[34] Analysis of DNA Sequence Data: Phylogenetic Inference

By DAVID M. HILLIS, MARC W. ALLARD and MICHAEL M. MIYAMOTO

Introduction

Comparison of biological attributes among organisms or genes in a meaningful manner requires an understanding of the evolutionary connections among the respective taxa or alleles. Thus the emphasis in systematic biology is on phylogeny, namely, the evolutionary history of lineages. Methods for inferring phylogeny from DNA sequences have proliferated greatly in the last few years. Unfortunately, decisions concerning which of many described methods will be used in a given study are rarely made by weighing the advantages and disadvantages of each approach; instead, issues of availability or historical inertia often dictate such choices. In part, this is because each method is advocated in a separate paper, so comparisons among methods are often difficult. Our goal in this chapter is to present a practical guide to selecting a set of methods for phylogenetic analysis of nucleic acid sequences. We focus on the assumptions, advantages, disadvantages, and limitations of the various approaches. Space does not permit a description of each of the algorithms, but many of these are described in an excellent review paper by Swofford and Olsen.[1]

There are five basic steps in the phylogenetic analysis of DNA sequences, although some of the steps are excluded or deemphasized by some investigators. A flowchart that includes these steps is presented in Fig. 1. The sequences under study must first be aligned so that positional homologs (the units of comparison) may be analyzed. Alignment may be straightforward if pairwise differences are small and most differences result from substitutions, but it becomes increasingly difficult as the sequences become more divergent and insertion/deletion events become more common. All phylogenetic analyses assume correct alignment of positional homologs.

Once sequences have been aligned, some assessment of the presence of phylogenetic signal is necessary. If all the sequences are identical, there is obviously no point in additional analysis. At the other extreme, the sequences may be so divergent that they have been randomized with respect

[1] D. L. Swofford and G. Olsen, *in* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), p. 411. Sinauer, Sunderland, Massachusetts, 1990.

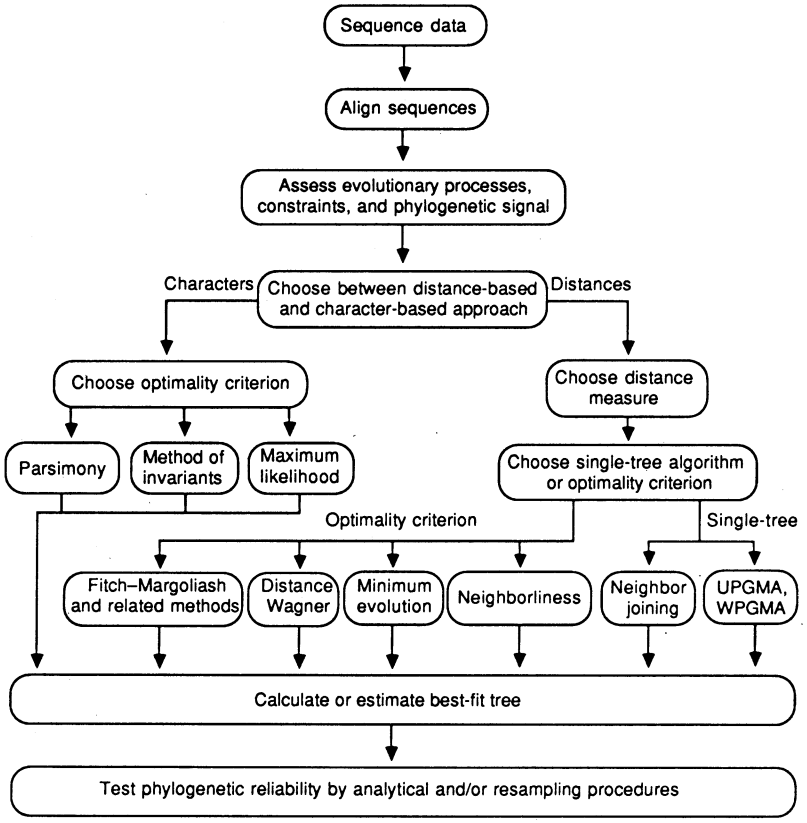

FIG. 1. Flowchart of steps from obtaining the sequence data to assessing the reliability of the final phylogenetic result.

to phylogenetic history. Analysis of the latter sequences will result in an inferred phylogeny, but the phylogenetic hypothesis might as well be selected at random. Many sequence analyses fall between these two extremes: some positions are highly conserved (perhaps invariant among the species), whereas other positions are randomized with respect to phylogenetic history (perhaps the third positions of codons). Thus, assessment of phylogenetic signal requires more than casual inspection of the sequences.

If phylogenetic signal is present in a matrix of sequences, then the third step is selecting a method of phylogenetic inference. Some of the following questions must be answered to make an informed choice among the

methods: (1) Are a few broad assumptions preferable to many detailed assumptions about evolution? (2) What parameters of sequence evolution have been examined for the sequences of interest? (3) How variable are rates of change among the study taxa? (4) Are the primary goals of the study to reconstruct accurate branch lengths, reconstruct branching relationships, examine details of character evolution, or some combination of the above? (5) Is combination or comparison of data sets (now or in the future) a goal of the study? (6) Is a particular analysis feasible given the size of the data set and the limitations of computer time?

Once a method has been selected and the appropriate software has been obtained, a strategy must be developed for finding the best tree under the selected optimality criterion. The number of distinct tree topologies (ignoring for the moment the infinite number of possible branch lengths) for even a modest number of taxa is very great.[2] For instance, with just 50 taxa, there are over 2.8×10^{74} distinct, labeled, bifurcating trees, or roughly 10,000 times as many trees as there are atoms in the universe! If one could develop a computer program capable of analyzing 1 trillion trees a second (well beyond the capability of any existing computer), it would still require 8.9×10^{54} years to evaluate all the possible trees for 50 taxa, or about 2×10^{45} times the age of the Earth. Therefore, methods must be selected that estimate the best-fit tree under these circumstances.

Finally, once a tree (or trees) has been obtained, some statement of confidence in the results is desirable. How much better is the tree obtained than the next-best alternative? How does the tree compare to a previous hypothesis of relationships? Which nodes of the tree are well-supported by the data, and which are not?

Alignment

Although alignment of DNA or RNA sequences is often quite simple among closely related taxa, it becomes very difficult as the sequences become more divergent. Alignment is one of the most troublesome aspects of phylogenetic analysis, and it is an area of intensive research and refinement. Details of the commonly used methods have been treated elsewhere in this series,[3] so our comments are limited to some of the practical aspects of producing aligned sequences for phylogenetic analysis.

Most methods of sequence alignment are designed for pairwise comparisons, although alignments among all taxa under study are necessary before phylogenetic analysis can begin. Many of the pairwise approaches

[2] J. Felsenstein, *Syst. Zool.* **27**, 27 (1978).
[3] R. F. Doolittle (ed.), this series, Vol. 183.

are variations on the algorithm described by Needleman and Wunsch,[4] in which matches are scored as positive (e.g., Ref. 1), mismatches as 0, and gaps (corresponding to insertion/deletion events) as negative. The negative scores for gaps can be weighted to account for the size of the gaps. Usually, gaps of more than one position are not weighted in direct proportion to the size of the gap, because it is likely that the adjacent nucleotides were inserted or deleted simultaneously in a single event. The penalty for gaps is typically greater than the positive score assigned to a match, but there are no clear guidelines for assigning the relative weights. One common approach is to assign a weight of -2 to gaps, so that gaps are introduced only if doing so results in the reduction of at least two substitutions.

Modifications of the Needleman–Wunsch criteria can also be used to align multiple sequences,[3] although no efficient algorithms exist to ensure optimal alignments beyond a relatively few sequences if insertion/deletion events are common or substitution rates are high. Therefore, most investigators restrict comparisons to regions in which alignments are relatively obvious. This has the effect of restricting analyses to regions that are likely to have the highest signal-to-noise ratio, because regions of difficult alignment are likely to be evolving at rates too high for effective phylogenetic analysis. Even in regions of high signal-to-noise ratio, however, alternative alignments are likely to be dependent on the weights assigned to gaps. As an example, consider the sequences of ribosomal DNA shown in Fig. 2. The upper alignment requires twenty-two substitutions (at twenty positions), and initial inspection might not indicate the necessity of introducing gaps. However, if matches are assigned a score of 1 and gaps are assigned a penalty between -1 and -3, the lower alignment is favored, which includes four gaps and seven substitutions (at six positions).

