even the most simple (but least general) distance transformations described above (Waddell, 1995). For example, when applied to simple, stationary models with equal base frequencies, the variance of the LogDet distance (Lockhart et al., 1994) becomes equal to that calculated by the usual variance formulas. Furthermore, four-taxon computer simulations (D.L. Swofford, P.O. Lewis, and P.J. Waddell, unpublished) show that when data are simulated according to any of the models in the GTR family (Figure 11), the minimum evolution method using LogDet distances leads to recovery of the correct tree about as often as using other distance measures—including the distance specific to the simulation model—for all but very short sequences (<200 bases).

The LogDet can be applied to amino acid sequences (Lake, 1994 gave a four-taxon example), or even using each of the 61 non-stop codons as character states. The variance of the LogDet may become more of a problem in these situations, so it may be useful to group some states together (e.g., into the six main amino acid classes). Another problem is that a state may be entirely absent in one or more of the sequences. In this case, the determinants of $\mathbf{F}_{xy}$ and of $\mathbf{\Pi}_x$ and/or $\mathbf{\Pi}_y$ will be zero (yielding an undefined distance when the log is taken). The best way to deal with this situation remains to be determined; possible solutions include removing the state from the $\mathbf{F}_{xy}$ matrix altogether (if the state is absent from all of the sequences), pooling this state with another, or set ting the corresponding elements of $\mathbf{F}_{xy}$ to some small value such as $1/(2N)$.

Lockhart et al. (1994) found that use of LogDet distances yielded more believable trees in three examples for which nucleotide composition was variable over taxa. However, a weakness of the standard LogDet transform in real applications is that it is no more robust to unequal substitution rates at different sites than are other distance measures (Barry and Hartigan, 1987b; Lockhart et al., 1994; Lake, 1994). Lockhart et al. (1994) reported that for some data sets, reasonable trees could be obtained only after eliminating sites that were uninformative according to the parsimony criterion, and suggested that inclusion of sites that were highly unlikely to change might be the cause of the problem. Unfortunately, unlike the less general distance transformations, LogDet distances cannot be directly modified to take account of a specific distribution of rates such as the gamma distribution.

Waddell (1995) has shown that by subtracting an appropriate proportion of invariant (constant) sites from the diagonal elements of $\mathbf{F}_{xy}$ (see "Accommodating Among-Site Rate Variation in Distance Corrections," above), LogDet distances can become nearly additive even if the true distribution of rates across sites follows a continuous distribution such as the gamma. Methods of estimating the proportion of invariable sites for maximum likelihood and other distance transformations perform well, whereas simple removal of parsimony-uninformative sites tends to be too severe. However, as base composition becomes more heterogeneous over taxa, sites with different rates of change also change base composition with respect to each other. Thus, it may be important to estimate base frequencies using only the constant sites, rather than the full data set, when calculating the proportion of sites to remove from the diagonal elements of $\mathbf{F}_{xy}$. Removing constant sites is helpful and may adequately correct for the problem of rate heterogeneity plus shifting base composition (Waddell, 1995), but a better strategy may be to classify sites into a few distinct rate classes, apply the LogDet transform to each, and sum these separate estimates to obtain the final distance.

WHICH SEQUENCE DISTANCE TRANSFORMATION IS BEST? As the above discussion indicates, distance analysis of sequence data requires choosing a distance transformation from a rather overwhelming number of possibilities. Ideally, we would always choose the most general distance available, as this distance has the smallest chance that assumptions corresponding to particular restrictions of the underlying model will be violated. Currently, this criterion would lead to a tradeoff between the LogDet/paralinear distance (which requires special treatment if there is substantial among-site rate variation) or the GTR (general time-reversible) distance with an appropriate correction for rate heterogeneity (Waddell and Steel, 1995). However, generality often comes at the price of increased variance, and many simulation studies have indicated that simpler distances based on models that are known to be violated may nonetheless perform better for phylogenetic inference than distances based on the same model being used to generate the data (e.g., see Nei, 1991 and references cited therein). For example, when sequences are relatively short, use of simple dissimilarity (*p*-distance) or the JC distance can lead to correct recovery of the true tree more often than the K2P distance, even when there is a fairly strong transition/transversion bias.

It is difficult to provide simple prescriptions for the choice of a distance measure (but see Kumar et al., 1993, for one such set of recommendations). In general, we believe that additional studies will confirm preliminary simulations that indicate little variance-inflation problem with LogDet/paralinear distances when all sites evolve at the same rate (see "Log-Determinant Distances," above). Because of their generality (including their robustness to base composition biases), log-determinant distances are probably preferable to other, more restricted, distances that do not incorporate corrections for among-site rate variation. Beyond that, we offer Kumar et al.'s (1993, p. 29) rule of thumb: "As a general rule, if two distance measures give similar distance values for a set of data, use the simpler one because it has a smaller variance." Of course, the longer the sequence length, the less variance considerations dominate the choice of a distance. With long sequences (e.g., >2000 bases), it may be more profitable to emphasize closer modeling of the substitution process than to worry too much about variance.

Transformation of Allozyme and Restriction Endonuclease Data to Distances

A large number of measures have been proposed for transforming allelic and genotypic frequency data to genetic distances (S. Wright, 1978); we will treat only a few of the more commonly used ones here. Historically, the most frequently used genetic distance has been that of Nei (1972, 1978). Let x_i and y_i be the frequencies of the ith allele at a particular locus in taxa X and Y, respectively. Nei's (1972) standard genetic distance can then be defined as

$$D_{\mathrm{N}} = -\ln\left(J_{\mathrm{XY}} / \sqrt{J_{\mathrm{X}} J_{\mathrm{Y}}}\right) \qquad \textbf{(24)}$$

where J_{X}, J_{Y}, and J_{XY} are the arithmetic means across loci of Σx_i^2, Σy_i^2, and $\Sigma x_i y_i$, respectively, with summations over alleles at each locus. Equation (24) gives a biased estimate when sample sizes are small; an unbiased estimate of the standard distance is obtained by replacing Σx_i^2 and Σy_i^2 with $(2n_{\mathrm{X}}\Sigma x_i^2 - 1)/(2n_{\mathrm{X}} - 1)$ and $(2n_{\mathrm{Y}}\Sigma y_i^2 - 1)/(2n_{\mathrm{Y}} - 1)$, respectively (Nei, 1978). D_{N} is intended to measure the number of codon substitutions per locus that have occurred after divergence between a pair of populations (taxa). However, this interpretation is valid only if the rate of gene substitution per locus is uniform across both loci and lineages, an assumption that is almost certainly unrealistic (Hillis, 1984) for any systematically informative data set. Hillis (1984) demonstrated that violation of the assumption of rate uniformity leads to a peculiar property of D_{N} when it is applied in systematic studies involving interspecific comparisons. He showed three hypothetical two-locus cases in which, for each case, two taxa had identical allele frequencies at one locus and shared no alleles at the second locus. However, due to different levels of polymorphism within the two taxa, D_{N} varied from 0.41 to 1.10. Hillis (1984) consequently recommended the following modification to D_{N} to alleviate the problems created by non-uniform rates of change:

$$D_{\mathrm{N}}^{*} = -\ln\left[\sum_{L}\left(\sum x_i y_i / \sqrt{\sum x_i^2 \sum y_i^2}\right)/L\right]$$

where L is the total number of loci; that is, the distance is computed from the arithmetic mean of the single-locus identities. (Although Hillis, 1984, did not specifically recommend it, an unbiased version of D_N^* could be obtained by a substitution equivalent to that for Nei's original distance.)

Nei's distances (in either their original form or as modified by Hillis, 1984) are non-metric in that they frequently violate the triangle inequality. Farris (1981) has heavily criticized it for this reason, arguing that when a distance measure is non-metric, it is meaningless to fit branch lengths under an additive-tree model in which branch lengths are interpreted as amounts of evolutionary change. Felsenstein (1984) countered that if branch lengths were interpreted as expected, rather than actual, amounts of change, Farris's objections were moot. While we do not wish to become entangled in this controversy (see also Farris, 1985, 1986a; Felsenstein, 1986), we basically agree with Felsenstein, without going so far as to recommend routine usage of Nei's distance. If Nei's model of evolution is appropriate (which is obviously open to question), then the nonmetricity of his distance is not in itself a reason to shun it.

Another widely used distance measure is that of J.S. Rogers (1972):

$$D_R = \frac{1}{L}\sum_{L}\sqrt{\sum\left(x_i - y_i\right)^2 / 2}$$

Rogers' measure has the virtues of simplicity and an easily interpretable geometric basis. Except for a scaling factor, it is simply the Euclidean distance between the allele frequency vectors for each locus of the two taxa being compared. However, Rogers' coefficient shares with Nei's the undesirable property of being too heavily influenced by within-taxon heterozygosity (S. Wright, 1978; Hillis, 1984); the distance between two taxa that are fixed for alternate alleles exceeds that between two taxa in which one or both are heteroallelic but have no alleles in common.

An alternative Euclidean measure that overcomes this limitation is the arc distance of Cavalli-Sforza and Edwards (1967), which is given by

$$D_{\text{arc}} = \sqrt{(1/L)\sum_{L}(2\theta/\pi)^2}$$

where $\theta = \cos^{-1}\sum\sqrt{x_i y_i}$. Thus, if no alleles are shared between a pair of taxa, the distance takes its limiting value of one regardless of the variability within either population. Perhaps more importantly, this distance incorporates an angular transformation of gene frequencies in an attempt to make the variances of the transformed frequencies independent of the ranges in which they fall. This transformation has the effect of standardizing the distance with respect to random drift, so that the rate of increase in genetic distance under drift is nearly independent of the initial gene frequencies. The Cavalli-Sforza and Edwards (1967) arc distance and its relative, the chord distance, thus incorporate some realistic assumptions about the nature of evolutionary change in gene frequencies without the undesirable properties of the Nei (1972, 1978) and the Rogers (1972) measures.

The simplest distance of all is the Manhattan distance (attributed to Prevosti by S. Wright, 1978), which for a single locus equals

$$D_M = \frac{1}{2}\sum|x_i - y_i|$$

An arithmetic mean is used to combine distances across loci. Unlike the Cavalli-Sforza and Edwards (1967) distances, this method gives equal weight to a given frequency difference, regardless of where it occurs on the scale from zero to one. It is not sensitive to intrataxon variability, however.

To transform restriction-site data to distances, Nei and Li's (1979) method for estimating the number of nucleotide substitutions that have occurred since divergence of a pair of taxa X and Y from a common ancestor is typically used. An estimate of the proportion of ancestral restriction sites that have remained unchanged until the present is given by

$$\hat{S} = 2n_{\text{XY}}/(n_{\text{X}} + n_{\text{Y}})$$

where n_{XY} is the number of identical sites shared by the two taxa, and n_{X} and n_{Y} are the total number of restriction sites in taxa X and Y, respectively. From this quantity we can estimate the mean number of substitutions per nucleotide site using either of the following:

$$d = -(\ln\,\hat{S})/r \qquad \textbf{(25a)}$$

$$d = -(3/2)\ln\left[\left(4\hat{S}^{1/(2r)} - 1\right)/3\right] \qquad \textbf{(25b)}$$

where r is the length of the endonuclease recognition sequence (usually 4 or 6). The first formula (25a) treats original restriction sites restored by back-mutations as new sites, and was first proposed by Upholt (1977). The second formula (more correctly) considers the reverted sites as identical to the original sites.

Li and Graur (1991) suggested estimating the proportion of nucleotide differences as

$$\hat{p} = 1 - \hat{S}^{1/r}$$

and then using the standard Jukes–Cantor distance transformation to estimate the number of nucleotide substitutions (i.e., substitute $\hat{p}$ for D in equation 20). A related method of estimating the number of nucleotide substitutions per site from restriction site data via maximum likelihood has been developed by J. Felsenstein (available as a test program "Restdist" from same location as PHYLIP; see Appendix). His method assumes a Kimura two-parameter model of evolution (i.e., equal base frequencies with a potentially different rate of transitions relative to transversions) and can include a correction for among-site rate variation according to a gamma distribution (see "Accommodating Among-Site Rate Variation in Distance Corrections," above). $\hat{S}$ is used to estimate the proportion of restriction sites that have been preserved by a pair of species, and $\hat{S}^{1/r}$ then represents the corresponding fraction of similarity at each of the r sites in the recognition sequence. The distance value that predicts this fraction of similar sites under the chosen model and parameter settings is then estimated by maximum likelihood.

The methods described above are appropriate when all restriction endonuclease recognition sites are the same length. For studies involving enzymes with different sizes of recognition sequences, more complicated methods developed by Nei and Tajima (1983) can be used, although we will not describe them here.

Nei and Li (1979) also addressed the problem of estimating nucleotide substitutions from restriction fragment data. However, these estimates are reliable only if the actual number of substitutions has been low (e.g., the samples are restricted to conspecific populations). Consequently, we will not describe their procedures for dealing with fragment data; the interested reader can consult their paper directly.

Immunological and Nucleic Acid Hybridization Data

When analyzing immunological measurements, it is usually assumed that, within certain limits, the measured immunological distance (ID) increases linearly with the number of amino acid differences in the proteins being compared. The constant of proportionality depends on the number of independent binding domains and on the fraction of amino acid changes that alter a domain sufficiently to inhibit antibody binding. Thus, there is significant uncertainty in the exact scaling. If we knew the scaling, we would apply a correction for superimposed amino acid replacements. This is of little practical importance, however, since the amount of divergence being measured is quite small, so any correction would also be small. We suggest equating evolutionary distance to the immunological distance—that is, assume that d = ID for each pair of proteins.

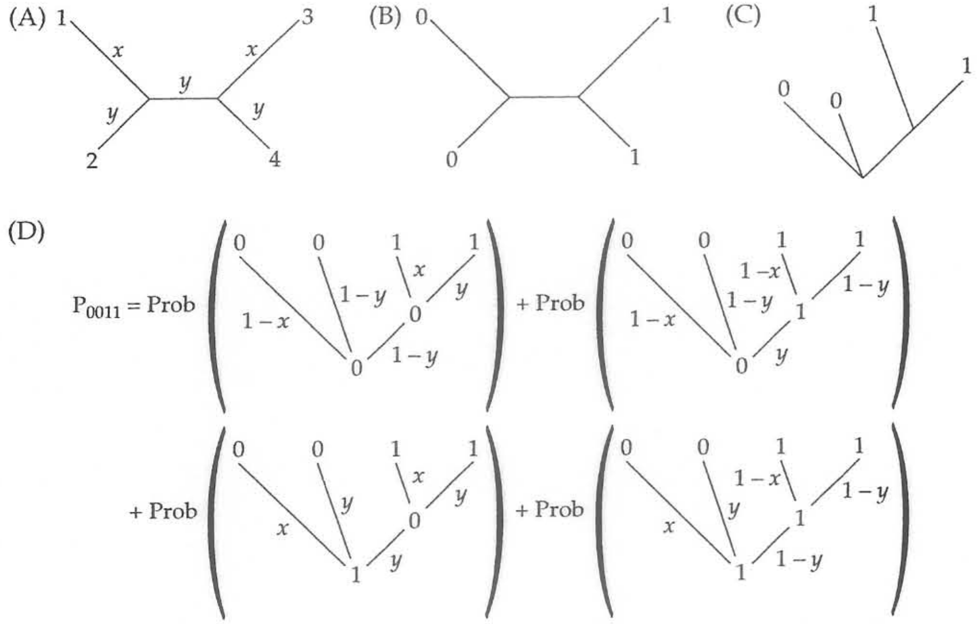
Hybridization data and their transformation to amount of difference in the DNAs are discussed extensively in Chapter 6. These data can be corrected for superimposed base changes by the methods discussed above.

Model-Based Corrections for Character Data: Hadamard Conjugation

The Hadamard conjugation, or spectral analysis (Hendy and Penny, 1993), offers another framework for taking superimposed changes into account. It will not be possible to provide a complete description and justification of this family of methods in the space available, so we will instead try to provide a clear explanation of the basic methodology. We begin by describing another model of character change introduced formally by Cavender and Felsenstein (1987). The Cavender–Felsenstein model is essentially a two-state equivalent of the Jukes–Cantor (1969) model. Each of the two states (0 and 1) are assumed to occur at equal frequency, and the probability of change from state 0 to state 1 is equal to the probability of change in the opposite direction. For example, this model might apply if we pool the purines (A and G) into one character state (0) and the pyrimidines (C and T) into another character state (1).

Revisiting the Felsenstein Zone

Consider the problem of calculating the probabilities of obtaining the various character patterns on a tree such as that shown in Figure 16A, which corresponds to one of the examples used by Felsenstein (1978a) to demonstrate the potential inconsistency of parsimony. Let P_{ijkl} represent the probability of each possible pattern, where i, j, k, and l are the states (0 or 1) found in taxa 1, 2, 3, and 4, respectively. These pattern probabilities can be determined using the same system described under "Calculating the Likelihood of a Tree." As an example, let us evaluate the probability that the pattern of Figure 16B (0011) will evolve under the conditions of the Cavender–Felsenstein model. We first note that because of the time-reversibility assumption, we can re-root the tree at



$$= (1/2)xy(1-x)(1-y)^2 + (1/2)y(1-x)^2(1-y)^2 + (1/2)x^2y^3 + (1/2)xy(1-x)(1-y)^2$$
$$= (1/2)[y(1-x)(1-y)^2(x+1) + x^2y^3]$$

(E) $P_{0011} + P_{1100} = 2P_{0011} = y(1-x)(1-y)^2(x+1) + x^2y^3$

(F) $P_{0101} + P_{1010} = 2P_{0101} = y^2(1-x)(1-y)(x+1) + x^2(1-y)^3$

Figure 16 Calculation of the probability of observing a given pattern of character states on a tree. (A) An unrooted tree for four taxa with probabilities of character differences x or y along each branch. (B) Tips of tree labeled by character states in the pattern of interest. (C) Tree re-rooted at an arbitrary internal node. (D) Calculation of the probability of the pattern shown in (B). (E) Calculation of expected proportion of characters that favor tree (A). (F) Calculation of expected proportion of characters that favor the tree grouping taxa 1 and 3.

an arbitrary internal node (Figure 16C), and then sum the probabilities of each of the four configurations of states at the two internal nodes (Figure 16D). That is, for each scenario, we multiply the prior probability of the basal state (= 1/2 in this case) times the product of the probabilities of the various changes (or non-changes) implied by each reconstruction. Because of the symmetry of the branch lengths used here, the probability (P_{1100}) of the other pattern that supports the tree of Figure 16A is equal to P_{0011}. Thus, the probability of a character pattern evolving that supports the true tree is

$$P_{0011} + P_{1100} = y(1-x)(1-y)^2(x+1) + x^2y^3 \qquad \textbf{(26)}$$

where x is the probability of a character-state change along the "long" branches and y is the corresponding probability for the "short" branches. Equation (26) is equivalent to one given by Felsenstein (1978a). A similar derivation reveals that the probability of a pattern evolving that supports the tree grouping taxa 1+3 and 2+4 is

$$P_{0101} + P_{1010} = y^2(1-x)(1-y)(x+1) + x^2(1-y)^3$$

Felsenstein (1978a) used these results to show that for many values of x and y ($x > y$), the probability of evolving character patterns that favor an incorrect tree exceeds that of patterns supporting the true tree. For example, if $x = 0.3$ and $y = 0.06$,

$$\begin{aligned} P_{0011} + P_{1100} &= 0.06(0.7)(0.94)^2(1.3) + \\ &\quad (0.3)^2(0.06)^3 = 0.048264 \\ P_{0101} + P_{1010} &= (0.06)^2(0.7)(0.94)(1.3) + \\ &\quad (0.3)^2(0.94)^3 = 0.077832 \end{aligned}$$

Thus, a sample of 1000 characters will, on average, contain 30 more characters favoring an incorrect tree than the true tree (78 versus 48).

The tedious strategy outlined in the above paragraph could be used to calculate the probabilities of any of the $2^4 = 16$ possible character patterns for the four terminal taxa. Furthermore, it could in principle be generalized to trees of any size. But as there are 2^T distinct character patterns and 2^{T-2} ways of generating each, this algebraic approach quickly becomes unmanageable.

(A) $$\mathbf{H}^{(1)} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{H}^{(n+1)} = \begin{pmatrix} \mathbf{H}^{(n)} & \mathbf{H}^{(n)} \\ \mathbf{H}^{(n)} & -\mathbf{H}^{(n)} \end{pmatrix}$$

(B) $$\mathbf{H}^{(2)} = \begin{pmatrix} \mathbf{H}^{(1)} & \mathbf{H}^{(1)} \\ \mathbf{H}^{(1)} & -\mathbf{H}^{(1)} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Figure 17 Definition of Hadamard matrices. A Hadamard matrix **H** is a square matrix whose entries are all 1 or –1, and with every row (and column) orthogonal to every other row (and column). (A) Basic form of a Hadamard matrix, and recursive formula for generating the next larger matrix. (B) Example calculation of a matrix with four rows and columns from the previous matrix with two rows and columns.

Calculating Character-Pattern Probabilities via the Hadamard Conjugation

Hadamard conjugation (Hendy and Penny, 1993) provides an alternative mechanism for obtaining the above pattern probabilities.* A Hadamard matrix (described by the nineteenth century mathematician of that name) is a matrix of 1's and –1's in a simple repeating pattern (Figure 17). For T taxa and two character states, we will use a Hadamard matrix containing $m = 2^{T-1}$ rows and columns. For the example discussed in the section above, $m = 8$, and the corresponding Hadamard matrix is

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

*This section assumes some familiarity with matrix algebra; see many statistics texts or any linear algebra text for introductions. Bulmer (1994, p. 293 ff.) provides an accessible overview for biologists. For now, note that the product of a matrix **A** and a vector **b**, denoted **Ab**, can be obtained as in the following example:

$$\mathbf{A} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} : \ \mathbf{Ab} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{pmatrix}$$

The inverse of matrix **A**, denoted $\mathbf{A}^{-1}$, is a matrix such that $\mathbf{AA}^{-1} = \mathbf{I}$, where **I** is an identity matrix that has 1's on the diagonal and 0 everywhere else. For example, if

$$\mathbf{A}^{-1} = \begin{pmatrix} j & k & l \\ m & n & o \\ p & q & r \end{pmatrix}$$

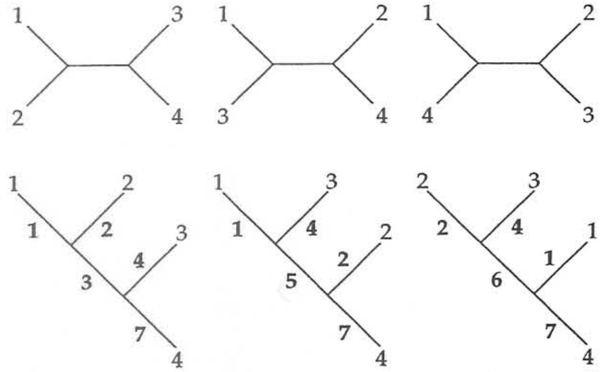
(see Figure 17 for an explanation of how these matrices are defined). The branch lengths x and y represent the probabilities that the character states at either end of a branch will be different ("observed differences") at a given site. We will store these values in an m-element vector $\mathbf{p}$ at a position determined by the indexing scheme shown in Figure 18. For our example, $\mathbf{p}$ is defined as

$$\mathbf{p} = \begin{pmatrix} p_0 \\ p_1 \\ p_3 \\ \vdots \\ p_{m-1} \end{pmatrix} = \begin{pmatrix} 0 \\ x \\ y \\ y \\ x \\ 0 \\ 0 \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0.3 \\ 0.06 \\ 0.06 \\ 0.3 \\ 0 \\ 0 \\ 0.06 \end{pmatrix}$$

The first element of this vector, p_0, is always set to 0. If the branch corresponding to a given index k does not exist in the tree, p_k is also set to 0. Thus, we have $p_5 = p_6 = 0$, because branch 5 (representing a partition separating taxa 1 and 3 from taxa 2 and 4) and branch 6 (representing a partition separating taxa 2 and 3 from taxa 1 and 4) do not exist on the tree.

