PHYLOGENETIC ANALYSIS IN MOLECULAR EVOLUTIONARY GENETICS

Masatoshi Nei
Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802; e-mail: NXM2@PSU.EDU

KEY WORDS: molecular phylogeny, tree-building methods, four-cluster analysis, linearized trees, ancestral proteins

ABSTRACT
Recent developments of statistical methods in molecular phylogenetics are reviewed. It is shown that the mathematical foundations of these methods are not well established, but computer simulations and empirical data indicate that currently used methods such as neighbor joining, minimum evolution, likelihood, and parsimony methods produce reasonably good phylogenetic trees when a sufficiently large number of nucleotides or amino acids are used. However, when the rate of evolution varies extensively from branch to branch, many methods may fail to recover the true topology. Solid statistical tests for examining the accuracy of trees obtained by neighbor joining, minimum evolution, and least-squares method are available, but the methods for likelihood and parsimony trees are yet to be refined. Parsimony, likelihood, and distance methods can all be used for inferring amino acid sequences of the proteins of ancestral organisms that have become extinct.

INTRODUCTION
Phylogenetic analysis of DNA or protein sequences has become an important tool for studying the evolutionary history of organisms from bacteria to humans. Since the rate of sequence evolution varies extensively with gene or DNA segment (17, 88, 142), one can study the evolutionary relationships of virtually all levels of classification of organisms (kingdoms, phyla, classes, families, genera, species, and intraspecific populations). Phylogenetic analysis is also
important for clarifying the evolutionary pattern of multigene families (4, 44, 93) as well as for understanding the adaptive evolution at the molecular level (15, 64, 143). This technique also gives much deeper insight into the mechanism of maintenance of polymorphic alleles in populations (34, 128).

Reconstruction of phylogenetic trees by using statistical methods was initiated independently in numerical taxonomy for morphological characters (120) and in population genetics for gene frequency data (13). Some of the statistical methods developed for these purposes are still used for phylogenetic analysis of molecular data, but in recent years many new methods have been developed. Felsenstein (31) and Swoford et al (124) reviewed various statistical methods from mathematical points of view. In this review, I discuss only recently developed methods or newly clarified statistical properties of previous methods, with emphasis on practical utilities rather than mathematical details or mathematical possibilities. Because of space limitation, I do not discuss the phylogenetic analysis of gene frequency data. Citation of the literature is also restricted to papers directly related to the subject.

**PHYLOGENETIC ANALYSIS OF DNA OR PROTEIN SEQUENCES**

It is now customary to consider the reconstruction of a phylogenetic tree as a statistical inference of a true phylogenetic tree, which is unknown. There are two processes involved in this inference: "estimation" of the topology (branching pattern of a tree) and estimation of branch lengths for a given tree topology. When a topology is known, statistical estimation of branch lengths is relatively simple, and one can use several statistical methods such as the least squares and the maximum likelihood methods. The problem is the estimation or reconstruction of a topology. When there are a sizable number of DNA or protein sequences (say 10), the number of possible topologies is enormously large (more than 1 million) (14), and it is generally very difficult to choose the correct topology among them.

In phylogenetic inference a certain optimization principle such as the maximum likelihood (ML) or minimum evolution (ME) principle is often used for choosing the most likely topology. The ML method is a well-established statistical method of parameter estimation; it gives the smallest variance of a parameter estimate when sample size is large. However, this is true only when a given probability space is considered. In the construction of phylogenetic trees, maximization of the likelihood is done for each topology separately by using a different likelihood function (28), and the topology with the highest (maximum) likelihood is chosen as an estimate of the true topology. However, since different topologies represent different probability spaces of parameters, it is not clear whether the highest likelihood tree is expected to be the true tree unless an infinite number of nucleotides are examined (88, 148, 150). Indeed, it is not difficult to find examples in which the ML method is inferior to other methods in obtaining the true tree, as is mentioned later. Note also that the regularity conditions (continuity and differentiability of the likelihood function) required for the asymptotic properties of ML estimators are not satisfied in phylogenetic reconstruction (148, 150). Some authors have suggested that topologies are parameters, but these parameters are not included in the likelihood function that is maximized.

Extending Cavalli-Sforza & Edwards' (14) idea, Rannala & Yang (97) attempted to estimate a topology under the assumption that a new species is formed following the birth-death process in statistics. In this case, a topology is treated as a random variable. Mathematically, this is a reasonable approach, but since the birth-death process is unlikely to describe the real speciation events (25), it is still unclear how useful their new approach is in real data analysis. Note also that the pattern of nucleotide substitution often changes with site and time, particularly when long-term evolution is considered (see later), and at this moment no study has been made on this problem.

A similar criticism applies to all other tree-building methods, though the nature of the criticism varies with the method. That is, the statistical foundation of topology estimation by any optimization principle is not well established. Nevertheless, computer simulations have shown that the optimization principles currently used generally work well under biologically realistic conditions.

The method of phylogenetic inference currently used in molecular phylogenetics can be classified into three major groups: distance methods, likelihood methods, and parsimony methods. Recently, Hendy and colleagues (53, 55, 56) proposed the use of the Hadamard conjugation for phylogenetic reconstruction (closest tree method). However, its practical utility is yet to be examined.

**Distance Methods**

In distance methods, an evolutionary distance is computed for all pairs of sequences, and a phylogenetic tree is constructed from pairwise distances by using the least squares, minimum evolution, or some other criterion. The evolutionary distance used for this purpose is usually an estimate of the number of nucleotides or amino acid substitutions per site, but other distance measures may also be used. There are a large number of distance methods for constructing phylogenetic trees (31, 88, 89), but those commonly used are based on the principles of least squares and minimum evolution.
LEAST-SQUARES (LS) METHODS  The principle of LS methods is to compute the minimum sum of squared differences between observed pairwise distances and estimated pairwise distances (patristic distances) (88) for a given topology and to choose a topology that shows the smallest minimum sum of squared differences. Cavalli-Sforza & Edwards (14) suggested that the ordinary or generalized LS methods can be used for distances computed from gene frequency data, whereas Fitch and Margoliash (37) used a weighted LS method. Later Bulmer (9) implemented and formalized the generalized LS method for DNA and protein sequence data.

However, LS methods often give negative branch lengths, and mainly for this reason the accuracy of the topology obtained is not particularly high (74, 111, 112, 121). One way to rectify this problem is to conduct the least squares estimation of branch lengths with the restriction of no negative branch lengths (14, 31). Bulmer (personal communication) and Gascuel (39) have shown that in the case of four sequences, this restricted LS method gives the same results as those obtained by the neighbor joining method, which is mentioned later. However, this does not seem to be the case when the number of sequences is greater than four, because neighbor joining also occasionally produces negative branches.

MINIMUM EVOLUTION (ME) METHODS  In this method, the branch lengths of a tree are estimated by a certain algorithm from pairwise distance data, and the total sum (S) of branch lengths is computed for each of the possible topologies. The topology that shows the smallest S value will then be chosen as the most likely tree (23). In this method, branch lengths are estimated either by Fitch & Margoliash's algorithm (110) or by the ordinary LS method (69, 102). Rzhetsky & Nei (102) presented a formal mathematical treatment of this method for DNA and protein sequence data and simplified the computational algorithm considerably. They (104) also presented a theoretical foundation of this method by showing that the expected value of S is smallest for the true topology when unbiased estimators of nucleotide or amino acid substitutions are used as distance measures. Of course, this does not mean that a tree with the smallest S value is expected to be the true tree unless a large number of nucleotides or amino acids are used.

