Syst. Biol. 42(3):247–264, 1993

SUCCESS OF PHYLOGENETIC METHODS IN THE FOUR-TAXON CASE

JOHN P. HUELSENBECK AND DAVID M. HILLIS

Department of Zoology, University of Texas, Austin, Texas 78712, USA

Abstract.—The success of 16 methods of phylogenetic inference was examined using consistency and simulation analysis. Success—the frequency with which a tree-making method correctly identified the true phylogeny—was examined for an unrooted four-taxon tree. In this study, tree-making methods were examined under a large number of branch-length conditions and under three models of sequence evolution. The results are plotted to facilitate comparisons among the methods. The consistency analysis indicated which methods converge on the correct tree given infinite sample size. General parsimony, transversion parsimony, and weighted parsimony are inconsistent over portions of the graph space examined, although the area of inconsistency varied. Lake's method of invariants consistently estimated phylogeny over all of the graph space when the model of sequence evolution matched the assumptions of the invariants method. However, when one of the assumptions of the invariants method was violated, Lake's method of invariants became inconsistent over a large portion of the graph space. In general, the distance methods (neighbor joining, weighted least squares, and unweighted least squares) consistently estimated phylogeny over all of the graph space examined when the assumptions of the distance correction matched the model of evolution used to generate the model trees. When the assumptions of the distance methods were violated, the methods became inconsistent over portions of the graph space. UPGMA was inconsistent over a large area of the graph space, no matter which distance was used. The simulation analysis showed how tree-making methods perform given limited numbers of character data. In some instances, the simulation results differed quantitatively from the consistency analysis. The consistency analysis indicated that Lake's method of invariants was consistent over all of the graph space under some conditions, whereas the simulation analysis showed that Lake's method of invariants performs poorly over most of the graph space for up to 500 variable characters. Parsimony, neighbor-joining, and the least-squares methods performed well under conditions of limited amount of character change and branch-length variation. By weighting the more slowly evolving characters or using distances that correct for multiple substitution events, the area in which tree-making methods are misleading can be reduced. Good performance at high rates of change was obtained only by giving increased weight to slowly evolving characters (e.g., transversion parsimony, weighted parsimony). UPGMA performed well only when branch lengths were close in length. [Phylogeny estimation; simulation; parsimony; Lake's invariants; UPGMA; neighbor joining; weighted least squares; unweighted least squares; tree space.]

Molecular systematists may choose among over 100 methods of phylogenetic estimation (Swofford and Olsen, 1990; Hillis et al., 1993). One of the goals of systematics research is to winnow this pool of methods, separating those that perform well from those that perform poorly. This testing procedure forms the basis for improving the trees that systematists produce; poor methods are discarded during this procedure, and better methods of phylogeny estimation can be incrementally improved in subsequent cycles. In this paper, we compare the effectiveness of methods of phylogenetic inference for molecular data under a wide variety of conditions and identify those conditions under which particular methods perform well or poorly.

Computer simulations of the efficiency of tree-making methods have become more sophisticated over the past two decades. In general, more recent computer simulations have examined a larger number of methods of phylogenetic inference under a larger number of evolutionary models of sequence evolution (Peacock and Boulter, 1975; Blanken et al., 1982; Tateno et al., 1982; Saitou, 1988; Sourdis and Nei, 1988; Jin and Nei, 1990; Nei, 1991). Previous computer simulations that have examined the performance of phylogenetic methods, however, have explored only a few model phylogenies and branch-length variations. This limitation in those studies is important because the relative performance of methods depends on the conditions under