Under the conditions of the model, the observed differences $\mathbf{p}$ can be converted to "ex-

(A)



(B) Index	Partition
0	∅, {1, 2, 3, 4}
1	{1}, {2, 3, 4}
2	{2}, {1, 3, 4}
3	{1, 2}, {3, 4}
4	{3}, {1, 2, 4}
5	{1, 3}, {2, 4}
6	{2, 3}, {1, 4}
7	{1, 2, 3}, {4}

Figure 18 (A) Indexing of partitions in the Hadamard conjugation. To label branches, root the tree arbitrarily at the highest numbered taxon. Label as 2^{i-1} the branch leading to each tip i. Label the remaining branches by the sum of the labels of the branches immediately above it. (B) Each branch defines a partition or split that is indexed by the branch's label. Note that the partition corresponding to any index k can be determined by decomposing it into its binary components. For example, with $T = 8$ the index $90 = 64 + 16 + 8 + 2 = 2^6 + 2^4 + 2^3 + 2^1$, corresponding to the partition {2,4,5,7},{1,3,6,8}.

then

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}\begin{pmatrix} j & k & l \\ m & n & o \\ p & q & r \end{pmatrix} = \begin{pmatrix} aj+bm+cp & ak+bn+cq & al+bo+cr \\ dj+em+fp & dk+en+fq & dl+eo+fr \\ gj+hm+ip & gk+hn+iq & gl+ho+ir \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Finally, multiplication of a matrix by a scalar (ordinary number) implies multiplication of every element of the matrix by the scalar:

$$2\mathbf{A} = 2\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} 2a & 2b & 2c \\ 2d & 2e & 2f \\ 2g & 2h & 2i \end{pmatrix}$$

pected total changes per site" **q** using the formula

$$q_i = -\frac{1}{2}\ln(1-2p_i) \tag{27}$$

where q_i is the expected number of changes per site along branch i. Note that this is just a special case of the general Poisson-correction formula (21) with $B = 1/2$. Define the *branch-length spectrum* $\boldsymbol{\gamma}(T)$ as

$$\boldsymbol{\gamma}(T) = \begin{pmatrix} -\sum_{i=1}^{m-1} q_i \\ q_1 \\ q_2 \\ \vdots \\ q_{m-1} \end{pmatrix} = \begin{pmatrix} -1.108040 \\ 0.458145 \\ 0.063917 \\ 0.063917 \\ 0.458145 \\ 0 \\ 0 \\ 0.063917 \end{pmatrix}$$

with q_1 through q_7 defined using equation (27). In some cases (e.g., simulation studies), it may be more convenient to start with the $\boldsymbol{\gamma}$ vector directly, in which case

$$p_i = \frac{1-e^{-2q_i}}{2} \tag{28}$$

Now let **s**(T) be the *expected sequence spectrum*—a vector where each element s_k is the predicted proportion of the characters supporting each possible bipartition of the taxa (division into two subsets; see Figure 18 for how bipartitions are indexed). For example, s_3 is equivalent to Felsenstein's (1978a) $P_{0011} + P_{1100}$, and s_5 is equivalent to $P_{0101} + P_{1010}$. The values of **s**(T) can be obtained using the following Hadamard conjugation:

$$\mathbf{s}(T) = \mathbf{H}^{-1}\exp[\mathbf{H}\boldsymbol{\gamma}(T)] \tag{29}$$

where the exponential function is applied separately to each element of **H**γ. Let us apply formula (29) to our example. First, the generalized distance vectors **ρ** and **r** are calculated as follows:

$$\boldsymbol{\rho} = \mathbf{H}\boldsymbol{\gamma} = \begin{pmatrix} 0 \\ -1.17196 \\ -0.38350 \\ -1.04412 \\ -1.04412 \\ -1.96041 \\ -1.17196 \\ -2.08825 \end{pmatrix} \tag{30}$$

$$\mathbf{r} = \exp(\boldsymbol{\rho}) = \begin{pmatrix} e^{\rho 0} \\ e^{\rho 1} \\ e^{\rho 2} \\ e^{\rho 3} \\ e^{\rho 4} \\ e^{\rho 5} \\ e^{\rho 6} \\ e^{\rho 7} \end{pmatrix} = \begin{pmatrix} 1 \\ 0.30977 \\ 0.68147 \\ 0.35200 \\ 0.35200 \\ 0.14080 \\ 0.30977 \\ 0.12390 \end{pmatrix} \qquad \textbf{(31)}$$

Each entry in $\boldsymbol{\rho}$ represents $-2\delta_i^*$, where δ_i^* is a *corrected generalized distance*. The exponential transformation then converts each ρ_i to an *observed generalized distance*, $r_i = 1 - 2d_i^*$. (They are called "generalized" distances because they represent the lengths of path sets that correspond not only to distances between pairs of taxa, but also to groups of non-intersecting paths involving even numbers of taxa.) The expected sequence spectrum $\mathbf{s}(T)$ for tree T is then obtained as follows:

$$\mathbf{s}(T) = \mathbf{H}^{-1}\mathbf{r} = \left(\frac{1}{m}\mathbf{H}\right)\mathbf{r} \qquad \textbf{(32)}$$

(The simple form of the inverse of a Hadamard matrix, shown above, is an important advantage of the method.) For our example,

$$\mathbf{s}(T) = \frac{1}{8}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0.309760 \\ 0.681472 \\ 0.352000 \\ 0.352000 \\ 0.140800 \\ 0.309760 \\ 0.123904 \end{pmatrix}$$

$$= \begin{pmatrix} 0.408712 \\ 0.177096 \\ 0.041928 \\ 0.048264 \\ 0.177096 \\ 0.077832 \\ 0.027144 \\ 0.041928 \end{pmatrix} = \begin{pmatrix} P_{0000} + P_{1111} \\ P_{1000} + P_{0111} \\ P_{0100} + P_{1011} \\ P_{1100} + P_{0011} \\ P_{0010} + P_{1101} \\ P_{1010} + P_{0101} \\ P_{1001} + P_{0110} \\ P_{1110} + P_{0001} \end{pmatrix} \qquad \textbf{(33)}$$

The probabilities corresponding to $P_{1100} + P_{0011}$ and $P_{0101} + P_{1010}$ (s_3 and s_5, respectively) correspond exactly to those calculated algebraically in the preceding section. Hadamard conjugation has strong advantages over such algebraic calculations, however. First, we have not only calculated the probabilities of characters supporting these two bipartitions, but all of the other bipartitions as well. Second, the method is general, and extends automatically to the calculation of expected character-state distributions even for more realistic evolutionary models and large trees. (Note, however, that the exponential growth of the size of the vectors—e.g., 524,288 elements for 20 taxa—puts a practical limit on tree size.) For instance, Hendy and Penny (1989) used this feature to show that parsimony could be inconsistent under a molecular clock; Bull et al. (1993a) used it to examine the consequences of combining different sources of data; and Charleston et al. (1994) used it for tree-selection simulations.

Invertibility of the Hadamard Conjugation

Although the prediction of pattern frequencies as outlined above can be useful, the power of the Hadamard conjugation in phylogenetic applications lies in its invertibility. Specifically, all of the above operations can be performed in the opposite direction: starting from an observed sequence spectrum $\hat{\mathbf{s}}$ (pattern frequencies observed in the data), we can work back to a conjugate spectrum $\hat{\boldsymbol{\gamma}}$, which is an estimate of the underlying branch-length spectrum $\boldsymbol{\gamma}(T)$. To demonstrate the inverse operations, suppose that the observed sequence spectrum $\hat{\mathbf{s}}$ corresponds exactly to $\mathbf{s}(T)$ as calculated in equation (33). Solving for $\mathbf{r}$ in equation (32) yields

$$\mathbf{r} = \mathbf{H}\hat{\mathbf{s}} = \begin{pmatrix} 1 \\ 0.309760 \\ 0.681472 \\ 0.352000 \\ 0.352000 \\ 0.140800 \\ 0.309760 \\ 0.123904 \end{pmatrix} \tag{34}$$

and since the log function is the inverse of the exponential,

$$\boldsymbol{\rho} = \ln(\mathbf{r}) = \begin{pmatrix} \ln r_0 \\ \ln r_1 \\ \ln r_2 \\ \ln r_3 \\ \ln r_4 \\ \ln r_5 \\ \ln r_6 \\ \ln r_7 \end{pmatrix} = \begin{pmatrix} 0 \\ -1.17196 \\ -0.38350 \\ -1.04412 \\ -1.04412 \\ -1.96041 \\ -1.17196 \\ -2.08825 \end{pmatrix} \tag{35}$$

Thus, the full Hadamard conjugation in this direction is

$$\hat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}\ln(\mathbf{H}\hat{\mathbf{s}}) \tag{36}$$

The conjugate spectrum $\hat{\boldsymbol{\gamma}}$ is evaluated as

$$\hat{\boldsymbol{\gamma}} = \mathbf{H}^{-1}\boldsymbol{\rho} = \left(\frac{1}{m}\mathbf{H}\right)\boldsymbol{\rho} = \begin{pmatrix} -1.108040 \\ 0.458145 \\ 0.063917 \\ 0.063917 \\ 0.458145 \\ 0 \\ 0 \\ 0.063917 \end{pmatrix} \tag{37}$$

Finally, use of formula (28) to convert our estimate of the expected number of changes ($\hat{\gamma}_i$) to the number of observed differences predicted for each branch ($\hat{p}_i$) leads to exact recovery of the original branch lengths ($x = 0.3$, $y = 0.6$).

Application to Real Data

Even if the assumptions of our evolutionary model were perfectly satisfied, we cannot expect the observed sequence spectrum $\hat{\mathbf{s}}$ to correspond exactly to the true spectrum $\mathbf{s}$, because the sequences obtained in an actual study represent a finite sample and therefore are subject to sampling error. To illustrate the use of Hadamard conjugation in practice, we will draw a sample of characters that have evolved according to our model; this sample will be used to represent a set of observed sequence data that have evolved according to the model.

The $\hat{\mathbf{s}}$ vector below shows the results of a random sample of 1000 characters (using the pseudorandom number generator in Mathematica®) according to the expected sequence spectrum in equation (33).

$$\hat{\mathbf{s}} = \begin{pmatrix} 0.418 \\ 0.168 \\ 0.053 \\ 0.048 \\ 0.174 \\ 0.076 \\ 0.030 \\ 0.033 \end{pmatrix}$$

As expected, parsimony analysis of this data set will choose an incorrect tree, as the 48 characters supporting the true tree ($\hat{s}_3$) are contradicted by 76 characters ($\hat{s}_5$) supporting the tree that groups taxa 1 and 3. Solution of equation (36) yields

$$\hat{\boldsymbol{\gamma}} = \begin{pmatrix} -1.05917 \\ 0.41125 \\ 0.09269 \\ 0.06348 \\ 0.43201 \\ 0.01716 \\ 0.00956 \\ 0.03303 \end{pmatrix}$$

If the transformed data conform to a treelike pattern, all but $2T-3$ of the elements in $\hat{\boldsymbol{\gamma}}$ will be close to zero (or negative, in the case of $\hat{\gamma}_0$), and the bipartitions corresponding to significantly positive elements will be compatible with a single tree. In this example, $\hat{\gamma}_5$ and $\hat{\gamma}_6$ are both close to zero. The remaining bipartitions are compatible (all but $\hat{\gamma}_3$ define an "uninformative" partition splitting a single terminal taxon from the remainder). Thus, the tree of Figure 16A is clearly specified by the corrected data.

CHOOSING A TREE With real data sets, the picture is seldom as clear as the above section suggests, and we must use one of several methods to choose an optimal tree based on the transformed data represented by the $\hat{\boldsymbol{\gamma}}$ vector. The closest tree procedure (Hendy, 1991) is one commonly recommended method. For a given tree τ containing K branches, it is straightforward to find a vector $\mathbf{q}(\tau)$ that minimizes the Euclidean distance from $\mathbf{q}(\tau)$ to $\hat{\boldsymbol{\gamma}}$. The squared distance can be obtained [without the need to form $\mathbf{q}(\tau)$ explicitly] using the formula

$$\Gamma^2(\tau,\hat{\boldsymbol{\gamma}}) = \sum_{e_i \notin e(\tau)} \hat{\gamma}_i^2 + \frac{\left(\hat{\gamma}_0 + \sum_{e_i \in e(\tau)} \hat{\gamma}_i\right)^2}{K+1} \qquad \textbf{(38)}$$

where the expressions $e_i \in e(\tau)$ and $e_i \in e(\tau)$ limit the summations to those branches (= edges) that are included in or absent from, respectively, the tree being tested (Hendy and Penny, 1993).

The *closest tree* is the one that minimizes the value of formula (38) over all possible trees, and can be found using (for example) a modification of the branch-and-bound algorithm of Penny and Hendy (1987). Note that some of the $\hat{\boldsymbol{\gamma}}$ values calculated using formula (37) (other than $\hat{\gamma}_0$) may be negative, although this did not happen in our example. Any tree that would include one of these branches is automatically rejected.

For the example in the above section, the squared distances of the three trees to $\hat{\boldsymbol{\gamma}}$ are:

((1,2),(3,4)):	5.05×10^{-4}
((1,3),(2,4)):	5.01×10^{-3}
((1,4),(2,3)):	5.41×10^{-3}

Thus, the first tree is the closest tree.

Another method for choosing a tree is corrected parsimony. The conjugate spectrum $\hat{\boldsymbol{\gamma}}$ can be thought of as a transformation of the original data matrix to a new data matrix containing 2^{T-1} characters (in the case of two states), each corresponding to the partitions associated with a row of $\hat{\boldsymbol{\gamma}}$. The elements of $\hat{\boldsymbol{\gamma}}$ are used as character weights, and a minimum-length tree under the weighted parsimony criterion is sought. As noted above, some elements of $\hat{\boldsymbol{\gamma}}$ may be negative due to lack of model fit or sampling error; these values are typically set to 0 before proceeding. Corrected parsimony is always consistent under the Cavender–Felsenstein model (Steel et al., 1993a), unlike standard parsimony. Corrected parsimony chooses the correct tree in our four-taxon example, because the weight of character patterns supporting the true tree ($\hat{\gamma}_3$) is greater than that of character patterns favoring alternative trees ($\hat{\gamma}_5$ and $\hat{\gamma}_6$). The simulation studies of Charleston (1994) suggest that corrected parsimony can be highly effective in some situations, and in general tends to outperform the closest tree and other methods described below.

An analogous method of corrected character compatibility also can be employed. This method searches for the largest weighted clique for the same data matrix and weights used for corrected parsimony. A clique is simply a set of mutually compatible characters that can all fit on the same evolutionary tree without homoplasy (e.g., Le Quesne, 1982; Estabrook, 1983). Standard graph theory algorithms exist for exact solution of the weighted clique problem (e.g., Bron and Kerbosch, 1973).

A final method is actually a hybrid of the closest tree and character compatibility approaches. Remember that when evolution proceeds exactly according to the model and there is no sampling error, $2T - 3$ of the elements in $\hat{\boldsymbol{\gamma}}$ will be positive; the remainder (except for $\hat{\gamma}_0$) will equal 0. Thus, for any particular tree, the squared deviations from 0 of the elements of $\hat{\boldsymbol{\gamma}}$ that correspond to bipartitions not found on the tree is a least-squares measure of the lack of fit:

$$\Delta^2(\tau, \boldsymbol{\gamma}) = \sum_{e_i \notin e(\tau)} \gamma_i^2 \tag{39}$$

Note that equation (39) is equal to the first term on the right-hand side of (38). The second term in (38), although different for each tree, appears not to contribute greatly to the discrimination among trees (Waddell, 1995), and dropping it from the optimality criterion allows us to use character compatibility methods to minimize (39). Specifically, after setting any negative values in $\hat{\boldsymbol{\gamma}}$ to 0, we square each element and find a maximum weighted clique; solution of this problem is then equivalent to minimizing the sum of squared deviations for the excluded partitions from their expected value of 0. This method seems especially promising when each $\hat{\gamma}_i$ is divided by its estimated sampling error before proceeding (yielding the vector $\hat{\boldsymbol{\gamma}}_{se}$), which gives a form of weighted least-squares tree selection (Waddell, 1995).

DATA EXPLORATION Apart from their use in estimating trees, spectral analysis methods are useful as aids in understanding the peculiarities of particular data sets. Strong contradictory signals in the $\hat{\boldsymbol{\gamma}}$ vector allow the data to reject the model, and we should explore the reasons that the correct data are not treelike if this occurs. Lack of fit to a tree may indicate that our model is too simple (e.g., we are not accounting adequately for rate heterogeneity across sites, or the substitution model is too restrictive). Alternatively, there may be multiple signals due to recombination or to non-independence among sites.

It is helpful to plot the inferred branch lengths ($\hat{\boldsymbol{\gamma}}$ values) divided by their estimated standard errors to see how much statistical support the "signals" really have (Waddell et al., 1994; Waddell, 1995). Another useful way of viewing the corrected sequences is to plot the magnitude of each signal in the conjugate spectrum against the sum of its pairwise incompatibilities with all other sequence patterns (a support/conflict spectrum; see Lento et al., 1995). These graphical representations of noise in the data set allow exploration of the factors responsible for conflicts in different regions of the tree and suggest which hypotheses of relationship should be subjected to further scrutiny. The paper by Lento et al. (1995) provides good examples of this approach.

Extension to Four Character States

Hadamard conjugations can be extended to handle all four bases as character states under a version of Kimura's (1981) K3ST model, which classifies substitutions into three types: type I = transitions; type II = transversions between A and C or G and T; and type III = transversions between A and T or C and G (see "Models of Evolution," above). The model is generalized to allow the probabilities of these events to be different for each branch of the tree. Under this model there are $4^T/4 = 4^{T-1}$ distinct sequence patterns (i.e., patterns such as AAGG, CCTT, GGAA, and TTCC are equivalent). These patterns are indexed using a modification of the binary coding for the two-state case (see Hendy et al., 1994; original derivation due to Székely et al., 1993) that define quadripartitions of the taxa (partitions into four or fewer subsets). Application of the Hadamard conjugation to this observed sequence spectrum using formula (36) corresponds to a correction for superimposed substitutions according to the generalized K3ST model. Within the corrected data (conjugate spectrum $\hat{\boldsymbol{\gamma}}$), there are three sets of 2^{T-1} entries, as for the two-state case. These elements correspond to the number of transitions, type II transversions, and type III transversions, respectively. The remaining elements are expected to be 0 under the model, again as in the two-state case. We can use closest tree, corrected parsimony or compatibility, or least-squares methods to select a tree from this spectrum.

Another promising way of treating four-state nucleotide data using three separate 2^{T-1} Hadamard conjugations has also been developed (Waddell and Hendy, 1995). These calculations give essentially the same results as the much more computationally expensive (order 4^{T-1}) approach of Hendy et al. (1994).

Subcases of the K3ST model can be handled by averaging the patterns in the observed data that are equivalent under the more restricted models (Waddell, 1995). For example, if we average the type II and type III transversions, we force the corrections to be made according to a generalized K2P model, and if we average all substitutions we obtain a generalized JC correction. This pooling of substitution types reduces stochastic errors if the simpler models are adequate.

Among-Site Rate Variation and Maximum Likelihood

The Hadamard conjugation can be modified to allow for unequal substitution rates across sites (Steel et al., 1993c; Waddell, 1995; Waddell and Penny, 1996b) in much the same way as the corrections are made for distances (e.g., G.J. Olsen 1987; Jin and Nei, 1990; see above). To estimate pattern probabilities assuming a gamma distribution, we need only replace the exponential function in the Hadamard conjugation (formula 29) with $[(\alpha - \mathbf{p})\ /\alpha]^{-\alpha}$, where α is the shape parameter. If going from observed sequence data to the corrected sequence spectrum using fomula (36), we replace the logarithm function with $\alpha(1 - \mathbf{r}^{-1/\alpha})$. For practically any distribution (e.g., the log-normal) the appropriate path-length correction can be estimated numerically (as in G.J. Olsen, 1987) if an analytic form does not exist as for the gamma distribution.

Recall that for any tree τ and branch-length spectrum $\boldsymbol{\gamma}$, we can obtain the associated vector of expected pattern frequencies $\mathbf{s}$ using formula (29). Since the log likelihood of the tree is given by

$$\ln L = \sum_i \hat{f}_i \ln s_i$$

where $\hat{f}_i$ is the frequency of sites with pattern i in the data, and s_i is the probability of this pattern under the model, Hadamard conjugation provides an alternative algorithm for maximum likelihood estimation. It is especially useful for maximum likelihood tree inference with among-site rate heterogeneity using continuous distributions such as gamma (Waddell, 1995; Waddell and Penny, 1996a). Although limited to the generalized K3ST model and its submodels, this approach can be much faster than that of Z. Yang (1993).

Statistics on the Corrected Sequences

It is straightforward to obtain the variance–covariance matrix of the corrected sequence data via the delta method approximation (Waddell et al., 1994). The simulations by Waddell et al. (1994) showed that the covariance matrix derived in this way gives nearly unbiased results, whereas bootstrap resampling tends to yield overestimates. As long as a pattern occurs five or more times in the observed data, it is reasonable to treat the corresponding corrected pattern (or branch length) as normally distributed, resulting in straightforward confidence intervals, or tests of the hypothesis that its true value is zero. The covariances of corrected patterns can also be thought of as covariances of tree branch length estimates. Generally, the more changes per site there are on the tree, the more strongly branch lengths become either positively or negatively correlated (Waddell et al., 1994). (These interdependencies tend to make the iterative search for a maximum likelihood solution slower.) Another conclusion from this study is that long branches, even when not biasing the topology of the tree, nonetheless cause a large increase in the variance of internal branch length estimates, reducing the reliability of tree selection. It is possible to estimate a confidence interval on transition:transversion ratios or the shape parameter of distributions used to model among-site rate variation (Waddell, 1995).

The Distance Hadamard

The last part of the Hadamard conjugation (from $\boldsymbol{\rho}$ to $\hat{\boldsymbol{\gamma}}$) can also begin from a matrix of pairwise distances (either corrected or uncorrected) (Hendy and Penny, 1993). We would like to estimate a branch-length spectrum (now called $\hat{\boldsymbol{\gamma}}_D$) and choose an optimal tree from this spectrum, analogously to the procedure used for sequence data. We input the distances at the level of the generalized distance vectors $\boldsymbol{\rho}$ (formula 35). However, a complication arises because these vectors include elements corresponding to path sets involving more than two taxa; see Hendy and Penny (1993) for a method of estimating these path-set lengths. The $\hat{\boldsymbol{\gamma}}_D$ vector resulting from formula (37) then serves as the basis for choosing a tree as described above.