A practical method of aligning multiple sequences is to align all pairs of taxa using the Needleman–Wunsch algorithm, then enter the sequences into a text processor for multiple alignment "by hand." A less desirable alternative is to align all taxa to a single reference taxon, but this may be necessary if the number of taxa is great. Another alternative is to use sequence similarity scores to determine the order of alignments (i.e., align the most similar pairs of taxa first). All pairwise alignments can be consulted for possible arrangements, and global alternatives can be evaluated using the Needleman–Wunsch criteria. It is important to establish *a priori* rules for weighting gaps, weighting sizes of gaps, and breaking ties so as not to bias the alignments. Currently, it is not feasible to ensure that the optimum alignment has been achieved unless the number of taxa are few or gaps are uncommon. For this reason, areas of questionable alignments

[4] S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).

```
                 2200                2220                2240
Mus              GTCAGCCAGGACTCTCTACCCGCTCACGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rattus           GTCAGCCAGGACTCTCTACCCGCTCACGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Homo             GTCAGCCAGGACTCTCTACCCGCTCGCGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rhineura         GTCAGCCAGGATTCTCTATCCGCTCGCGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Cacatua          GTCAGCCAGGATTCGCTATCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Xenopus          GTCAGCCAGGATTCTCTACCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAGC
Rhyacotriton     GTCAGCCAGGATTCTCTATCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Typhlonectes     GTCAGCCAGGATTCTCTATCCGCTCGCGGCAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Latimeria        GTCAGCCAGGATTCTCTACCCGCTTGCGGCAAGGCTTCCCTGCCCGCTACCGGAGGCAGC
Cyprinella       GTCAGTCCAGGATTCCTACCCGCTGGCGGTCAAGCCTTCCCTCCGGCTACCGGAGGCAGC
                       * ** ** **   *      **    ** ** * *  **  *           *

                 2200                2220                2240
Mus              GTCAG-CCAGGACTCTCTACCCGCTCACGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rattus           GTCAG-CCAGGACTCTCTACCCGCTCACGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Homo             GTCAG-CCAGGACTCTCTACCCGCTCGCGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Rhineura         GTCAG-CCAGGATTCTCTATCCGCTCGCGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAAC
Cacatua          GTCAG-CCAGGATTCGCTATCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Xenopus          GTCAG-CCAGGATTCTCTACCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAGC
Rhyacotriton     GTCAG-CCAGGATTCTCTATCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Typhlonectes     GTCAG-CCAGGATTCTCTATCCGCTCGCGG-CAAGCCTTCCCTGCCCGCTACCGGAGGCAAC
Latimeria        GTCAG-CCAGGATTCTCTACCCGCTTGCGG-CAAGGCTTCCCTGCCCGCTACCGGAGGCAGC
Cyprinella       GTCAGTCCAGGATTC-CTACCCGCTGGCGGTCAAGCCTTCCCT-CCGGCTACCGGAGGCAGC
                       *       *  *   *     **   *    *       *              *
```

FIG. 2. Alignment of a section of the 28 S ribosomal RNA gene of 10 species of vertebrates. The upper alignment requires no gaps but does require twenty-two substitutions (at twenty positions, marked with asterisks). The lower alignment is favored if gaps are weighted between -1 and -3, which encompasses the usual range of weighting for gaps. The lower alignment requires four gaps and seven substitutions. [Adapted from "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), p. 368. Sinauer, Sunderland, Massachusetts, 1990.]

are often removed from consideration prior to phylogenetic analysis, because of the likelihood that positional homology has not been correctly established.[1]

A worthwhile but little-used approach to alignment is to combine alignment and phylogenetic analysis in an iterative process.[5] This results in an alignment and a phylogeny that are both "best fits" according to an optimality criterion such as parsimony. Unfortunately, it is computationally intensive to carry out such an analysis for even a few sequences simultaneously. An initial phylogenetic estimate may be used to break ties in multiple-alignment methods: if two alignments are equally good according to the defined weighting criteria, the alignment that requires fewer changes on the initial phylogenetic estimate is preferred. However, if this tie-breaking procedure is used, it is critically important to use an unbiased method of breaking ties in initial alignments. If this is not done, then preconceived ideas of relationships may be favored at the expense of globally optimal solutions. The danger in such iterative procedures is that

[5] D. Sankoff, C. Morel, and R. J. Cedergren, *Nature (London)* **245,** 232 (1973).

they may be highly sensitive to the initial alignments, and any bias that favors a particular alignment/phylogeny is likely to be magnified through the iterations.

Assessing Phylogenetic Signal

Many molecular systematists assume that obvious alignment of sequences is sufficient to justify the use of these sequences in phylogenetic analysis. However, sequences can be aligned with ease if a majority of the positions are invariant and the remaining positions are randomized by high mutation rates. Analysis of such sequences may produce an optimal tree with a given method, but if the variable sites are truly randomized, there is no reason to expect that the optimal tree is a good estimate of phylogeny. In addition, it is often desirable to partition regions of DNA sequences that are evolving at different rates (e.g., introns versus exons) and to identify the regions that are most likely to be informative for a given problem. Thus, some means of assessing phylogenetic signal in a given set of sequences is necessary.

Many investigators have used pairwise comparisons of the sequences to evaluate the potential phylogenetic importance of their data. For example, the transition/transversion ratios for sequence pairs can be compared to those expected for the sequences at equilibrium, given the observed base compositions (i.e., relative frequencies of A, C, G, and T).[6] DNA sequences that are largely free of homoplasy (parallel fixations and reversals) will have transition/transversion ratios greater than those for sequences that are saturated by change, but similar to those observed for closely related taxa (which remain highly structured). In a similar way, pairwise divergences have been used to assess the potential phylogenetic value of DNA sequences, by plotting percent divergence against time.[7] In such plots, regions of sequences or categories of character state transformations (e.g., transitions) that are saturated by change do not show a significant positive relationship with time. Both the transition/transversion ratio and sequence divergence are influenced by homoplasy, and, as such, both can provide insights into the potential phylogenetic value of sequence data.

Another way of detecting the presence of phylogenetic signal in a given data set is to examine the shape of the tree-length distribution that results from a parsimony analysis of all possible trees or a random subset of all

[6] A. Larson, *in* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, eds.), p. 221. Oxford Univ. Press, New York, 1991.

[7] M. M. Miyamoto and S. M. Boyle, *in* "The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis" (B. Fernholm, K. Bremer, and H. Jörnvall, eds.), p. 437. Elsevier Science Publ., Amsterdam, 1989.

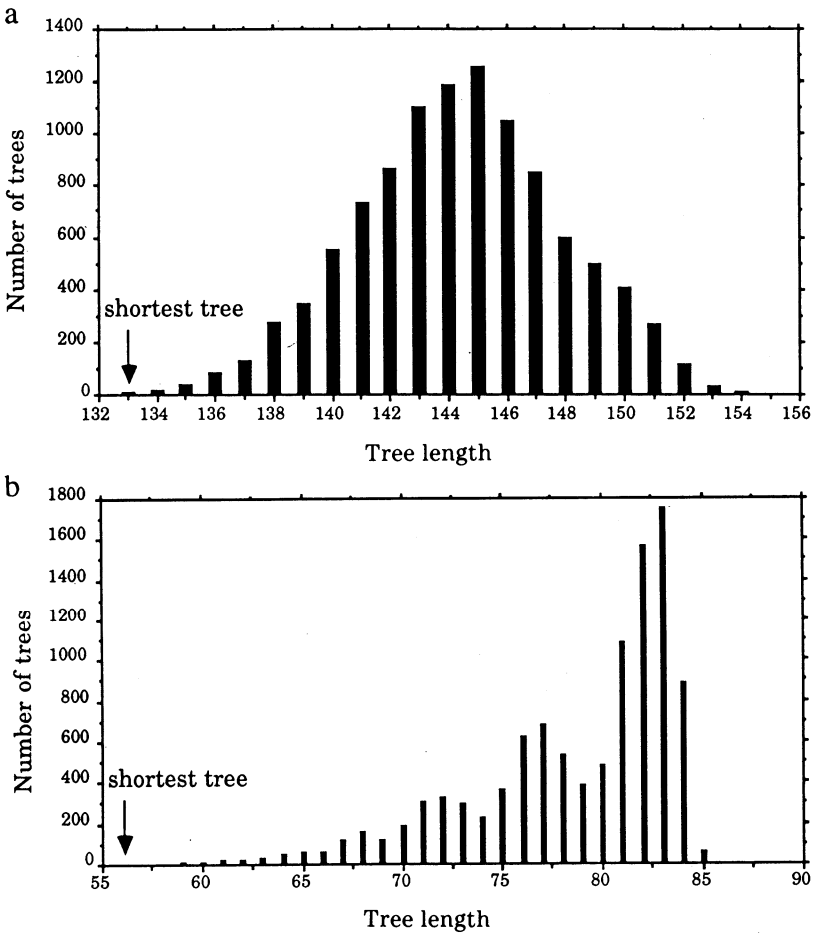

FIG. 3. (a) Nearly symmetrical tree-length distribution, based on an analysis of α-hemoglobin sequences from eight orders of mammals. Such distributions indicate that little or no phylogenetic signal is present in the data set. (b) Strongly skewed tree-length distribution, based on an analysis of a α-crystallin sequences from eight orders of mammals. This distribution indicates that the data are significantly nonrandom and, therefore, potentially informative about phylogeny. (Based on data from Refs. 8 and 9.)

trees.[8-10] Distributions that are close to symmetrical (Fig. 3a) indicate little or no structure in a data set; random sequences produce nearly symmetrical tree-length distributions. A strongly left-skewed tree-length distribution

[8] W. M. Fitch, *Syst. Zool.* **28,** 375 (1979).

(Fig. 3b) is an indication of the presence of correlated characters, which are expected if phylogenetic signal is present (the correlation is the result of the shared history of the taxa). If there is no indication that the data are more structured than random sequences, there is little point in pursuing further phylogenetic analysis of the data. Skewness is measured by the g_1 statistic, and tables of critical values of this statistic[11] for various numbers of taxa and characters should be consulted to test for the presence of nonrandom sequence variation. Skewness is calculated automatically in exhaustive and/or random-tree searches of two phylogenetic analysis software packages (PAUP and MacClade; see Implementation, below).