Kidd & Sgaramella-Zonta (69) suggested that the total branch lengths [L(\bar{S})] be computed by summing the absolute values of all branch lengths under the conjecture that there are no negative branches for the true topology. However, L(\bar{S}) does not have a nice statistical property that permits the fast computation of S values and the statistical tests as developed by Rzhetsky & Nei (102, 104). Note also that in the presence of statistical errors, some branch lengths may become negative by chance even for a correct topology (119). Furthermore, if one wants to have an ME tree without negative branches, a better way would be to estimate branch lengths by the least squares method under the constraint of nonnegative branches.

Although the ME method is statistically appealing, it requires a large amount of computational time to examine all different topologies if the number of sequences (m) is greater than 10. For this reason, Rzhetsky & Nei (102, 105) suggested that the neighbor joining (NJ) tree (see below) be first constructed and then a set of topologies close to this NJ tree be examined to find a tree with a smaller S value (temporary ME tree). A new set of topologies close to this temporary ME tree (excluding previously examined topologies) are now examined to find a tree with an even smaller S value. This process will be continued until no tree with a smaller S is found, and the tree with the smallest S is regarded as the ME tree. The theoretical basis of this strategy is that the ME tree is generally identical or close to the NJ tree when m is relatively small (102, 110) and thus the NJ tree can be used as a starting tree when m is large. They (105) also suggested that a special type of bootstrapping could be used for generating topologies for examination. Kumar (77) devised a new algorithm to obtain an ME tree, extending the NJ algorithm to examine many potential ME trees. This algorithm does not examine all topologies, but computer simulation has shown that it almost always examines the true tree even if m is quite large.

FOUR-CLUSTER ANALYSIS  In phylogenetic analysis, it is often important to establish the evolutionary relationships of four groups of organisms. For example, the evolutionary relationships of animals, plants, fungi, and protists have been studied for many decades, yet we do not have a definitive answer, partly because each group contains so many different kinds of organisms (5, 45, 116, 139). In most methods of phylogenetic analysis the number of organisms to be included is limited because of computational difficulties. For this reason, only a few representative organisms are used from each group, but this procedure often gives erroneous conclusions (2).

The four-cluster analysis (101) is an application of the theory of the ME method (104) and can handle a large number of species from each group of organisms as long as each group is known to be monophyletic, and it does not require any information regarding the branching order of organisms within groups. Let A, B, C, and D be the four monophyletic groups or clusters, and suppose that A, B, C, and D contain nA, nB, nC, and nD sequences, respectively. In this case, there are three possible unrooted trees of clusters, i.e. T1 = ((AB)(CD)), T2 = ((AC)(BD)), and T3 = ((AD)(BC)), and one of them must be correct. This correct tree is expected to have the smallest sum of branch lengths. Let S1, S2, and S3 be the sums of branch lengths for trees T1, T2, and T3. To compute S1, S2, and S3, we have to know the phylogenetic
relationships of all sequences within clusters, but what we need is to compute
the differences $S_1 - S_2$, $S_1 - S_3$, and $S_2 - S_3$. These differences can be computed
by a simple algorithm without knowing the phylogenetic relationships within
clusters, and the statistical significance of each difference can be tested.

This technique was applied to resolve the branching pattern of animals,
plants, fungi, and a group of protists, using ribosomal RNA genes. It was con-
cluded that animals and fungi are significantly closer to each other than to the
others (78).

NEIGHBOR JOINING (NJ) METHOD This method (112) is a simplified version of
the ME method for inferring a bifurcating tree. In this method, the $S$ value
is not computed for all or many different topologies, but the examination of
different topologies is imbedded in the algorithm, so that only one final tree is
produced. Computation of $S$ starts with a star phylogeny, in which all interior
branches are assumed to be 0. This tree is clearly incorrect, so the $S$ value ($S_O$)
is much higher than the $S$ for the true tree. The next step is to compute $S_{ij}$ for
a tree in which sequences $i$ and $j$ are paired and are separated from the rest of
the sequences that still form a star tree. If $i$ and $j$ are the neighbors connected
by only one node (e.g. sequences 1 and 2 in Figure 1A), then $S_{ij}$ is smaller
than $S_O$. Therefore, computing $S_{ij}$'s for all pairs of sequences and choosing the
smallest $S_{ij}$, we can identify a pair of neighbors. Once this pair is identified,
they are combined as a single unit and treated as a single sequence in the next
step. This process is continued until all multifurcating nodes are resolved into
bifurcating ones. In practice, any distance measures are subject to stochastic
effects, so that the NJ tree obtained may not necessarily be the true tree.

According to computer simulations (77, 102), this method nearly always
produces the same topology as that of the ME tree when the extent of sequence
differences is sufficiently large and the number of nucleotides examined is large
(>500). When the latter condition is not satisfied, however, the NJ tree can be
considerably different from the ME tree (104), yet the difference in $S$ between
the NJ and ME trees is usually statistically nonsignificant. Furthermore, using
computer simulation, Kumar (77) showed that when the NJ and ME trees are
different, the latter tree is not necessarily closer to the true tree (topology). When
a large number of closely related sequences are used as in the case of Vigilant
et al.'s data (138), both NJ and ME trees will give only a rough evolutionary
relationship of the sequences, and there is no point in spending excessive efforts
to find the ME tree (77). At any rate, the high efficiency of the NJ method in
obtaining the ME tree or the true tree seems to stem from the fact that in each
step of sequence clustering the principle of minimum evolution is applied, and
the repeated application of this principle would reduce the effects of sampling
errors in topology construction.

DISTANCE MEASURES Theoretically, if the total number of substitutions be-
tween any pair of sequences is known, all the above distance methods produce
the correct phylogenetic tree (additive tree). In practice, however, this number is
almost always unknown, and thus many different methods for estimating this
number have been proposed, some being quite sophisticated (6, 79, 107, 124,
146, 154). Examples are Kimura's (70) and Hasegawa et al.'s (51) methods.
These methods are certainly useful for correcting for parallel and backward mu-
tations and are expected to give better estimates of the number of substitutions
per site ($d$) than the simple proportion of nucleotide differences ($p$) distance.
However, these estimates usually have a larger variance than the uncorrected
$p$ distance. Partly for this reason, the $p$ distance or a simple distance measure
such as the Jukes-Cantor distance (67) tends to produce the correct topology
more often than sophisticated distance measures when the rate of nucleotide
substitution is more or less constant in all evolutionary lineages and the num-
ber of nucleotides examined is not very large (108, 111, 112, 121, 127). In
this case, therefore, it is preferable to use $p$ distance for topology construction
rather than a more sophisticated distance. However, the advantage of $p$
distance over more sophisticated distance measures diminishes when the number
of nucleotides examined is large and the evolutionary rate varies extensively
with evolutionary lineage, and in this case it is better to use unbiased distance
measures (79, 112).
The efficiency of distance measures in obtaining the correct tree depends on at least two factors: the linear relationship with the number of substitutions and the standard error or coefficient of variation of the estimate of the distance measure. For Kimura’s (70) model of nucleotide substitution, several authors have attempted to produce better distance measures than the original estimator (43, 114, 127), but the utility of these distance measures with actual data has not been tested.

Many distance measures for estimating the number of nucleotide substitutions per site \( d \) become inapplicable when the distance is very large, because they usually involve logarithmic terms in the mathematical formula and the arguments of the logarithms often become negative. This problem can be avoided by expanding the logarithmic terms into an infinite series, but the variance of the distance obtained in this way seems to be quite large when the sequence divergence is high (106, 126). However, phylogenetic trees are generally constructed with pairwise distances whose values are rather small (say \( d < 0.5 \)). In this case, the use of \( p \), Jukes-Cantor, and Kimura distances is usually sufficient for topology construction. However, if \( d \) is large or if there is evidence that the substitution rate varies extensively among sites and with evolutionary lineage, the gamma distance (65, 79) is expected to produce better trees. For some special sets of data, even more complicated distance measures may give better results. For example, DNA sequences of the control (D-loop) region of mitochondrial DNA evolve in a complicated way (73), and a special model (131) has been developed for analyzing data for this region.

