Annu. Rev. Genet. 1996. 30:371–403

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

0066-4197/96/1215-0371$08.00

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 Swofford 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

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.

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 & 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 [$L(\hat{S})$] 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, $L(\hat{S})$ 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 m_A, m_B, m_C, and m_D sequences, respectively. In this case, there are three possible unrooted trees of clusters, i.e. $T_1 = ((AB)(CD))$, $T_2 = ((AC)(BD))$, and $T_3 = ((AD)(BC))$, and one of them must be correct. This correct tree is expected to have the smallest sum of branch lengths. Let S_1, S_2, and S_3 be the sums of branch lengths for trees T_1, T_2, and T_3. To compute S_1, S_2, and S_3, 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 concluded 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

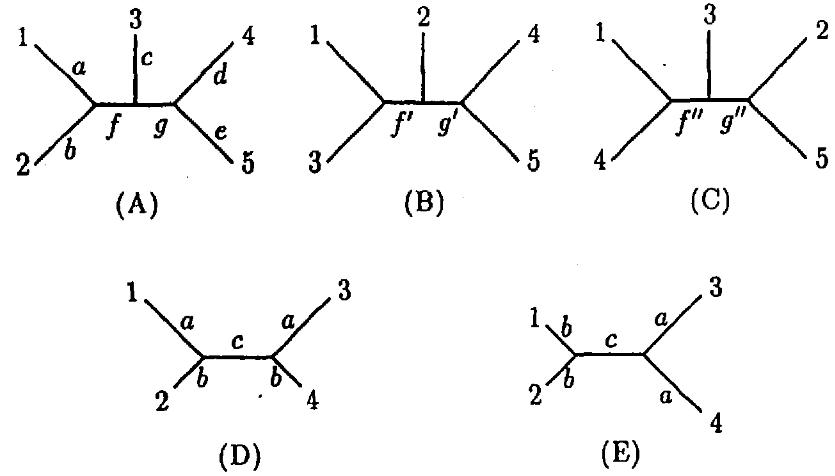


Figure 1 (*A*)–(*C*), three different unrooted trees for five sequences; *a*–*g*, expected branch lengths; (*D*)–(*E*), two unrooted trees for four sequences with different branch lengths. These two model trees were used for computer simulation of which the results are presented in Table 1.

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

[redacted]

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 nuseveral 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 considered, and the nucleotides (or amino acids) of ancestral sequences for a hypothetical 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 is done

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 of the normal deviate $Z = \hat{b}/s(\hat{b})$. 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 $s(\hat{b})$ 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 (P_C) that $\hat{b} > 0$ is computed by using the Z test, and if the probability is higher than 95% or 99%, then $\hat{b}$

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 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.

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 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

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

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 $b = c(0 \leq p \leq 0.75; 0 \leq$ corrected distance $d \leq \infty)$ 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, Tateno 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 Tateno 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_C 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

Table 1 Percent probabilities of obtaining the correct tree topology

Number	Tree D						Tree E					
Nucleotides	NJ			MP		ML	NJ			MP		ML
(*n*)	*p*	JC	K2	UW	W		*p*	JC	K2	UW	W	
100	44	68	72	47	64	76	98	73	74	88	96	64
200	41	79	81	52	80	84	100	83	82	97	99	76
300	43	87	88	59	80	92	100	88	86	98	100	82
500	35	94	95	62	89	97	100	96	94	100	100	90
800	29	96	96	63	94	98	100	98	96	100	100	94
1000	35	99	99	66	98	100	100	99	99	100	100	96

In both trees *D* and *E* in Figure 1, $a = 0.4$, $b = 0.1$, and $c = 0.05$ were assumed. Sequences data were generated by using Kimura's (70) two-parameter model with a transition/transversion rate ratio of 2, and the method of simulation was the same as that of Nei et al (91).

Abbreviations: NJ, neighbor-joining method; *p*, *p* distances; JC, Jukes-Cantor distance; K2, modified Kimura distance (32); MP, maximum parsimony method; UW, unweighted; W, weighted; ML, maximum likelihood method.