Figure 4 shows why tests for structured data are important, and that

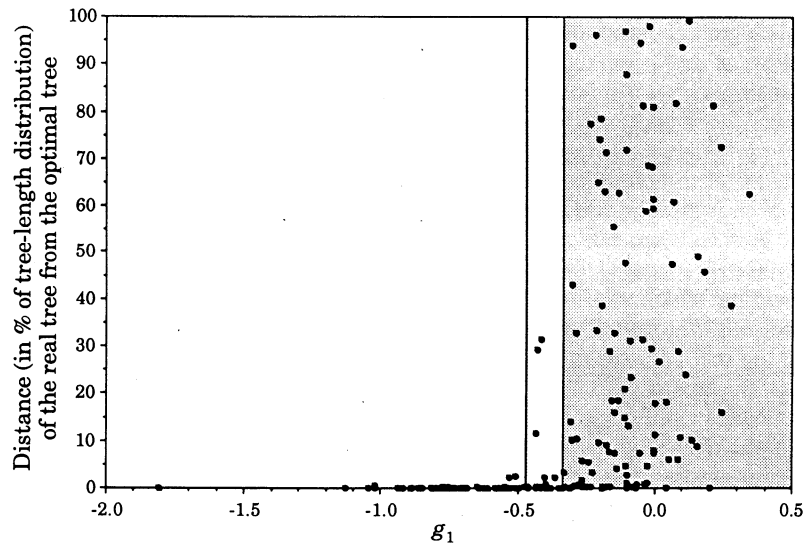

FIG. 4. Relationship between length of the correct tree and skewness of the tree-length distribution in simulated phylogenies. The optimal (most parsimonious) tree is likely to be the correct tree only in analyses of data sets that produce tree-length distributions which are significantly more skewed than expected from random data. The shaded regions correspond to the 95% (dark) and 99% (light) confidence limits for g_1 (the skewness statistic) for random sequence data. (Adapted from Ref. 12).

[9] M. Goodman, J. Czelusniak, and G. W. Moore, *Syst. Zool.* **28,** 379 (1979).
[10] D. M. Hillis, *in* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, eds.), p. 278. Oxford Univ. Press, New York, 1991.
[11] D. M. Hillis and J. P. Huelsenbeck, *J. Hered.* **83,** 189 (1992).

skewness of tree-length distributions is a useful indicator of data sets that are likely to be phylogenetically informative. In simulated phylogenies (in which the true tree is known),[12] data sets that produce significantly skewed tree-length distributions are also likely to produce the correct tree topology in phylogenetic analysis (in this case, using parsimony). However, data sets that produce distributions not significantly different from those obtained from random data (because of high mutation rates) are unlikely to yield trees that resemble the true phylogeny.

Other tests for assessing phylogenetic signal using trees involve repeatedly randomizing characters within data matrices, rather than comparing a given data set to results obtained from random sequences.[13,14] These methods are thus less sensitive to base-compositional or other biases, because the original data are randomized among taxa. However, they also require much greater computational time and are thus less suited for initial assessments of phylogenetic signal than for assessment of confidence in results (see below).

Choosing a Method of Phylogenetic Inference

Assumptions

The first aspect of choosing a method of phylogenetic inference is deciding which assumptions and models one is willing to accept. The choice is important because whenever assumptions of a model are not met by the real patterns of nucleotide substitutions, errors may be introduced into the tree construction. Models of evolutionary processes must reflect biological reality, and the extent to which they fulfill this goal will influence the phylogenetic inferences they provide. It is not always predictable which assumptions, when violated, will affect the phylogenetic estimate. Many different kinds of macromolecules exist, and models that do not take into account this huge amount of variability are bound to fail for some molecular systems. As discussed in the previous section, there is no reason to believe that all molecules or all regions of a single molecule will reflect phylogenetic history. To carefully practice phylogenetic inference one must know more about the specific molecule being examined before it is used to reconstruct phylogeny. It is best if the evolutionary models are chosen based on the molecules that are being studied.

Assumptions are directly related to the evolutionary process of nucleotide substitution. One must decide what general assumptions are acceptable for the particular molecule being examined and then choose among the models of phylogenetic inference by the assumptions incorporated in the different tree construction procedures. Most methods share certain assumptions: the characters are evolving independently; the comparisons involve orthologous genes; positional homology has been inferred correctly; and, in many cases, the nucleotide changes examined are neutral.[1,15]

[12] J. P. Huelsenbeck, *Syst. Zool.* **40,** 257 (1991).
[13] J. W. Archie, *Syst. Zool.* **38,** 239 (1989).
[14] D. P. Faith and P. S. Cranston, *Cladistics* **7,** 1 (1991).

Sequence data are naturally a character-based information source comprising four bases (e.g., A, C, G, and T for DNA) and gaps (insertions/deletions). Multiple mutation events (Fig. 5) can effectively randomize a particular nucleotide position with respect to phylogenetic history. There are 12 possible ways that bases can be substituted, and phylogenetic approaches differ in their treatment of these (Fig. 6). The various methods of phylogenetic inference differ in their assumptions of the pattern of evolutionary change. The observed mutations are analyzed using an explicit or implicit model of nucleotide substitution. Thus, whether one uses a character-based or a distance-based approach to phylogenetic reconstruction, assumptions about the evolutionary process must be made. For a general approach, the most realistic models are limited to a few assumptions, well supported by available evidence. This is important because the more assumptions made, the more likely some of them will be incorrect for the specific macromolecule being examined.

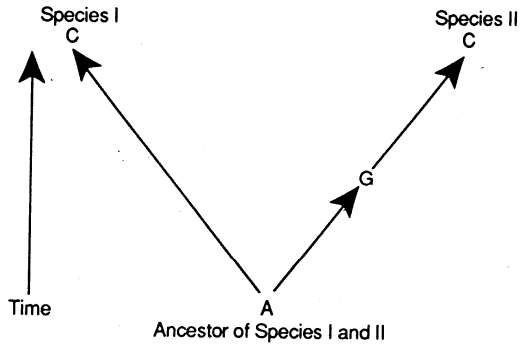

FIG. 5. Example of unobserved multiple substitution events between two distantly related species (I and II). Although three mutations have occurred since the common ancestor, species I and species II show no differences.

[15] J. Felsenstein, *Annu. Rev. Genet.* **22,** 521 (1988).

		N			
		A	C	G	T
M	A	w	a	b	c
	C	g	x	d	e
	G	h	i	y	f
	T	j	k	l	z

FIG. 6. The 12 possible pathways of substitution (base *M* to base *N*) for the four nucleotides of DNA (A, C, G, and T). Lowercase letters *(a–l)* represent the individual frequencies of each substitution. The symbols *w–z* therefore correspond to the probability of a base remaining unchanged [e.g., $w = 1 - (a + b + c)$].

Weighting

By examining the way in which nucleotide substitutions are weighted by the various phylogenetic methods, one can understand the evolutionary assumptions that must be accepted if one is to utilize any particular method. Nucleotide substitutions can be subdivided either across sites or across mutations. A natural division of nucleotide site change in protein-coding genes is based on codon structure. One can treat sites preferentially by whether the mutation causes the amino acid to change (nonsynonymous change) or remain the same (synonymous change). Further subdivision exists in the base position of the codon (first, second, or third position) and the number of codons that code for the same amino acid (redundancy, multiplicity class, or degenerate sites). The amino acid code is not universal; thus, different rules may be needed in these cases with regard to redundancy. Codon structure is only relevant to gene sequences that are translated.

The 12 possible types of nucleotide substitution can be treated differently (assuming nonsymmetry of change, e.g., the frequency of A to C does not equal that for C to A) or treated equally, or any combination of these substitutions can be grouped. One obvious division of base substitutions is to treat transitions (changes of purine to purine or pyrimidine to pyrimidine) separately from transversions (change of purine to pyrimidine or vice versa). Insertion/deletion events can also be treated as a separate type of mutation. Additionally, nucleotide substitutions can be preferentially treated by a combination of position and mutation (e.g., transversions occurring in the first and second codon positions).

There are several large classes of DNA sequences which are not translated, including those for structural RNAs [ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs)], pseudogenes, and repetitive DNAs [e.g., short and long interspersed repeated sequences (SINES and LINES)]. Ribosomal

RNA and tRNA genes are constrained by the secondary structures of their products. Their double-stranded (stem) and single-stranded (loop) regions could be treated as a major division of nucleotide site change.[16] Multigene families and repetitive DNAs are influenced by several different mechanisms including unequal crossing-over, gene conversion, and sequence transposition which could affect the types and rates of nucleotide substitution observed. At present few general guidelines are available to account for the particular structure and molecular evolution of these macromolecules.

A Priori and A Posteriori Weighting

All methods weight nucleotide change either equally or selectively, before the analysis *(a priori)* and/or after an initial phylogeny is built *(a posteriori)*. When phylogeneticists pick and choose among the available characters by weighting them, they must justify their rationale; otherwise, one is left with subjective interpretations and considerable confusion for phylogenetic inference. Although there is a great diversity in the way that one can subdivide molecular mutation (see above), it is often unclear which types of characters will consistently provide the greatest signal for phylogenetic inference. Any *a priori* selection may bias the results in favor of a preconceived notion of the evolutionary process. However, weighting of characters can also be approached by *a posteriori* methods, in which one judges the relative importance of characters by the levels of homoplasy observed (rather than expected as done *a priori*). In *a posteriori* weighting, an approximation of a phylogeny is first made, homoplasy is then measured for each character on the tree, and weighting is applied to characters based on the amounts of homoplasy observed. Possible weighting schemes (both *a priori* and *a posteriori*) are limited only by the finite ways in which macromolecules are organized, the possible mutations by which nucleotides can change, and the infinite number of relative weights that could be applied to these substitutions.