**Maximum Likelihood (ML) Methods**

**DNA Likelihood Method** The idea of using an ML method for phylogenetic inference was first presented by Cavalli-Sforza & Edwards (14) for gene frequency data, but they encountered a number of problems in implementing the method. Later, considering nucleotide sequence data, Felsenstein (28) developed an algorithm for constructing a phylogenetic tree by the ML method. The first model of nucleotide substitution used was rather simple and did not take into account the transition/transversion bias, which is often observed in actual data. This deficiency was later removed by using a more realistic model involving five parameters (31). A number of authors developed more general models (33, 51, 131, 146), and some of them have been implemented in computer programs (1, 32, 92, 149).

However, for obtaining the true topology, sophisticated models do not necessarily give better results than simple models such as the Jukes-Cantor model, though the likelihood value of an ML tree is almost always higher for the former models than for the latter (40, 100). In fact, computer simulations have shown that under certain circumstances a simple model gives a higher probability of obtaining the true tree than a complicated model, even if sequence evolution has occurred following the latter model (150). Of course, this result was obtained in a simulation with four sequences, and it is not clear what will happen if more than four sequences are used. However, this problem is the same as that for distance methods discussed above, and it emphasizes the difficulty of topology construction mentioned earlier. It should also be noted that the pattern of nucleotide substitution varies from site to site (132) and with evolutionary time (3, 49, 76, 132), particularly when long-term evolution is considered. At this moment, it is not clear how these factors affect the topology estimation by ML.

A serious problem with ML is the computational time. Even if the number of sequences is about ten, it requires an enormous amount of computational time. Olsen et al (92) developed a faster algorithm, but it still requires a large amount of time if many topologies are to be examined. Saitou (109) and Adachi & Hasegawa (1) took a different approach to tackle this problem using a new algorithm. Their algorithm (star-decomposition algorithm) is essentially the same as that of the NJ method, except that the ML principle is used in finding neighbors instead of the ME principle. In Adachi & Hasegawa’s (1) computer software MOLPHY, the star-decomposition (SD) tree is regarded as a first potential ML tree from which trees with higher ML values are searched for by using other algorithms such as local branch rearrangement. In practice, however, the SD algorithm seems to be quite efficient in obtaining the true tree, and the relationship between the SD and exhaustive search algorithms for ML trees may be similar to that between the NJ and ME algorithms (100; T Sintnikova & M Nei, unpublished results).

**Protein-Likelihood Method** When the DNA sequences are relatively closely related to one another, DNA likelihood methods seem to work well. However, if they are distantly related and encode protein sequences, many complications arise because the rate of synonymous substitution is generally much higher than that of nonsynonymous substitution and the transition/transversion bias exists. The relative frequencies of the four nucleotides at third codon positions also varies considerably with sequence (63, 76), suggesting that the stationary model of nucleotide substitution is not appropriate. By contrast, the evolutionary change of protein sequences does not suffer very much from these problems and seems to be much simpler than that of DNA sequences when long-term evolution is considered. Noting this property, Kishino et al (72) proposed a protein-likelihood method by using Dayhoff et al’s (18) empirical transition matrix for 20 different amino acids. Adachi & Hasegawa (1, 3) extended this method by using various transition matrices including Jones et al’s (66) matrix for nuclear proteins and their own for mitochondrial proteins. They applied these methods to various sequence data and obtained reasonably good trees for
several groups of vertebrate organisms (11, 12). Analyzing mitochondrial gene
data for 11 vertebrate species, Russo et al. (100) showed that protein sequences
are more reliable than DNA sequences for obtaining the correct phylogeny.

Maximum Parsimony (MP) Methods

In MP methods, a given set of nucleotide (or amino acid) sequences are consid-
ered, and the nucleotides (or amino acids) of ancestral sequences for a hypothet-
ical topology are inferred under the assumption that mutational changes occur
in all directions among the four different nucleotides (or 20 amino acids). The
smallest number of nucleotide substitutions that explain the entire evolutionary
process for the given topology is then computed. This computation
for all other topologies, and the topology that requires the smallest number of
substitutions is chosen to be the best tree (22, 35, 47).

If there are no multiple substitutions at each site, MP is expected to generate
the correct topology as long as enough parsimony-informative sites (79) are examined. In practice, nucleotide sequences are often subject to backward and
parallel substitutions, and this introduces uncertainties in phylogenetic inference. When the true tree has a special type of topology and branch lengths, MP
may generate an incorrect topology even if an infinite number of nucleotides
are examined (27). This can happen even if the rate of nucleotide substitution
is constant for all evolutionary lineages (35, 129, 157). Furthermore, in parsim-
ony analysis it is difficult to treat the phylogenetic inference in a statistical
framework because there is no natural way to compute the means and variances
of minimum numbers of substitutions obtained by the parsimony procedure.
However, under certain circumstances, MP is quite efficient in obtaining the
correct topology (89). Note also that MP is the only method that can easily take
care of insertions and deletions of nucleotides, which sometimes give important
phylogenetic information.

Weighted Parsimony One factor that makes MP inefficient is the transition/transversion bias and the heterogeneity of substitution rate among different
nucleotide sites. In the control region of mitochondrial DNA, the transition/transversion ratio (R) is as high as 15 in humans (138), and the rate
heterogeneity as measured by the inverse (1/\( \lambda \)) of the gamma parameter (65)
seems to be as high as 6.7 (73, 131, 140). In this case, nucleotide sites with
transitional changes or high substitution rates are not very informative for phy-
logenetich construction when relatively distantly related sequences are used. The
reason is that at these sites multiple substitutions are likely to have occurred,
and this will introduce noise in phylogenetic inference.

One way to reduce this noise is to give higher weights to transitional changes or
slowly evolving sites and lower weights to transitional changes or
fast evolving sites (26, 124). In this case, the tree length no longer gives an
estimate of the minimum number of nucleotide substitutions, but this method
substantially improves the probability of obtaining the correct topology (60,
91). One problem with this approach is that we do not know the actual R
value for the data set under investigation. In this case, it is possible to use a
so-called dynamically weighted parsimony method (113, 141). In this method,
a probable R value is first used to generate an MP tree, and then a new R
value is estimated from the tree obtained. This new R value is then used to
generate a new MP tree. (In practice, all different nucleotide pairs are weighted
differently.) This process is repeated until a stable MP tree (or trees) is obtained.
This is a time-consuming method and does not guarantee the convergence of
an MP tree. Nevertheless, computer simulations have shown that this method
substantially improves the probability of obtaining the correct tree when R is
high (134). A similar weighted method can also be used to take into account
the variation of substitution rate among different sites (38, 141).

STATISTICAL TESTS OF PHYLOGENETIC TREES

During the past two decades, many authors have studied the statistical methods
for testing the reliability of the tree obtained. Some of these studies have been
reviewed by Felsenstein (31), Li & Gouy (83), and Li & Zharkikh (84). Here
I present a brief summary of recent studies on the subject. Statistical tests of
phylogenetic trees can be divided into two categories: a test of reliability of a
tree obtained and a test of topological differences between two or more different
trees obtainable from the same data set.