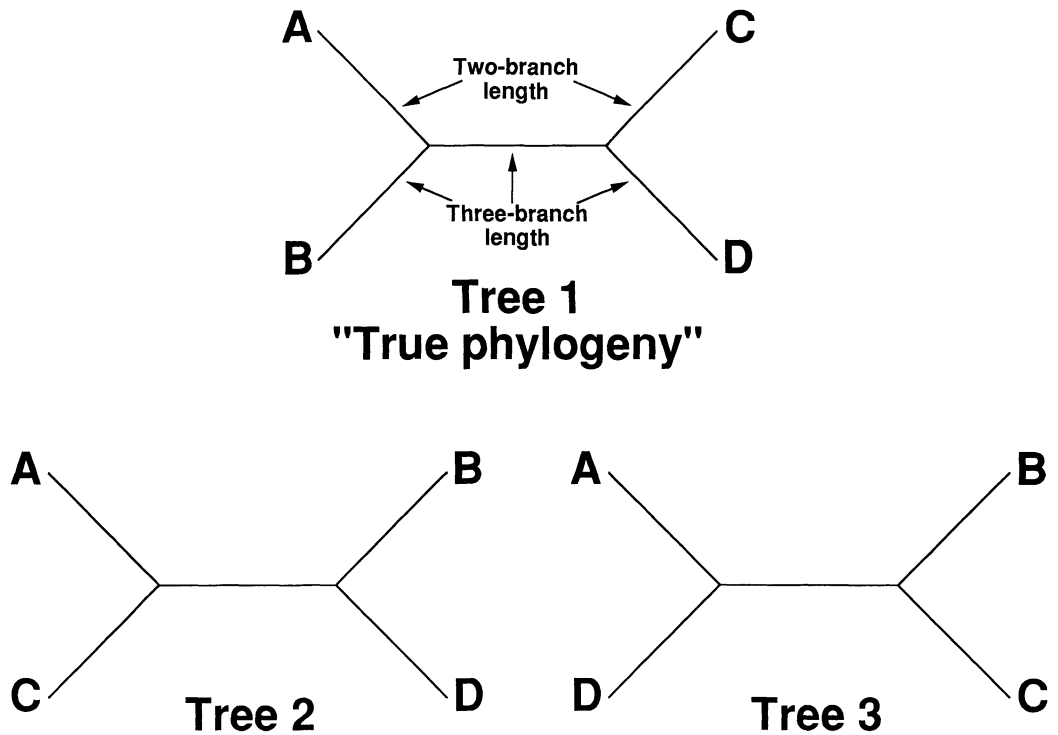


FIGURE 1. For the unrooted four-taxon (A, B, C, D) tree analyzed in this study, the lengths of two sets of branches were varied independently. The internal branch and two peripheral branches were varied together (=three-branch length), as were the remaining two peripheral branches (=two-branch length). Tree 1 represents the simulated or "true" phylogeny, whereas trees 2 and 3 represent the remaining possible phylogenies.

which the simulation was performed; the results of previous computer simulations indicate that one of the most important determinants of the performance of tree-making methods is relative branch lengths.

This study extends earlier simulation studies of the performance of tree-making methods by examining numerous methods under a wide variety of conditions. The performance of methods was examined using both consistency and simulation analysis. In particular, branch lengths were varied in such a way that a large portion of the "tree space," or possible branch-length values, could be explored. This exhaustive approach depicts the relative performance of methods of phylogenetic inference in a fair and informative manner for a given number of taxa under a specified set of conditions. Furthermore, this approach highlights the strengths and weaknesses of tree-making methods and may serve as a basis for the a priori selection of a particular method.

METHODS

Model Trees

In this study, we analyzed unrooted four-taxon trees. The lengths for two sets of branches were varied independently: two of the peripheral branches and the internal branch were equal in length, and the remaining two peripheral branches were equal in length (Fig. 1). The length of a branch represents the percentage of characters that would be expected to change between nodes (see Table 1 for definitions of terms). Very few sites change along the length of a short branch, whereas many sites change along the length of a long branch. When the product of the substi-

tution rate and time is infinite, then 75% of the characters would be expected to differ between the endpoints of a single branch for a four-character-state system, such as was examined in this study.