When one chooses weights for particular changes, it is usually on the basis of the assumed prevalence of the mutation, with more rapidly evolving changes given less weight and more conservative ones greater weight. One must be willing to accept the hypothesis that where more mutations are taking place there will also be greater chances for homoplasy. One assumes that within a "conservative" gene region, all mutations have occurred slowly, incorporating little homoplasy, thereby making these characters more reliable.

[16] S. A. Gerbi, *in* "Molecular Evolutionary Genetics" (R. J. MacIntyre, ed.), p. 482. Plenum, New York, 1985.

Evolutionary processes that should be closely monitored include whether one mutation is favored over another and whether one gene region evolves differently from another. This can take the form of differential mutation rates for different gene positions or mutations. To add to the complexity, observed differential rates are known for various organisms over the same positions and mutations.[17] Despite this complexity, Nei[18] has stated that "the pattern of evolutionary change is well understood," but, in our opinion, only in broad terms. If selection is invoked, this further complicates understanding the evolutionary process to the extent that Lewontin[19] has remarked that "sequences with significant but intermediate constraints on amino acid replacements are in principle unanalyzable." For phylogeneticists to have a hope of accurately reconstructing history, one must be able to make some predictions about the processes by which macromolecules evolve. Choosing a macromolecule to address a specific phylogenetic question and choosing an appropriate inference method with which to analyze the comparative sequence data are two manifestations of this problem; the latter is addressed in our next sections.

Phylogenetic Reconstruction

Deciding which evolutionary processes are affecting the molecule under study is only the first step toward the resolution of a phylogeny. Numerous alternative methods are currently available (Fig. 1), and each makes different assumptions about the molecular evolutionary process. Some methods are more general and may apply to a wider range of macromolecules and phylogenetic questions, whereas others are restricted to specific types of phylogenetic problems. To assume that any one method can solve all problems is naive, given the complexity of genomes and their evolution. The various phylogenetic methods are interconnected, and we have provided one interpretation of their linkages (Fig. 1). The assumptions, weaknesses, and strengths of each are discussed in more detail below.

Distances and Sequence Divergence

Phylogenetic analyses of sequences can be conducted by analyzing discrete characters (i.e., the nucleotides themselves) or by making pairwise comparisons of whole sequences (the distance approach). Deciding whether to use a distance-based or a character-based method depends on

[17] R. J. Britten, *Science* **321,** 1393 (1986).
[18] M. Nei, *in* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, eds.), p. 90. Oxford Univ. Press, New York, 1991.
[19] R. C. Lewontin, *Mol. Biol. Evol.* **6,** 15 (1989).

the assumptions one is willing to accept and the goals of the study. If one chooses a distance method for phylogenetic inference, then an assumption that a single coefficient of sequence similarity or dissimilarity provides an accurate measure of evolutionary divergence has been accepted. Distance-based approaches may incorporate the various types of change in estimating a single divergence value, leading to a matrix of all pairwise comparisons of the taxa studied. In one sense, the transformation of nucleotide sequence variation to a distance value reduces the available information. However, others have noted that distance approaches may use more of the available information than some character methods such as parsimony procedures which rely only on "phylogenetically informative positions" (see Table III) and ignore variation unique to single taxa.[20] An estimate of nucleotide divergence often uses a model of substitution to "correct" for (unobserved) multiple substitution events occurring between the more divergent pairs of taxa (Fig. 5). Weighting of substitutions is usually done *a priori.* One can sort the various estimates of divergence (Table I)[21] by the number of parameters incorporated in the algorithms to calculate these divergence values. The more complicated models attempt to use numerous parameters of substitution in their calculations.

Both the type and the position of a mutation have been incorporated into the parameters of divergence values. For example, a one-parameter model treats all nucleotide substitutions as equal, whereas a two-parameter model subdivides nucleotide change into transitions and transversions. Three-, four-, six-, and twelve-parameter models have been proposed[22,23] although one could envision seven- to eleven-parameter models depending on which classes of nucleotide change are grouped. The simplest estimates of nucleotide difference count up the total number of substitutions (sometimes including gaps) and divide by the number of base pairs examined, making no attempt to "correct" the distance value.[24,25] For closely related taxa (when substitution events are relatively low, sequence differences $<10\%$) different estimates of uncorrected and corrected divergence have been shown to give similar values. As distance increases, so does the underestimation of divergence by many methods. When divergence values are very large between taxa, all estimates become suspect.[26] Proponents of

[20] D. Penny, M. D. Hendy, and M. A. Steel, *in* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, eds.), p. 155. Oxford Univ. Press, New York, 1991.
[21] T. Gojobori, E. N. Moriyama, and M. Kimura, this series, Vol. 183, p. 531.
[22] C. Lanave, G. Preparata, C. Saccone, and G. Serio, *J. Mol. Evol.* **20,** 86 (1984).
[23] C. Saccone, C. Lanave, G. Pesole, and G. Preparata, this series, Vol. 183, p. 570.
[24] M. M. Miyamoto, J. L. Slightom, and M. Goodman, *Science* **238,** 369 (1987).
[25] M. Nei, "Molecular Evolutionary Genetics." Columbia Univ. Press, New York, 1987.
[26] T. Gojobori, K. Ishii, and M. Nei, *J. Mol. Evol.* **18,** 414 (1982).

TABLE I
ESTIMATES OF DIVERGENCE[a]

Method	Comments	Refs.[b]
Nucleotide-based methods		
Uncorrected approaches	Only account for observed differences	
p (difference)	Accounts for observed substitutions and/or gaps	*1, 2*
Corrected approaches	Attempt to account for unobserved parallel substitutions and reversals in addition to observed differences. For *d* (divergence), corrections are added for multiple substitution events by adopting some distribution of nucleotide change (e.g., Poisson distribution) and by weighting substitutions	*2*
One-parameter	All nucleotide substitutions treated as equal	*3*
Two-parameter	Transitions treated differently than transversions	*4*
Three-parameter	Two classes of transversion with all transitions treated equally	*5*
Four-parameter	Two classes of transversion and two classes of transition	*6, 7*
Six-parameter	Four classes of transversion and two classes of transition	*8, 9*
Codon-based methods		
Unweighted pathway	Synonymous and nonsynonymous changes, with each further subdivided into three categories of nucleotide substitution	*10*
Miyata and Yasunaga's weighted pathway	Pathways from one amino acid to another are weighted by biochemical similarity of amino acid replacement	*11*
Nei and Gojobori's unweighted pathway	Takes into account codon position (first, second, or third) and whether change is synonymous or nonsynonymous (unweighted version of Miyata and Yasunaga's method)	*2, 12*
Li, Wu, and Luo's weighted pathway	Changes weighted as nondegenerate, 2-fold degenerate, or 4-fold degenerate sites, and by expected to observed frequencies of base pair mutations	*13*
Four-parameter	Multiplicity classes for 2-, 3-, 4-, and 6-fold degenerate codon groups are used to estimate number of synonymous substitutions per codon. Several options are available for representing constraints on amino acid replacements	*14*

[a] Formulas for each estimate can be obtained from Gojobori *et al.*[21]

[b] *Key to references: (1)* D. L. Swofford and G. Olsen, *in* "Molecular Systematics" (D. M. Hillis and C. Mortiz, eds.), p. 411. Sinauer, Sunderland, Massachusetts, 1990 (see p. 428 for generalized formulas); *(2)* M. Nei, *in* "Molecular Evolutionary Genetics" (M. Nei, ed.), p. 64. Columbia Univ. Press, New York, 1987; *(3)* T. H. Jukes and C. R. Cantor, *in* "Mammalian Protein Metabolism III" (H. N. Munroe, ed.), p. 21. Academic Press, New York, 1969; *(4)* M. Kimura, *J. Mol. Evol.* **16,** 111 (1980); *(5)* M. Kimura, *Proc. Natl. Acad. Sci. U.S.A.* **78,** 454 (1981); *(6)* F. Tajima and M. Nei, *Mol. Biol. Evol.* **1,** 269 (1984); *(7)* N. Takahata and M. Kimura, *Genetics* **98,** 641 (1981); *(8)* T. Gojobori, K. Ishii, and M. Nei, *J. Mol. Evol.* **18,** 414 (1982); *(9)* M. Hasegawa, H. Kishino, and T. Yano, *J. Mol. Evol.* **22,** 160 (1985); *(10)* F. Perler, A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson, *Cell (Cambridge, Mass.)* **20,** 555 (1980); *(11)* T. Miyata and T. Yasunaga, *J. Mol. Evol.* **16,** 23 (1980); *(12)* M. Nei and T. Gojobori, *Mol. Biol. Evol.* **3,** 418 (1986); *(13)* W.-H. Li, C.-I. Wu, and C.-C. Luo, *Mol. Biol. Evol.* **2,** 150 (1985); *(14)* R. C. Lewontin, *Mol. Biol. Evol.* **6,** 15 (1989).

distance encourage the use of "appropriate distance measures." Divergence measures that are often regarded as "inappropriate" include approaches which add gaps to the calculation (owing to the difficulties in alignment and modeling of these mutations), equally weighted procedures (except when used for closely related taxa) which generally do not "correct" for multiple substitution, and methods that attempt to take selection factors into account. The simplest of the "appropriate" divergence measures is the two-parameter method of Kimura.[27] This model appears to estimate divergence as well as the more complicated algorithms, over a broad range of divergences, and without the need for additional specifics about the evolutionary process.