Reliability of an Estimated Tree

Interior Branch Tests One way of knowing the reliability of an estimated
tree is to examine the reliability of each interior branch. This is particularly
appropriate for trees constructed by distance methods. Consider the tree for
five sequences given in Figure 1(A). In the case of five sequences, there are 15
possible unrooted bifurcating trees, and each tree is composed of five exterior
branches and two interior branches. Suppose one obtains tree (A) in Figure 1
by some tree-building method. The reliability of this tree (topology) is assured
if the two interior branch lengths \( f = 0 \) and \( g = 0 \) are different from 0 and positive.
Therefore, by testing the null hypothesis of \( f = 0 \) and \( g = 0 \), we can establish
the validity of the tree. In general, the null hypothesis of an interior branch
length \( b = 0 \) can be tested by computing the standard error \( s(\hat{b}) \) of an estimate
\( \hat{b} \) of \( b \). Since \( \hat{b} \) is known approximately to follow the normal distribution even
when the number of nucleotides examined is as small as 100 (102), the null
hypothesis of \( b = 0 \) can be tested by examining the statistical signifi-

of the normal deviate \( Z = \frac{b}{\sigma(b)} \). This test is called the interior-branch test (119).

This type of test was first used by Nei et al (90) for a UPGMA tree and then by Li (81) for an unrooted tree when the number of sequences is four or five. Later Rzhetsky & Nei (102, 104) developed a fast algorithm for computing \( \sigma(b) \) using the ordinary least-squares approach and made it possible to use this test for a large number of sequences. This method requires a specific model of nucleotide substitution, and the test seems to be robust about the substitution models unless the extent of sequence divergence is high (134).

In this method, the confidence probability \( (PC) \) that \( b > 0 \) is computed by using the Z test, and if the probability is higher than 95% or 99%, then \( b \) is considered to be significantly positive. One theoretical problem concerning Rzhetsky & Nei's (102) method was that \( PC \) was computed without considering the estimation error of the topology obtained. If we take into account this error in the computation of confidence probability, the actual confidence probability \( (PC') \) can be smaller than the original \( PC \) value (119). In practice, however, the difference between \( PC \) and \( PC' \) is small if we consider the region of \( PC \geq 0.99 \).

Sifnikova (117) produced an approximate formula \( PC' = 3PC - 2 \), where \( PC > 2/3 \). This is a useful formula for computing \( PC' \), because the computation of \( PC' \) is much simpler than that of \( PC' \).

Another interior branch test that is applicable to distance trees is Rodrigo & Dopazo's (98) bootstrap test. This test is different from Felsenstein's bootstrap test, which is mentioned below, and is intended to examine the reliability of each interior branch for a given topology. As in the case of Felsenstein's test, the same number of nucleotides as that of the original sequence is randomly sampled with replacement for each set of sequences, and the lengths of all branches are estimated by a given tree-building method for a given topology, which was obtained by the original sequence data. This process is repeated many times for the same topology. Therefore, the length of an interior branch varies from replication to replication. We count the number of replications in which a given interior branch takes a positive value \( (b > 0) \), and the proportion of these among the entire set of replications is used as an estimate of the confidence probability \( (PC) \) of the interior branch. The advantage of this method is that it requires no particular substitution models and thus is applicable for a wide variety of situations (117). The computational time required is also much shorter than the analytical method when the number of sequences is large. However, when the number of nucleotides examined is small, this may give biased estimates of \( PC' \).

A number of authors (31, 124) have suggested that the null hypothesis of \( b = 0 \) could be tested by the likelihood ratio test, because the trees with \( b = 0 \) and with \( b \geq 0 \) are nested. In the case of phylogenetic trees, however, this test does not seem to be justified (42, 150, 151), and computer simulations (40, 134) have shown that the test may give strong statistical support for a wrong topology. The DNAML program in PHYLIP computes the confidence interval of a branch length, but this interval also does not seem to be reliable (134).

**Felsenstein's Bootstrap Test** One of the most commonly used tests of the reliability of an inferred tree is Felsenstein's (30) bootstrap test. In this test, the reliability of an inferred tree is examined by using Efron's (24) bootstrap resampling technique. A set of nucleotide sites is randomly sampled with replacement from the original set, and this random set is used for constructing a new phylogenetic tree. This process is repeated many times, and the proportion of replications in which a given sequence cluster (sequence partitioning: e.g. sequences 1 and 2 vs others in Figure 1A) appears is computed. If this proportion \( (PB) \) is high (say, \( PB > 0.95 \)) for a sequence cluster, this cluster is considered to be statistically significant. The null hypothesis of this test as implemented in MEGA (79) is the same as that of the interior branch test. In recent years, a number of authors (57, 119, 155, 156, 158) have shown that this test is generally very conservative except for high \( PB \) values close to 1. Zharkikh & Li (158) then invented a method to rectify the conservativeness of \( PB \), but it is not yet clear whether this correction method is applicable to the cases of many sequences.

Although Felsenstein's bootstrap test can be very conservative, I believe that it is a useful method for evaluating the statistical reliability of an inferred tree. The actual pattern of nucleotide substitution is very complicated (76, 146) and often changes with site and evolutionary time (132). Therefore, it is better to use a conservative test for examining trees for distantly related sequences. However, we have to be cautious even with this conservative test, because any tree-building method may generate an incorrect tree almost consistently for a given data set, as is mentioned later. Furthermore, when a tree is produced for closely related sequences, one may use mathematically more rigorous methods such as the interior-branch test (119) or the Zharkikh & Li test (158). Note that Felsenstein's (30) original bootstrap method is slightly different from the method described above and is for testing a consensus tree generated by repeated resampling. The null hypothesis of this test is not clearly specified, but it should be similar to the one mentioned above.

**Tests of Topological Differences**

**Minimum Evolution Trees** The second class of tests is a comparison of two topologies in terms of the quantity used for an optimization process of phylogenetic inference. Previously we mentioned that the minimum evolution tree is a tree with the smallest sum \( S_M \) of branch lengths. In practice, however there may be several other trees whose \( S \) is greater than \( S_M \) but is not significantly
different from the latter. These trees are potentially correct trees, and thus one may want to keep them until other data are obtained to identify the true tree. 

Rzhetsky & Nei (102) developed a statistical method for testing the difference \( D = S_B - S_A \) in \( S \) between two topologies. This test is equivalent to the test of the lengths of the interior branches at which the two topologies are different. Suppose that topology \( (A) \) in Figure 1 is the correct tree and \( (B) \) and \( (C) \) are incorrect ones. When unbiased distance estimators are used, the expectation \( [E(D)] \) of \( D = S_B - S_A \) for topologies \( (A) \) and \( (B) \) is given by \( f/2 \), where \( f \) is the true length of the left interior branch of topology \( (A) \) (102). Therefore, if \( D \) is significantly greater than 0, we can establish the validity of topology \( (A) \). However, if \( D \) is significantly smaller than 0, topology \( (B) \) will be the correct one. Similarly, comparison of topologies \( (A) \) and \( (C) \) gives the expectation of \( D (\approx S_C - S_A) \) equal to \( 3(f + g)/4 \) (102). Therefore, the null hypothesis for the test of \( D \) has a clear-cut biological meaning. Using this test, one can identify topologies that are not significantly different from the ME tree.

Rzhetsky & Nei's (102) method for testing \( D \)'s is dependent on the mathematical model on which a particular distance measure is based and requires an intensive computation when \( m \) is large. Nei (89) suggested that the hypothesis \( E(D) = 0 \) can be tested by a bootstrap method. In this test \( S \) is computed for a given pair of topologies \( (i \) and \( j) \) for each sequence resampling, and \( D_{ij} = S_i - S_j \) is computed. If this is repeated many times, we can compute the mean \( D \) and its standard error. Therefore, we can test the null hypothesis of \( E(D) = 0 \). When there are several potentially correct trees, \( D_{ij} \) can be computed for all pairs of \( i \) and \( j \) by using the same set of resampled sequences.