Simulations and analytic calculations have shown that the variances of entries in the $\hat{\boldsymbol{\gamma}}_D$ vector resulting from this approach are lower than the $\hat{\boldsymbol{\gamma}}$ values from the usual Hadamard conjugation under the same model, because the distance method of estimating path-set lengths involving more than two taxa has lower variance (Waddell, 1995). Consequently, tree selection using this vector tends to be more reliable (Charleston, 1994). However, the distance Hadamard does not seem to be as sensitive as the Hadamard conjugation at detecting violations of the model's expectations. The studies of Lento et al. (1995) and Lockhart et al. (1995b) suggest that this method is a useful exploratory tool when trying different distance transformations, although more study is needed on how directly a pattern from the distance Hadamard can be treated as evidence for specific sequence patterns.

Lake's Method of Invariants

Rationale

As discussed earlier in this chapter, the presence of more than one long, unbranched lineage in an analysis can lead to systematic error in the absence of perfect compensation for superimposed substitutions. In the context of parsimony, the homoplasies along the long branches can overwhelm the informative character changes along the internal branch(es) of the tree (see Figure 8 and the section "Parsimony and Inconsistency").

Ideally, we would like to distinguish informative changes from homoplasies. In parsimony and maximum likelihood analyses, the addition of new sequences whose branch points subdivide the longest lineages (i.e., representation of taxa that are specifically related to the most divergent taxa already in the tree) will tend to accomplish this goal. The effect is illustrated in Figure 19 where adding sequences A′ and B′ to the tree would reduce the effects of homoplasies along the branches leading to A and B. Of course, the practical utility of this approach requires that appropriate taxa exist, that their identities are known, and that the corresponding sequence data exist or can be generated. A second method of reducing the effects of homoplasy is to confine the analysis to the most conserved sequences (both on the basis of the overall conservation of the molecule and

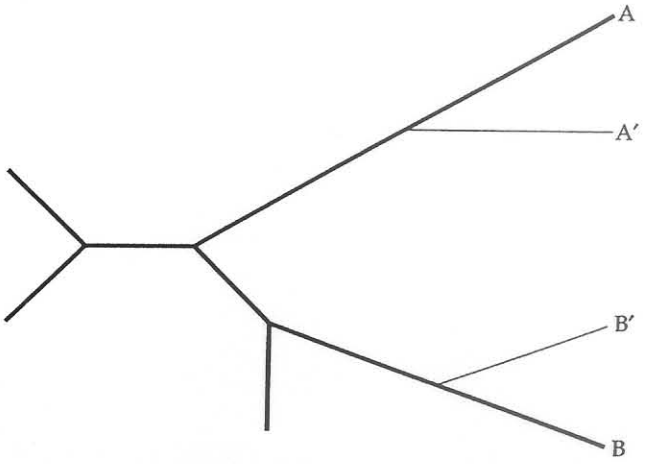


Figure 19 Adding new taxa to a parsimony or maximum likelihood tree to reduce the effects of homoplasy. Given the unrooted tree shown in heavy lines, the long lineages leading to A and B would have the greatest tendency to artifactually group due to parallel or convergent changes in sequence. Adding taxa A′ and B′ would reduce this effect by subdividing the long lines.

by selecting the most conserved portions of the molecule). In distance-based analyses, estimates of the superimposed substitutions (which include the homoplasies) can also be included.

Lake (1987a) suggested an alternative method, which he called *evolutionary parsimony*, for analyzing the branching pattern linking four nucleotide sequences. The analysis can be derived from the following assumptions: (1) substitutions at a given sequence position are independent; (2) a balance exists among specific classes of transversions (a sufficient condition for this balance is that transversions are equally likely to yield each of the two possible substitution products, so that C is equally likely to change to A or G, etc.); and (3) insertions or deletions can be safely ignored. An advantage of the method is that it does not assume anything about rate equality over sites; each site is free to evolve at a different rate than all other sites.

If the assumptions are satisfied, then parallel transversions in the two branches of a tree produce equal numbers of similar (type 1 in Figure 20) and dissimilar (type 2 in Figure 20) nucleotides. Thus, the net effect of peripheral branch transversions could be cancelled if the type 2 events were subtracted from the type 1 events. A complete accounting of possible transversions and transitions yields the scoring system in Table 2.

Methodology

Lake's method can be described by the following sequence of steps:

1. Choose a quartet of aligned sequences; call them A, B, C, and D.
2. Find the alignment positions in which two sequences have purines and two have pyrimidines.
3. Consider the three possible groupings of sequences (see Figure 21): AB/CD (A with B, C with D), AC/BD and AD/BC. Call these branching patterns X, Y, and Z, respectively.
4. Using the sequence positions at which sequences A and B are *both* purines or *both* pyrimidines (and sequences C and D are both of the opposite class of base), use the rules in Table 2 to count the number of positions that support and the number that counter branching order X. Call these totals X^+ and X^-, respectively. Similarly, find the support (Y^+) and countersupport (Y^-) for branching order Y, using the sequence positions at which sequences A and C have the same class of base, and B and D have the opposite class. Finally, find the support (Z^+) and countersupport (Z^-) for branching pattern Z. If the counting has been done correctly, the total of X^+, X^-, Y^+, Y^-, Z^+, and Z^- will be equal to the total number of positions with two purines and two pyrimidines, as found in the second step.
5. The net supports for branching patterns X, Y, and Z are

$$X = X^+ - X^- \quad (40a)$$
$$Y = Y^+ - Y^- \quad (40b)$$
$$Z = Z^+ - Z^- \quad (40c)$$

The support for two of the branching patterns should be near zero, while the remaining branching pattern may or may not be supported by a significantly non-zero score.

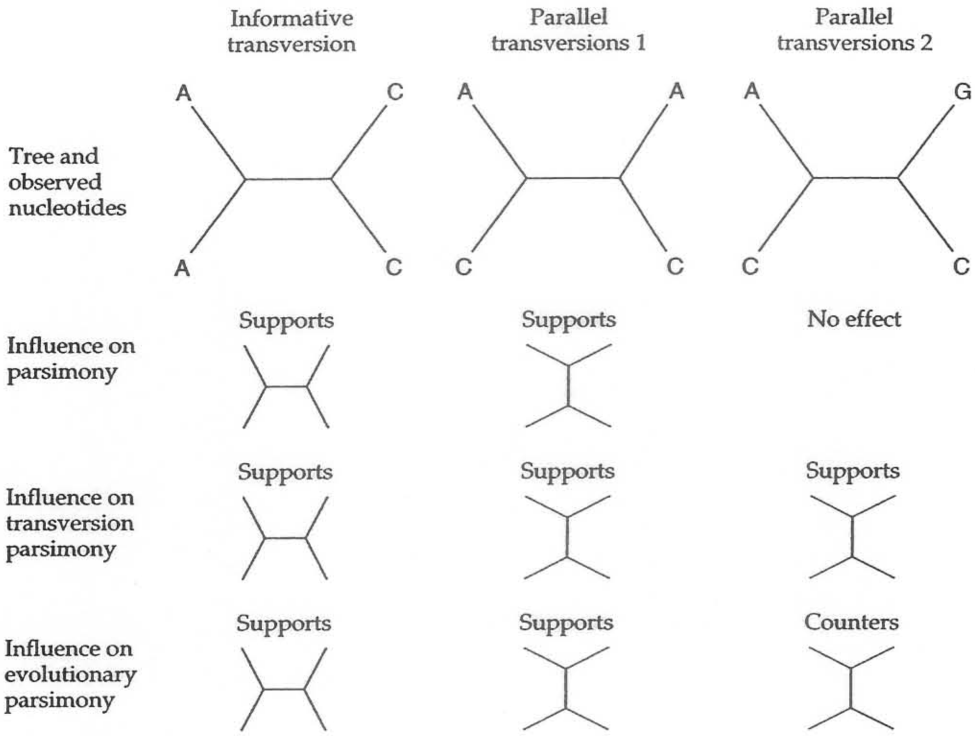


Figure 20 Nucleotide substitution patterns and their effects on different methods of phylogenetic tree inference. The first pattern, informative transversion, represents the effect of a single nucleotide substitution that is in the internal (central) branch of the tree. It is an example of the informative characters upon which parsimony depends. Transversion parsimony and Lake's method of invariants rely entirely upon transversions for informative events. The second pattern portrays a possible outcome of two peripheral branch transversions. Because the results are indistinguishable from the first pattern (two A's and two C's), all methods will mistake this as support for an incorrect phylogeny. The third pattern illustrates the possibility that independent transversions in two peripheral branches will yield different nucleotides. The pattern is uninformative to traditional parsimony (two substitutions would be required regardless of the assumed branching order). Transversion parsimony will consider this pattern to be support for the incorrect tree since the outcome looks like a central branch transversion (in an incorrect tree) combined with a peripheral branch transition (which is ignored). Lake's method treats this third pattern as an estimator of multiple substitutions in peripheral branches and subtracts it from the support for the incorrect tree.

6. Lake (1987a) suggested that statistical significance be evaluated by a one degree of freedom χ^2 test:

$$\chi_X^2 = X^2/(X^+ + X^-)$$
$$\chi_Y^2 = Y^2/(Y^+ + Y^-)$$
$$\chi_Z^2 = Z^2/(Z^+ + Z^-)$$

Therefore the outcome of interest is two values of χ^2 that do not differ significantly from zero and one value that does. Holmquist et al. (1988a) correctly pointed out that the χ^2 approximation is inadequate when counts are low and recommended the use of the exact binomial test instead.

NEGATIVE VALUES The net support of a tree can be negative and yet significant (e.g., χ is negative and χ_X^2 is significantly large). Lake (1987a) suggested that this result could be interpreted as positive evidence for the corresponding branching pattern, if no other pattern has significant support. However, significantly negative values should be viewed with extreme caution, because

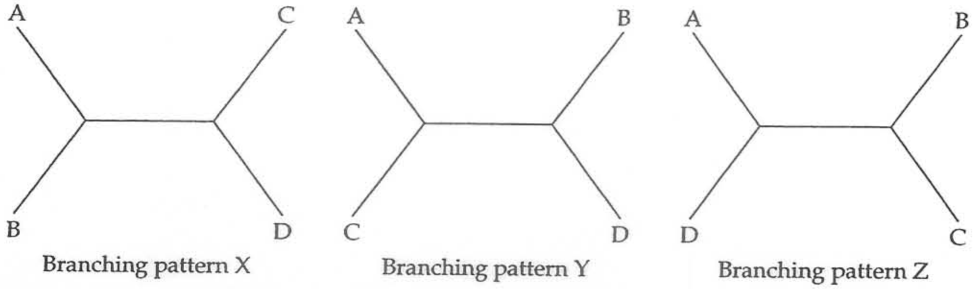


Figure 21 The three unrooted branching patterns with four sequences.

such outcomes are most likely to be the result of selective pressure or some other non-random process.

TRANSITIONS AND TRANSVERSIONS The phylogenetic information provided by Lake's method is based entirely on transversion substitutions, so positions with two purines and two pyrimidines are required. If there are no transversions, there will be no signal. On the other hand, transition substitutions decrease the signal. In particular, peripheral branch transitions convert informative (supportive) positions into countersupport, suggesting that the method might be particularly sensitive to the ratio of transitions to transversions. If transitions are indeed substantially more frequent than transversions, then it is difficult to accumulate a sufficient number of transversions to infer the branching pattern without having the signal randomized by transitions (see W.-H. Li et al., 1987b). As noted above, generalized parsimony (character-state weighting), transversion parsimony, and transversion-based distance methods provide alternative methods of coping with a high transition:transversion ratio. Under many conditions, these methods are much more efficient than Lake's method at finding the correct tree (Hillis et al., 1994b).

Interestingly, transversion parsimony (as defined in this chapter, which differs from Lake's use of the term) applied to four sequences seeks the tree, X, Y or Z, with the largest value of $X^+ + X^-$, $Y^+ + Y^-$, and $Z^+ + Z^-$. By examining the equations in (40), it can be seen that transversion parsimony uses the same data but *adds* the terms that look like a peripheral branch transition (and a central branch transversion) rather than subtracting them as does Lake's method.

Performance

Despite its intuitive appeal, the drawback of Lake's method is inefficiency. Especially when rates of change are high, simulation studies suggest that it requires vastly more data to achieve the same probability of inferring the correct phylogeny as other methods. For example, in four-taxon simulations using the K2P model under long-branch-attraction conditions, Hillis et al. (1994b) found that Lake's method required about 10^8 nucleotides before its probability of selecting the correct tree exceeded 1/3 (= the probability of a randomly chosen tree). Maximum likelihood analysis, on the other hand, achieved 95% success at only 5000 nucleotides under the same conditions. Lake's method can be consistent under conditions in which maximum likelihood (as currently implemented) is inconsistent, so given enough data, it remains a potentially useful method. Unfortunately, "enough data" may be vastly more than the amount available.

Rooting Revisited

Most of the methods discussed above do not specify the location of the root. If, as is generally the case, a rooted tree is desired, the root must be located using extrinsic information. As mentioned above, the most commonly used method is to include one or more taxa that are assumed to lie cladistically outside of a presumed monophyletic group. We recommend including more than one outgroup taxon as a means of testing the assumption of ingroup monophyly. If there is a single branch on the unrooted tree that partitions the ingroup taxa from the outgroup taxa (e.g., Figure 22A), then the tree is consistent with the assumption of ingroup monophyly. If, on the other hand, there is no such branch (Figure 22B), then we have rejected the monophyletic ingroup hypothesis (at least in a non-statistical sense). Of course, this test is one-sided: the existence of a branch that partitions the assumed ingroup versus outgroup taxa is no guarantee that the root does not lie somewhere within the ingroup. But at least the attempt to reject the hypothesis of ingroup monophyly failed, and one can feel somewhat more confident about the assumption for that reason.

Rooting is frequently the most precarious step in any phylogenetic analysis. In particular, connecting a distant outgroup to a tree can be very problematical, as there may be so many changes along the branch connecting·the ingroup to the outgroup that the sequences have become effectively randomized. In the worst case, this can lead to spurious "long branch attraction" effects (see the section on "Parsimony and Inconsistency"), with artifactual rooting along longer ingroup branches (Hendy and Penny, 1989; Miyamoto and Boyle, 1989; W.C. Wheeler, 1990b; D.R. Maddison et al., 1992). For this reason, it is often preferable to be satisfied with an unrooted tree than to include a highly divergent outgroup taxon in the analysis. An alternative strategy (Nixon and Carpenter, 1993) is to perform an analysis of only the ingroup taxa first, and connect an outgroup taxon to the resulting unrooted tree secondarily (Lundberg, 1972). Although the location of the root may still be suspicious, at least the distant outgroup will not confound the estimation of the (unrooted) relationships of the ingroup.

The choice of outgroup taxa can exert a strong effect on the analysis, so the outgroup(s) must be chosen carefully. It is especially important to choose outgroups that minimize the impact of long branches (i.e., it is much more important to break up long sister-group lineages than to increase the sampling density of more distant clades). A.B. Smith (1992) provides an excellent discussion of these and related issues.

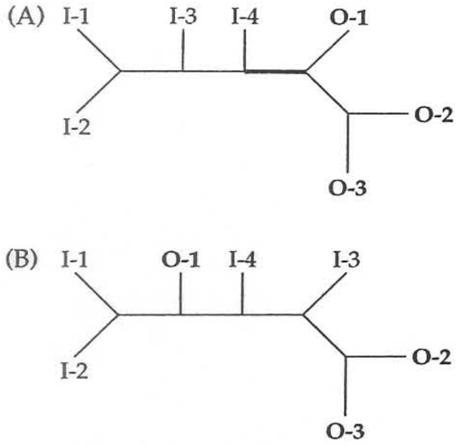


Figure 22 Use of multiple outgroup taxa to infer the location of the root of a tree. (A) The branch indicated in bold partitions the ingroup taxa from the outgroup taxa, yielding an unambiguous root for the ingroup portion of the tree. (B) No single branch partitions the ingroup taxa from the outgroup taxa. The data do not support the assumption of ingroup monophyly.

SEARCHING FOR OPTIMAL TREES

As emphasized above, methods that have explicit optimality criteria (e.g., maximum parsimony, additive-tree distance methods, and maximum likelihood) separate the problem of evaluating a particular tree under the selected criterion from that of finding the optimal tree(s). Most of our presentation to this point has dealt with the former problem; in this section, we address the latter. For data sets of small to moderate size (8–20 taxa, depending on the criterion), exact methods that guarantee the discovery of all optimal trees may be used. For larger data sets, exact solutions require a prohibitive amount of computing time; consequently, approximate methods that do not guarantee optimality must be used.

Exact Algorithms

Exhaustive Search

The conceptually simplest approach to the search for optimal trees is to evaluate every possible tree. Assuming that exact methods exist for evaluating a particular tree, we need only a method for enumerating all possible (strictly bifurcating) trees in

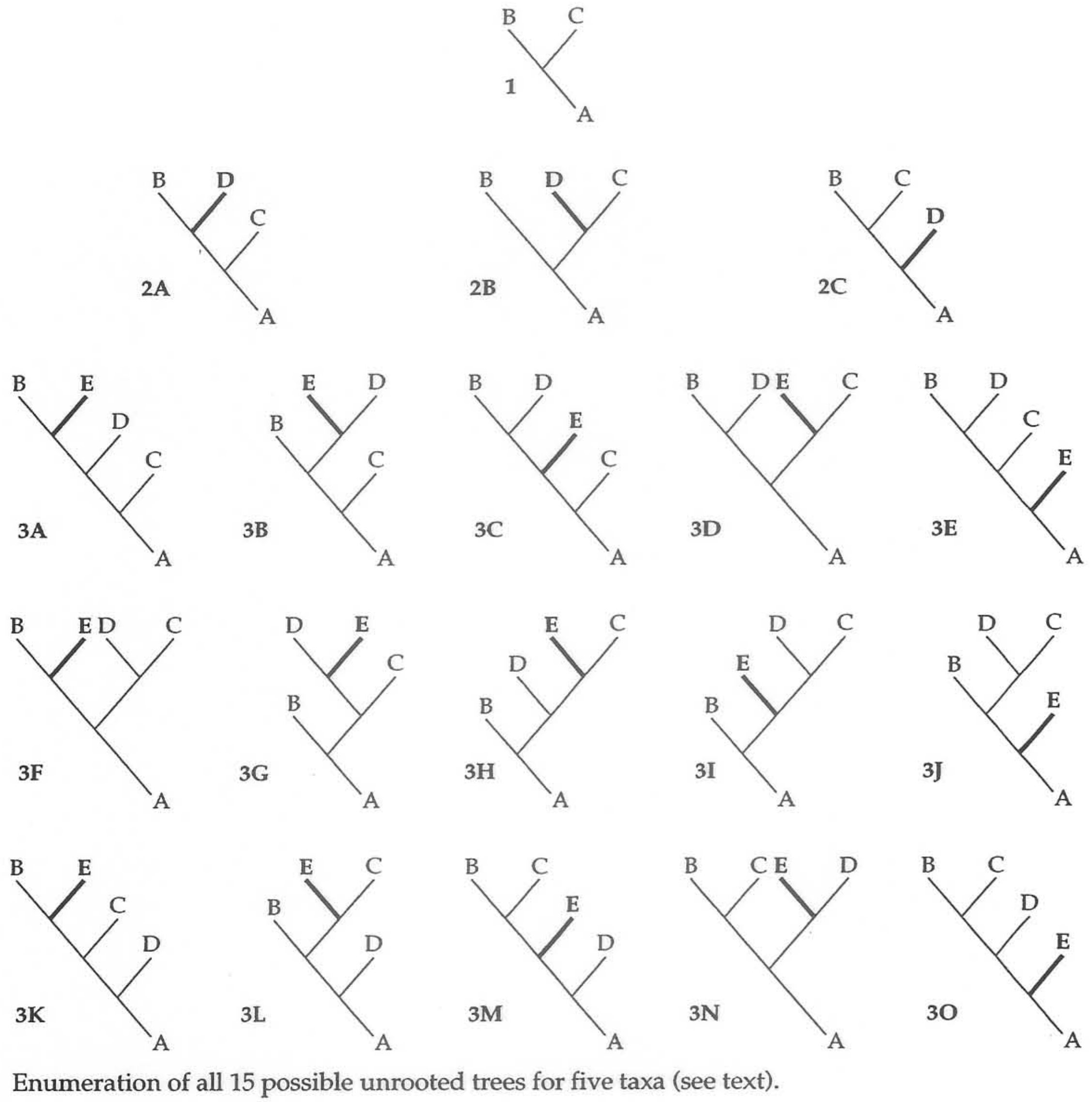


Figure 23 Enumeration of all 15 possible unrooted trees for five taxa (see text).

order to find a globally optimal solution. A simple algorithm, outlined in Figure 23, can be used to perform this enumeration. Initially, we connect the first three taxa in the data set to form the only possible unrooted tree for these taxa (Figure 23, row 1). In the next step, we add the fourth taxon to each of the three branches of the three-taxon tree, thereby generating all three possible unrooted trees for the first four taxa (Figure 23, row 2). We continue in a similar fashion: adding the *i*th taxon to each branch of every tree (containing $i-1$ taxa) generated during a previous step. Thus, for example, row 3 of Figure 23 contains all 15 possible trees for the first five taxa, obtained by adding the fifth taxon to each of the five possible branches for the three trees obtained at the four-taxon stage. This makes clear the rationale for expression (1) for counting the number of possible unrooted bifurcating trees for T taxa: for each of the possible trees for $i-1$ taxa, there are $2(i-1)-3 = 2i-5$ branches to which the *i*th taxon can be connected. Note that the order of addition is immaterial; we could have just as easily chosen taxa at random for next addition at each step.