Clustering Algorithms for Distance Data

Methods for clustering distance data can be broken down into those that rely on clocklike mutation rates and those that are less sensitive to this assumption. Two commonly used algorithms that rely on clocklike behavior are the unweighted and weighted pair group methods using arithmetic means (UPGMA and WPGMA, respectively). These algorithms assume that the data are ultrametric (i.e., clocklike), a property that is often not satisfied by sequence data. Owing to this assumption, UPGMA and WPGMA have largely been replaced by alternative methods that do not rely on this assumption (Table II). The neighbor joining method does not depend on ultrametric data, although it may rely on the assumption of additivity[1,18] (i.e., the evolutionary distance between any two taxa is equal to the sum of the branches that join them). Nonreliance on ultrametric data is a desirable quality for any algorithm as many molecules are not clocklike, even among closely related taxa. The neighbor joining method, because of its ability to handle unequal rates, its connection to minimum-length trees (see below), and its ease of calculation with regard to both topology and branch lengths, has become a popular approach for analyzing sequence distances.

Multiple-tree methods (Table II) rely on a defined criterion of optimality, unlike the single-tree algorithms above which give an answer (usually a single tree topology), but do not select it according to some objective measure of fit or provide a method for ranking alternatives. Optimality is defined as an objective quantity, measuring the conformity of the original data (distance or character) to a tree. Each algorithm is designed to find the best tree given its optimality criterion and different searching methods are employed for this purpose.[1] If one does not agree with the criterion to be optimized, then neither the approach nor the searching method for identi-

[27] M. Kimura, *J. Mol. Evol.* **16,** 111 (1980).

TABLE II
CLUSTERING ALGORITHMS USING DISTANCE DATA

Method	Comments	Refs.[a]
Single-tree algorithms	Methods that provide single topology by following specific series of steps. No specific optimality criterion is used to select tree	
UPGMA and WPGMA	Groups taxa in order of decreasing similarity (or increasing dissimilarity); assumes constant rate of evolution	*1*
Neighbor joining method	Heuristic approach for estimating minimum evolution phylogeny (see below)	*2*
Multiple-tree algorithms	Methods that use optimality criterion to compare alternative topologies and to select final tree	
Fitch–Margoliash method	Best tree chosen as that which maximizes fit of observed (original) versus tree-derived (patristic) distances, as measured by % standard deviation statistic. Branch lengths are determined by linear algebraic calculations of observed distances among three different taxa interconnected by common node	*3*
Distance Wagner procedure	Assumes that patristic distances must be greater than or equal to observed distances (e.g., negative branch lengths are not permitted). Best tree chosen is that with shortest overall tree length	*4*
Neighborliness	Maximizes the four point condition for an additive tree (see text); different quartets of taxa are examined one at a time when five or more taxa are represented	*5, 6*
Minimum evolution	Computes sum of all branch lengths for each tree, considering all possible topologies, and chooses phylogeny which minimizes total overall length. A Fitch–Margoliash approach is used to calculate branch lengths	*7, 8*

[a] *Key to references: (1)* P. H. A. Sneath and R. R. Sokal, "Numerical Taxonomy." Freeman, San Francisco, 1973; *(2)* N. Saitou and M. Nei, *Mol. Biol. Evol.* **4,** 406 (1987); *(3)* W. M. Fitch and E. Margoliash, *Science* **155,** 279 (1967); *(4)* J. S. Farris, *Am. Nat.* **106,** 645 (1972); *(5)* S. Sattath and A. Tversky, *Psychometrika* **42,** 319 (1977); *(6)* W. M. Fitch, *J. Mol. Evol.* **18,** 30 (1981); *(7)* L. L. Cavalli-Sforza and A. W. F. Edwards, *Am. J. Hum. Genet.* **19,** 233 (1967); *(8)* N. Saitou and T. Imanishi, *Mol. Biol. Evol.* **6,** 514 (1989).

fying the optimal tree will satisfy its opponents. In turn, the algorithms used to calculate optimality and to search for optimal trees limit the ability of the investigator to satisfy the original criterion (see below). Quantitative phylogeneticists are continually upgrading their software to improve the speed and accuracy of finding the optimal solution.

Several criteria of optimality are used for building distance trees (Table II). Each approach permits unequal rates and assumes additivity. The Fitch–Margoliash method minimizes the deviation between the observed pairwise distances and the path length distances for all pairs of taxa on a tree. This fit is measured by percent standard deviation (one calculation of

the least-squares method). For the distance Wagner method, the observed distance values impose a minimum bound on the branch lengths, thereby ensuring that the tree-derived distances are always greater than or equal to the original ones. The selected topology becomes the one of minimum length, where length is determined as the sum of lengths over all branches of the tree. The neighborliness method compares the distances *(d)* of the three possible groupings of four taxa (A, B, C, D). Under the assumptions of the four-point condition, two relationships must be satisfied for A–B and C–D to be clustered: (i) $d(A, B) + d(C, D) < d(A, C) + d(B, D)$; and (ii) $d(A, B) + d(C, D) < d(A, D) + d(B, C)$. For larger phylogenetic problems, four-taxon comparisons are conducted for all possible subclusters, and paired taxa are clustered by their arithmetic means. Minimum evolution is the exhaustive implementation of the neighbor joining method (a single-tree heuristic approach). Branch lengths are optimized using the Fitch–Margoliash method and are added to determine the overall length of the tree. The tree with the minimal overall length is then selected.

Character-Based Approaches

Rather than reducing all of the individual variation to a single divergence value, character-based methods treat each substitution separately. By counting each mutation event, one determines the relationships among organisms by the distribution of mutations observed (Table III). These methods are preferred for studying character evolution, for combining

TABLE III
CHARACTER-BASED METHODS OF TREE CONSTRUCTION

Method	Comments	Refs.[b]
Parsimony	Selects phylogeny that minimizes number of evolutionary changes for data set. Approach relies on phylogenetically informative characters (i.e., those with two or more states shared by two or more taxa)	*1*
Maximum likelihood	Calculates probability of data set, given particular model of evolutionary change and specific topology	*2*
Method of invariants[a]	Counts number of transversion events supporting phylogeny after adjusting for homoplastic change. Designed for four-taxon problems where homoplasy is expected to be abundant (e.g., for distantly related taxa with unequal rates of evolution)	*3, 4*

[a] Commonly known as evolutionary parsimony.

[b] *Key to references: (1)* See review by D. L. Swofford and G. Olsen, *in* "Molecular Systematics" (D. M. Hillis and C. Mortiz, eds.), p. 411. Sinauer, Sunderland, Massachusetts, 1990; *(2)* J. Felsenstein, *J. Mol. Evol.* **17,** 368 (1981); *(3)* J. A. Lake, *Mol. Biol. Evol.* **4,** 167 (1987); *(4)* R. Holmquist, M. M. Miyamoto, and M. Goodman, *Mol. Biol. Evol.* **5,** 217 (1988).

multiple data sets or sequentially adding data, and for inferring ancestral genotypes. All sequence information is retained through the analyses; no information is lost in the conversion to distances. Therefore, character-based methods are often preferred when they are feasible. The chief disadvantages are the greater computational time they require and the greater difficulty of correcting for multiple substitutions.

Parsimony

Parsimony is the principle of logic that simple explanations should be preferred over more complex explanations. In the context of phylogenetic inference, the most parsimonious tree is the tree that requires the fewest evolutionary changes to explain the data. Parsimony remains the most popular character-based approach for sequence data. This popularity is due to its logical simplicity, its ease of interpretation, its prediction of both ancestral character states and amount of change along branches, the availability of efficient and powerful programs for its implementation, and its flexibility in terms of maximizing weighting strategies and conducting character analyses. Parsimony procedures search for the phylogeny that minimizes the number of evolutionary events required to explain the original data. Parsimony, which permits unequal rates, assumes that homoplasy occurs at levels that do not interfere with phylogenetic inference. When more than one of the taxa in a study is connected to the tree by an excessively long branch and rates of mutation are relatively high, parsimony procedures can be expected to converge onto the wrong tree even as more data are added.[28] However, inconsistency under these conditions is a property of many tree-building procedures. The excessive homoplasy may be avoided by assigning greater weight to the more conservative sites or gene regions (e.g., functional domains) and/or by giving more weight to the slower types of nucleotide change (e.g., transversions). With parsimony procedures, a wide range of weighting schemes is possible (Table IV).

Weighting strategies must be considered for parsimony analyses as they are for all phylogenetic methods. Weighting is practiced even when weights are not specified, in that all changes are uniformly counted. Thus, no attempt to weight nonetheless carries an assumption about the evolutionary process. Current parsimony programs are well designed to implement sophisticated weighting schemes, and, as such, weights based on at least general patterns of molecular evolution are encouraged (e.g., first and second codon positions versus third), as long as they are explicitly presented and defended.

[28] J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).