Our analysis examined all tree space under the constraints of two sets of branch lengths. We chose to constrain the analysis to an exhaustive examination of four taxa with two branch lengths to limit computation expense. The branches for which change was varied concurrently were chosen because previous work suggested that methods of phylogenetic inference have difficulty estimating the true phylogeny under certain branch-length inequalities (Felsenstein, 1978). These branch-length inequalities are encountered in the two-branch-length situation of this study.

The model of sequence evolution employed a substitution matrix (**M**):

	G	A	T	C
G	c	x	z	y
A	x	c	y	z
T	z	y	c	x
C	y	z	x	c

where x, y, and z represent the substitution rate from one base to another and c is the probability of no change ($c = 1 - x - y - z$) (see Swofford and Olsen, 1990; Nei, 1991). Three different models of character evolution were examined in this study (Fig. 2): a Jukes–Cantor (1969) model, a two-parameter (Kimura, 1980) model, and a modified two-parameter model. The Jukes–Cantor model of evolution assumes that all substitution events are equally probable. Under the Jukes–Cantor model, the probability of a substitution occurring is

$$\frac{3(1 - e^{-4\alpha t})}{4}, \quad (1)$$

whereas the probability of no change occurring is

$$\frac{1 + 3e^{-4\alpha t}}{4}, \quad (2)$$

where α is the substitution rate and t is time (Jukes and Cantor, 1969; Swofford and Olsen, 1990). With reference to matrix **M**, $\alpha = x = y = z$ for the Jukes–Cantor model.

The Kimura two-parameter model of sequence evolution treats transitions separately from transversions. Under the Kimura model, the probability of a transition occurring is

TABLE 1. Definitions of the terms used in this paper.

Term	Definition
Accuracy, performance, success	Terms used interchangeably in this paper to describe the frequency with which a tree-making method correctly identifies the true branching relationships
Branch length	The percentage of characters that are expected to change from one end of a branch to the other
Consistency	A consistent phylogenetic method is one that converges on the true tree as the sample size becomes infinite
Felsenstein zone	A term restricted to phylogenetics that describes a general set of conditions under which phylogenetic methods are inconsistent
Substitution rate	The number of substitutions per unit time that occur along a branch of the model phylogeny
Optimality criterion	An objective function that is used to evaluate a given tree; the tree that maximizes or minimizes the function is chosen as the best estimate of phylogeny
Clustering algorithm	A method that adds taxa to a growing tree according to some rule
Three-branch length	The length of the internal branch and two opposing peripheral branches on the model phylogeny
Two-branch length	The length of the remaining two peripheral branches of the model phylogeny
Tree space	The various combinations of branch-length conditions possible for a given set of trees

$$\frac{1 - 2e^{-2(\alpha+\beta)t} + e^{-4\beta t}}{4}, \quad (3)$$

the probability of a transversion occurring is

$$\frac{1 - e^{-4\beta t}}{2}, \quad (4)$$

and the probability of no change occurring is

$$\frac{1 + 2e^{-2(\alpha+\beta)t} + e^{-4\beta t}}{4}, \quad (5)$$

where α is the rate of transitions and β is the rate of transversions (Kimura, 1980; Swofford and Olsen, 1990). With reference to matrix **M**, $\alpha = x$ and $\beta = y = z$.

The last model of evolution examined was a modified Kimura model. The probability of G ↔ C and A ↔ T changes under the modified Kimura model is an equation of the same form as Equation 3, the probability of G ↔ A, G ↔ T, C ↔ A, and C ↔ T changes is an equation of the same form as Equation 4, and the probability of no change occurring is an equation of the same form as Equation 5. With respect to the mutation matrix **M**, now $\alpha = y$ and $\beta = x = z$. Trees were generated using the modified Kimura model of evolution to examine the consequences of violating the assumptions of the Kimura two-parameter distance correction, and these trees represent a model that assumes a difference in the rate of change between G-C and A-T pairs as compared with other types of changes.

Trees were constructed by calculating the probability of the different types of nucleotide substitutions that occur for branches of a given length on the four-taxon model tree. The probabilities derived from the above equations were used to determine character change in the consistency and computer simulation analyses.