Evaluation of expression (1) for several possible values of T quickly reveals why exhaustive

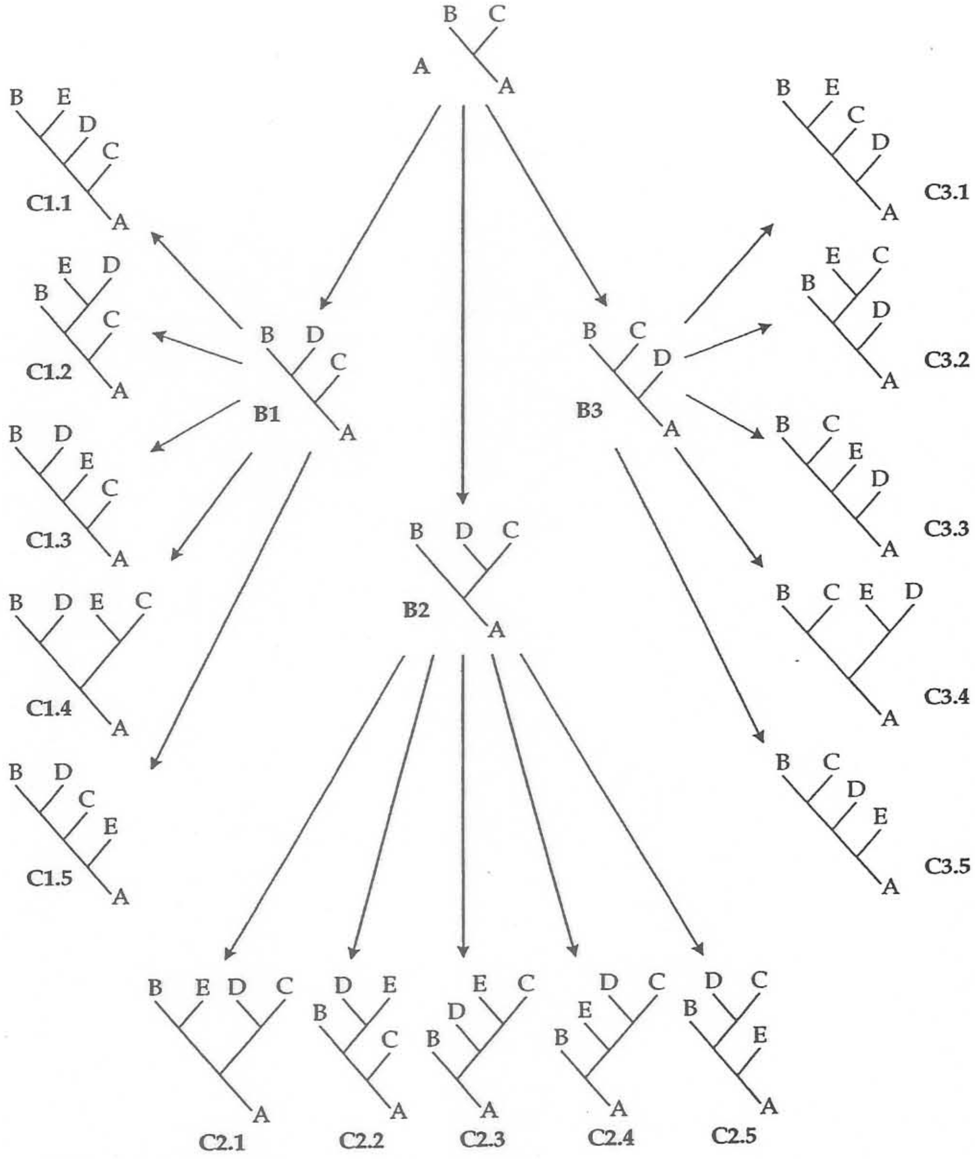


Figure 24 Search tree for branch-and-bound algorithm (see text).

search procedures are useful only for small numbers of taxa. There are 945 possible unrooted trees for only 7 taxa, over 2 × 10^6 trees for 10 taxa, and over 2 × 10^{20} possible trees for 20 taxa (Felsenstein, 1978b; see Table 2 in Chapter 12). Thus, exhaustive enumeration of all possible trees typically is feasible only for 11 or fewer taxa (34,459,425 trees).

Branch-and-Bound Methods

Fortunately, an exact algorithm for identifying all optimal trees that does not require exhaustive enumeration is available for any criterion whose value is known to be non-decreasing as additional taxa are connected to a tree. The branch-and-bound method, frequently used to solve problems in combinatorial optimization, was first applied to evolutionary trees by Hendy and Penny (1982). The branch-and-bound method closely resembles the exhaustive search algorithm described above. In this procedure, we traverse a search tree in a depth-first sequence, as illustrated in Figure 24. The root of the search tree (A) contains the only possible tree for the first three taxa. We first construct one of the three possible trees obtained by connecting taxon 4 to tree A, yielding tree B1. Then, to this tree, we connect taxon 5, yielding tree C1.1. (If there were more than five terminal taxa, we would continue to join additional taxa in this manner until a tree containing all *T* taxa had been completed.) Now, we backtrack one node on the search tree (i.e., back to tree B1) and generate the second tree resulting from the addition of taxon 5 to tree B1 (= tree C1.2). When all five of the trees derivable from tree B1 (C1.1–C1.5) have been constructed, we backtrack all the way to tree A of the search tree and take the second path away from this node, leading to tree B2. As before, all five trees derivable from tree B2 (C2.1–C2.5) are constructed in turn. Then we backtrack once again to tree A and proceed down the third path, toward trees C3.1–C3.5. Eventually we will have constructed all of the possible trees, culminating with tree C3.5. If the score of each tree containing all five taxa were evaluated at the time of its construction, then the search would be an exhaustive one equivalent to that described in the above section. However, a branch-and-bound search differs by eliminating parts of the search tree that only contain suboptimal solutions.

Let *L* represent an upper bound on the optimal value of the chosen optimality criterion. (We assume that we want to minimize this criterion, just as we minimize the tree length under a parsimony criterion or minimize the sum of squared deviations in an additive-tree distance method.) For the present, we can obtain *L* by evaluating a random tree; if we know that a tree of score *L* exists, then the score of the optimal tree(s) cannot exceed this value. As we are moving along a path of the search tree toward its tips (containing all *T* taxa), if we encounter a tree whose score exceeds *L*, then there is no need to proceed further along this path; connecting additional taxa cannot possibly decrease the score. Thus, we can dispense with the evaluation of all (phylogenetic) trees that descend from this node in the search tree and immediately backtrack and proceed down a different path. By cutting off portions of the search tree in this manner, we can greatly reduce the number of trees that must actually be evaluated.

If we reach the end of a path on the search tree and obtain a tree whose score is equal to the upper bound *L*, then this tree is a candidate for optimality. If this score is *less* than *L*, then this is the best tree found so far, and we have improved the upper bound on the score of the optimal tree(s). This improvement is important, as it may enable other search paths to be terminated more quickly. When the entire search tree has been traversed, all optimal trees will have been identified.

The branch-and-bound method is extremely effective for many criteria, permitting exact solutions for 20 or more taxa, depending on the efficiency of the implementation, the speed of the available computer, and the "messiness" of the data. The method can be used to search for optimal trees under parsimony, maximum likelihood, and additive distance criteria in programs such as PAUP* (see Appendix).

The above presentation of the branch-and-bound method, although correct, is an oversimplification of the algorithms actually used in state-of-the-art computer programs. Refinements in the algorithm that greatly speed the computations usually are implemented. These refinements, designed to promote earlier cut-offs in the traversal of the search tree, include: (1) using **heuristic methods** (discussed below) to obtain a near-optimal tree whose score is used as the initial upper bound; (2) designing the search tree so that divergent taxa are added early, thereby increasing the length of the initial trees in the search path; and (3) using pairwise incompatibility to improve the *lower* bound on the length that will ultimately be required by trees descending from a tree at a given node of the search tree. These methods are discussed in more detail in Hendy and Penny (1982) and Swofford (1996).

An obvious question may have occurred to the reader at this point. Since the branch-and-bound method requires evaluation of all trees as its worst possible case, why would we ever want to perform an exhaustive search? In fact, if we were interested only in the optimal trees, the branch-and-bound algorithm would indeed be the preferred means of finding them. However, exhaustive searches permit the researcher to examine the frequency distribution of tree lengths.

It is often useful to know, for example, whether there are few or many near-optimal trees, or where some tree of prior interest lies in the distribution of tree lengths. In addition, with very noisy data, the time spent evaluating bounds can exceed the time spent evaluating the extra trees.

Heuristic Approaches

When a data set is too large to permit the use of exact methods, optimal trees must be sought via heuristic approaches that sacrifice the guarantee of optimality in favor of reduced computing time. The task of searching for an optimal tree by approximate methods is somewhat analogous to the plight of a myopic pilot who loses his glasses when forced to parachute from his airplane into a mountainous region. He suspects that there is a manned outpost at the top of the highest peak in the area, and he must somehow grope his way there to have any hope of rescue. Simply walking uphill from the point of his landing will not necessarily lead to his goal, since he may not have started on a slope of the highest peak. Suppose that he reaches a summit and finds no outpost. Two possibilities remain: (1) he is, in fact, at the top of the highest peak, but was wrong about the existence of the outpost; or (2) he has climbed the wrong hill. Although rather absurd, the analogy is quite appropriate.

Heuristic tree searches generally operate by hill climbing methods. An initial tree is used to start the process; we then seek to improve the tree by rearranging it in a way that improves its score under our chosen optimality criterion (e.g., minimum length). When we can find no way to further improve the tree, we stop. Like the downed pilot, however, we generally have no way of knowing whether we ended up at the top of the highest hill—we do not know whether we have arrived at a global or merely a local optimum.

The details of heuristic search procedures vary considerably from one implementation to the next. In addition, better methods are often invented. Consequently, we prefer to leave the specifics to the documentation of the computer program used to perform the search, and will concentrate on more general concepts.

Stepwise Addition

The most commonly used method for obtaining a starting point for further rearrangement is by stepwise addition of taxa to a growing tree. First, three taxa are chosen for the initial tree. Next, one of the unplaced taxa is selected for next addition. Each of the three trees that would result from joining the unplaced taxon to the tree along one of its (three) branches is evaluated, and the one whose score is optimal is saved for the next round. In this next round, yet another unplaced taxon is connected to the tree, this time to one of the five possible branches on the tree saved from the previous round. The process terminates when all taxa have been joined to the tree.

Of course, the above description is oversimplified in that several decisions are required, none of which has a straightforward answer. Which three taxa should be used initially? How do we decide which unplaced taxon to connect to the tree next? One approach is to simply add the taxa in the same order in which they are presented in the data matrix, starting with the first three and sequentially adding the rest. This strategy, for example, is the one used in Felsenstein's (1993) PHYLIP package. Another approach, optionally available in Swofford's PAUP*, is to check all triplets of taxa and start with the one that yields the shortest tree. At each successive step, all remaining unplaced taxa are considered for connection to every branch of the tree, and the taxon–branch combination that requires the smallest increase in tree length is chosen. Still another approach, suggested by Farris (1970), is to pre-specify an addition sequence based on each taxon's distance to a reference taxon (called a hypothetical ancestor by Farris, but it could just as well be any taxon in the data matrix). Unfortunately, there seems to be no strategy that works best for all data sets; the best approach is to try as many alternatives as possible, each of which may potentially provide a different starting point for branch swapping (see below).

Algorithms like this are referred to as "greedy algorithms" by computer scientists. Like the nearsighted pilot who is unable to scan the horizon and must simply proceed up the nearest hill, these methods choose the solution that looks best given

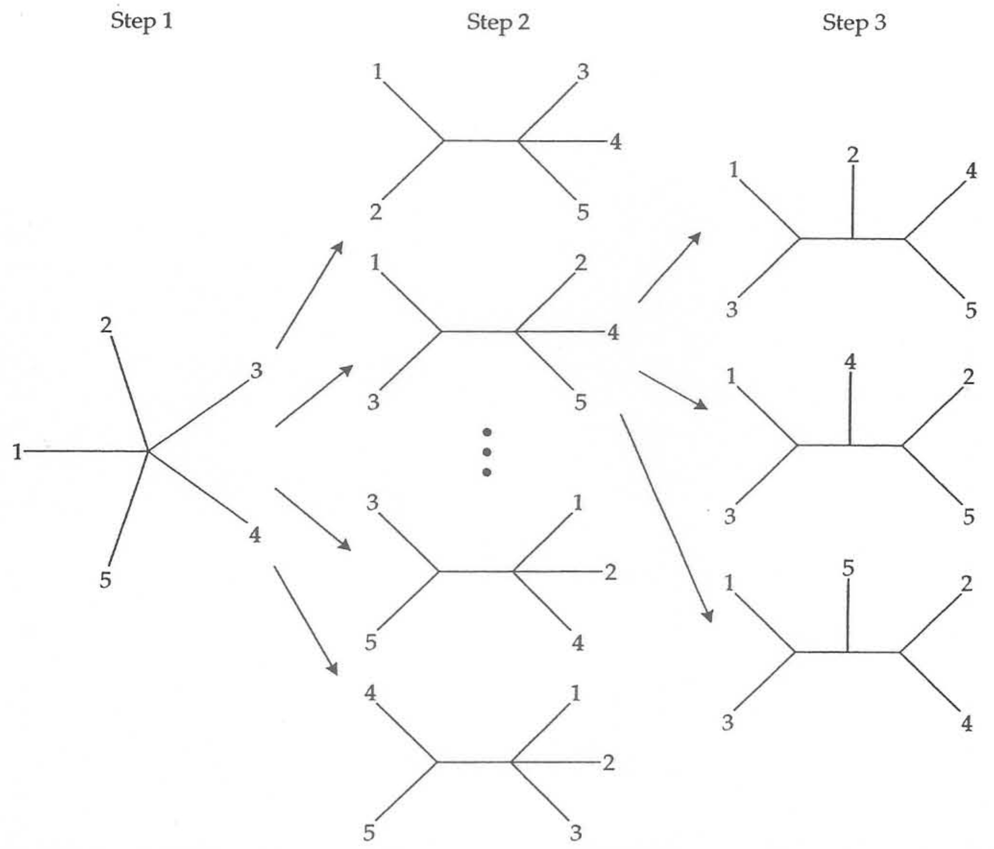


Figure 25 Heuristic tree selection using star decomposition method. At each step, the optimality criterion is evaluated for each possible joining of a pair of lineages leading away from the central node. The best tree found during each step becomes the starting point for the next step.

the current situation rather than attempting to see more broadly into the future. Thus, one placement of a taxon may be best given the taxa currently on the tree, but that placement may become suboptimal upon the addition of subsequent taxa. Once a decision has been made to connect a taxon to a certain point, however, we must usually accept the consequences of that decision for the remainder of the stepwise addition process, perhaps ending up in a local optimum as a result.

Star Decomposition Methods

An alternative to stepwise addition is the star decomposition method, a divisive pairwise clustering method (see "Cluster Analysis," below). The algorithm can be used with any criterion that can be evaluated on a non-binary (polytomous) tree. To begin, we connect all of the terminal taxa connected in a "star tree" containing a single internal node (Figure 25, step 1). Next, we evaluate the optimality criterion for all possible trees that can be constructed by joining two of the terminal nodes into a new group (Figure 25, step 2). The tree from this stage that scores best according to the criterion is saved for the next step. Each time we form a new group, we reduce by one the number of branches connected to the central node. The process continues until the step in which all generated trees are binary (Figure 25, step 3), and we choose the best of these (again according to the chosen optimality criterion).

The most commonly used star decomposition method is the neighbor-joining algorithm of Saitou and Nei (1987; see below). Saitou (1990), Adachi and Hasegawa (1992), and Z. Yang (1995) have also implemented the method for both DNA and protein maximum likelihood. Star decompo-

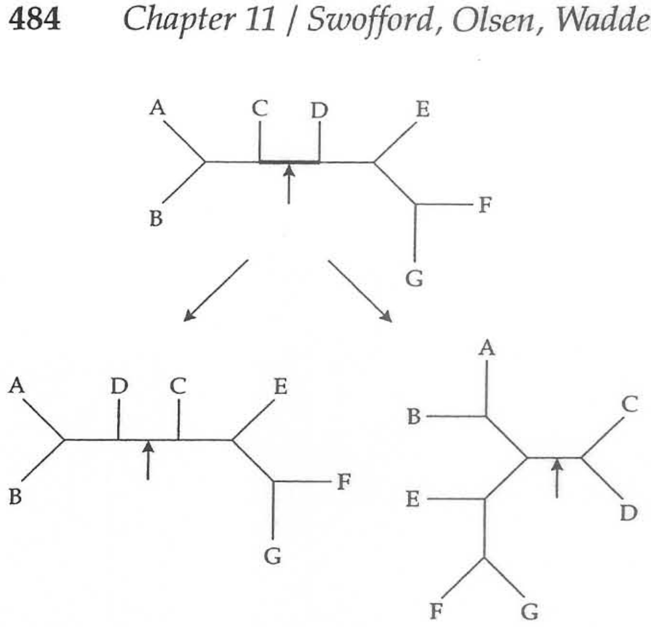


Figure 26 Branch swapping by nearest-neighbor interchanges (NNIs). Each interior branch of the tree defines a local region of four subtrees connected by the interior branch. Interchanging a subtree on one side of the branch with one from the other constitutes an NNI. Two such rearrangements are possible for each interior branch.

sition, like stepwise addition, is a greedy algorithm that is prone to entrapment in local optima.

Branch Swapping

Because of the excessive greediness and susceptibility to local optima problems, stepwise addition and star decomposition algorithms generally do not find optimal trees unless the number of taxa is small or the data are very clean. However, it may be possible to improve the initial estimate by performing sets of predefined rearrangements, a technique commonly referred to as branch swapping. In general, any one of these rearrangements amounts to a "stab in the dark," but the hope is that if a better tree exists, one of the rearrangements will find it. Examples of three kinds of rearrangements used in current branch-swapping algorithms are shown in Figures 26 through 28.

Of course, the globally optimal tree(s) may be several rearrangements away from the starting tree. If a rearrangement is successful in finding a better tree, a round of rearrangements is initiated on this new tree. As long as each round of rearrangements is successful in finding an improved tree (according to its score under the optimality criterion), then we will eventually arrive at the global optimum. However, if the intermediate trees would require us to pass through trees that are inferior to the one(s) already obtained, we will once again find ourselves trapped in a local optimum unless an option is provided for branch swapping on suboptimal trees (e.g., the "KEEP" option in PAUP*; Swofford, 1993, 1996). A related problem concerns plateaus on the optimality surface. It may be the case, for example, that an optimal tree lies several rearrangements away from the current tree, and that these rearrangements all correspond to trees having equal scores under the optimality criterion. If the intermediate trees are discarded because they are "not better," then the optimal tree will not be found. A few programs do not retain equally good trees because they have no protection against cycling (alternation between two trees, each of which can be rearranged to yield the other); these programs will not be effective if plateaus are encountered, since they are unable to traverse the plateau.

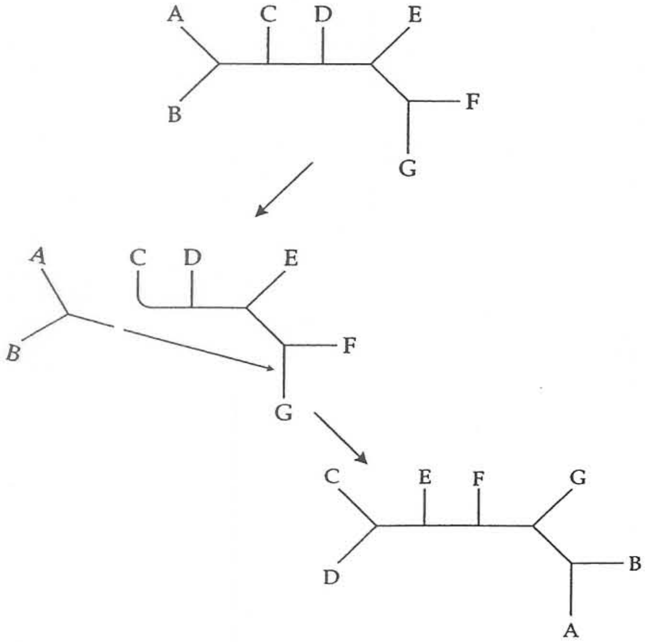


Figure 27 Branch swapping by subtree pruning and regrafting. A subtree is pruned from the tree (e.g., the subtree containing terminal nodes A and B as indicated). The subtree is then regrafted to a different location on the tree. All possible subtree removals and reattachment points are evaluated.

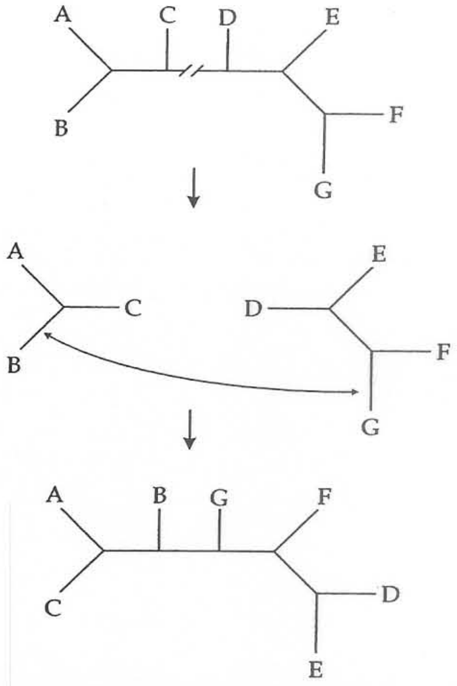


Figure 28 Branch swapping by tree bisection and reconnection. The tree is bisected along a branch, yielding two disjoint subtrees. The subtrees are then reconnected by joining a pair of branches, one from each subtree. All possible bisections and pairwise reconnections are evaluated.

Testing for Convergence

Because of the limitations of heuristic approaches, some way of evaluating the success of the chosen method in obtaining a globally optimal solution is needed. The obvious strategy in this regard is to begin from different starting points and ask whether the same result is always obtained. For example, a set of random sequences for the addition of taxa can be used to generate initial trees for input to branch swapping. Since, for reasonably noisy data at least, the starting trees will vary depending on the addition sequence, convergence to a common optimal tree (or set of trees) is encouraging. (A more extreme approach—using random trees rather than random addition sequences—could be adopted; however, the starting trees are, on average, so far from the optimal trees that this strategy seems to be less effective.) Even if rearrangements of different starting trees do not converge to the same end point, the use of several starting trees is a good idea; if multiple peaks on the optimality surface exist, we will be more likely to find them.

Alternatives to Hill Climbing

Even when greedy algorithms like stepwise addition or star decomposition are followed by branch swapping, entrapment in local optima can still occur. Fundamentally, any search heuristic consists of pseudorandomly perturbing (rearranging) the current solution until either the resulting solution is acceptable, or a stopping criterion is satisfied. The criteria of acceptability are what separate the heuristic search methods from each other: the nature of the perturbations used is problem-dependent.

We can think of the "goodness" of a solution t_i as some function $z(t_i)$, for each step i. Thus, in hill climbing, t_{i+1} is acceptable if $z(t_{i+1}) \geq z(t_i)$: our myopic pilot will never go anywhere that takes him downhill, only uphill or across. In simulated annealing (Van Laarhoven and Aarts, 1987) a new solution is accepted if $z(t_{i+1}) \geq z(t_i)$, as in hill climbing, but even if $z(t_{i+1}) < z(t_i)$, then the procedure will accept the new solution with a certain probability, as follows:

$$\text{Prob}\left[\text{accepting solution } z(t_{i+1})\right] = \begin{cases} 1 & \text{if } z(t_{i+1}) \geq z(t_i) \\ e^{-k[z(t_i+1)-z(t_i)]} & \text{otherwise} \end{cases}$$

where k is a parameter that can vary over time.

In the Great Deluge method (Dueck, 1990; Dueck and Scheuer, 1990), the probability of accepting a new solution t_i is 1 if $z(t_{i+1}) > w_i$, where w_i is a bound that increases slowly with time, so that if t_{i+1} is accepted, then $w_{i+1} = w_i + c[z(t_{i+1}) - (t_i)]$. The constant c is usually about 0.01 to 0.05. These methods of determining the acceptability of a new solution offer an efficient means of improving the performance of heuristic searches (M. Charleston, personal communication), and there are many other variants, including the use of a "tabu list" (Glover, 1989) that prevents the search from revisiting any solutions it has just tried (the list usually contains about 5 to 10 solutions).

Algorithmic and Other Methods

The methods for tree searching described in the above sections are appropriate when an optimality criterion that can be evaluated for any given tree is chosen. The problem is then reduced to finding an optimal tree given the chosen criterion. The methods described below do not cleanly fit into this framework, either because they are defined solely on the basis of an algorithm or because the task of finding an optimal tree cannot be cleanly separated from that of evaluating a specific tree.

Cluster Analysis

Cluster analysis is a family of related techniques for representing similarity or distance data (we will use distances) in the form of an ultrametric tree (Sneath and Sokal, 1973). If the data themselves are ultrametric, then the representation on the tree will be exact. It should be obvious that if the distance data themselves are not ultrametric, then they *cannot* be fit exactly to such a tree, and therefore errors might be introduced.