TABLE IV
METHODS OF CHARACTER WEIGHTING

Weighting method	Comments[a]	Refs.[b]
Uniform weighting	All characters and changes are given equal weight	
Nonuniform weighting	Selective weighting of particular characters and/or changes	
Across positions	Emphasizes structural/functional differences between gene regions or base positions	
Codon positions	Selective weighting of first, second, and third codon positions in translated genes, because of redundancy of genetic code. A general rule is that third-codon positions are under less selective constraint than first and second and, as such, are more likely to change than the latter	
Stems and loops	Selective weighting of double-stranded (stem) versus single-stranded (loop) regions of structural RNAs (tRNA and rRNA), reflecting constraints on stem regions to maintain secondary structure through base pairing	*1*
Within positions	Emphasizes mutational bias	
Transitions versus transversions	Weighting of transition bias which is most evident in vertebrate mitochondrial DNA but apparent in other systems as well. The general rule is that transitions occur more frequently than transversions and, as such, deserve less weight than the latter	
Relative substitution frequencies	The 12 possible substitutions (Fig. 6) are weighted differently according to relative frequencies. Different combinations of the 12 substitutional types can be recognized, with transition/transversion categorizations representing one extreme (see above)	
Weighting by base composition	Weighting schemes based on either observed or expected base compositions of sequences being examined. This approach assumes that base frequencies reflect substitutional frequencies	
Synonymous versus nonsynonymous change	Unlike synonymous mutations, nonsynonymous changes alter primary sequence of a polypeptide and, as such, are under greater selective constraint and occur less frequently. Nonsynonymous mutations therefore warrant greater weight	
Within and across positions	Refers to weighting for both positional effects and mutational bias. A large number of combinations are possible	
Successive approximations	*A posteriori* weighting of characters according to levels of homoplasy, as judged with an initial estimate of topology. Subsequent reiterations of weighting and tree construction are performed until topology stabilizes. Dynamic weighting uses successive approximations approach to weight sequence data both across and within positions	*2–4*

[a] See Swofford and Olsen[1] for review.

[b] *Key to references: (1)* M. J. Dixon and D. M. Hillis, *Mol. Biol. Evol.* **10**, 256 (1993); *(2)* J. S. Farris, *Syst. Zool.* **18**, 374 (1969); *(3)* D. Sankoff and R. J. Cedergren, *in* "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparisons" (D. Sankoff and J. B. Kruskal, eds.), p. 253. Addison-Wesley, London, 1983; *(4)* P. L. Williams and W. M. Fitch, *in* "The Hierarchy of Life: Molecules and Morphology in Phylogenetic Analysis" (B. Fernholm, K. Bremer, and H. Jörnvall, eds.), p. 453. Elsevier Science Publ., Amsterdam, 1989.

Maximum-Likelihood Methods

Statistical models have been developed for character-based nucleotide change. By considering each site separately, one determines the likelihood of these changes in the data, given a particular topology and model of molecular evolution. The maximum likelihood method therefore depends heavily on the model chosen and on how well it reflects the evolutionary properties of the macromolecule being studied. Because of questions about the accuracy of the models, coupled with the computational complexities of the approach, maximum-likelihood methods have not received the attention that they probably deserve. The more recent versions of maximum likelihood rely on models of evolution that are quite sophisticated, taking into account the possibility of unequal rates of change among lineages, site-specific rate variability, and the random distribution and/or clustering of variable sites. However, it remains unclear whether these more complex models will be better at phylogenetic inference than the more general ones since the former demand specific insights about the evolutionary process (information which is not typically available).

Methods of Invariants

Phylogenetic inference methods have been developed for selecting the correct topology when large amounts of homoplasy exist, as when rate heterogeneities occur among distantly related branches. By relying on a few specific patterns of nucleotide variation that represent the most conservative changes, one can avoid the abundant homoplasy while recognizing signal. In evolutionary parsimony, quantities called "operator invariants" are calculated for the three possible topologies of four taxa. Each invariant reflects specific patterns of shared transversions corrected for homoplastic similarity. These calculations are based on the variable positions with two purines and two pyrimidines. Zero-value invariants represent cases in which random multiple mutation events have canceled each other out, and, as such, a chi-squared (χ^2) or binomial test is used to identify the correct topology as the one with an invariant significantly greater than zero. Evolutionary parsimony assumes that the transversion rates between the two types of transversions for each given base are equal (e.g., the frequency of A to C is the same as A to T). A recent modification of the procedure has been proposed which corrects such inequalities by taking into account base compositional differences.[29]

The method of invariants recognizes that there are 36 patterns of transitions/transversions (called spectral components) for four taxa. Different methods of phylogenetic inference rely on various combinations of these components, with 12 used to calculate the three operator invariants of evolutionary parsimony. Unlike parsimony, the method of invariants does not construct intermediate ancestors, and it is limited to direct comparisons of only four taxa at any one time. When more than four taxa are considered, all possible quartets are typically analyzed and a composite tree constructed from the individual results. Relatively long sequences are needed by the procedure to obtain enough transversions for its statistical tests.

[29] A. Sidow and A. C. Wilson, *J. Mol. Evol.* **31**, 51 (1990).

Searching for Optimal Trees

Once an optimality criterion has been selected, it is necessary to calculate or estimate the best tree for the given criterion. For relatively few taxa (up to as many as 20 or 30, depending on the level of homoplasy present in the data), it is possible to use exact algorithms that will be certain to find the optimal tree. For greater numbers of taxa, one must rely on heuristic algorithms (i.e., useful and efficient algorithms that approximate the exact solutions but may not give the optimal solution under all conditions). When heuristic algorithms are used, it is always a possibility that a better solution exists. Therefore, the use of heuristic algorithms should be described exactly so that alternative procedures can be explored by other workers who wish to search for better solutions.

Variations of two exact algorithms (algorithms that will always find the optimal solution) are commonly used. The first is to search exhaustively through all possible tree topologies for the best solution(s). This method is computationally simple for 9 or fewer taxa (for which there are $\leq$135,135 labeled, unrooted, bifurcating trees) and is only moderately time-consuming for 10 or 11 taxa (2,027,025 and 34,459,425 trees, respectively).[2] For 12 taxa, the evaluation becomes laborious (654,729,075 trees), and for 13 or more taxa ($\geq$13,749,310,575 trees) the calculations are usually impractical. The chief advantages of exhaustive searches are (1) the optimal tree(s) is always found and (2) all other possibilities can be ranked with respect to the optimal solution(s).

If an exhaustive search is impractical for a given data set, another exact algorithm can be used that is generally much faster, namely, the branch-and-bound algorithm.[30] Most implementations of this algorithm calculate an initial upper bound for a tree (using one of the heuristic methods described below) and then search exhaustively along paths that lead to all possible trees by sequentially adding taxa. If the upper bound is reached

[30] M. D. Hendy and D. Penny, *Math. Biosci.* **59**, 277 (1982).

before all the taxa have been added to the tree, then no more trees along that search path need be examined, because any further addition of taxa could only increase the length of the tree. If all taxa are added and the upper bound has still not been reached, then a shorter tree has been found. The upper bound is reset to this new score, and the search is continued. In this way, the shortest tree can always be found, even though many trees (all of which must be longer than the upper bound) are never examined by the algorithm.

Although the branch-and-bound algorithm will always find the shortest tree, it cannot rank suboptimal solutions if implemented as above (its usual form). However, the best implementations of this algorithm (e.g., as in the Phylogenetic Analysis Using Parsimony or PAUP program[31]) allow an investigator to save all trees that are shorter than or equal to a specified bound. In this way, it is possible to look at the lower end of a tree-length distribution, even for relatively large numbers of taxa, and thereby rank all alternatives near the optimal solution(s).

If the exact algorithms described above are not feasible for a given data set (the limitation is usually number of taxa), then various heuristic approaches can be tried. The heuristics used should be described in sufficient detail that they can be replicated, and so that alternative searches can be attempted. It is also worthwhile to discuss the number of alternative solutions examined, to give a sense of the thoroughness of the search.

Most heuristic techniques start by finding a reasonably good estimate of the optimal tree(s) and then attempting to find a better solution by examining structurally related trees. The initial tree is usually found by a stepwise addition algorithm.[1,32] These algorithms add taxa sequentially to a tree, in each step adding the new taxon at the optimal place in the growing tree. Once a taxon is added to the tree, however, the tree is constrained for the next round of addition. Therefore, it is likely that the solution that is optimal when only a few taxa are joined together will not be globally optimal for these taxa when the tree is complete (hence, the "inexact" nature of stepwise addition algorithms). The various stepwise addition algorithms differ primarily in the order in which taxa are added to the tree. The simplest (and usually least efficient) algorithms simply add taxa in the order in which they appear in the matrix. Other implementations base the addition of sequences on their distance to a reference taxon or on the number of steps they add to the growing tree.[1]

After an initial tree has been obtained (either by stepwise addition or user input), it can often be improved by examining related topologies by a family of procedures known as branch swapping. Several alternatives are commonly implemented and are described by Swofford and Olsen.[1] All involve rearranging branches of the initial tree to search for a shorter alternative (or one equal to or shorter than a specified limit). Because any of these methods may be the most efficient under certain circumstances, it is often necessary to try as many options as are available to be reasonably sure of finding the optimal solution.

Another strategy that may be used for large data sets is to reduce the number of possible topologies by constraining the analyses to look at a subset of trees with exact algorithms. This approach is useful if a study is designed to address specific questions. For instance, assume the relationships among 10 families of angiosperms are in debate, but no one questions the monophyly of each of the 10 families. Also assume that orthologous sequences are available for three species of each of the 10 families. Under such circumstances, it may be desirable to conduct at least one analysis in which the 10 families are each constrained to be monophyletic, because an exact solution is thereby possible (there are only $\sim 1.2 \times 10^{11}$ trees if the 10 families are constrained to be monophyletic but $\sim 8.7 \times 10^{36}$ trees if they are not). Of course, the prior hypotheses of monophyly are being assumed rather than tested, but this is appropriate under the conditions described.