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 the 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 *Nd4l*) almost always produced incorrect trees regardless of the method and algorithm

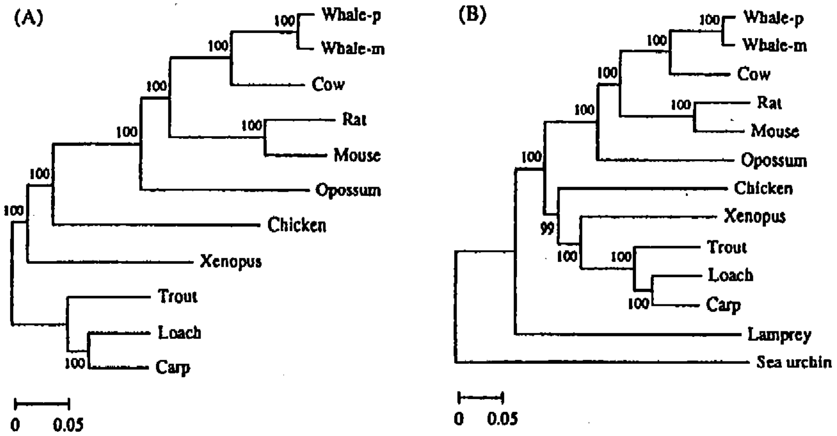


Figure 2 (*A*) Known phylogeny for 11 vertebrate species. The total amino acid sequences of 13 coding genes of mitochondrial DNA produced the correct phylogeny with a bootstrap value of 100% for each interior branch. (*B*) When the lamprey and sea urchin sequences were added, an incorrect topology was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

used. This result clearly indicates that some genes are more suitable than others in phylogenetic inference and that all the tree-building methods tend to produce the same topology whether the topology is correct or not. Similar results were obtained by Kumazawa & Nishida (80). Since only 13 genes were studied, it was difficult to evaluate the relative efficiencies of the different tree-building methods. In general, however, sophisticated methods such as the ML method with Jones et al's substitution model were no better than simple methods such as NJ with *p* distance or ML star-decomposition algorithm. Similar results were obtained by Cao et al (11, 12). These results suggest that the pattern and the rate of amino acid substitution vary with a group of organisms (also with evolutionary time) and thus sophisticated mathematical models do not necessarily generate better results.

However, a surprising result was obtained when the lamprey and sea urchin sequences were added to the 11 sequences in Figure 2(*A*): a clearly wrong tree [Figure 2(*B*)] was obtained by all tree-building methods even when all genes were used, and a bootstrap test showed strong statistical support for this wrong tree! The reason for this is unclear, but the unusually slow rate of evolution of fish genes and the change in the pattern of amino acid substitutions with site and time (76) seem to be contributing factors.

Empirical studies of a few cases of known phylogenies have shown that when the sequences used are relatively closely related the correct phylogeny is generally obtained as the number of codons or nucleotides increases but that the topology of distantly related sequences may well be incorrect even when a large number of codons or nucleotides are used and a bootstrap test may give strong statistical support for it.

THE MOLECULAR CLOCK AND LINEARIZED TREES

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 (R_C) with that for the case of no such assumption (R_N) 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 χ^2 distribution asymptotically with $m - 2$ degrees of freedom. Goldman (42) questioned the χ^2 approximation of the test statistic, but Yang et al's (151) simulation study suggests that the χ^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

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 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

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

maximum likelihood over neighbor joining. *Mol. Biol. Evol.* 12:843–49

62. Huelsenbeck JP, Hillis DM. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42:247–64

63. Hughes AL, Nei M. 1989. Nucleotide substitution at major histocompatibility

tioning genes. *Proc. Natl. Acad. Sci. USA* 90:3009–13

76. Kumar S. 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* 143:537–48

77. Kumar S. 1996. A stepwise algorithm for

90a. Nei M, Takahata N, eds. 1995. *Current Topics on Molecular Evolution*. Penn. State Univ., USA, and Grad. Univ. Adv. Stud., Hayama, Jpn.

91. Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. *Science*

program package for inferring and testing minimum-evolution trees. *Comput. Appl Biosci.* 10:409–12

106. Rzhetsky A, Nei M. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varie

of phylogenetic trees. *Mol. Biol. Evol.* 12:319–33

120. Sokal RR, Sneath PHA. 1963. *Principles of Numerical Taxonomy.* San Francisco: Freeman

121. Sourdis J, Krimbas C. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.* 4:159–66

122. Sourdis J, Nei M. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.* 5:298–311

123. Swofford DL. 1993. *PAUP: Phylogenetic Analysis Using Parsimony.* Champaign, IL: IL Natl. Hist. Surv.

124. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. DM Hillis, C Moritz, BK Mable, 407–514. Sunderland, MA: Sinauer. 2nd ed.

125. Tajima F. 1993. Simple methods for testing molecular clock hypothesis. *Genetics* 135:599–607

126. Tajima F. 1993. Unbiased estimate of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 10:677–88

127. Tajima F, Takezaki N. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Mol. Biol. Evol.* 11:278–86

128. Takahata N. 1993. Allelic genealogy and human evolution. *Mol. Biol. Evol.* 10: 2–22

129. Takezaki N, Nei M. 1994. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *J. Mol. Evol.* 39:210–18

130. Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized tree. *Mol. Biol. Evol.* 12:823–33

131. Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512–26

132. Tanaka T, Nei M. 1989. Positive Darwinian selection observed at the variable region genes of immunoglobulins. *Mol. Biol. Evol.* 6:447–59

133. Tateno Y, Nei M, Tajima F. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.* 18:387–404

134. Tateno Y, Takezaki N, Nei M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* 11:261–77

135. Templeton AR. 1983. Phylogenetic inference from restriction cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* 37:221–44

136. Thomas RH, Hunt JA. 1993. Phylogenetic relationships in *Drosophila*: a conflict between molecular and morphological data. *Mol. Biol. Evol.* 10:362–74

137. Uyenoyama M. 1995. A generalized least-squares estimate of the origin of sporophytic self-incompatibility. *Genetics* 139:975–92

138. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–7

139. Wainright PO, Hinkle G, Sogin ML, Stickel SK. 1993. Monophyletic origins of metazoa: an evolutionary link with fungi. *Science* 260:340–42

140. Wakeley J. 1993. Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. *J. Mol. Evol.* 37:613–23

141. Williams PL, Fitch WM. 1990. Phylogeny determination using dynamically weighted parsimony method. In *Methods in Enzymology*, ed. RF Doolittle, pp. 615–26. San Diego, CA: Academic

142. Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu. Rev. Biochem.* 46:573–639

143. Wistow G. 1993. Lens crystallins: gene recruitment and evolutionary dynamism. *Trends Biochem. Sci.* 18:301–6

144. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc. Natl. Acad. Sci. USA* 87:4576–79

145. Wu C-I, Li W-H. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* 82:1741–45

146. Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39:105–11

147. Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–14

148. Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.* 43:329–42

149. Yang Z. 1995. *PAML: Phylogenetic Analysis by Maximum Likelihood.* University Park: Inst. Mol. Evol. Genet., Penn. State Univ.