The \( D \) test mentioned above is clearly related to the interior branch test mentioned earlier, but the exact relationship remains uncertain. One might speculate that if every interior branch of the ME tree is significant, the \( D \) test will also establish that \( S_B \) is significantly smaller than \( S \) for any other tree. If this is the case, the interior branch test or the Felsenstein bootstrap test would be simpler than the \( D \) test in finding a reliable tree.

Some authors (124) criticized the NJ method as producing only one final tree, rather than several potentially correct trees. This criticism is valid. However, if all the interior branches of a NJ tree are statistically supported, there will be no need to consider other alternative trees. By contrast, if some of the interior branches are not statistically supported, one may consider the alternative trees that can be generated by changing the branching pattern for each nonsignificant interior branch. This approach would be simpler than the enumeration of alternative trees by using the \( D \) test mentioned above. Another approach to this problem is to construct a condensed tree (79) in which all weakly supported interior branch-lengths are reduced to 0. This tree is conceptually similar to a consensus tree (123), which is constructed for MP trees.

ML TREES One might think that a simple test for the difference in topology between an ML and a suboptimal tree would be to use the standard log likelihood ratio (LR) test. Unfortunately, this cannot be done because all bifurcating trees have the same number of degrees of freedom (31). Kishino & Hasegawa (71) suggested that the difference in log likelihood value between an ML tree and a suboptimal tree be tested by using the variance of the difference in single-site log likelihood between the two trees. However, these authors have not specified the null hypothesis of this test in relation to the topological differences as given in Figure 1. Without knowing what is being tested, it is difficult to interpret the results of the test. Clearly, a detailed study of the theoretical basis of the test is necessary.

As mentioned earlier, the number of possible topologies is enormously large even when the number of sequences is about 10. If an unlimited amount of computer time is available, it is possible to conduct a large number of bootstrap resamplings and evaluate the probability of occurrence of each topology. If a particular topology occurs with a high probability (say \( P > 0.95 \)), one may conclude that this topology is most likely to be the correct one. In practice, it is usually unnecessary to consider all topologies because most topologies have an extremely low probability of occurrence. Furthermore, it is often possible to identify a small number (say \( 10 \)) of potentially correct topologies on the basis of other biological information. In these cases, one can evaluate the relative bootstrap probability values of the potentially correct topologies. However, even if we consider a small number of potentially correct topologies, Felsenstein's bootstrap procedure requires a prohibitive amount of computer time for ML trees.

To cope with this problem, Kishino et al (72) developed two approximate methods of computing the bootstrap probability of the \( i \)-th topology \( (P_i) \). In the first method, the MND method, the log likelihoods \( \ell(1) \), \( \ell(2) \), \( \ldots \), and \( \ell(k) \) for topologies 1, 2, \ldots, and \( k \) are assumed to follow a multivariate normal distribution, of which the means, variances, and covariances are estimated by Kishino & Hasegawa's method (71). One can then choose a random set of \( \ell(1) \), \( \ell(2) \), \( \ldots \), \( \ell(k) \) from this distribution and then determine the topology that has the highest likelihood. This topology is the ML tree in this set of sample. This process is repeated many times, and the probability \( \{P(i)\} \) that the \( i \)-th topology is the ML tree is determined. If a particular topology has a \( P(i) \) value of 95% or higher, then this topology is assumed to be the correct topology. In the second method, the RELL method, the log likelihood at each site \( \ell(j) \) is computed for all \( k \) topologies, and for each set of bootstrap-resampled sites, the sum \( \sum \ell(j) \) of log likelihoods for individual sites is computed for each topology. If this process is repeated many times, one can determine \( P(i) \)'s.

The above test procedures, particularly the latter one, seem to be more reasonable than Kishino & Hasegawa's earlier test (71). However, this method has
one problem, which is how to choose a set of potentially correct trees, particularly when the number of sequences is large. Kishino et al (72) proposed an ad hoc procedure to solve this problem, but its utility is still untested.

MP TREES As mentioned earlier, it is difficult to develop any parametric test for MP trees because of the nonrandom nature of "minimum numbers of substitutions." Templeton (135) suggested a nonparametric test for comparing two topologies that is similar to Kishino & Hasegawa's (71) test for ML trees. However, the null hypothesis of this test is also unclear in relation to the topologies to be compared. Probably the best way of testing the reliability of inferred MP trees would be Felsenstein's bootstrap test, though one has to be cautious about the possibility of inconsistency of MP methods (see below).

MERITS AND DEMERITS OF DIFFERENT TREE-BUILDING METHODS

Criteria of Comparison

Because there are many different tree-building methods, one is naturally interested in the merits and demerits of different methods. There are several different criteria for comparing different tree-building methods. Important ones are (a) computational speed, (b) consistency as an estimator of a topology, (c) statistical tests of phylogenetic trees, (d) probability of obtaining the correct topology, and (e) reliability of branch length estimates.

The computational speed of each tree-building method can be measured relatively easily, though it depends on the algorithm used. According to this criterion, the NJ method is superior to most other tree-building methods which are currently in use. This method can handle a large number of sequences \((m > 100)\) even with a personal computer, and the application of bootstrap tests is easy. The orthodox MP, LS, ME, and ML methods examine all possible topologies searching for the MP, LS, ME, and ML trees, respectively. Since the possible number of topologies rapidly increases with \(m\) (14), it is difficult to use these methods when \(m\) is large. In the case of ME, however, simplified algorithms (77, 104, 105) seem to be as efficient as the exhaustive search in obtaining the correct tree. It is hoped that similar simplified algorithms will be developed for other methods as well. [In the case of MP methods the branch and bound method (54) may be used when \(m \leq 20\).] Note that a vast majority of tree topologies are clearly incorrect when \(m\) is large, and there is no need to examine all these trees. The algorithm suggested by Rzhetsky & Nei (104, 105) may be used for identifying MP and ML or suboptimal MP or ML trees rapidly.

A tree-building method is said to be a "consistent estimator" if the method tends to give the correct topology as the number of nucleotides used \((n)\) approaches infinity (27). The NJ, ME, and LS methods are a consistent estimator if unbiased estimates of nucleotide substitutions are used as distance measures (19, 102, 112), and so is the ML method when the correct model of nucleotide substitution is used (148). By contrast, MP is often inconsistent, as mentioned earlier. In practice, however, \(n\) is usually of the order of hundreds to thousands, and in this case even NJ, ME, LS, and ML may fail to produce the correct tree with a relatively high probability when MP fails (60, 62, 115). Therefore, consistency is not always a useful criterion for comparing the efficiencies of different tree-building methods.

We have already discussed statistical tests of phylogenetic trees for several different tree-building methods. At present, the statistical methods for testing NJ and ME trees are well established. Solid statistical tests are also available for trees obtained by the generalized LS method (9, 137). In the case of ML methods, however, there seem to be many complications, as mentioned above. The best method for testing MP trees is probably Felsenstein's bootstrap test, as long as the cases of inconsistency are avoided (30).

The probability of obtaining the correct topology is probably the most important criterion for comparing different tree-building methods, but this is also the most difficult problem to study. During the past 15 years, many authors have studied this problem, yet we do not have a clear-cut answer, as is discussed below. Another important criterion for comparing different methods is the reliability of branch length estimates. Once the correct topology is obtained for a given data set, this problem can be studied relatively easily. Theoretically, ML, LS, NJ, and ME are expected to give more reliable estimates of branch lengths than MP. At present, MP (and sometimes ML) trees are almost always presented without branch length estimates. This practice is regrettable because it gives a distorted picture of a phylogenetic tree. Since computer programs are available for estimating branch lengths of MP trees (123), any tree should be provided with branch length estimates.