Assessing Confidence in Results

Testing How Well Sequence Data Support Trees

Once the data are collected and a topology constructed, it is necessary to evaluate the reliability of those data and the supported tree. It is important to keep in mind that even randomly generated data can lead to a single, best result. Therefore, several methods exist for testing the robustness of the final topology using analytical and resampling procedures (Table V).

A problem in assessing confidence when there are more than four taxa is that the number of trees available for testing increases dramatically. Thus, few methods can reliably compare the more complex phylogenies. As the size of individual trees increases, the stringency of the tests increases as well. It is generally recommended that subsets of taxa be examined instead, from within the more complex topologies, focusing on specific major questions targeted before the analysis (see below). Alternatively one could limit the comparisons to just those topologies deemed plausible for biological reasons (e.g., a previous hypothesis of relationship based on independent data).

[31] D. L. Swofford, "PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0." Illinois Natural History Survey, Champaign, Illinois, 1990.
[32] J. S. Farris, *Syst. Zool.* **34**, 21 (1970).

TABLE V
METHODS FOR ASSESSING CONFIDENCE IN RESULTS

Method	Comments	Refs.[a]
Analytical techniques		
For parsimony procedures		
Wilcoxon rank-sum test, sign test, winning sites method	Determines whether significant character support exists for one tree relative to a second. Wilcoxon rank-sum test allows one to assign mutations different weights (i.e., transversions favoring one tree are given greater importance than transitions). For six or fewer taxa and no ordering as above, Wilcoxon rank-sum test reduces to simpler sign test. In winning sites method, binomial test is used to determine whether a greater number of phylogenetically informative positions (*sensu* parsimony) supports one tree versus a second	*1–3*
Confidence limits without clock	Assumes worst-case scenario for four taxa (two unrelated taxa with fast rates of evolution, with other two and common stem experiencing virtually no change). Under these conditions, two unrelated taxa are expected to share 3/16 of their positions by chance alone. Thus, to be statistically significant, a tree must be supported by more than 3/16 of its characters	*4*
Confidence limits with clock	Here, polytomy (star phylogeny) for four taxa is taken as worst-case situation. Thus, probability that a phylogenetically informative site supports a tree is same for all three resolutions of polytomy, 1/3	*5*
Williams/Goodman confidence limits	Similar to approach just described, except that a clock is not assumed. Method is based on a worst-case situation whereby support for correct tree is $\geq 1/3$ and $\leq 2/3$ for the two incorrect topologies combined	*6*
For evolutionary parsimony	A chi-square or binomial test is used to determine which phylogenetic invariants deviate significantly from zero and which do not	*7*
For maximum likelihood		
Likelihood ratio test	Ratio of likelihood scores for selected tree and star phylogeny is treated as a chi-square statistic with one degree of freedom. Alternatively, standard normal test of the mean and variance of the difference of their likelihood scores can be used to compare one tree to another	*2, 8, 9*
For distance approaches		
Branch length variances	An internal branch length is considered significant only if its length plus or minus two standard errors exceeds zero	*10–12*
Resampling techniques	Characters of original data set are randomly sampled and a tree is produced from new matrix. Many resampled matrices are analyzed (usually ≥ 100). Frequency of replication of a group is taken as measure of its statistical reliability or, at least, its stability	
Booststrapping	Characters are randomly sampled with replacement, leading to new data set of same size as original	*13, 14*

TABLE V *continued*

Method	Comments	Refs.[a]
Jackknifing	Characters are randomly sampled without replacement, leading to new data set smaller than original one. Jackknifing of taxa is sometimes done instead of characters	*15, 16*

[a] *Key to references: (1)* A. R. Templeton, *Evolution* **37,** 221 (1983); *(2)* J. Felsenstein, *Annu. Rev. Genet.* **22,** 521 (1988); *(3)* E. M. Prager and A. C. Wilson, *J. Mol. Evol.* **27,** 326 (1988); *(4)* J. Felsenstein, *in* "Statistical Analysis of DNA Sequence Data" (B. Weir, ed.), p. 113. Dekker, New York, 1983; *(5)* J. Felsenstein, *Syst. Zool.* **34,** 152 (1985); *(6)* S. A. Williams and M. Goodman, *Mol. Biol. Evol.* **6,** 325 (1989); *(7)* J. A. Lake, *Mol. Biol. Evol.* **4,** 167 (1987); *(8)* H. Kishino and M. Hasegawa, *J. Mol. Evol.* **29,** 170 (1989); *(9)* J. Felsenstein, *J. Mol. Evol.* **26,** 123 (1987); *(10)* M. Nei, J. C. Stephens, and N. Saitou, *Mol. Biol. Evol.* **2,** 66 (1985); *(11)* M. Hasegawa, H. Kishino, and T. Yano, *J. Mol. Evol.* **22,** 160 (1985); *(12)* W.-H. Li, *Mol. Biol. Evol.* **6,** 424 (1989); *(13)* J. Felsenstein, *Evolution* **39,** 783 (1985); *(14)* D. M. Hillis and J. J. Bull, *Syst. Biol.* **42,** 182 (1993); *(15)* S. Lanyon, *Syst. Zool.* **34,** 397 (1985); *(16)* D. Penny and M. Hendy, *Mol. Biol. Evol.* **3,** 403 (1986).

Groups to be evaluated need to be specified *a priori,* otherwise problems of multiple testing can lead to an unreasonably high probability of accepting some group as significantly supported. In addition, the following tests assume that each nucleotide substitution is independent and derives from a large sample, assumptions which often are not met. Despite these limitations, many systematists have argued for the importance of placing phylogenetic inference in a statistical framework and for improving the "primitive state" of testing its reliability.[33]

Analytical Methods

Analytical procedures for testing phylogenetic reliability operate by comparing the support for one tree to that for another, under the assumption of randomly distributed data. These methods have been extensively developed for parsimony procedures, with one of the earliest approaches using the Wilcoxon rank-sum test to compare the number of unique changes favoring one topology over a second. When fewer than six taxa are considered, this test reduces to the simpler sign test and binomial test (the latter being the winning sites method of Prager and Wilson[34]).

Another approach for testing parsimony results has been to compare the support for the best tree against that expected for a worst-case situation. If no molecular clock is assumed, then the worst-case scenario for four taxa occurs when two unrelated lineages evolve randomly and rapidly, coupled

[33] W.-H. Li and M. Gouy, *in* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, eds.), p. 249. Oxford Univ. Press, New York, 1991.
[34] E. M. Prager and A. C. Wilson, *J. Mol. Evol.* **27,** 326 (1988).

with virtually no change in the other two lineages or the central branch. Under these conditions, one expects two unrelated taxa to share 3 of 16 nucleotide sites by random chance alone. If a molecular clock is assumed, the worst-case situation becomes a trichotomy, with the probability of a phylogenetically informative site (*sensu* parsimony) supporting any one resolution being 1 in 3. Tables have been calculated for each of these cases, summarizing the number of unique changes and extra steps needed to favor statistically one tree over another, relative to the availability of sequence data and phylogenetically informative positions. A recent development of the above tests is the Williams–Goodman approach which does not assume a molecular clock. Instead, this approach assumes that the correct tree will be supported by one-third or more of the informative positions, whereas the two incorrect topologies together will be supported by a total of two-thirds or fewer of the informative sites. Each of the above procedures is largely restricted to four taxa, although at least one heuristic test has been developed to extend this type of approach to five taxa or more.[33] The null model used by these tests assumes equal support for the trees being compared. A significant departure from this expectation implies that more support, greater than expected by chance alone, exists for the best tree relative to the alternatives.

The method of invariants[35] uses a χ^2 test or binomial test to determine which phylogenetic invariants deviate significantly from zero and which do not. Significant departure of an invariant from zero indicates that the associated topology is well supported. In likelihood techniques, one tests the significance of the internal branch length of a tree for four taxa against the null model of an unresolved trichotomy. Here, the logarithm of the ratio of the maximum likelihood scores for the best tree and unresolved phylogeny is treated as a χ^2 statistic, with one degree of freedom. This approach can be extended to more than four taxa. Alternatively, one can compare two maximum likelihood trees in a heuristic way by the mean and variance of the difference of their likelihood scores.[36] Analytical tests of tree reliability for distance approaches have also been developed, with the most popular relying on tests of the variances of internal branch lengths. If an internal branch length plus or minus two times its standard error is greater than zero, then it is considered well-supported at the $\alpha = 0.05$ level.

Resampling Techniques

Resampling procedures estimate the reliability of a phylogenetic result by bootstrapping or jackknifing the characters of the original data set. In bootstrapping, a new data set of the original size is created by sampling the

available characters with replacement (Fig. 7). Thus, some characters become represented more than once, others only once, and others not at all in each bootstrapped data set. In contrast, jackknife methods randomly drop one or more data points (or taxa[37]) at a time, thereby creating smaller data sets by sampling without replacement. In either case, a phylogeny is then reconstructed from the resampled data set, and the replication of individual nodes is tallied. The frequency at which a node reappears among different permutations is taken as a measure of its reliability or, at least, of its relative stability (but see below).