The method of cluster analysis is conceptually simple. The raw data are provided as a table of distances between all pairs of taxa. Call d_{ij} the distance between taxa i and j. The tree is constructed by linking the least distant pairs of taxa, followed by successively more distant taxa, or groups of taxa. When two taxa are linked, they lose their individual identities and are subsequently referred to as a single cluster. Initially, each taxon constitutes its own cluster. At each stage in the process, as two clusters are merged into one, the number of clusters declines by one. The process is complete when the last two clusters are merged into a single cluster containing all of the original taxa.

The steps of the method are as follows:

1. Given a matrix of pairwise distances, find the clusters (taxa) i and j such that d_{ij} is the minimum value in the table.
2. Define the depth of the branching between i and j (l_{ij}) to be $d_{ij}/2$.
3. If i and j were the last two clusters, the tree is complete. Otherwise, create a new cluster called u.
4. Define the distance from u to each other cluster (k, with $k \neq i$ or j) to be an average of the distances d_{ki} and d_{kj}.
5. Go back to step 1 with one less cluster; clusters i and j have been eliminated, and cluster u has been added.

The variants are primarily in the details of step 4. The most commonly used clustering method is **UPGMA** (unweighted pair group method using arithmetic averages), in which the averaging of the distances in step 4 is based on the total number of taxa in the clusters. That is, if cluster i contains T_i taxa, and cluster j contains T_j taxa, then $d_{ku} = (T_i\, d_{ki} + T_j\, d_{kj})/(T_i + T_j)$. If the simple average $[d_{ku} = (d_{ki} + d_{kj})/2]$ is used instead, the technique is called **WPGMA** (weighted PGMA). Other variants include using the maximum distance [$d_{ku} = \max(d_{ki}, d_{kj})$, called complete linkage], or the minimum distance [$d_{ku} = \min(d_{ki}, d_{kj})$, called single linkage]. These alternatives all give the same results when the data are ultrametric, but they can differ in their inferences when the data are not ideal.

An example of using UPGMA to infer a tree of five taxa (5S rRNA sequences) is given in Figure 29. The figure presents the upper right half of the pairwise distance matrix at each stage of the cluster analysis. Starting with the first table, the smallest distance, the 0.1715 substitutions per sequence position separating Bsu and Bst, is indicated in bold face. Thus, the first inferred branching unites these taxa at a depth of 0.1715/2 = 0.0858. These two taxa are merged into a cluster in the next table, and their distances to all other taxa are averaged. For example, the distance from the Bsu-Bst group to Lvi is (0.2147 + 0.2991)/2 = 0.2569. The smallest distance in the second table joins the Bsu-Bst cluster with Mlu at a depth of 0.1096 (= 0.2192/2). The distances of the Bsu-Bst-Mlu cluster to the other taxa are then computed by the unweighted method. For example, the distance to Lvi is (2 × 0.2569 + 0.3943)/3 = 0.3027. Notice that this value is identical to (Bsu:Lvi + Bst:Lvi + Mlu:Lvi)/3, where A:B is the distance from taxon A to taxon B. Each taxon in the original data table contributes equally to the averages, which is why the method is called *unweighted*. The

	Bsu	Bst	Lvi	Amo	Mlu
Bsu	—	**0.1715**	0.2147	0.3091	0.2326
Bst		—	0.2991	0.3399	0.2058
Lvi			—	0.2795	0.3943
Amo				—	0.4289
Mlu					—

	Bsu-Bst	Lvi	Amo	Mlu
Bsu-Bst	—	0.2569	0.3245	**0.2192**
Lvi		—	0.2795	0.3943
Amo			—	0.4289
Mlu				—

	Bsu-Bst-Mlu	Lvi	Amo
Bsu-Bst-Mlu	—	0.3027	0.3593
Lvi		—	**0.2795**
Amo			—

	Bsu-Bst-Mlu	Lvi-Amo
Bsu-Bst-Mlu	—	**0.3310**
Lvi-Amo		—

Figure 29 Cluster analysis (UPGMA) of 5S rRNA evolutionary distance estimates. Abbreviations correspond to Figure 15. Each table represents the pairwise distances (estimated nucleotide substitutions per sequence position) for one round of clustering (only the upper right half of the symmetrical matrix is shown). The minimum distance value in each table is in bold. The corresponding pair of taxa (or clusters) are merged into a single cluster in the next table. The bold distance value is twice the depth of the branch point separating the clusters merged. A diagram of the inferred tree is in Figure 15B.

smallest distance in the third table unites Lvi and Amo at a depth of 0.1398. The distance between the Bsu-Bst-Mlu and Lvi-Amo clusters is then (3 × 0.3027 + 3 × 0.3593)/6 = 0.3310. Thus the implied root of the tree joins these two clusters at a depth of 0.1655. The complete tree is shown in Figure 15B.

Note that cluster analysis cannot join two taxa (sometimes called operational taxonomic units or **OTUs**) unless at least one pairwise distance links them. Thus, missing data within a group can force one or more members out of the group in the inferred tree, a problem discussed in greater detail under "Similarity and Distance Data."

Cluster analysis has historically been very popular for several reasons. First, the principal assumption is that the data are approximately ultrametric. This assumption is of course a very strong one, but it is seductive to believe that a single stringent assumption can be satisfied more easily than a long list of (what might be) less restrictive assumptions. Second, the idea of grouping the taxa that are least different, regardless of any finer points of consideration, has a strong intuitive appeal. The extreme of this view is the phenetic perspective in which it is asserted that nothing but the extent of similarity matters biologically and that consideration of the historical branching order is of purely secondary interest. A third reason is the availability of programs to do cluster analysis and the relative speed of the calculations, thereby enabling large numbers of taxa to be analyzed.

As emphasized above, simple cluster analysis has drawbacks. First, it is just an algorithm (or family of algorithms) with no objective definition of what constitutes an optimal tree when the data are not ideal. In particular, because genes do not diverge uniformly in all organisms or organelles (Chapters 8, 9, and 12), systematic errors are likely to be introduced into cluster analysis reconstructions. Finally, alternative, rapid methods are available that will work for all additive trees, not just those that are ultrametric.

Algorithmic Methods for Additive Trees

A variety of algorithmic methods related to cluster analysis have been proposed that will correctly reconstruct additive trees, whether the data are ultrametric or not. These methods fall into three primary categories. Those in the first category transform any additive distance matrix into an ultrametric matrix and then use cluster analysis to infer the tree. They include the transformed distances method of W.-H. Li (1981), the present-day ancestor method of Klotz and Blanken (1981), and, in a less obvious sense, the neighbor-joining method of Saitou and Nei (1987). The second category comprises methods that form the clusters consistent with the largest fraction of taxon-quartets, using a relaxed definition of additivity for a four-taxon tree. These methods include those of Sattath and Tversky (1977) and Fitch (1981). Methods of the third class, which includes the distance Wagner method (Farris, 1972), build an additive representation of the tree by sequential addition of taxa. The transformed distance approaches all have a computational complexity that is proportional to T^3; therefore, any problem that is tractable with standard cluster analysis can also be solved with these methods. We present a version of the neighbor-joining method below.

Unlike cluster analysis, additive-tree methods yield unrooted trees, which are adequate for some purposes. If a root is to be placed, however, it must be based on an ancillary criterion. Usually, one or more taxa that are assumed to lie outside a monophyletic group of interest are included in the analysis. The location at which these taxa join the tree defines the root with respect to the ingroup. Another method, midpoint rooting, depends on an assumption of rate uniformity that is somewhat weaker than assuming a molecular clock across the entire tree: if the two most divergent lineages have evolved at the same rate, then the appropriate root is at the midpoint of the path connecting these taxa.

THE NEIGHBOR-JOINING METHOD Neighbor joining (Saitou and Nei, 1987) is conceptually related to traditional cluster analysis, but removes the assumption that the data are ultrametric. In practical terms, it does not assume that all lineages have diverged equal amounts. However, it does assume that the data come close to fitting an additive tree, so correction for superimposed substitutions is important for data that might include lineage-to-lineage differences in average rate.

The neighbor-joining algorithm is a special case of the star decomposition method described earlier. In contrast to cluster analysis, neighbor joining keeps track of nodes on a tree rather than taxa or clusters of taxa. The raw data are provided as a distance matrix, and the initial tree is a star tree. A modified distance matrix is constructed in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes (conceptually, this adjustment has the effect of normalizing the divergence of each taxon for its average clock rate). The tree is constructed by linking the least-distant pair of nodes as defined by this modified matrix. When two nodes are linked, their common ancestral node is added to the tree and the terminal nodes with their respective branches are removed from the tree. This pruning process converts the newly added common ancestor into a terminal node on a tree of reduced size. At each stage in the process, two terminal nodes are replaced by one new node (corresponding to an internal node on the final tree). The process is complete when two nodes remain, separated by a single branch.

The steps of the method (modified from Studier and Keppler, 1988) are as follows:

1. Given a matrix of pairwise distances (**d**), for each terminal node i calculate its net divergence (r_i) from all other taxa using the formula

$$r_i = \sum_{k=1}^{N} d_{ik} \qquad \textbf{(41)}$$

where N is the number of terminal nodes in the current matrix. Note the assumption that $d_{ii} = 0$, otherwise the summation would need to skip over $k = i$.

2. Create a rate-corrected distance matrix (**M**) in which the elements are defined by

$$M_{ij} = d_{ij} - (r_i + r_j)/(N-2) \qquad \textbf{(42)}$$

for all i and with $j > i$ (the matrix is symmetrical, and the case of $i = j$ is not interesting). Only the values i and j for which M_{ij} is minimum need be recorded; saving the entire matrix is unnecessary.

3. Define a new node u whose three branches join nodes i, j, and the rest of the tree. Define the lengths of the tree branches from u to i and j:

$$v_{iu} = d_{ij}/2 + (r_i - r_j)/[2(N-2)]$$

$$v_{ju} = d_{ij} - v_{iu}$$

4. Define the distance from u to each other terminal node (for all $k \neq i$ or j)

	Bsu	Bst	Lvi	Amo	Mlu	R	R/3
Bsu	—	0.1715	0.2147	0.3091	0.2326	0.9279	0.3093
Bst	−0.4766	—	0.2991	0.3399	0.2058	1.0163	0.3388
Lvi	−0.4905	−0.4356	—	**0.2795**	0.3943	1.1876	0.3959
Amo	−0.4527	−0.4514	**−0.5689**	—	0.4289	1.3574	0.4525
Mlu	−0.4972	−0.5535	−0.4221	−0.4441	—	1.2616	0.4205

Lvi to node 1 distance = 0.2795/2 + (0.3959 − 0.4525)/2 = 0.1114
Amo to node 1 distance = 0.2795 − 0.1114 = 0.1681

	Bsu	Bst	Mlu	Node 1	R	R/2
Bsu	—	0.1715	0.2326	**0.1222**	0.5263	0.2631
Bst	−0.3701	—	0.2058	0.1798	0.5571	0.2785
Mlu	−0.3856	−0.4278	—	0.2719	0.7103	0.3551
Node 1	**−0.4278**	−0.3856	−0.3701	—	0.5739	0.2869

Bsu to node 2 distance = 0.1222/2 + (0.2631 − 0.2869)/2 = 0.0492
node 1 to node 2 distance = 0.1222 − 0.0492 = 0.0730

	Bst	Mlu	Node 2	R	R/1
Bst	—	0.2058	**0.1146**	0.3204	0.3204
Mlu	−0.5116	—	0.1912	0.3970	0.3970
Node 2	**−0.5116**	−0.5116	—	0.3058	0.3058

Bst to node 3 distance = 0.1146/2 + (0.3204 − 0.3058)/2 = 0.0646
node 2 to node 3 distance = 0.1146 − 0.0646 = 0.0500

	Mlu	Node 3
Mlu	—	0.1412
Node 3		—

Mlu to node 3 distance = 0.1412

Figure 30 Neighbor joining of 5S rRNA evolutionary distance estimates. The data and abbreviations are as in Figure 29. Each table presents the pairwise distance values input to the round of analysis (upper right half of the matrix). The rightmost two columns present the row totals for the uncorrected distances (the row being defined based on the full symmetrical matrix; see equation 41) and the total divided by the number of terminal nodes minus two. The rate-corrected pairwise distances as defined by equation (42) are given in the lower left half of the matrix. The minimum corrected distance value in each table and the corresponding uncorrected pairwise distance are shown in bold. The corresponding pair of taxa (or clusters) are removed from the matrix and replaced by their common ancestral node in the next table and distances based on equation (43). The inferred tree is diagrammed in Figure 15A.

$$d_{ku} = \left(d_{ik} + d_{jk} - d_{ij}\right)/2 \qquad \textbf{(43)}$$

5. Remove distances to nodes *i* and *j* from the data matrix, and decrease *N* by 1.

6. If more than two nodes remain, go back to step 1. Otherwise, the tree is fully defined except for the length of the branch joining the two remaining nodes (*i* and *j*). Let this remaining branch be

$$v_{ij} = d_{ij}$$

Each step has generated one internal node and has estimated the lengths of two of the branches connected to that node. The tree can be drawn from these data.

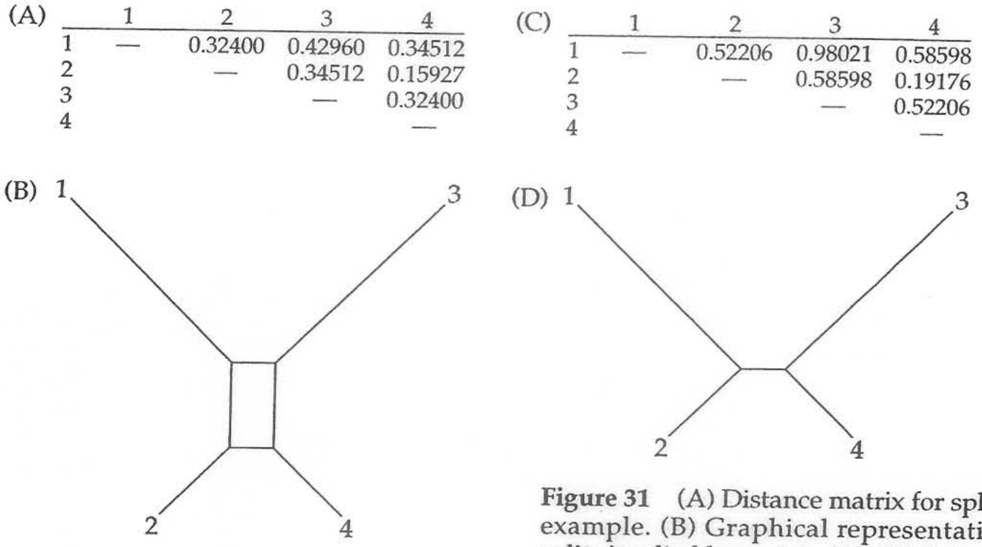
An example of using neighbor joining to infer a tree of five taxa is given in Figure 30. The data are the same as in the cluster analysis example in Figure 29. The pairwise distance estimates are in the upper right triangle of each matrix (ignoring the last two columns). The distance matrix row totals [*r* from equation (41)] and *r*/(*N*–2) are given in the last two columns. The rate-corrected distances are in the lower left triangle of the table. For example, the corrected Bsu to Bst distance is 0.1715 – (0.3093 + 0.3388) = –0.4766. A general property of these corrected distances is that they are negative; therefore, finding the minimum distance means finding the most negative value. In the first table, the minimum value is the –0.5689 relating Amo and Lvi. Both this value and the corresponding uncorrected distance, 0.2795, are in boldface. Thus, Amo and Lvi are joined to one another and to the rest of the taxa through a new node, called node 1 in this example. The two lines below the table illustrate the calculation of the branch lengths from the two taxa to the node. Amo and Lvi are then removed from the distance table, and the distances from node 1 to the remaining taxa are calculated using equation (43). For example, the Bsu to node 1 distance is (0.2147 + 0.3091 – 0.2795)/2 = 0.1222. The second table, which now relates only four terminal nodes, is treated just as the first table. Looking at the corrected distances, we find two pairs with the lowest value, –0.4278. This is not a coincidence: if Bsu and node 1 are sister nodes, then Bst and Mlu must also be sister groups. (If this observation is unclear, try drawing the unrooted tree of four taxa.) The remaining arithmetic will yield identical trees regardless of which of these two pairs are joined at this step. In this example, node 2 is added to the tree, joining Bsu, node 1, and the rest of the tree. The branch lengths from Bsu and node 1 to node 2 are calculated below the table. The third table eliminates Bsu and node 1, and adds node 2. In this table, which relates three peripheral nodes, all three rate-corrected distances are identical. As in the previous step, this result is not a coincidence: only one possible unrooted tree can link three taxa. The choice of the pair to be joined is arbitrary; the ultimate outcome will be the same. Adding node 3 to the tree so that it links Bst and node 2 to the rest of the tree (which is only Mlu at this point) gives one more pair of branch lengths and a "tree" containing node 3 and Mlu. Their pairwise distance is used directly as the length of the segment joining them. The tree is completed. The results are shown in Figure 15A.

As the neighbor-joining algorithm seeks to represent the data by an additive tree, it can assign a negative length to a branch. Kuhner and Felsenstein (1994) modified the algorithm so that when a negative branch length occurred, it was set to zero, and the difference was transferred to the adjacent branch length so that the total distance between an adjacent pair of terminal nodes was unaffected. This change does not alter the topology of the tree found by the algorithm; it just guarantees non-negativity of branch lengths (e.g., for interpreting branch lengths as estimated numbers of substitutions).

Neighbor joining is classified as an algorithmic method because it constructs only one tree and does not explicitly optimize any objective function (the branch-length estimates from neighbor joining are not, in general, optimal for the minimum evolution criterion). We believe that it should be thought of as a means of getting a starting tree for more thorough searches using branch swapping under the minimum evolution or other additive-tree criteria, not as a method for choosing a final tree.

SPLIT DECOMPOSITION All of the methods described above will select a tree regardless of how non-treelike the data appear. When the data do not conform to a treelike model, criterion-based methods may provide some indication of a problem, for example, by discovering some nearly optimal trees that are quite different in topology. Algorithmic methods such as neighbor joining provide little or no indication that the data do not conform to the model. Split decomposition (Bandelt and Dress, 1992) is a method for graphically representing trends in distance data. The method detects well-supported groupings when they occur, but also identifies conflicting (incompatible) groups that may also have strong support in the data. These conflicts can arise from sources such as inadequate correction for superimposed changes in the distance transformation, convergence driven by natural selection, or reticulate evolution. We will not give a complete description of this method, but will outline the basic ideas using a simple example.

The method is based on the four-point metric (formula 10) (Buneman, 1971), which states that if taxa i, j, k, and l (a quartet) are related by a tree $((i, j), (k, l))$ and the distances are tree-additive, then the minimum sum will be $d_{ij} + d_{kl}$, while the larger sums $d_{ik} + d_{jl}$ and $d_{il} + d_{jk}$ will be equal. With real data (i.e., imperfectly additive distances), the relationship $d_{ik} + d_{jl} = d_{il} + d_{jk}$ will not hold. Al-



(A)

	1	2	3	4
1	—	0.32400	0.42960	0.34512
2		—	0.34512	0.15927
3			—	0.32400
4				—

(C)

	1	2	3	4
1	—	0.52206	0.98021	0.58598
2		—	0.58598	0.19176
3			—	0.52206
4				—

Figure 31 (A) Distance matrix for split decomposition example. (B) Graphical representation (network) of splits implied by matrix (A). (C) Poisson-corrected distance matrix. (D) Correct tree inferred from matrix (C).

though we could hope that $d_{ij} + d_{kl} < d_{ik} + d_{jl}$ and $d_{ij} + d_{kl} < d_{il} + d_{jk}$, which forms the basis of the Sattath–Tversky (1977) and Fitch (1981) "neighborliness" methods, even this relationship will usually be violated by some quartets. Split decomposition adopts the working assumption that at the very least, $d_{ij} + d_{kl}$ will not be the largest of the three sums. Usually, phylogenetic methods assume that if $d_{il} + d_{jk}$ exceeded both other sums, then there is no support in the data for the tree $((i,l), (j,k))$. However, we can also ask whether there is relatively unambiguous support for one of the other two trees. For example, if $d_{ij} + d_{kl}$ and $d_{ik} + d_{kl}$ are nearly equal, but both are distinctly smaller than $d_{il} + d_{jk}$, conflicting support is evident. The closer one of these two sums approaches $d_{il} + d_{jk}$, the more consistent is the support for the tree corresponding to the other sum.

We illustrate this procedure using the hypothetical example of Figure 31. The distances in Figure 31A are the observed or uncorrected distances that would be expected from the example used to illustrate the Hadamard conjugation (i.e., calculated using the relationship $d = (1 - r)/2$; see equation 31). The three relevant distance sums are:

$$d_{12} + d_{34} = 0.64800$$

$$d_{13} + d_{24} = 0.58887$$

$$d_{14} + d_{23} = 0.69024$$

Thus, we reject the tree ((1,4),(2,3)) and calculate an *isolation index* representing support for each of the other partitions (splits) as

$$S_{12,34} = [(d_{14} + d_{23}) - (d_{12} + d_{34})] / 2 = 0.04224$$

$$S_{13,24} = [(d_{14} + d_{23}) - (d_{13} + d_{24})] / 2 = 0.10137$$

The observation that support for the ((1,2),(3,4)) split is nearly half that of the support for the ((1,3),(2,4)) split suggests that there is conflicting support for two different groupings in the data set. This conflict is represented by drawing the tree as a network showing the amount of support for each of the two supported groupings (Figure 31B). A standard tree-building method such as neighbor joining would, in contrast, select the tree ((1,3),(2,4)) but give no indication of the support in the data set for the alternative tree ((1,2),(3,4)).

In this example, we can identify the cause of the conflict as failure to account for superimposed changes, which in this case would cause selection of an incorrect tree using neighbor joining or other additive-tree methods. However, using the standard Poisson correction (equation 27), we can obtain the corrected distance matrix shown in Figure 31C. (Note that the elements of this matrix are equal to one-half of the appropriate elements in the corrected generalized distance vector **ρ** of equation 30.) For this corrected matrix, we have

$$d_{12} + d_{34} = 1.0441$$

$$d_{13} + d_{24} = 1.1720$$

$$d_{14} + d_{23} = 1.1720$$

When split decomposition is performed using the corrected distances, the box in Figure 31B indicating conflicting support disappears because $[(d_{14} + d_{23}) - (d_{13} + d_{24})] / 2 = 0$, and the correct tree is inferred (Figure 31D).

For a tree of more than four taxa, the deviation from the additive four-point metric condition is measured for all possible subsets of four taxa. Bandelt and Dress (1992) showed that only a certain number of the implied splits can be portrayed on a planar graph (the *split decomposable* portion); the proportion which cannot is referred to as the *split-prime residue*. Bandelt and Dress (1992) suggested that the majority of the random noise contained in a data set is transferred to the split-prime residue (which also contains some systematic biases that are only locally uniform in their direction). Remaining random noise and systematic error is retained in the split-decomposable component and is observed on the resulting network as incompatibilities between splits (or unresolved nodes; see below).

When using split decomposition on substantial numbers of taxa, the resulting graph often appears more like an unresolved tree than a network with many boxes. Distant outgroups, for example, can show large random fluctuations and also different systematic biases, tending to hide the information on ingroup systematic bias (as all three quartet relations may be optimal depending on which taxa are used). When this happens, localized (but possibly strong) systematic error is lost in the split-prime residue and the graph loses both "boxiness" and resolution. One solution to this problem is to look for systematic errors by restricting the analysis to smaller subsets of taxa (4–10).