**Probability of Obtaining the Correct Topology**

In most cases of phylogenetic reconstruction, we never know the true phylogeny for real data under investigation, so it is difficult to study this problem empirically. However, if we use an appropriate mathematical model, we can simulate the evolutionary changes of DNA sequences following a given model tree. We can then reconstruct a tree by various methods using the artificially generated present-day sequences and compare the topology of the tree obtained with that of the model tree. If this process is repeated many times, we can estimate the probability of obtaining the correct topology \((P_T)\), and this probability can be used for comparing the efficiencies of different tree-building methods (7, 89, 95, 133).
THEORETICAL STUDY When the number of sequences examined \( m \) is small (four or five), it is possible to evaluate \( P_T \) analytically for the NJ, LS, and MP methods (111, 119, 155). These studies have shown that when the evolutionary rate is more or less constant for all four or five sequences, NJ has a slightly higher \( P_T \) value than MP, which in turn has a somewhat higher \( P_T \) than Fitch & Margoliash’s (37) LS method (111). Both the ordinary and generalized LS methods are inferior to the ME method in obtaining the correct topology (103). This inferiority seems to be partly due to the fact that the LS methods often generate negative branches, as mentioned earlier. However, analytical evaluation of \( P_T \) is very difficult when \( m \) is large, and the conclusion obtained from these studies may not apply to a wide variety of situations. No study has been made for ML trees even in the case of \( m = 4 \). For this reason, comparison of \( P_T \) among different methods is usually done by computer simulation.

COMPUTER SIMULATION If we use computer simulation, \( P_T \)’s can be estimated for a variety of evolutionary conditions. Thus, a large number of simulation studies have been done during the past 15 years. The results obtained before 1990 have been summarized by Nei (89), but there are many recent studies (40, 48, 50, 60, 61, 74, 91, 102, 115, 148, 150). It is not easy to summarize these studies because different authors considered different evolutionary models and used different computer algorithms.

One of the most popular model trees used in computer simulation is the unrooted tree of four sequences in the form given in Figure 1(D), where \( a, b, \) and \( c \) represent the expected number of nucleotide substitutions per site. When \( a = b = c \) and \( a \) is greater than 0.1 but smaller than 0.5, almost any tree-building method produces the correct topology if \( n \) is greater than 100. Therefore, this model tree is not useful for discriminating the efficiencies of different methods. For this reason, many authors have assumed \( a > b \). If we use the Jukes-Cantor model of nucleotide substitution, the MP method becomes inconsistent when \( b = c = 0.05 \) and \( a \geq 0.394 \) (134). Therefore, MP always fails to produce the correct tree when a large number of nucleotides is used. However, NJ and ML usually recover the correct tree in this case if \( a < 0.5 \).

Some authors (59) have used cases of an extremely high degree of sequence divergence \( a = 2.83; p \) distance \( = 0.65, \) and \( b = c = 0.05; p = 0.05 \) to show a superiority of ML methods. However, such divergent sequences are almost never used in practice because of the difficulty of sequence alignment. Therefore, such a study is not biologically meaningful. For the same reason, a large part of computer simulations conducted by Huelsenbeck (60) also do not seem to be biologically meaningful (91). Although he considered the complete two-dimensional space for \( a \) and \( b = c (0 \leq p \leq 0.75; 0 \leq \text{corrected distance} \leq \infty) \) for the sake of completeness, actual data used for phylogenetic analysis fall into a relatively small portion of the space near the origin (108). When \( b \) and \( c \) are of the order of 0.05 and \( 0.1 < a < 0.5 \), MP is generally less efficient than NJ, which is in turn less efficient than ML (48, 50, 60, 134). However, when \( a, b, \) and \( c \) are all of the order of 0.01 \( \sim 0.025 \) and \( n \) is about 1000, all three methods reconstruct the true tree quite easily (134).

Note that the comparison of different tree-building methods is not always straightforward when a complicated model of nucleotide substitution is used, because appropriate computer programs are not always available. Thus, Tateo et al (134) compared the robustness of MP, NJ, and ML using available computer programs for the case where the substitution rate varies among nucleotide sites following the gamma distribution. Since the computer program for ML was not available, their comparison of NJ and ML was not adequate. Using a newly developed ML algorithm with the gamma distribution (147), Huelsenbeck (61) attempted to rectify Tateo et al’s inadequate comparison between NJ and ML. However, he used a continuous gamma distribution for NJ but a discrete version for ML. Although this difference would not affect the final conclusion significantly, it illustrates a difficulty in computer simulation. This problem is compounded by the fact that for the NJ or ME methods, biased distance measurers often give a higher \( P_T \) value than unbiased distances.

The model tree \( (D) \) in Figure 1 obviously does not cover all possible types of trees for four sequences. The model tree \( (E) \) is different from tree \( (D) \) in that two long branches with length \( a \) are now neighbors and two short branches with length \( b \) are also neighbors. Interestingly, this model tree gives different relative \( P_T \) values compared to those for tree \( (D) \). Some results for the two trees are given in Table 1. In tree \( (D) \), ML gives the highest \( P_T \) value among the three methods ML, MP, and NJ, and NJ with \( p \) distance shows the lowest value. In tree \( (E) \), however, ML gives the lowest \( P_T \), whereas NJ with \( p \) distance gives the highest. Furthermore, both unweighted and weighted MP show much higher \( P_T \)’s than ML. These results were obtained apparently because in parsimony and NJ with \( p \) distance short branches tend to attract each other. Yang (150) has also shown that even when the evolutionary rate is constant, ML can be inferior to unweighted MP. These results indicate the difficulty of obtaining a general conclusion about the relative efficiencies of different tree-building methods, even for the simplest case of \( m = 4 \).

A number of simulation studies have been done for the cases of six or more sequences, although it is difficult to consider more than a dozen sequences. When \( m \) is very large, the interior branch lengths become very small if we want to make the most divergent sequence pair biologically reasonable \( (d \leq 1.0) \). For this reason, \( P_T \) becomes very low for any method, and an enormous amount
than MI? However, the differences in $dT$ among the cases were very small. Note that the above comparison was done with very special types and ML, respectively when $dT$ is for the high divergence case were 0.68, 0.67, and 0.54 for MP, NJ, and ML, respectively. Therefore, on the basis of $dT$ values, ML is better than NJ, which is in turn slightly better than MP. However, the differences in $dT$ among the three different methods are very small. Note that the above comparison was done with very special types of computer time is required (77, 133). The model trees considered usually represent the case of constant rate or its modifications (50, 110, 112, 121, 122). In some of these studies (110, 121), the exhaustive search of MP or ML trees was not done because of an excessive computer time required, but the true topology was always included. Therefore, the simulations were somewhat more favorable for MP or ML than for NJ. In general, these simulation studies have shown that ML is as good as or better than NJ, which is in turn often better than MP. However, the number of these studies is quite limited, and it is difficult to extrapolate these results to other cases.