Resampling procedures for testing phylogenetic reliability (particularly, the bootstrapping approach) are currently popular, primarily because of the wide availability of powerful and efficient algorithms for their implementation. These approaches have largely been used in conjunction with parsimony procedures, but they can be used in combination with other methods as well (e.g., bootstrapping of sequence positions prior to a distance analysis using the neighbor joining method). The interpretation of bootstrap proportions varies among authors; the values provide unbiased but highly imprecise estimates of repeatability (the probability that the result would be found again given a new sample of characters from the same distribution) and biased, but usually conservative, estimates of phylogenetic accuracy (the probability that the result represents the true phylogeny).[38] However, the degree of bias in the accuracy of estimates varies from node to node in a given tree, as well as from study to study, so bootstrap proportions are not directly comparable with each other. Nonetheless, they are sometimes used as relative measures of confidence among nodes within a single phylogenetic estimate.[39]

Faith and Cranston[14] have developed a procedure (the cladistic permutation tail probability) that randomizes the assignment of character states to taxa at individual sites, while retaining the original configuration of variation at each position. A tree is then constructed and the process repeated to yield a distribution of tree lengths given separate randomizations of the data. The length of a tree is then compared to this distribution to test its departure from lengths expected from randomness. Trees with original overall lengths less than or equal to the shortest 5% of the randomized trees are taken to have significant cladistic structure.

---

[35] J. A. Lake, *Mol. Biol. Evol.* **4**, 167 (1987).
[36] H. Kishino and M. Hasegawa, *J. Mol. Evol.* **29**, 170 (1989).
[37] S. Lanyon, *Syst. Zool.* **34**, 397 (1985).
[38] D. M. Hillis and J. J. Bull, *Syst. Biol.* **42**, 182 (1993).
[39] M. J. Sanderson, *Cladistics* **5**, 113 (1989).

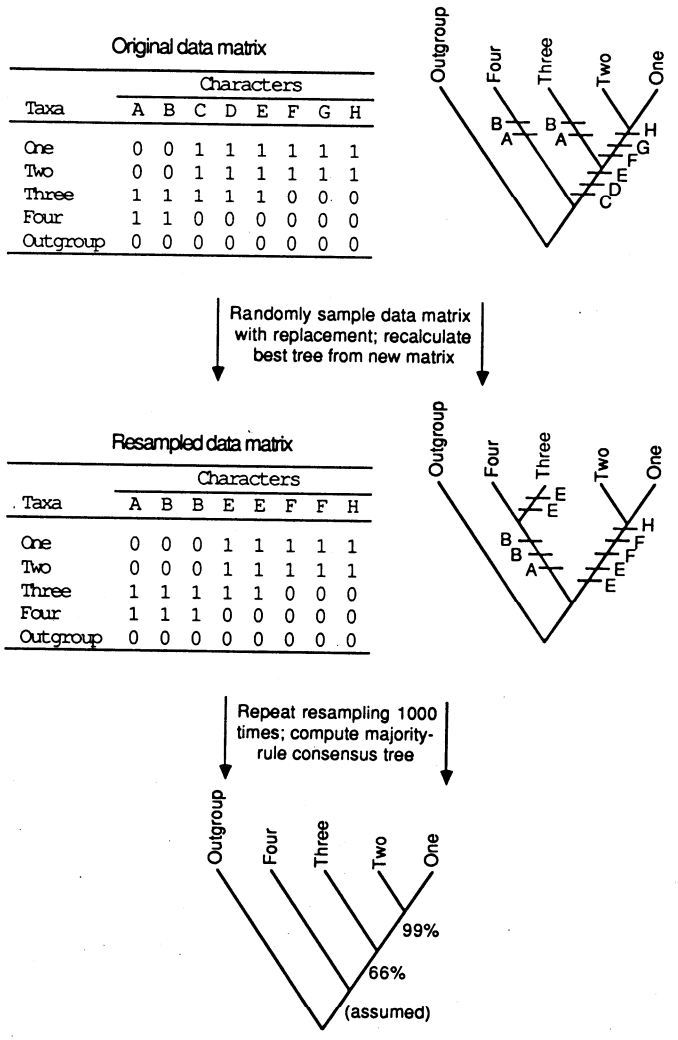

FIG. 7. Bootstrap analysis among characters in a parsimony analysis. The tree to the right of each matrix is the most parsimonious tree for that matrix. The final results of the bootstrap analysis are shown in the tree at the bottom. The number of times each branch was supported in the bootstrap replication is shown as a percentage. Outgroup rooting carries the assumption of ingroup monophyly, so no confidence interval can be assigned to the branch that unites the ingroup.

Statistical versus Phylogenetic Significance

A statistically significant result does not guarantee that an accurate reflection of the phylogeny has been achieved. Rather, it only suggests that a specific topology is strongly supported by a particular method of analysis of the sequence data. This discrepancy can occur because of two different reasons: (1) the statistical significance reflects systematic error that caused the chosen procedure to converge on the wrong topology; and (2) the current topology is correct, but it constitutes a gene tree, which may differ from that for the species owing to allelic polymorphism and lineage sorting, gene duplications or conversion, horizontal transfer, or other molecular evolutionary phenomena. Because of these two possibilities, one cannot accept a tree as correct, given only one set of sequences, even if the current results are considered statistically significant.

The ultimate criterion for determining phylogenetic reliability rests instead on tests of congruence among independent data sets representing both molecular and nonmolecular information.[40,41] Different character types and data sets are unlikely to suffer from the same evolutionary biases, and, as such, congruent results supported by each are more likely to reflect convergence onto the single, correct tree. In the absence of *a priori* knowledge of the truth, congruence remains the final arbiter. Systematists have always relied on concordance in this way to test their hypotheses, and it is therefore not surprising that a similar role for congruence in molecular phylogenetics is starting to emerge as well.

Implementation

No single computer program or package will allow an investigator to conduct all of the analyses described herein. We do not address alignment programs, which were the subject of a recent volume in this series.[3] Many programs for phylogenetic analysis have been described in the literature and are available by writing the authors of the original papers. The following programs are widely used and easily available, either for free or for a small fee. This list is not exhaustive but provides a starting point for conducting most of the analyses described in this chapter.

The PAUP (Phylogenetic Analysis Using Parsimony) program was written by David L. Swofford (Smithsonian Institution, Washington, D.C.; program available from Illinois Natural History Survey, 607 E. Peabody Drive, Champaign, IL 61820). It is a highly versatile, interactive program

[40] D. M. Hillis, *Annu. Rev. Ecol. Syst.* **18**, 23 (1987).
[41] M. M. Miyamoto and J. Cracraft, *in* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, eds.), p. 3. Oxford Univ. Press, New York, 1991.

for character-based analyses that allows a wide variety of weighting schemes and modifications of parsimony and methods of invariants. It also conducts bootstrapping, random sampling of trees, analyses of tree-length distributions, consensus analyses, and character analyses, and it has routines for producing camera-ready output of trees. The full-featured program currently is available only for Macintosh computers; an earlier version that lacks many of the advanced features is available for MS-DOS machines (an update is planned). Some versions are also available as C source code for use on workstations and mainframes.

Hennig86 was written by James S. Farris (American Museum of Natural History, New York, NY 10024). It is a fast and effective parsimony program. It is often faster than PAUP but has many fewer features and options. However, Hennig86 does contain a routine for successive approximation *a posteriori* character weighting.

Phylip (Phylogenetic Inference Package) was written by Joseph Felsenstein (Department of Genetics, SK-50, University of Washington, Seattle, WA 98195). The package includes a diverse collection of programs, including routines for calculating estimates of divergence and programs for both distance-based and character-based phylogenetic analyses. The parsimony programs are much slower and less efficient than in PAUP or Hennig86, but Phylip implements many methods that are not widely available elsewhere (e.g., maximum likelihood, many of the distance-based approaches). The package is distributed in Pascal source code or is available in precompiled versions for most computers.

The MacClade program, written by Wayne P. Maddison and David R. Maddison [Department of Ecology and Evolution (WPM) and Department of Entomology (DRM), University of Arizona, Tucson], is another program for parsimony analyses. However, it is primarily designed for interactive tree manipulation and studies of character evolution, rather than for finding most-parsimonious trees. It contains many features especially designed for analysis of DNA sequences and numerous features for the production of camera-ready tree output. It is completely compatible with PAUP, so the two programs are effectively used in combination. MacClade is available only for Macintosh computers; it is available from Sinauer Associates (Sunderland, MA 01375).

NJTREE, UPGMA, and TDRAW were written by Li Jin, J. W. H. Ferguson, N. Saitou, and J. C. Stephens (contact Li Jin, Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston, P.O. Box 20334, Houston, TX 77225). These programs build neighbor joining and UPGMA trees. They are written in FORTRAN-77; precompiled versions are available for MS-DOS computers.

NJDRAW, NJBOOT, and related programs and available from M. Nei and T. S. Whittam (Institute of Molecular Evolutionary Genetics, Penn State University, 328 Mueller Laboratory, University Park, PA 16802-5303). These programs are available precompiled for MS-DOS computers. They are used for computing various DNA distances and for constructing and testing neighbor joining trees.

ANCESTOR, WTSUBS, AUTSUBS, and ALLTOPS were written by P. L. Williams and W. M. Fitch (contact W. M. Fitch, Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717). These FORTRAN programs are available in uncompiled form or precompiled for MS-DOS computers. The various programs are designed for generating ancestral states, choosing among tree topologies, and performing various aspects of dynamically weighted parsimony procedures.

Acknowledgments

Our work has been supported by the National Science Foundation (DEB 91-22823 and DEB 92-21052 to D.M.H.; BSR 88-57264 and BSR 89-18606 to M.M.M.). We thank Cliff Cunningham, Winston Hide, David Swofford, Elizabeth Zimmer, and the Smithsonian Molecular Phylogenetics Discussion Group for comments on the manuscript.