Because it is based on distances and not characters, split decomposition by itself does not allow one to detect when conflicting splits are due to events such as horizontal transfer of DNA or recombination. Such claims should be evaluated with more sensitive character- and sequence-based methods (e.g., Stephens, 1985; Hein, 1990a, 1993). A more straightforward use of the method is in the choice of a distance transformation (e.g., allowing more substitution parameters, unequal rates across sites, and/or unequal base compositions). Split composition can give some idea of whether these transformations are improving the "treelikeness" of the graph or making it worse (visualized as a more "boxy" network; e.g., see Lockhart et al., 1995b).

Split decomposition analysis will not necessarily detect some kinds of departures from predictions of a model, again because we are starting from distances rather than characters. For example, unlike the Hadamard conjugation, split decomposition will not recognize an excess of patterns supporting all three four-taxon trees, as would happen if there were more superimposed changes than the model predicts. Like the Hadamard conjugation, we need a means of determining whether a conflicting "signal" is really present or is simply due to sampling error causing inequality of $d_{ik} + d_{jl} = d_{il} + d_{jk}$ by chance. Unfortunately, this question has received little attention, but with small data sets it is possible to determine analytically how many standard deviations separate the three sums of distances. A bootstrap approach (see below) to assessing the reliability of features in the split decomposition graph is also feasible, but will probably be conservative. The relationship between split decomposition and the distance Hadamard is not well understood; both methods should be considered useful because they give different insights.

METHODS BASED ON A RELAXED FOUR-POINT METRIC

The methods of Sattath and Tversky (1977) and Fitch (1981) are also based on a relaxation of the four-point metric condition of Buneman (1971). However, they are based on a somewhat stricter criterion than split decomposition. These methods operate by creating a similarity matrix $\mathbf{s}_{ij}$ that counts the number of times each pair of taxa *i* and

j satisfy the conditions $d_{ij} + d_{kl} < d_{ik} + d_{kl}$ and $d_{ij} + d_{kl} < d_{il} + d_{jk}$ over all pairs (k, l). This matrix forms the basis for a cluster analysis. We begin by choosing the pair (i, j) for which s_{ij} is maximal, and form the corresponding cluster. These two taxa are merged into a single object and distances are recalculated as in UPGMA. The quartet-based scoring of pairs of taxa is then repeated, and the cycle continues until all taxa have been clustered. (The Sattath–Tversky and Fitch methods differ slightly in the details of how averaging is performed in preparation for the next clustering cycle.)

The Sattath and Tversky (1977) and Fitch (1981) methods have not been widely used. Furthermore, simulations by Charleston (1994) indicate that these methods (and other transformed distance methods, such as that of W.-H. Li, 1981) are less effective in identifying the correct tree than methods such as neighbor joining or closest tree (applied to the distance Hadamard). They are also more computationally intensive (requiring time proportional to T^5, as opposed to T^3 for neighbor joining).

DISTANCE WAGNER AND RELATED METHODS The conceptual perspective of Fitch–Margoliash methods and neighbor joining is that the estimated pairwise distances are to be fit to an additive tree, with some of the estimates (observations) being greater than the true values and some of them being smaller than the true values. An alternative view is one in which the sequence (or other) differences are not corrected for superimposed changes and thus provide lower bounds for the actual evolutionary distance. In this framework, the length of the path connecting any pair of taxa must equal or exceed the corresponding observed distance. In analogy to character-based parsimony, the desired tree is the one that minimizes the total of all branch lengths in the tree, while using the pairwise distances as lower bounds on the path-length distances. Beyer et al. (1974) and Waterman et al. (1977) have described exact methods for accomplishing the desired minimization on a given tree. Farris's (1972) distance Wagner algorithm can be thought of as a heuristic approach to the same problem. Modifications to the distance Wagner procedure have subsequently been proposed by Swofford (1981) and Tateno et al. (1982). As with neighbor joining, if the experimentally determined distances are additive, then the optimal solution will always be found. However, when the fit is not exact, the behavior is not intuitively obvious.

RELIABILITY OF INFERRED TREES

Systematic Versus Random Error

In any statistical analysis, two kinds of error (systematic and random) need to be distinguished. We define **random error** as deviation between a parameter of a population and an estimate of that parameter, due strictly to a limited sample size used to make the estimate. By definition, random error disappears in infinite samples. In contrast, **systematic error** is deviation between a parameter of a population and an estimate of that parameter, due to incorrect assumptions in the estimation method. Systematic error persists (and may intensify) as sample sizes increase and become infinite.

Throughout this chapter, we have discussed various conditions under which systematic error arises in phylogenetic analyses. In general, systematic error occurs when the evolutionary process violates the assumptions of a phylogenetic method in a critical way. Under these conditions, a bias may be introduced into the evaluation of alternative phylogenies, favoring some branching patterns and decreasing the support for others. If the bias becomes sufficiently great, it may overcome the legitimate support for the correct tree and lead the researcher to an incorrect conclusion. Because the effect is systematic, the addition of more data will tend to solidify the incorrect conclusion (and the method is said to be *inconsistent* or *positively misleading* under these conditions; Felsenstein, 1978a). For a mistake to occur in phylogenetic estimation of the branching order as a result of systematic error, the magnitude of the bias must exceed the valid support for the correct tree. Furthermore, the bias must be in the direction of an erroneous tree, as it is possible for systematic bias to increase apparent support for the historically correct tree. Thus, the presence of a bias does not necessarily lead to wrong answers, but it does cast doubt upon the validity of the inference process.

Even if evolution occurred exactly as assumed by a particular analytical method, an incorrect tree may be inferred with finite data due to chance events (which introduce random error). For example, convergent substitutions might be expected to occur (in a given situation) only once per 100 nucleotide sites, but because of sampling error we might observe three convergent substitutions in a single sample of 100 nucleotide sites. This type of error occurs even when the presumed model is correct. By analogy, the observation of 20 consecutive "heads" in a coin-tossing experiment might lead us to conclude that the coin is two-headed, but of course this outcome has a finite probability of occurring (approximately 10^{-6}) even if the coin is fair. In inferential statistics, we generally choose a certain probability (typically 0.05) below which an outcome is improbable enough (assuming that random error accounts for the deviation) to warrant rejection of a null hypothesis.

Random error does not necessarily produce a random effect on the outcome of an analysis, however. For instance, for many methods of calculating pairwise distances, small distances and large distances are affected differently by sampling error. Under some conditions, this leads to a sample-size-dependent bias in methods that are nonetheless consistent for the model under consideration (see Hillis et al., 1994b for an example). In other words, even if a method is consistent and will lead to the correct tree if given an infinite amount of data, it nonetheless may be biased with finite data, even if its assumptions are met perfectly.

Realistically, both random and systematic error are expected in any given study. Random error occurs in any finite data set (since the expected proportions of different character patterns are real numbers), so the sensitivity of the results to the presence of random error needs to be assessed. Because systematic error is expected when the assumptions of a method are violated, the assumptions should be tested, the effects of potential sources of bias should be explored, and methods should be used to reduce the effects of systematic error in the analysis.

Systematic Error

Conditions That Lead to Systematic Error

Fortunately, the situations likely to lead to systematic error under most of the methods we have described are relatively well understood. We have discussed some of these conditions in the sections describing each of the methods; here we present a brief review for the major classes of analysis.

GENERAL ASSUMPTIONS Almost all methods assume that the characters analyzed are vertically inherited (rather than horizontally acquired). This assumption is usually met for molecular data, and so probably only rarely introduces systematic error into molecular systematic studies. The other general assumption of most methods is that characters are independent with respect to probability of change. If, for example, a change in one nucleotide position makes a change in a second position more likely, then this assumption is violated (see Wheeler and Honeycutt, 1988, and Korber et al., 1993 for examples). If methods do not explicitly account for this nonindependence, it may lead to systematic error.

PARSIMONY If the number of actual sequence changes per sequence position in a macromolecule is always small (zero or one), then parsimony will correctly reconstruct the phylogeny given enough data (Felsenstein, 1978a). As the number of changes increases, the proportion of those changes that are homoplastic (parallel, convergent, or reversed) increases. If the tree is relatively dense (i.e., branch lengths are short enough so that the expected number of changes on any one branch is small), these homoplastic changes usually will be detected as such. However, parsimony analyses do not detect multiple changes on long unbranched lineages, thereby creating the potential for bias if a mixture of long and short branches are present in an analysis (Felsenstein, 1978a).

ADDITIVE-TREE TECHNIQUES The additive-tree techniques discussed in this chapter are free of systematic error if the distance data are additive (satisfy the four-point condition) and no distance values between sister taxa are missing from the data matrix. This internal consistency of the technique places the burden of accuracy on the estimation and transformation of the distance data as opposed to the actual tree inference procedure. Specifically, the model used to correct for superimposed changes must reflect the underlying evolutionary processes. To the extent that it does not, additive-tree methods are susceptible to systematic error.

MAXIMUM LIKELIHOOD If the model of evolution used to evaluate the likelihood of given trees does not reflect the actual evolutionary processes, then maximum likelihood analyses will be subject to systematic error. In general, maximum likelihood appears to be more robust to violations of its assumptions than are additive-tree methods (Huelsenbeck, 1995b). In principle, maximum likelihood models can be made arbitrarily complex to account for particular evolutionary processes, but the cost in terms of computational time may be severe. Moreover, complex models may be more sensitive to random error than are simple models (because more parameters need to be estimated from the same amount of data).

CLUSTER ANALYSIS If the assumption of ultrametricity is satisfied and no distance values between sister taxa are missing from the data matrix, cluster analysis will be free of systematic error. However, if two lineages are not equally distant from a third, more diverged lineage (i.e., if the pairwise distances are not ultrametric), then systematic error will be introduced. As pointed out above, satisfaction of the three-point condition establishes that the distances are ultrametric. In practice, this condition is rarely satisfied by real data.

Recognizing Systematic Error

There is no foolproof method for identifying artifacts in phylogenetic trees that result from systematic error. There are, however, a few techniques that can help in evaluating the extent of systematic error, and for assessing the expected effects of identified systematic error.

TESTS OF MODEL FIT Often there are tradeoffs between model complexity (which provides consistency under a wide range of conditions) and both computational complexity and sensitivity to random error. Therefore, in using a method that assumes an explicit model of evolution, it is important to choose a model that is complex enough to explain the observed data, but not so complex as to be subject to impractically long computations or require impractically large data sets. Choosing a model, therefore, requires a test to compare the fit of one model of evolution against another for a particular data set. Furthermore, we need to know if the best model provides an adequate explanation of the observed data. Reeves (1992) and Goldman (1993a,b) have described tests for this purpose.

To compare two models of evolution, Goldman (1993a) suggested using the likelihood ratio test statistic, δ:

$$\delta = 2(\ln L_1 - \ln L_0)$$

where $\ln L_1$ is the log likelihood under the more complex (parameter-rich) model and $\ln L_0$ is the log likelihood under the simpler model. This statistic will always take on a value greater than or equal to zero because the likelihood under the complex model will always be equal to or higher than the likelihood under the simple model. To test whether the more complex model provides a significantly better explanation of the observed data, Goldman (1993a) suggested that the null distribution of the statistic δ be determined using simulation. The tree and the parameters of the model are estimated under the null hypothesis that the simpler model of evolution is correct, and this estimated tree and parameterized model are then used to simulate many replicate data sets of the same size as the original. Maximum likelihood scores are then calculated under both the simple and complex models to produce a null distribution for the test statistic δ. If δ (from the original data) is greater than 95% of scores from the simulated data, then the simpler model of evolution is rejected. Note, however, that rejection of the null hypothesis only indicates that the simpler model is inadequate to explain the observations; it does not necessarily indicate that the more complex model is adequate. The more complex model is now the null model and is subject to further testing.

Typically, one can conduct tests to see if a given parameter that can be added to a model provides a significant improvement in the optimality score. For instance, many models assume a difference in the probabilities of transitions and transversions. To test if this parameter (transition: transversion ratio) is necessary, one could test the Kimura two-parameter model against the Jukes–Cantor one-parameter model of DNA substitution (see the section on "Maximum Likelihood"). In this example, the log likelihood under the Kimura model would be ln L_1 and the log likelihood under the Jukes–Cantor model would be ln L_0.

To test the adequacy of a given model of evolution, Goldman (1993a) suggested that the log likelihood under the multinomial distribution (ln L_1) be tested against the model of interest (ln L_0). This test is very stringent, however, and under a wide variety of circumstances the model of interest will be rejected as an "adequate" explanation of the observed data. This does not mean that the model is inadequate to provide a reasonable estimate of phylogeny, but it does mean that the model fails to provide a perfect description of the underlying evolutionary processes. Since we never expect models of evolution to be correct in every detail, the test is perhaps best used to estimate how far the assumed model deviates from the underlying processes. The greater the deviation, the more attention one should pay to discovering those aspects of the evolutionary process that have not been adequately incorporated into the model.

In applying the likelihood ratio test, the number of tests being conducted needs to be considered. For example, in comparing the likelihoods of ultrametric trees (i.e., assuming a "constant clock") to trees in which a given lineage is allowed to change at a different rate, it is tempting to perform the test on the most deviant lineages (those with the greatest and/or least total length in a rooted additive tree). Alternatively, some authors have simply varied the assumed rate for each branch or subtree, one after the other. In either case, the approach amounts to multiple hypothesis testing, and lowers the significance below that of a single likelihood-ratio test with the same value of δ.

Another approach for testing model fit has been proposed by Rzhetsky and Nei (1995). They derive linear invariants that are independent of evolutionary time and phylogeny and reflect the constraints on a restricted model relative to more general time-reversible models. By testing whether the deviations of these invariants from their expected values are greater than would be expected by chance if a particular model were true, a test of whether that model is applicable to a particular data set is obtained. Goldman's (1993a) method has some theoretical advantages, but Rzhetsky and Nei's (1995) method is much more computationally feasible.

ASSESSING THE EFFECT OF A POTENTIAL BIAS In some cases, a model of evolution may be adequate for the majority of taxa, but not applicable to all taxa. For instance, if a model incorrectly assumes that the same equilibrium base frequencies exist in all lineages, then systematic error will be introduced into the analysis. The problem may be particularly severe if the differences in base composition do not follow phylogenetic lines. If base composition is affected by ecological or physiological factors, then the potential for convergence in base composition exists. For instance, Pettigrew (1994) argued that the metabolic constraints of flying bias the base composition of microchiropterans (the mostly small, echolocating bats) and the megachiropterans (flying foxes and their relatives) toward a higher AT content, and that this bias misleads phylogenetic analyses of many different mitochondrial and nuclear genes (an effect he called the "flying DNA hypothesis"). He argued that the numerous studies that support the monophyly of these two bat groups (e.g., Bennet et al., 1988; Adkins and Honeycutt, 1991; Mindell et al., 1991; Ammerman and Hillis, 1992; Bailey et al., 1992; Stanhope et al., 1992) can all be explained by this base compositional bias. Instead of bat monophyly, Pettigrew (1986, 1991a,b, 1994) has argued (primarily on the basis of neuroanatomy) that megachiropterans are more closely related to primates than to microchiropterans. Therefore, two different explanations have been presented for the apparent support from DNA sequences for bat monophyly: either the two "bat" groups are phylogenetically related, or the results are accounted for by systematic bias. Van Den Bussche et al. (1996) have tested Pettigrew's flying DNA hypothesis for the relevant data sets through simulation, and have shown that the support for bat monophyly cannot be explained on the basis of base composition bias alone. Even if Pettigrew's phylogenetic hypothesis is correct, and every substitution in the two bat lineages went to an A or a T, then the bias would still not be sufficient to explain the observed support for bat monophyly. Furthermore, analyses that are better at taking different base composition among lineages into account (such as LogDet analyses) still support bat monophyly. Therefore, the analyses show that the particular bias is not a sufficient explanation for the data. This does not mean that the data have no systematic bias, but it does mean that the hypothesized bias is not an explanation for the results in this case.

In other cases, base composition does have a demonstrable effect on phylogenetic analyses (see Rzhetsky and Nei, 1995, for a test to detect significant base compositional differences). For instance, Leipe et al. (1993) and Hasegawa and Hashimoto (1993) have suggested that early eukaryote evolution is especially difficult to analyze because of unequal base composition (e.g., the *Giardia* genome is about 70% G+C, whereas the average microsporidian genome is 35% G+C). Observed distances, transformed distances, standard parsimony, and current maximum likelihood models all support *Giardia* as the sister lineage to other eukaryotes with high bootstrap support [based on Gouy and Li's (1989) small-subunit rRNA data set]. However, phylogenetic analysis of LogDet distances shows equal support for either *Giardia* or microsporidians as the sister group to the remaining eukaryotes. Furthermore, if invariant sites are taken into account, then the support shifts strongly in favor of microsporidians as the most basal eukaryotic lineage (Waddell, 1995).

Similar tests can be conducted to examine the potential effects of any hypothesized systematic bias. For example, both Gouy and Li (1989a) and Olsen and Woese (1989) have argued that if the tree of life proposed by Lake (1988)—in which Archaea is paraphyletic or polyphyletic—were correct, a systematic bias due to "attraction" of long branches would not be sufficient to yield the trees observed by the former groups (in which Archaea is monophyletic). Gouy and Li (1989a) and Olsen and Woese (1989) interpreted these results as grounds to reject the proposal of Lake as being inconsistent with their observations, a conclusion that is contested by Lake (1990b).

SENSITIVITY TO SPECIFIC TAXA IN THE TREE If the data and tree inference technique were ideal, analyzing any two subsets of taxa would yield congruent trees (i.e., the trees would be identical after pruning taxa absent from one or both trees). In practice this is not the case. (Otherwise, finding optimal trees would be almost trivial, since constructing a tree by sequential addition of taxa would always lead directly to the globally optimal tree, regardless of the order of addition.) Both systematic and random error can distort the tree so that the inferred branching order is dependent on the taxa included. Because the total error contains both systematic and random components, variation with the sampling of taxa does not necessarily indicate an effect of systematic error, but it is suggestive. Most sources of systematic error are expected to increase with branch length; therefore, if the changes in tree topology are specific to the most diverged taxa, then there is again reason to suspect that systematic error is having a significant effect on the analysis.

Lanyon (1985) described a jackknife method that evaluates taxon stability by computing T trees, each time leaving out one taxon. By computing a strict consensus of these trees using a method that allows different subsets of taxa to be contained on each of the rival trees, the investigator can determine which relationships are consistent. Felsenstein (1988a) suggested that this method may not have the properties of a statistically valid jackknifing procedure, but it nonetheless provides a useful index of which groups are most stable to taxon selection.

CONTRIBUTION OF INDIVIDUAL TAXA TO THE OPTIMALITY CRITERION If the placement of a particular taxon is problematic (due to systematic error), removal of that taxon from the analysis will frequently make a disproportionate change in a measure of tree quality, such as the least-squares criterion in a distance tree, the estimated homoplasy of a tree derived by parsimony, or the overall likelihood ratio statistic δ. However, such measures are correlated with the number of taxa in an analysis, so one must confirm that the change in a given statistic is significantly greater than would be predicted by the removal of an average taxon (in many cases this will require a simulation study).

INFERENCES BASED ON DIFFERENT MOLECULES Phylogenetic relationships inferred from two or more different molecules should, in theory, be congruent if the molecules had the same overall history. If the inferred relationships are different, the reasons for the differences should be investigated (Bull et al., 1993b). It is important to avoid confusing differences between the optimal trees with the conclusion that the results are significantly incongruent: the former might simply be due to random errors in one or both trees, whereas the latter asserts the existence of a significant conflict. One method for deciding between these two possibilities is to fit each data set to the tree(s) derived from the other data set(s). Most modern programs allow the input of user-defined trees for evaluation under a particular optimality criterion. For example, suppose tree 1 is optimal for data set A and tree 2 is optimal for data set B. If tree 2 is nearly as good as tree 1 for data set A, and if tree 1 is nearly as good as tree 2 for data set B, then there is no real conflict, just inadequate information. This result can sometimes occur even though the two trees differ substantially in their topologies!

If a conflict cannot be explained by random error associated with finite sampling, then one of the following possible explanations should be considered: the inadvertent use of non-orthologous genes (e.g., a tree with mouse and rabbit α-globin and rat β-globin; paralogy); reticulation of lineages due to hybridization or lateral gene transfer (xenology); or the presence of significant levels of systematic error (leading to inconsistent conditions) in one or both trees.

NONPARAMETRIC APPROACHES Nonparametric tests may provide an additional source of guidance in evaluating a tree inferred from distance data (or for which pairwise distance estimates can be generated from the character data). In practice, the usefulness of these tests is dependent on the details of the tree inferred, and in many circumstances the tests may not be able to distinguish alternatives. An illustration of a case in which they might be useful is provided by the trees in Figure 32. A comparison of the paths from A and B to D yields the expectation that $d_{AD} > d_{BD}$ for all three trees. Let us assume that this trend is significantly supported by the data (for example, the trend is verified by bootstrap samplings of sequence positions). If we now consider the relationships of C to A and B, we expect that $d_{AC} > d_{BC}$ in trees 1 and 3 (an expectation that could also be true of a minor variant of tree 2), whereas $d_{BC} > d_{AC}$ is only consistent with tree 2. Again, we can examine the data directly to see if one of these inequalities is significantly supported. In particular, if we observe that $d_{BC} > d_{AC}$, then we must conclude that trees 1 and 3 are incorrect, leaving tree 2 by elimination. Yet, if tree 2 were historically correct, systematic error could have biased the tree inference procedure to group the long branches leading to C and D, leading to the incorrect choice of tree 1. The reason that it is possible to infer tree 2 from the data and yet to find certain distances significantly inconsistent with that tree lies in the particular ratios of branches and in the fact that the latter test does not need to examine the most underestimated distance (i.e., that separating C and D). In contrast, the tree inference procedures discussed would include the distance from C to D

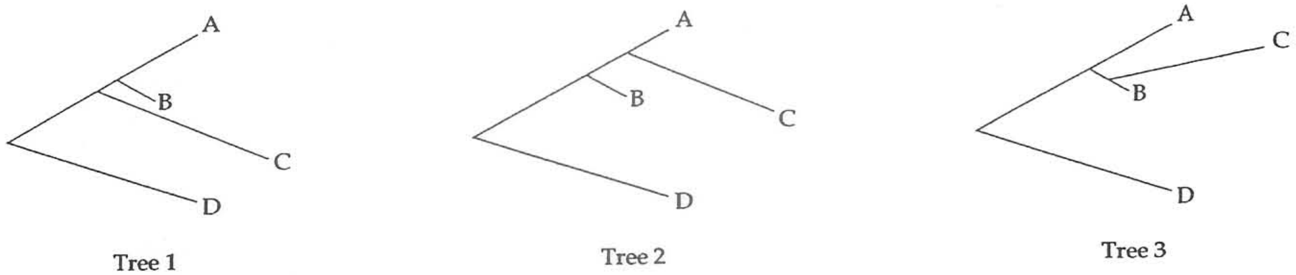


Figure 32 Three alternative trees relating four taxa that can be distinguished by a non-parametric test on the distance data. See text.

(directly or indirectly) and potentially be misled by this value.

Reducing Systematic Error

Several strategies are available to minimize systematic error and its effects on a phylogenetic analysis.