Consistency Analysis

A consistent method is one in which an estimated parameter converges on the true value of the parameter as the sample size becomes infinite. For phylogenetic methods, a consistent method is one that estimates the correct tree if the sample size is sufficiently large. The consistency of 16 different methods of phylogenetic inference under three models of evolution was examined using a combined analytical/simulation approach. The consistency of each method was determined in several steps.

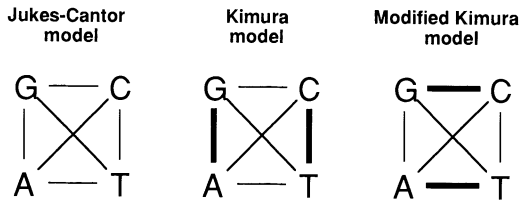


FIGURE 2. Three different models of character evolution were examined in this study: a Jukes–Cantor one-parameter model, a Kimura two-parameter model, and a modified Kimura model of evolution. The Jukes–Cantor model of evolution assumes that all substitution events are equally probable. The Kimura two-parameter model assumes that all transitions are equally proable and all transversions are equally probable. The modified two-parameter model of evolution assumes that G ↔ C and A ↔ T changes are equally probable and that G ↔ A, G ↔ T, C ↔ A, and C ↔ T changes are equally probable.

1. Given a four-taxon tree with known branch lengths and model of evolution, the probability of different substitution events occurring was calculated using the equations given above.
2. We then calculated the probability of observing each of the 256 combinations of four nucleotides that can be assigned to the tips of the tree. This vector of probabilities contains information on the proportion of the time that each combination of nucleotides would be expected to appear, given the assumption of infinite numbers of character data.
3. The tree that would be chosen given the vector of probabilities calculated at step 2 was determined. If the tree chosen represented the true phylogeny, then the method was consistent under the specified branch-length conditions and model of evolution. The method was inconsistent if the incorrect phylogeny was chosen.

4. This procedure was repeated for all of the branch-length combinations examined in this study.

In a simple example to illustrate the procedure outlined above, the probability for just 1 of the 256 combinations of nucleotide assignments to tips of the branches is determined under a Jukes–Cantor model of evolution and for branches that are all 10% in length. A branch length of 10% means that on average 10% of the characters are expected to change between the ends of the branch. Consider the model tree shown in Figure 3, where T_1, T_2, T_3, and T_4 represent the nucleotide states assigned to the tips of the tree and i and j represent the nucleotide states assigned to the internal nodes of the tree. One of the 256 combinations of nucleotides is one in which T_1 and T_2 are assigned G and T_3 and T_4 are assigned C. There are 16 possible assignments of nucleotides to the internal nodes of the tree (nodes i and j). The probability of observing each combination of base pairs at the tips of the four-taxon tree under a given substitution model is given by the summation

$$\sum_{i=1}^{4}\sum_{j=1}^{4} P(i, T_1)P(i, T_2)P(j, T_3)P(j, T_4)P(i, j),$$

where $P(i, T_k)$, $P(j, T_k)$, and $P(i, j)$ are the probability of observing specific nucleotides at the ends of each branch; 1, 2, 3, and 4 represent the nucleotides G, A, T, and C, respectively; and the peripheral branch tips take the value of 1, 2, 3, or 4. In this example, the probability of a substitution occurring is 0.1 (0.0083 for each of the 12 substitutions) and the probability of no change occurring along the length of the branch is 0.9 (0.225 for each of the four possible ways no change would occur, i.e., G ↔ G, A ↔ A, T ↔ T, or C ↔ C). The probability of observing G at nodes T_1 and T_2 and C at nodes T_3 and T_4 is 2.30×10^{-5} for this example. This process would be repeated for the remaining 255 possible combinations of nucleotide assignments.