A somewhat different type of simulation was conducted by Kuhner & Felsenstein (74). They generated a model tree of 10 sequences following the branching process in statistics in each replication, and the sequence data generated according to this model tree were used to reconstruct a phylogenetic tree. The topology of this tree was then compared with that of the model tree. The topological difference between the model tree and the estimated tree was measured by the number of nonidentical sequence partitions between the two trees being compared ($dT$) (96). They considered a case of low divergence with an expected value of the root-to-tip branch length equal to 0.0193 and a case of high divergence with an expected value of 0.193. The average $dT$ values for the low divergence case with a constant rate were 1.95, 1.82, and 1.64 for MP, NJ, and ML, respectively when $n = 1,000$, whereas $dT$'s for the high divergence case were 0.68, 0.67, and 0.54 for MP, NJ, and ML, respectively. Therefore, on the basis of $dT$ values, ML is better than NJ, which is in turn slightly better than MP. However, the differences in $dT$ among the three different methods are very small. Note that the above comparison was done with very special types of model trees that seem to have had very short interior branches occasionally. (None of the model trees used was published.) Therefore, many inferred trees should have had multifurcating nodes, yet the authors did not treat them as such; they accepted whatever resolution of the multifurcation a particular computer algorithm produced. Here again, we see an example where the comparison of different methods is algorithm-dependent. Note that ML algorithms often give zero branch lengths even if the true tree is apparently bifurcating (16).

Despite many recent computer simulations, the interpretation of the results is not as straightforward as was originally expected, and more careful studies are needed to know the relative efficiencies of different methods. However, it is now clear that any method is not almighty, and there are situations in which one method is more efficient than others in obtaining the true tree and that, unless the evolutionary rate varies drastically with evolutionary lineages, all the three methods considered here generally give the same or similar topologies (110). Computer simulations have also indicated that one of the most important factors is the number of nucleotides or amino acids used per sequence and that if this number is small, one cannot produce reliable trees.

**Tests Based on Known Phylogenies** Although it is generally difficult to know the true topology in real data analysis, there are a few such cases. One is a phylogenetic tree experimentally produced by artificial mutagenesis with T7 phages (58). However, this type of experiment produces only one or a few replications, so it is difficult to compare different methods statistically. Furthermore, the pattern of nucleotide changes produced by mutagens seems to be somewhat unusual (8). It is thus unclear whether we can extrapolate the results obtained from these experiments to real cases.

However, there are few instances in which the phylogenetic tree for a group of organisms is firmly established on the paleontological and morphological bases. One such example is given in Figure 2(A). The complete nucleotide sequence of mitochondrial DNA (mtDNA) has recently been published for the 11 vertebrate species given in this figure. MtDNA in these species contains 13 protein-coding genes, the number of shared codons for each gene varying from 52 to 582. A phylogenetic tree was reconstructed for each of these genes and for the entire set of genes (3682 codons), and the trees obtained were compared with the true tree (100). In this study, amino acid sequences rather than nucleotide sequences were used, because the former produced more reliable trees.

When all 13 genes were used, all tree-building methods (NJ, ML, and MP) produced the correct tree irrespective of the algorithm used. A few genes (usually large genes) such as Nd5, Cytb, and Co3 also produced the correct or nearly correct topology. However, some genes (e.g. Co2, Nd1, Nd3, and Nd4t) almost always produced incorrect trees regardless of the method and algorithm.
sequences were added. The reason for this is unclear, but the unusually slow rate of fish genes and the change in the pattern of amino acid substitutions with site and time (76) seem to be contributing factors.

However, a surprising result was obtained when the lamprey and sea urchin sequences were added. An incorrect topology was produced with high bootstrap values.

Figure 2 (A) Known phylogeny for 11 vertebrate species. The total amino acid sequences of 13 coding genes of mitochondrial DNA produced the correct phylogeny with a bootstrap value of 100% for each interior branch. (B) When the lamprey and sea urchin sequences were added, an incorrect topology was produced with high bootstrap values.

The molecular clock is one of the most important concepts in molecular evolutionary genetics, yet it has been controversial for many decades (21, 36, 82). Strictly speaking, the rate of nucleotide or amino acid substitution would never be constant over the entire evolutionary process because nucleotide or amino acid substitution is a complicated process that is dependent on the evolutionary stability and functional changes of genes. Therefore, if we study a large number of nucleotides or amino acids and the extent of sequence divergence is sufficiently large, we would surely be able to detect the heterogeneity of evolutionary rate. Yet, the extent of rate heterogeneity is usually moderate when relatively closely related sequences are used, so that one can use an approximate clock to obtain rough estimates of times of divergence between sequences from molecular data. Actually, a number of molecular evolutionists (75, 136) have attempted to estimate divergence times even when the molecular clock fails.

To use a molecular clock for estimating divergence times, however, it is important to test the applicability of a clock for the data set under consideration. If a molecular clock does not hold, we must identify and eliminate the sequences that deviate significantly from the assumption of rate constancy. After elimination of these sequences, we can reestimate the branch lengths of the tree for the remaining sequences under the assumption of rate constancy. A tree constructed in this way is called a linearized tree and can be used for estimating the divergence time of any pair of sequences provided that the rate of substitution can be estimated from other sources such as fossil records or geological dates (130). In this case, the test of a molecular clock need not be very strict, because the estimates of divergence times obtainable are generally very rough. Actually, we may retain certain important sequences even if they evolve significantly faster or slower than the average, unless they distort the tree substantially.

A commonly used test of the molecular clock is the relative rate test for three sequences (36, 87, 125, 145), but this test is not appropriate for our purpose. We need a test that is applicable for many sequences simultaneously. Felsenstein (29) suggested that for trees constructed by distance methods the test be done by comparing the least-squares residual sums obtained under the assumption of rate constancy ($\hat{R}_c$) with that for the case of no such assumption ($\hat{R}_w$) using
Fisher's F test. When the ordinary or weighted least-squares method is used to compute $R_C$ and $R_N$ (Fitch and Kitch programs in the PHYLIP package), it is implicitly assumed that pairwise distance estimates are independently and normally distributed. Normality may not be seriously violated, but pairwise distances are positively correlated because of the tree-like relationships of the sequences. Therefore, this is not a rigorous statistical test (31).

The hypothesis of rate constancy can also be tested by computing the likelihood values with and without the assumption of rate constancy (31). Twice the difference of the log likelihood values between the two cases is expected to follow the $\chi^2$ distribution asymptotically with $m - 2$ degrees of freedom. Goldman (42) questioned the $\chi^2$ approximation of the test statistic, but Yang et al's (151) simulation study suggests that the $\chi^2$ approximation is acceptable in most cases.

Takezaki et al (130) presented two simple methods of testing rate constancy specifically designed to identify sequences evolving excessively fast or slow: the two-cluster and the branch-length tests. In these methods, the root of the tree is first established by using outgroup sequences as in the case of Figure 2(A), where the fish genes can be regarded as outgroups. The two-cluster test examines whether the difference in average branch length between two clusters of sequences created by an interior node is statistically significant or not. This test is done by computing the standard error of the difference between the two average branch lengths and applying the $Z$ test. If this $Z$ test shows that one branch length is significantly different from the other, the one that is more different from the average root-to-tip distance for all sequences is eliminated. In the branch-length test, the root-to-tip branch length ($y$) is computed for all sequences, and the difference between the $y$ value for a particular sequence and the average ($\bar{y}$) for all sequences is computed. This difference is then subjected to a statistical test to identify the sequences that evolve significantly faster or slower.

The deviation of $y$ from $\bar{y}$ can also be tested by Uyenoyama's (137) generalized least-squares method. Since the generalized least-squares estimates of branch lengths have a smaller variance than the ordinary least-squares estimates, this test is expected to be more powerful than the above methods. However, the application of this method to the case of a large number of sequences is difficult because it requires a large amount of computer time. For our purpose, we do not need a very powerful method to detect the rate heterogeneity, because we are interested only in approximate rate constancy.