CHANGING THE ASSUMPTIONS One obvious way to reduce the chance of having systematic error lead to inconsistency is to change the assumptions of the analysis to better match the observed data (e.g., see "Tests of Model Fit," above). One example has already been given: if base composition is thought to vary significantly among taxa, then pairwise distances can be corrected using the LogDet transformation. However, the source of the systematic error may not always be so obvious, or a method may not have been devised for dealing with an identified bias. The following techniques may be useful in these cases.

REMOVING LONG BRANCHES A practical consideration in the inference of trees from pairwise distance data is that the effects of systematic error are expected to be worse with larger than with smaller distances. As noted in the discussion of the Fitch–Margoliash technique, pairwise distance methods include all measurements in the calculations as though they were independent. Therefore, having many long distances in a tree will tend to compound errors. In order to work around this problem, the use of outgroup sequences should be kept to a minimum when using a pairwise distance method. However, the substitution of different outgroup taxa, one or two at a time, can still be used to evaluate the reliability in the position of the root.

Ironically, the effect of multiple outgroups in parsimony is almost exactly the opposite. The use of multiple species of an outgroup taxon will tend to divide the longest branch in the tree, thereby decreasing its tendency to attract other long branches (Felsenstein, 1978a; Hendy and Penny, 1989; A.B. Smith, 1994). To be most effective, however, additional outgroups should be chosen so as to divide long branches reasonably evenly; adding an extremely close relative of a very distant outgroup will gain little. Of course, the benefits of adding additional taxa are not limited to the outgroup. Long branches (sparse regions) within the ingroup can also contribute to systematic error, and multiple substitutions are more easily detected in dense regions. A somewhat paradoxical phenomenon results. With large numbers of taxa, correctly inferring every aspect of the true topology is extremely difficult, but if we were interested in the relationships of only, say, four taxa, we would be much better off to compute a tree for 20 taxa (interspersed among the four of interest) and prune 16 of them from the tree than to compute the tree for only the four taxa.

ELIMINATING UNRELIABLE DATA Another practical consideration concerns the fact that a branch is long because a large number of substitutions have occurred in the sequences being compared. Limiting an analysis to those sequence regions in which positional homology is most certain tends to exclude the most variable positions in sequences, thereby shortening branches and decreasing the sensitivity of the analysis to multiple substitutions. This concept can be pushed further: if hypervariable regions can be identified in a set of sequences, then they might be eliminated from the analysis, even if their positional homology is not in doubt. This phenomenon provides one motivation for character weighting.

Subjective elimination of data is sometimes criticized as being too arbitrary (e.g., Gatesy et al., 1993). Although we share the concerns of these authors, we take the position that data are excluded from the moment one chooses a particular gene, set of genes, or gene region to use in a systematic study. Most researchers would agree that certain genes are evolving at an inappropriate rate for the level of a study, and would avoid those genes in an attempt to minimize saturation effects and other problems (see, e.g., Simon et al., 1994). It seems unreasonable to argue that just because sequence data have been obtained (perhaps even accidentally) for a region that is evolving too rapidly to be reliable in a study, we are forced to retain them at all costs. It is unrealistic to think that subjectivity in a molecular systematic study can be entirely avoided—for example, one could almost always sequence additional taxa relevant to a question, and it is a subjective decision when to stop. We believe that the benefits of excluding clearly unreliable regions—however subjectively determined—outweigh the dangers.

The above paragraph notwithstanding, we look forward to the development of methods that allow a more objective assessment of which positions in a sequence are worth retaining. One promising approach is the elision method of W.C. Wheeler et al. (1995), which attempts to identify stable versus unstable alignment regions by asking which positions align consistently over a wide range of alignment parameters.

CHARACTER WEIGHTING Obviously, all characters are not equally informative with respect to the evolutionary history of the taxa under study. Some characters are both informative and reliable; they are telling us the truth about their past. Other characters may be reliable but uninformative: although they are not actively misleading us, they are not telling us anything very useful either. The reason that phylogenetic analysis is so difficult lies in a third category of characters: those that are misinformative. These observations lead us to the rationale for character weighting. If we could somehow deduce which characters were in fact the unreliable ones, the task of reconstructing evolutionary trees would be greatly simplified, because we could minimize their influence in the analysis by giving them less weight.

Identification of unreliable characters is also an effective way to avoid systematic error. By assigning lower weight to the characters that either violate the assumptions of a method or are known to predispose the method to inconsistency, we can minimize the likelihood that systematic error will occur. For instance, parsimony methods are much more likely to be consistent if character change is low, and consequently work best if the events being minimized (i.e., homoplastic changes) are in fact the rare events. If the rapidly evolving characters are recognized as such and given little weight in the analysis, the problem of attraction of long branches due to chance convergences will be minimized. Unfortunately, beyond the use of alignment difficulty as a criterion for macromolecular sequences, methods for assessment of character reliability have received little attention.

One extreme form of weighting is the elimination of characters, as discussed above. By assigning one set of characters the maximum weight (unity) and another set of characters the minimum weight (zero), we essentially assert that there are two classes of characters, one comprising characters that, at least on an *a priori* basis, are all equally reliable, the other containing characters that are worthless for the analysis in question. If we believed that characters actually behaved in this way, we would use a method of analysis known as character compatibility (Felsenstein, 1981b), which searches for the largest "clique"—a set of mutually compatible characters that can all fit on the same evolutionary tree without homoplasy (e.g., Le Quesne, 1982; Estabrook, 1983). Compatibility methods are no longer in widespread use, probably because of their implicit adherence to an unrealistic model that asserts that once a character has been excluded from the largest clique, it no longer conveys any useful information whatsoever.

An approach that uses compatibility as an objective weighting criterion (rather than to infer phylogeny directly) was developed by Penny and Hendy (1985, 1986). Sharkey (1989), apparently unaware of the work of Penny and Hendy, described a related approach, but limited to binary characters. The strategy of these workers is to count the observed number of incompatibilities (O_j) between each character (j) and each other character. (For methods to test the pairwise compatibility of unordered multistate characters, see Estabrook and Landrum, 1975; Fitch, 1975, 1977; Sneath et al., 1975.) To convert this number to a weight, Penny and Hendy recommended computing the number of incompatibilities expected by chance (E_j) if the distribution of states for each character were independent of that for other characters (i.e., free of any non-independence imposed by their evolution on a common phylogeny). Penny and Hendy (1985) tested several weighting functions, but seem to have settled on the simple relationship

$$w_j = \max[1-(O_j / E_j), 0]$$

Thus, a character that is compatible with all other characters is assigned the maximum weight (unity), whereas a character that is incompatible with as many characters as would be expected by chance alone is assigned zero weight. More importantly, characters that fall between these two extremes are assigned intermediate weights. (Note that if the observed number of incompatibilities actually exceeded the expected number, a negative weight would be assigned unless the weights are constrained to be non-negative.) This method of weighting thus uses hierarchical structure in the data to assign weights, but does not base weights on any specific tree. Unfortunately, these methods remain relatively untested.

Another approach to character weighting is to estimate optimal weights by successive approximation (Farris, 1969). An initial set of weights (perhaps uniform weights) is used to obtain an initial estimate of the tree. From some measure of the fit of the characters to this tree, a new set of weights is derived, which are then used to estimate a second tree. The iterative rederivation of weights and recomputation of trees continues until the solution stabilizes (i.e., the tree derived from a new set of weights is identical to the tree that was used to derive those weights). Farris (1969) used reweighting functions based on the consistency index (Kluge and Farris, 1969), defined as r_j/l_j where r_j is the range of character j (defined as the minimum number of steps that the character would require on any possible tree) and l_j is the length required by the character on the tree at hand. Thus, characters that change the minimum possible number of times have perfect consistency (1.0), whereas characters that change more often have lower consistencies (approaching zero in the limit). Farris also noted that more extreme forms of weighting might be more effective than the use of the consistency index in successive weighting procedures.

One danger inherent in any successive approximations (*a posteriori*) approach is the likelihood of the search becoming trapped in a local optimum that depends on the starting tree (see also Neff, 1986). It is easy to see that a character that is inconsistent with the initial tree and downweighted as a result will have less influence in the second iteration than it did in the first. But there are some trees on which the character would have been perfectly consistent, and would therefore have been given maximum weight. Farris (1969) tested the effectiveness of his successive approximations method by adding random noise to a data set containing otherwise perfectly compatible characters and testing whether the noisy characters were in fact the ones assigned little weight in successive iterations (they were). We suggest that one not become overconfident upon seeing this kind of result, however, as characters in real data sets do not fall cleanly into "completely reliable" versus "random noise" categories. Nonetheless, recent model-based simulation studies and studies of well-supported phylogenies (J. McGuire and J. Huelsenbeck, personal communication) indicate that successive approximation approaches can be effective, although not as they are usually implemented. McGuire and Huelsenbeck observed little or no improvement in accuracy over the initial parsimony estimate when successive weighting was performed using a character's average consistency index across all of the most-parsimonious trees as the reweighting criterion. However, they found that successive weighting did increase the accuracy of the estimated phylogeny when used with more extreme forms of weighting (such as the inverse of the total number of character-state changes raised to the tenth power) and when the best observed index value for a character across all of the most-parsimonious trees is used (as suggested by Campbell and Frost, 1993).

Another problem with successive weighting approaches similar to Farris's (1969) method is that there is no objective criterion for comparing any two trees (D.R. Maddison, 1990). That is, if a tree is found to be optimal by the successive approximations algorithm, one cannot say how much worse (if at all) an alternative tree is. Goloboff (1993) has developed a method for weighting characters based on their implied homoplasy that avoids this limitation by defining a weighting function and optimality criterion that can be evaluated for any tree and compared across trees. The idea is promising, although the method needs to be more thoroughly evaluated.

Simon et al. (1994) have written an excellent review of character-weighting strategies that is both more data-oriented and more comprehensive than the discussion here; readers are urged to consult their paper for additional insights into issues concerning weighting in distance and character-based contexts.

CHARACTER-STATE WEIGHTING In character weighting, entire characters (e.g., nucleotide positions in a gene) are weighted differentially. In contrast, character-state weighting provides different weights for different character-state transformations within a character (see the section on "Generalized Parsimony"). Differential character-state weighting provides a mechanism for increasing both the consistency and the efficiency of parsimony analyses when the relative probabilities of character-state transformations differ, especially at high rates of evolution (Huelsenbeck and Hillis, 1993; Hillis et al., 1994a,b). The method works by giving greater weight to rare changes, which are less likely to be homoplastic (especially at overall high rates of character change) and hence more likely to be reflective of phylogenetic history (Williams and Fitch, 1989).

Character-state weights can be implemented by use of the step matrices described in the section on "Generalized Parsimony." Several methods have been proposed for determining appropriate weights. If we knew the actual probabilities for each type of transformation (e.g., for DNA sequence data, A → C, A → G, A → T, etc.), then an appropriate transformation would be

$$C_{i \to j} = -\ln P_{i \to j}$$

where $C_{i \to j}$ is the cost of a state change from state i to state j and $P_{i \to j}$ is the relative probability that state i will change to state j across a given branch or tree (Felsenstein, 1981c; W.C. Wheeler, 1990a). If the entire probability matrix of state changes is converted in this way into a step-matrix of change costs (including the diagonals, which represent the probability that a state will *not* change), then the most-parsimonious reconstructions of ancestral states represent maximum Bayesian probability estimates for these states (D.R. Maddison, 1990; Maddison and Maddison, 1992).

How can the relative probability matrix of state changes be estimated? If we can assume a constancy of processes across characters, then it is possible to estimate the probability matrix from the observed data. For instance, with DNA sequences, we might assume that the relative probabilities of substitutions are affected in the same way across sites by exposure to the same mutagens and repair mechanisms. Given this assumption, one way to estimate relative probabilities of change is to base the calculation on the ratio of expected to observed changes in all pairwise comparisons of the sequences, taking the relative base frequencies of each base into account (Thomas and Beckenbach, 1989; Knight and Mindell, 1993). However, the various pairwise comparisons are not evolutionarily independent, so the calculations will be biased by the underlying phylogeny. One way to account for this non-independence is to reconstruct all most-parsimonious ancestral states in an initial estimate of the tree, and then use this information to produce a change-and-stasis matrix (Maddison and Maddison, 1992). Of course, the reconstruction requires an initial tree, which (if estimated by parsimony) requires an initial matrix of change costs, so the estimate may be biased by the initial assumptions. In practice, the relative frequencies of the various changes usually are not biased greatly by the initial tree, and subsequent rounds of tree estimation can also involve reweighting of the character-state changes until a stable solution is reached (a procedure called dynamic weighting by Williams and Fitch, 1989). Alternatively, matrices of change costs can be estimated for several alternative hypotheses to examine directly the extent to which the starting tree biases the estimates of change costs.

One problem with the above approach is that the most-parsimonious reconstructions are not the only changes possible. Ideally, the relative probability matrix should be based on summed probabilities across all possible character-state histories. This can be accomplished using maximum likelihood (e.g., Z. Yang, 1994a; Z. Yang et al., 1994). But if the relative probability matrix is estimated using likelihood methods, then what is the advantage of using weighted parsimony methods over an explicit likelihood estimation procedure to estimate the tree? One of the principal advantages is one of computation time: complex maximum likelihood models typically constrain an investigator's ability to search tree space thoroughly, so only a very small portion of the potential solution space can be explored. Weighted parsimony procedures often provide a close approximation to the likelihood solutions, and the calculations are much faster. Thus, one strategy is to estimate the relative probability of change matrix using likelihood, and then use weighted parsimony to explore the solution space as thoroughly as computational limits permit. Once optimal or near-optimal solutions have been found (under the weighted parsimony criterion), they can be used as input trees and evaluated under the likelihood model. Given a fixed and finite amount of computation time, this procedure often finds better solutions under the likelihood criterion than does a direct search of tree space under the likelihood criterion, at least for moderately large data sets.

Step matrices used for character-state weighting can be either symmetric (e.g., the cost of a change from A to G will equal the cost of a change from G to A) or asymmetric (in which case the reciprocal costs will differ, with Dollo parsimony being the most extreme form). Under most circumstances, the reciprocal costs should be symmetric, so that any part of the tree can be rooted (by inclusion of an outgroup, for instance) without changing the length of the tree. If asymmetrical step matrices are used, then the various rootings of a tree will differ in tree length, so rooted trees must be examined to determine the tree length of the potential solutions. Since small asymmetries in the estimated matrix are expected from random error associated with finite sample sizes, one would not want to root the tree on the basis of this random error alone. However, if the asymmetries of change among states are strong and obvious (as with some RNA viruses; Moriyama et al., 1991), then the use of an asymmetrical step matrix may be justified (e.g., see Hillis et al., 1994a).

The assumption of constant substitutional processes operating across sites can be violated for any number of reasons, including dependence on the state of neighboring bases (Randall et al., 1987; Schaaper and Dunn, 1987), codon usage in protein-coding genes (W.-H. Li et al., 1985b), strand bias (Wu and Maeda, 1987; Thomas and Beckenbach, 1989), mutation bias (Loeb and Preston, 1986), secondary structural constraints (Gerbi, 1985; Dixon and Hillis, 1993; Tillier and Collins, 1995), and other non-phylogenetic sources of covariation among sites (Fitch and Markowitz, 1970; Korber et al., 1993). Therefore, in some situations, it may be necessary to divide the data set (e.g., into first, second, and third positions of codons) for the purpose of computing separate step matrices to provide differential weighting of state changes among the various sites in the sequence.

Random Error

The only way to avoid random error is to obtain an infinite amount of data; this practice will guarantee the correct result as long as the method is consistent. This option is unrealistic, however, so it is important to maximize the extraction of phylogenetic information by using the most efficient methods that are applicable to the available data. In any case, methods must be used to estimate the sensitivity of the results to finite sampling. Penny and Hendy (1986), Felsenstein (1988a), Li and Gouy (1991), Hillis et al. (1993a), and Li and Zharkikh (1995) have presented reviews of the many methods available. Here we present and discuss a few of the more common methods.

Testing for Hierarchical Structure

Even if a data set were constructed by randomly assigning character states to taxa, some random covariation would be expected due to the stochastic nature of the sampling process. This random covariation would lead phylogenetic reconstruction methods to prefer some trees over others even though true hierarchical structure in the data was absent. Thus, it is worthwhile to ask whether a data set contains more hierarchical structure than would be expected purely by chance.

One way to assess the non-randomness of hierarchical structure is through permutation tests, which provide a means for approximating the distribution of a test statistic under a given null hypothesis by permuting (randomizing) the observed data. In a phylogenetic context, permuted data sets are created by randomizing character states among taxa, while holding the total number of occurrences of any state constant. Thus, any correlation among character states that results from actual phylogenetic structure is destroyed. By comparing the null distribution of a test statistic generated from a series of permuted data sets with the observed value of the statistic from the original data, one can determine whether the null hypothesis of no phylogenetic structure can be rejected. If the test statistic does not lie in the extreme (say 5%) tail(s) of the null distribution, then there is a reasonably good chance that it could have arisen by chance in the absence of meaningful hierarchical structure, and further analysis of the data would seem ill-advised. It is important to remember, however, that although significant hierarchical structure may be due to phylogenetic signal, other sources of structure (such as base compositional bias, or convergence) may also lead to rejection of the null hypothesis. Statistics that have been used in permutation tests include tree length and measures of character consistency or homoplasy (Archie, 1989a,b; Faith, 1990, 1991; Faith and Cranston, 1991).

An alternative approach to permutation is the examination of shape of the distribution of tree lengths for either all possible trees or a random sample of them (Hillis, 1991; Hillis and Huelsenbeck, 1992). Fitch (1979, 1984) observed that data sets with little or no hierarchical structure tended to produce relatively symmetric tree-length frequency distributions. Hillis and Huelsenbeck (1992) showed that as the amount of hierarchical structure was increased, these distributions became more left-skewed. The degree of skewness can be quantified using the standard g_1 statistic. For n trees of length T, g_1 is calculated as

$$g_1 = \frac{\sum_{i=1}^{n}(T_i - \overline{T})^3}{ns^3}$$

where s is the standard deviation of the tree lengths (Sokal and Rohlf, 1981). Strong skewness can be misleading, however, as very localized structure can lead to highly asymmetric tree-length frequency distributions. (For example, a purely random data set can produce a highly skewed tree-length distribution if one taxon is duplicated, as trees consistent with the monophyly of the duplicated pair will be much shorter than the remaining trees.) Hillis (1991) suggested a procedure for detecting those groupings most responsible for the observed structure by calculating the g_1 statistic after successive restrictions of the sample space of trees. He used random character states (rather than permuted states from the observed matrix) to estimate the null distribution. The latter approximation is computationally much faster than permutation (in fact, it need only be calculated once for a given number of taxa and characters), but it is sensitive to deviations in the frequencies of the observed character states.

Tests for Comparing Two Trees

Many tests have been described to compare two hypothesized trees: Is tree A significantly better (under a given optimality criterion) than tree B, or are the differences within the expectations of random error? Such tests have been devised for each of the major optimality criteria.

PARSIMONY The first analytical tests for parsimony were devised by Cavender (1978, 1981), who studied the case of a four-taxon tree. Felsenstein (1985b) extended these results to include an assumption of a constant molecular clock, and Steel et al. (1993b) extended Felsenstein's test to take into account unequal nucleotide frequencies among the taxa. Li and Zharkikh (1995) noted that these tests could, in principle, be extended to more than four taxa, but that the tests are expected to have very low power. Therefore, we concentrate here on related heuristic tests that can be used with any number of taxa.

Templeton (1983b) devised a nonparametric test for comparing two trees. The test utilizes a Wilcoxan ranked sums test of the relative number of steps required by each character on each of the respective trees. If the characters are uniformly weighted and require no more than one additional change on either of the trees, then the test can be simplified into the "winning sites" test of Prager and Wilson (1988). This simple test compares the number of characters that favor each of the two trees and tests the results against a binomial distribution. Under the assumption that random noise will be equally likely to favor either of the two trees, the test asks whether the support for one hypothesis is significantly better than would be expected from random variation among the characters. Although the assumption is usually not met exactly (because the size of the relevant subgroups in the two trees is expected to differ), the effect of the violation is likely to be small and the test gives an easy approximation of the probability that the observed difference is due to random error.

Kishino and Hasegawa (1989) devised a parametric test for comparing two trees, under the assumption that all nucleotide sites are independently and identically distributed. This test uses the difference in lengths of the two trees (D) as a test statistic, where $D = \sum D_{(i\ )}$ and $D_{(i)}$ is the difference in the minimum number of nucleotide substitutions on the two trees at the ith informative site. The expectation for D (under the null hypothesis that the two trees are not significantly different) is zero, and the sample variance of D is

$$s_D^2 = \frac{n}{n-1}\sum_{i=1}^{n}\left[D_{(i)} - \frac{1}{n}\sum_{k=1}^{n} D_{(k)}\right]^2$$

where n is the number of informative sites. The null hypothesis that $D = 0$ can be tested with a paired t-test with $n - 1$ degrees of freedom, where

$$t = \frac{D/n}{s_D/\sqrt{n}}$$

If there is no *a priori* reason to suspect that tree 1 is better than tree 2, the test should be two-tailed. If there is an *a priori* reason to suspect that one tree is better than the other (for instance, if one tree is the optimal tree found in a search, and it is being compared against nearby suboptimal trees), then the expectation for D is no longer zero. For this reason, the test is strictly valid only when the two trees being compared are selected on an *a priori* basis.

DISTANCE TESTS Rzhetsky and Nei (1992a, 1993) proposed a test for comparing two trees under the minimum evolution criterion. In this test, D is the difference in the sum of the branch lengths for the two trees as estimated by the least-squares method, and the variance of D is either estimated by bootstrapping (Nei, 1991) or computed analytically (Rzhetsky and Nei, 1992a). Rzhetsky and Nei suggested a search strategy for solutions under the minimum evolution criterion by comparing the neighbor-joining approximation to all trees that differ from the neighbor-joining tree by up to four symmetric-difference distance units (d_{SD}), and accepting all trees that are not significantly worse than the neighbor-joining tree. They restricted the comparisons to trees within 4 d_{SD} because studies based on six taxa showed that it is unlikely for the optimal solution to be any more distant from the neighbor-joining tree under these conditions, at least if the number of characters examined is large. However, this search strategy is likely to miss many solutions that are equal to or better than the neighbor-joining tree if there are greater numbers of taxa. For instance, one of us (DLS) has found more than 27,000 trees that are equal to or better than the neighbor-joining tree (under the minimum evolution criterion) for the distance matrix examined by S.B. Hedges et al. (1992b; based on the data of Vigilant et al., 1991). All but 345 of these equal or better solutions are more than 4 d_{SD} from the neighbor-joining tree, and better solutions are as much as 30 d_{SD} from the neighbor-joining tree. Therefore, a neighbor-joining estimate (with a search of nearby trees) is a poor substitute for a thorough search of tree space for near-optimal solutions. If the number of taxa is very small (the conditions under which this search strategy is likely to be successful), an exact search (exhaustive or branch-and-bound) is computationally simple and will always find the optimal solutions.

An alternative approach to testing the difference between two trees is to use a measure called the generalized least-squares sum of squares, which is similar to a weighted least-squares measure but takes covariances between distances (e.g., shared branches in the tree) into account. This statistic can be compared against a χ^2 distribution (see Bulmer, 1991 for examples).