The consistency of each method was examined under three models of evolution:

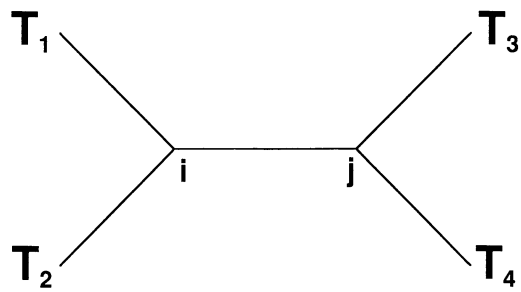


FIGURE 3. The unrooted tree used as the model phylogeny for the consistency study. T_1, T_2, T_3, and T_4 represent the terminal taxa and i and j represent the internal nodes.

(1) equal probabilities of all nucleotide changes, (2) transition : transversion bias in the ratio of five transitions for every transversion, and (3) a mutation bias in which G ↔ C or A ↔ T changes are five times more probable than other changes.

Simulated Phylogenies

In addition to comparing the consistency of phylogenetic methods, six simulation analyses were performed (Table 2). Simulated sequences for the four terminal taxa were constructed in several steps.

1. A random string of nucleotides was generated for an internal node of the unrooted tree. All nucleotides had an equal probability of appearing in the random string.
2. The probabilities of the different possible nucleotide changes given the branch lengths of the model tree were determined using the equations discussed above.
3. Using these probabilities, thresholds between 0 and 1 were constructed where the intervals between thresholds represent different nucleotide changes.
4. A pseudorandom number between 0 and 1 was used to determine which event occurred between nodes of the model tree for a single site.

In a simple example to illustrate the simulation process, begin with a single branch of the model tree, say the internal branch, with length of 10%. In this example, one

TABLE 2. Conditions under which the six different simulation analyses were performed.

Analysis	Number of variable characters	Mutation model
I	10	Jukes–Cantor
II	100	Jukes–Cantor
III	500	Jukes–Cantor
IV	100	Kimura (5:1 transition : transversion bias)
V	100	Kimura (10:1 transition : transversion bias)
VI	100	Modified Kimura (5:1 G ↔ C and A ↔ T : other changes bias)

end of the branch is occupied by the nucleotide G, and a Jukes–Cantor model of sequence evolution is used to describe changes along the length of the branch. The probability of a change occurring along the length of the branch is 0.1, and the probability of no change occurring along the length of the branch is 0.9. These probabilities can be used to construct thresholds between 0 and 1. In this example, the threshold values are 0–0.033 for G → A changes, 0.034–0.067 for G → C changes, 0.068–0.1 for G → T changes, and 0.101–1.0 for G → G (no change). A pseudorandom number is used to determine which of these possible events occur. If, for example, the pseudorandom number is 0.891, then no change occurs at this site and both tips of the branch are occupied by a G. If, however, the pseudorandom number is 0.012, then a change from G to A occurs (one end of the branch is occupied by a G and the other end of the branch by an A at this site). This process is repeated for all of the sites and branches of the model tree.

In each simulation analysis, sequence strings were standardized based on the total number of variable positions, although invariant positions were recorded because of their effect on the various distance measures. The model of evolution was changed for each analysis (Table 2).

One hundred independently constructed trees were examined for each combination of branch lengths for 16 tree-making methods. Analyses I–VI represent over 3 million simulated trees.

Methods Examined

The performances of eight commonly used methods of phylogenetic inference were examined: parsimony (Farris et al., 1970; Fitch, 1971), transversion parsimony (see Swofford and Olsen, 1990), weighted parsimony (Sankoff, 1975), Lake's method of invariants (Lake, 1987), UPGMA (Sokal and Michener, 1958), neighbor joining (Saitou and Nei, 1987), a weighted least-squares criterion (Fitch and Margoliash, 1967), and an unweighted least-squares criterion (Cavalli-Sforza and Edwards, 1967). For the distance methods (UPGMA, neighbor joining, and least-squares criteria), three different distances were used: similarity, the Jukes and Cantor (1969) one-parameter correction, and Kimura's (1980) two-parameter correction. In total, 16 commonly used phylogenetic methods were examined (four discrete data methods and 12 distance data methods). One important method of phylogenetic inference, the maximum-likelihood method (Felsenstein, 1981), was not examined in this simulation because of the computational expense of the likelihood algorithm. Future simulations will examine the performance of the maximum-likelihood method for the graph space examined in this study.