150. Yang Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* 42:294–307

151. Yang Z, Goldman N, Friday AE. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* 44:384–99

152. Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–50

153. Zhang J, Nei M. 1996. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* In press

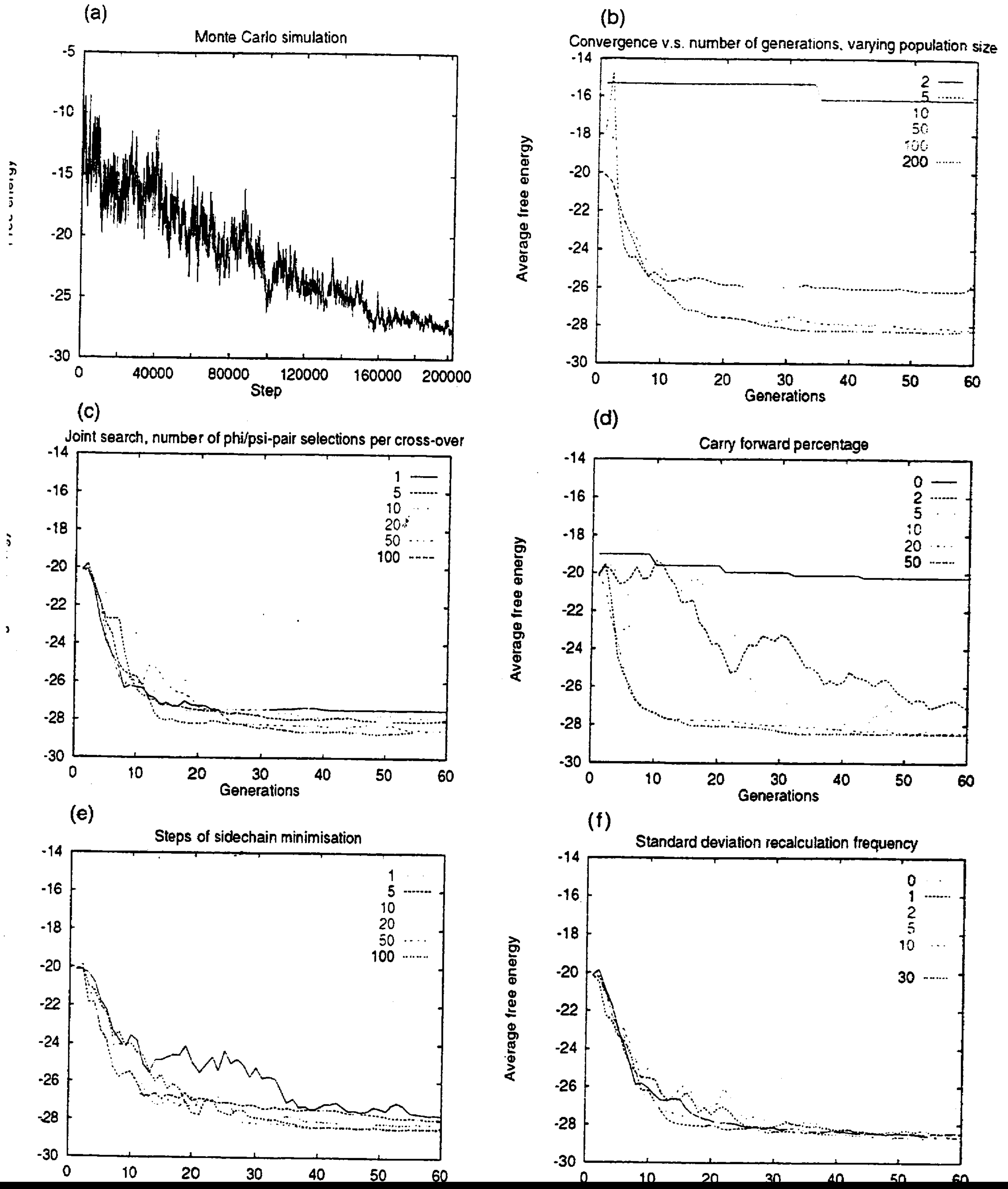
154. Zharkikh A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* 39:315–29

155. Zharkikh A, Li W-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Mol. Biol. Evol.* 9:1119–47

156. Zharkikh A, Li W-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. *J. Mol. Evol.* 35:356–66

157. Zharkikh A, Li W-H. 1993. Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. *Syst. Biol.* 42:113–25

158. Zharkikh A, Li W-H. 1995. Estimation of confidence in phylogeny: complete-and-partial bootstrap technique. *Mol. Phyl. Evol.* 4:44–63


(a)
Monte Carlo simulation
-5
-10
-15
-20
-25
-30
0
40000
80000
120000
160000
200000
Step
(b)
Convergence v.s. number of generations, varying population size
Average free energy
2
5
10
50
100
200
-14
-16
-18
-20
-22
-24
-26
-28
-30
0
10
20
30
40
50
60
Generations
(c)
Joint search, number of phi/psi-pair selections per cross-over
1
5
10
20
50
100
Generations
(d)
Carry forward percentage
Average free energy
0
2
5
10
20
50
Generations
(e)
Steps of sidechain minimisation
1
5
10
20
50
100
(f)
Standard deviation recalculation frequency
Average free energy
0
1
2
5
10
30

[redacted]