LIKELIHOOD If one tree is a subset of a second, more fully resolved tree, then the two hypotheses can be compared with a standard likelihood ratio test, using twice the difference of the log likelihoods of the two trees as a test statistic (δ). This statistic is compared against the χ^2 distribution, with the degrees of freedom equal to the difference in the number of parameters of the two hypotheses (in this case, the number of additional branches in the more fully resolved tree). Unfortunately, we would usually like to compare two trees that are not subsets of one another. In a strict sense, the likelihood ratio test is invalid under these conditions, because the number of parameters in the two hypotheses is equal, so we have zero degrees of freedom. Felsenstein (1988a) has suggested that in cases where two tree topologies differ by a single branch rearrangement, we could test one topology against the other by pretending that there was one degree of freedom and using the likelihood ratio test.

Other approaches have been used to estimate the significance of a difference in log likelihoods. One is the application of the Kishino and Hasegawa (1989) test (discussed above, under the parsimony criterion). An alternative is to generate the expected distribution of δ (rather than assuming a χ^2 distribution) through simulation of the null hypothesis (i.e., the tree with the lower likelihood). The likelihood analysis already provides the expected branch lengths given the topology of the null hypothesis, under an explicit model of character evolution. Thus, this parameterized tree can be simulated under the assumed model of evolution, and the simulated data sets can be analyzed under the maximum likelihood criterion. The expected distribution of differences in log likelihood scores (or twice the differences, if the standard test statistic is maintained) between the optimal tree and null tree can then be generated under the assumption that the null hypothesis is true. If the difference in the test statistic for the trees being compared exceeds 95% of the simulated differences, then the two trees are significantly different at $p < 0.05$, and the null hypothesis can be rejected. An example of this approach (which could be used with any optimality criterion) is presented in Chapter 12. The primary limitation to its implementation is the computation time involved, which can be considerable when the data sets are large and the optimality criterion is maximum likelihood.

Assessing the Reliability of Individual Branches

In many situations, it is desirable to assess the reliability of the individual internal branches of an estimated tree. Many methods have been suggested for this purpose. For instance, several methods have been proposed for testing whether a particular internal branch length is significantly greater than zero in an additive-distance tree (see Li and Gouy, 1991). Here we describe two nonparametric approaches that have been widely used for testing the degree of support for particular branches.

DECAY/SUPPORT INDICES AND T-PTP TESTS In parsimony, a useful index of support for a monophyletic group may be obtained by calculating the difference in tree lengths between the shortest trees that contain versus lack that group (K. Bremer, 1988). This statistic has been referred to as the *decay index* (Donoghue et al., 1992) or the *support index* (K. Bremer, 1994). A difficulty with this index is that it is not clear how large a value must be for the group to be considered well supported. Faith (1991) extended permutation approaches to test for the monophyly of a given group of taxa. His *a priori* T-PTP (topology-dependent permutation tail probability) test uses as a test statistic the difference in the lengths of the shortest trees in which a particular group is non-monophyletic and monophyletic, respectively. This statistic is equivalent to the support/decay indices described above, suggesting that it might provide a useful means of assessing their significance. The null distribution of the test statistic is determined by evaluating the corresponding length differences of trees calculated from permuted data sets. Faith's *a posteriori* T-PTP test uses the same test statistic as the *a priori* T-PTP test but uses a different method for generating the null distribution. After permutation of the data matrix, one calculates the length difference for all groups of the same size as the group of interest and picks the greatest length difference between the shortest tree in which the group of interest is non-monophyletic and the shortest tree in which the group is monophyletic. Unfortunately, these tests are sensitive to structure in the data set that is unrelated to the specific hypothesis of monophyly being evaluated (Thorne et al., 1996). Simulations of Faith's topology-dependent cladistic permutation tail probability (T-PTP) tests (Huelsenbeck et al., 1995; Thorne et al., 1996) demonstrate that it does not accurately test for monophyly of the specified group, so the question of how to assess the significance of a support/decay index remains unanswered.

NONPARAMETRIC RESAMPLING METHODS The bootstrap and the jackknife (Efron, 1982; Efron and Gong, 1983; Efron and Tibshirani, 1993) can be used to estimate the variance associated with a statistic for which the underlying sampling distribution is either unknown or difficult to derive analytically. These methods are called *resampling techniques* because they operate by estimating the variance of the sampling distribution by repeatedly resampling data from the original data set. Under certain reasonable assumptions (Efron, 1982), the variance of the statistic of interest can be approximated from the distribution of the sample estimate over replications of the resampling process. These resampling methods were first used in a phylogenetic context by Mueller and Ayala (1982), Felsenstein (1985a), and Penny and Hendy (1985).

The bootstrap and the jackknife differ in the way in which resampling is performed. In the bootstrap, data points are sampled randomly, with replacement, from the original data set until a new data set containing the original number of observations is obtained. Thus, some data points will not be included at all in a given bootstrap replication; others will be included once, and still others twice or more. For each replication, the statistic of interest is computed. The jackknife, on the other hand, resamples the original data set by dropping k data points at a time and recomputing the estimate from the remaining $n - k$ observations (see R.G. Miller, 1974). We describe bootstrapping here because it is much more commonly used in phylogenetic applications, but much of the discussion applies to jackknifing as well.

Figure 33 illustrates the bootstrapping procedure in a phylogenetic context. History (the true phylogeny) has given us one actual distribution of characters among taxa for any given data set of interest. The ideal way to examine the effects of random error would be to replay the evolutionary tape many times; this would allow us to examine sampling variance in our data directly (see Figure 33A). However, this is not possible due to the singularity of evolutionary history. Instead, bootstrapping allows us to generate a series of pseudosamples (by resampling the unique data set with replacement; Figure 33B), which we can use in place of the actual samples to estimate sampling variance. Typically, the pseudosamples are

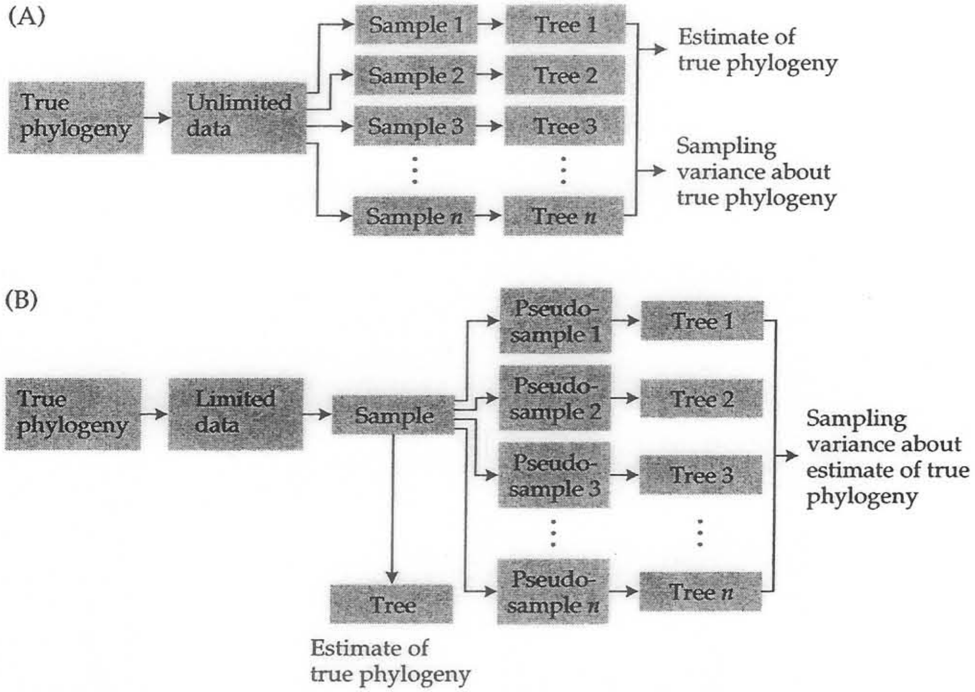


Figure 33 (A) If phylogenies were repeatable experiments, it would be possible to generate many independent samples of characters for a given gene and taxa of interest. In this case, the sampling variance about the true phylogeny could be calculated directly from estimates based on these independent samples. (B) Because phylogenies are not usually repeatable, it is not possible to draw more than one sample of characters for a given gene and taxa of interest. Therefore, bootstrapping is used to generate pseudosamples from the unique sample, and sampling variance is calculated from estimates based on these pseudosamples.

analyzed individually, and the proportion (*P*) of the pseudosamples that support a given internal branch on a tree is recorded.

How many pseudosamples must be generated to obtain a precise estimate of *P*? The sampling variance of *P* follows the binomial distribution, such that $\sigma^2 = P(1 - P)/n$, where n is the number of pseudosamples (S.B. Hedges, 1992). For instance, if we draw 100 pseudosamples, the sample variance of *P* ranges from a maximum of 0.0025 (when *P* is 50%) to a minimum of 0 (when *P* is 0 or 100%). However, this just tells us how similar the estimate of *P* is likely to be to what we would obtain if we could analyze an infinite number of pseudosamples. It does not tell us anything about the interpretation of *P*.

Felsenstein (1985a) originally suggested that *P* could be used as a measure of repeatability, or the probability that a specified internal branch would be found in an analysis of a new, independent sample of characters (assuming we could replay the evolutionary tape). More recently, Felsenstein and Kishino (1993) have suggested that *P* can be interpreted as a measure of accuracy, or the probability that the specified branch is contained in the true tree (assuming that the phylogenetic method is consistent).

Hillis and Bull (1993) examined these two interpretations of bootstrap proportions, using both simulated and known experimental phylogenies. They found that bootstrap proportions provide relatively unbiased, but highly imprecise, estimates of repeatability. They also found that bootstrap proportions provide biased estimates of accuracy (a result that was also found analytically by Zharkikh and Li, 1992a,b, for four-taxon trees both with and without a molecular clock). When the phylogenetic method is consistent, bootstrapping gives underestimates of accuracy at high bootstrap values, and overestimates of accuracy at low bootstrap values. The extent of the bias depends (at least) on the number of taxa, the number of characters, and the location of the internal branch in the tree (Hillis and Bull, 1993; Zharkikh and Li, 1995; Li and Zharkikh, 1995).

Two corrections have been proposed to recalibrate bootstrap proportions to account for this bias. Rodrigo (1993) proposed using an iterated bootstrap (Hall and Martin, 1988). This involves bootstrapping each of the pseudosamples obtained in the first round of bootstrapping, and thus is computationally very intensive. Zharkikh and Li (1995) showed that a simpler correction can be obtained with just two rounds of bootstrapping on the original sample (with one set of pseudoreplicates the same size as the original data matrix, and the other set of pseudoreplicates with reduced character matrices). The estimates from the two sets of pseudoreplicates can be combined along with a correction for sample size to produce a corrected estimate of phylogenetic accuracy. The simulations of Zharkikh and Li (1995) indicate that this complete-and-partial bootstrap technique can be effective for reducing the bias of bootstrap proportions, at least if the number of informative characters in the original data set is large (≥100).

As with other methods, for a valid test using bootstrapping the null hypothesis should be specified in advance. Otherwise, we run into a multiple-tests problem similar to the one arising in *a posteriori* comparison of means following an analysis of variance: inflation of the type I error rate above the nominal level. (The problem may be circumvented to some degree if the researcher interprets the frequency in which a group appears in replicate trees as an index of support rather than as a statistical statement, but this interpretation is far from satisfactory.) If we are interested in testing more than one internal branch or if we are unable to pre-specify the branch(es) of interest, we can adjust the significance level to allow for the fact that we are conducting more than one test (e.g., by dividing the significance level by the number of tests implied). However, if the branches of interest cannot be pre-specified, the number of potential branches is often so large that an almost hopelessly low alpha level would be required in order to maintain an overall type I error rate of, say, 0.05.

Another concern is the assumption that the sequence positions are changing independently of one another. To the extent that this is not true, the pseudosamples will be too large, and the bootstrap values will be higher than they would be otherwise. It is also important to note that the bootstrap can only assume that the data at hand are representative of the underlying distribution and thereby estimate the variation that would be obtained by sampling additional data from that distribution. If the data are not representative or if the reconstruction method makes an inconsistent estimate of the phylogeny, then bootstrapping will not remove this bias.

Bootstrapping and jackknifing can be used either with methods that operate on characters directly or with methods in which character data are first transformed into distances. In character-based methods, weighting vectors corresponding to the number of times each character is sampled can be constructed and input to the analysis. For distance methods, the resampling is conducted prior to calculation of the distance matrix; each replication is then performed using a different input matrix corresponding to the replicate sample. However, an additional source of bias exists with methods that make non-linear transformations of sequence data (including distance corrections). Under these conditions, the bootstrap will (in expectation) overestimate the variance of the corrected data (e.g., Waddell et al., 1994), which leads to conservative tests of significance .

Finally, the bootstrap replicates should be evaluated under an optimality criterion rather than just a tree-building algorithm. Otherwise, any bias of the algorithm will artificially inflate the bootstrap proportions. Imagine, for example, an algorithm that clustered taxa solely on the basis of their input order in the data matrix. Even with no data, such an algorithm would find the same tree for every pseudoreplicate. However, the resulting 100% bootstrap proportions would bear no relation to any measure of phylogenetic accuracy.

APPENDIX: PROGRAMS AND SOFTWARE PACKAGES AVAILABLE FOR CONDUCTING PHYLOGENETIC AND POPULATION GENETIC ANALYSES

Some of this information was extracted from a file compiled by J. Felsenstein and distributed as part of the PHYLIP documentation in the file **main.doc**. That file should be consulted for recent updates on availability and information about new programs.

Program/Package (author)	Operating system or source code	Applications	Availability
ABLE (J. Dopazo)	DOS	To implement a form of parametric bootstrapping in conjunction with PHYLIP	By anonymous ftp from ftp.cnb.uam.es (in directory software/molevol)
CAIC (A. Purvis and A. Rambaul)	Macintosh OS	For comparative analysis of independent contrasts, with partially or fully resolved trees	By anonymous ftp from evolve.zps.ox.ac.uk (in directory packages/CAIC)
CLADOS (K. Nixon)	DOS	Mapping characters and manipulation of trees	Contact K. Nixon, L. H. Bailey Hortorium, Cornell University, Ithaca, New York 14853 USA
CLINCH (K. Fiala and G. Estabrook)	DOS and FORTRAN source code	Compatibility analysis	By anonymous ftp from muse.bio.cornell.edu (in directory pub/software/clinch)
ClustalW (D. Higgins, J. Thompson, and T. Gibson)	Macintosh OS, DOS, C source code	Primarily for sequence alignment, but includes the neighbor-joining algorithm and bootstrapping	By anonymous ftp from ftp.embl-Heidelberg.de (in directory pub/software) or ftp.bio.indiana.edu (in directory molbio/align)
Component (R. Page)	Windows	Tree comparison and consensus methods for coevolutionary and biogeographic analyses	Contact L. Timpson at emt@nhm.ic.ac.uk or an order form is available on the World Wide Web at http://evolve.zps.ox.ac.uk/Rod/cpw.html
COMPROB (C. Meacham)	Pascal source code	To compute the probability that characters would be compatible in random data	Contact C. Meacham at meacham@violet.berkeley.edu
DNArates (G. J. Olsen)	C source code	Site-by-site maximum likelihood estimation of the rate of nucleotide substitution from a sequence alignment and a tree	From the World Wide Web at http://rdpwww.life.uiuc.edu, or by anonymous ftp from rdp.life.uiuc.edu (in directory pub/RDP/programs/fastDNAml)
Evomony (J. A. Lake)	DOS	For Lake's method of invariants (Lake, 1987a)	Contact J. A. Lake at lake@uclaue.mbi.ucla.edu
FastDNAml (G. J. Olsen)	C source code (can be compiled for parallel processing)	A faster adaptation of DNAml from PHYLIP (version 3.3) for use on workstations, mainframes, or supercomputers (including parallel machines)	From the World Wide Web at http://rdpwww.life.uiuc.edu, or by anonymous ftp from rdp.life.uiuc.edu (in directory pub/RDP/programs/fastDNAml)

(continued)

Program/Package (author)	Operating system or source code	Applications	Availability
MALIGN (W. C. Wheeler and D. Gladstein)	Macintosh OS, DOS, Unix, and C souce code	Simultaneous alignment of multiple sequences and construction of parsimony trees. Code for implementation on parallel architectures is available	Contact W. C. Wheeler (Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024-5192, USA) or source code is available by anonymous ftp from ftp.amnh.org
MARKOV (G. Pesole and C. Saccone)	FORTRAN source code	To compute distance measures and substitution matrices under a stationary Markov model of DNA substitution. Bootstrapping is included to assess the reliability of the results	Contact C. Lanave at lanave@vaxba0.ba.it
MEGA (S. Kumar, K. Tamura, and M. Nei)	DOS	Calculation of nucleotide and protein pairwise distances, and calculation of trees using the neighbor-joining and UPGMA algorithms. Also searching capabilities under the parsimony criterion using stepwise addition, local branch-swapping, or branch-and-bound algorithms. Includes bootstrapping and tests for comparing the length of two additive trees	Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802 USA (imeg@psuvm.psu.edu)
METREE	DOS	Search for trees under minimum evolution criterion; with standard errors and significance tests	Contact M. Nei (same address as MEGA)
Molevol (W. Fitch)	DOS, FORTRAN source code	A package of about 20 programs for estimating parsimony and distance trees, dynamic weighting, alignment, searching for secondary structure, and other analyses of molecular data	Contact W. Fitch at wfitch@daedalus.bio.uci.edu
MOLPHY (J. Adachi and M. Hasegawa)	C source code	A package of programs for maximum likelihood analyses with either nucleotide (NUCML) or protein (PROTML) sequences, basic statistics of nucleotide (NUCST) and protein (PROTST) sequences, and neighbor-joining analysis (NJDIST)	By anonymous ftp from sunmh.ism.ac.jp
MUST and 3S (H. Philippe)	DOS	Sequence management, analysis of taxon sampling effects, and estimation of appropriate sequence lengths for a given analysis	Contact H. Philippe at hp@bio4.bc4.u-psud.fr
NONA (P. Goloboff)	DOS	For parsimony analyses using Hennig86 data file format but with no limit on the number of taxa and characters	Contact P. Goloboff at Department of Entomology, American Museum of Natural History, Central Park West at 79th St., New York, New York 10024

(continued)

Program/Package (author)	Operating system or source code	Applications	Availability
ODEN (Y. Ina)	C source code	For distance matrix analyses on nucleotide or protein sequences	By anonymous ftp from bioslave.uio.no (in directory pub/oden)
PAML (Z. Yang)	C source code	A package mostly for maximum likelihood analyses with either nucleotide or protein sequences. Includes programs for reconstruction of ancestral sequences and conducting analyses of multiple genes (baseml, codonml) and simulating trees (mcml) under maximum likelihood. Also includes a parsimony program (pamp) for estimating substitution matrices, intersite variability of rates of evolution, and ancestral states	By anonymous ftp from ftp.bio.indiana.edu (in directory molbio/evolve)
PARBOOT (P. Roux and T. Littlejohn)	C source code	For parallel processing of bootstrapped data sets in conjunction with PHYLIP	By anonymous ftp from megasun.bch.umontreal.ca
PAUP* (D. L. Swofford)	Macintosh OS, DOS, Unix, VAX/VMS	For finding and evaluating trees under the minimum evolution, DNA maximum likelihood, and parsimony (including generalized parsimony) criteria. Includes branch swapping, branch-and-bound, and exhaustive searches. Reliability of trees may be assessed with permutation tests, decay/support indices, bootstrapping, invariant tests, or maximum likelihood scores. Includes extensive pairwise distance calculations, consensus techniques, and reconstruction of ancestral states using parsimony and likelihood methods	Sinauer Associates, Sunderland, Massachusetts 01375 USA (orders@sinauer.com)
Pee-Wee (P. Goloboff)	DOS	For parsimony analyses using character weights determined by their homoplasy during tree search	Contact P. Goloboff at Department of Entomology, American Museum of Natural History, Central Park West at 79th St., New York, New York 10024
PHYLIP (J. Felsenstein)	DOS, Windows, Macintosh OS, C source code	A package of 30 programs, including parsimony, methods of invariants, maximum likelihood (for nucleotide, protein, and restriction site data), distance methods, and compatibility analysis. Searching by stepwise addition, branch swapping, and the branch-and-bound algorithm for some methods. Includes bootstrapping, tree drawing, assessment of independent contrasts, various statistical tests of trees, and consensus analysis	By anonymous ftp from evolution.genetics.washington.edu (in directory pub/phylip) or from the World Wide Web site: (http://evolution.genetics.washington.edu/phylip.html)

(continued)

Program/Package (author)	Operating system or source code	Applications	Availability
Random Cladistics (Mark Siddall)	DOS	For conducting permutation tests, bootstrapping, or jackknifing in conjunction with Hennig86	By anonymous ftp from zoo.utoronto.ca/pub (random.doc and random.exe)
RAPDistance (J. S. Armstrong, A. J. Gibbs, R. Peakall, and G. Weiller)	DOS, Windows	For computing distance matrices in RAPD analyses	By anonymous ftp from life.anu.edu.au (in directory pub/RAPDistance)
REAP (D. McElroy, P. Moran, E. Bermingham, and I. Kornfield)	DOS	Estimation of sequence divergences, nucleotide and restriction site diversity; tests for heterogeneity of allele frequencies using randomization methods	Contact D. McElroy (mcelrdm@wkuvx1.wku.edu)
Relatedness (K. F. Goodnight)	Macintosh OS	For calculation of relatedness statistics from allele frequencies	Contact K. F. Goodnight, Department of Ecology and Evolutionary Biology, Rice University, Houston, Texas 77252 (keithg@whittaker.rice.edu)
RESTSITE (J. C. Miller)	DOS	Manipulation of restriction site data and estimation of sequence divergences; neighbor joining	Contact J. C. Miller, Whitehead Institute, 9 Cambridge Center, Cambridge, Massachusetts 02142
RSVP (K. Rice)	C source code	For calculating distance matrices and measures of variability from restriction map data	By anonymous ftp from oeb.harvard.edu (in directory rice)
The Siminator (J. Huelsenbeck)	C source code	For simulation of data under several models of nucleotide substitution for use in parametric bootstrap analyses	By anonymous ftp from onyx.si.edu or from the World Wide Web at http://mws7.biol.berkeley.edu/john/john.html
Splits (R. Wetzel and D. Huson)	Macintosh OS	For conducting split decomposition analyses	Contact D. Huson at huson@mathematik.unibielefeld.de
TreeAlign (J. Hein)	C source code	Simultaneous construction of trees (with approximate parsimony or distance methods) and alignment of multiple sequences	By anonymous ftp from ftp.bio.indiana.edu (in directory molbio/align)
TREECON (Y. van de Peer)	DOS, Windows	For distance methods with molecular data sets. Includes bootstrapping and tree drawing capabilities	By anonymous ftp from uiam3.uia.ac.be
VOSTORG (A. Zharkikh and A. Rzhetsky)	DOS	Alignment of sequences and calculation of parsimony and distance trees. Other programs available at this address (by A. Zharkikh, in the directory zharkikh/bootstrap/double-bootstrap) conduct full-and-partial bootstrap analyses	By anonymous ftp from hgc6.sph.uth.tmc.edu
WINAMOVA (L. Excoffier)	DOS, Windows	Analysis of genetic structure of populations using an analysis of variance approach	By anonymous ftp from acasun1.unige.ch (in directory pub/amova)