Once excessively fast- or slow-evolving sequences are eliminated, one can construct a linearized tree under the assumption of rate constancy, using the method described by Takezaki et al (130). Linearized trees have been constructed to estimate the times of divergence for various pairs of Drosophila species using the alcohol dehydrogenase gene sequences and the geological estimates of the times of formation of the Hawaiian islands (99, 130). These studies indicate that when many sequences (about 40 sequences) are used, the time estimates remain nearly the same even if some sequences that evolved significantly slower or faster (1% level) than the average are included. The same method has also been used for estimating the times of origin of the orders of placental mammals and birds (52).

**PERSPECTIVES**

In this review, I have discussed recent developments in phylogenetic analysis that are biologically important. I have emphasized that the statistical foundation of phylogenetic inference is not well established for any tree-building method and that there is an urgent need to clarify this foundation. However, computer simulations and a few empirical studies suggest that currently used methods such as the NJ, ME, MP, and ML methods generate reasonably good phylogenetic trees (topologies) when a sufficiently large number of nucleotides or amino acids are used. In general, all these methods produce the same or similar trees unless the evolutionary rate extensively varies with evolutionary lineage. When the evolutionary rate varies with evolutionary lineage, MP tends to be less efficient than other methods in obtaining the true topology, but if the extent of rate heterogeneity is very high, all methods may fail to identify the true topology. MP methods also tend to give more biased estimates of branch lengths than others. However, MP methods have an advantage over other methods in that they can easily utilize information generated by insertions/deletions.

In practice, the number of nucleotides or amino acids used is sometimes quite small, and the sequences analyzed are closely related. In this case, any tree-building method would make some errors in topology construction, and the use of a simple method would be sufficient for obtaining rough evolutionary relationships of sequences. By contrast, when the extent of sequence divergence is very high and the pattern of nucleotide or amino acid substitution remains nearly the same for all sites during the entire evolutionary time, ML methods are expected to give better results in topology estimation than other methods. The only problem is whether the substitution pattern remains the same or not for a long evolutionary time. If this pattern varies with site and changes with time, the advantage of ML methods over other methods would decline, because other methods require less rigid assumptions about the substitution pattern than ML methods. Another complicating factor is sequence alignment. When there are many deletions and insertions, it is often very difficult to have a reliable
sequence alignment, and a relatively small difference in sequence alignment often has a profound effect on the phylogenetic tree reconstructed. Yet, few studies have been made on the sensitivities of different tree-building methods to the alignment differences. At this moment, the relative efficiencies of different tree-building methods remain unclear, particularly when various biological factors are considered.

Our current knowledge of the relative efficiencies of different tree-building methods is largely based on computer simulation and some theoretical considerations. However, our ability to simulate the real process of DNA sequence evolution is limited, and in the future the relative efficiencies should be studied by using actual sequence data for known phylogenies. Fortunately, as the amount of sequence data increases, the number of known phylogenies for which such a study can be done is increasing. Therefore, we will probably be able to know the empirical relative efficiencies in the near future.

In the past two decades many investigators have studied the molecular phylogeny of three basic forms of life, archaebacteria, eubacteria, and euukaryotes, but the results obtained are still conflicting (20, 46, 68, 144). To resolve this problem, Golding & Gupta (41) examined the relationships of gram-positive bacteria, gram-negative bacteria, archaebacteria, and euukaryotes using 24 different protein sequences, and showed that different proteins generate different but statistically significant topologies. On the basis of this result, they hypothesized that euukaryotes evolved by fusion of an archaebacterium and a gram-negative bacterium. While this hypothesis is interesting and should be pursued in more detail, it should be kept in mind that current statistical tests are not very reliable for examining the evolutionary events in such ancient times.

This proviso applies even to phylogenetic trees of different phyla, classes, and families. Note that we could not reconstruct the correct phylogeny even for vertebrates and invertebrates by using mtDNA (Figure 2B). Therefore, great caution is necessary in interpreting molecular phylogenies for distantly related organisms. Since our statistical methods depend on so many simplifying assumptions, it is not always clear whether we are uncovering the genealogical history of early evolution or merely describing the extent of functional differentiation of genes that have been highly conserved. (Note that there are only two amino acid differences between calf and pea in a sequence of 105 amino acids of histon H4 (88). Do they represent stochastic changes of amino acid sequences?) In this type of study, we should use both quantitative and qualitative characters (68). Here we need some principles of the cladistic approach as well as the statistical approach. It is also important to examine a large number of genes and analyze them simultaneously (20).

In this article, I have been concerned primarily with statistical inference of phylogenetic trees. However, phylogenetic analysis of DNA or protein sequences is also useful for understanding the mechanism of evolution as mentioned in the Introduction. One approach to this problem is to infer the amino acid sequences of proteins in ancestral organisms from sequence data of extant organisms and to study how each amino acid substitution has changed the function of genes in the evolutionary process. This approach was suggested as early as in 1963 by Pauling & Zuckerkandl (94), but not until recently was it used very often because of difficulty in predicting the function of ancestral genes by this approach alone.

In recent years statistical methods for inferring ancestral amino acid sequences have been refined, and we can use any of the parsimony (35, 47, 85, 86), likelihood (152), and distance (153) approaches for this purpose. Computer simulations have shown that the likelihood and distance methods give more accurate inference than does the parsimony method, but all methods give reasonably good results when extant sequences are relatively closely related (153). Once the ancestral sequences are inferred, it is possible to reconstruct the ancestral proteins by site directed mutagenesis. Jermann et al (64) used this technique to study the evolutionary change of catalytic activity of ribonucleases in artiodactyls in relation to the evolution of ruminant digestion. Similarly, Chandrasekharan et al (15) reconstructed an inferred chymase (an enzyme belonging to the serine protease family) of an ancestral organism of mammals and showed that the catalytic activity of the ancestral enzyme was higher than that of the chymase of current mammals.

This type of study will be conducted more often in the future, because both statistical and biochemical techniques are now available. It is no longer a dream to reconstruct proteins of many extinct organisms by using these techniques and to study how morphological and physiological characters have evolved. In the study of evolution the phylogenetic analysis of DNA or protein sequences will play a more important role in the future.

ACKNOWLEDGMENTS

I thank Arndt von Haeseler, Blair Hedges, John Huelsenbeck, Sudhir Kumar, Andrey Rzhetsky, Tanya Sitnikova, Naoko Takezaki, and Ziheng Yang for their comments on an earlier version of this paper. I am also grateful to Naoko Takezaki for her help in making the illustrations. This work was supported by grants from NIH (GM20293) and NSF (DEB-9520832).

maximum likelihood over neighbor joining. Mol. Biol. Evol. 12:843–49
75. Kumada Y, Benson DR, Hillenmann D, Hosted TJ, Rocchelot DA, et al. 1993. Evolution of the glutamine synthetase gene, one of the oldest existing and func-
80. Kumazawa Y, Nishida M. 1996. Phylogenetic utility of mitochondrial transfer RNA genes for deep divergence in animals. See Ref. 90a, pp. 23–35
116. Deleted in proof
Figure 1. Parameter optimization simulations performed on the Bar-1 fragment from Barnase. a, MC simulation trajectory from which initial GA population was drawn; b, effect of population size; c, effect of joint search; d, fraction of conformations carried forward to the next generation; e, effect of side-chain free energy minimization; f, effect of \( \Delta G_{\text{total}} \) standard deviation recalculation frequency. Simulations were performed relative to a standard set of parameters: population size, 50; number of generations, 60; number of pairs applied to joint, 50; \( \sigma_{\text{local}} \) recalculation frequency, 60; fraction of conformations carried forward to the next generation, 10; rounds of side-chain minimization, 50; \( T \), 300 K. The key in each figure shows the parameter values used.