There are several caveats concerning the treatment of the different methods. For UPGMA, an ultrametric method that produces a rooted tree, any rooted tree that was consistent with the simulated unrooted tree used in the analysis was treated as correct. In other words, UPGMA was treated leniently with respect to its ability to retrieve the true phylogeny to facilitate comparison with the other methods that do not need to specify a root. Some debate exists about the best way to treat negative patristic distances, which may be obtained for the weighted and unweighted least-

squares method (Kidd and Sgaramella-Zonta, 1971; Olsen, 1988; Swofford and Olsen, 1990). In this study, patristic distances were calculated using the equations from Kidd and Sgaramella-Zonta (1971), and negative branch lengths were set to 0. One trial set of simulations was performed that allowed negative branch lengths, and the performance of the least-squares methods was considerably worse than when these branch lengths were set to 0.

RESULTS

The results from the consistency and simulated analyses were plotted with length 1 (=three-branch length) as the abscissal value and length 2 (=two-branch length) as the ordinal value. Figure 4 shows, in a general sense, the branch lengths in different parts of the graph space. The diagonal across Figure 4 represents equal branch lengths.

Consistency Analysis

Figure 5 shows the results from the consistency examinations of the tree-making methods, i.e., how different estimation methods (A–P) perform under three different models of evolution (I–III). White areas represent areas in which the methods are consistent, whereas black areas represent areas of the graph space in which the methods are inconsistent. A method is consistent if it converges on the correct answer as more data are added. In Figure 5, areas of consistency represent combinations of branch lengths that result in the correct tree and areas of inconsistency represent combinations of branch lengths that result in an incorrect tree (tree 2 from Fig. 1 is chosen). Felsenstein (1978) first showed that the parsimony method is inconsistent under a Camin–Sokal model of evolution (Camin and Sokal, 1965) even if a Camin–Sokal model of evolution is used as a description of character change on the tree. Felsenstein (1978) predicted that parsimony would be positively misleading when the internal branch and two opposing peripheral branches are very small and the other two branches are very long. We refer to this area where methods perform

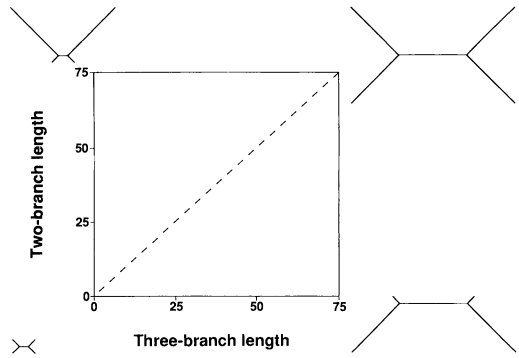


FIGURE 4. The results of the simulations were plotted with the three-branch length on the abscissa and the two-branch length on the ordinate. Different areas of the graph space represent trees with different branch lengths. Change along branches was varied from 1% internodal difference in 1% increments to the maximum length possible (=75% for four-character states). These axes apply to Figures 5–8.

inconsistently as the Felsenstein zone. DeBry (1992) extended consistency analyses by examining the consistency of four phylogenetic methods for the five-taxon case.

Figure 5 shows that parsimony, transversion parsimony, and weighted parsimony all have regions of inconsistency. However, when transitions are evolving at a higher rate than transversions and transversions are weighted more heavily (i.e., transitions are completely discounted or given a reduced weight, as is the case with transversion parsimony or weighted parsimony, respectively), the area of inconsistency becomes slightly smaller. Lake's method of invariants is consistent over all of the graph space examined in this study when the model of evolution matches the assumptions of the invariants method: (1) substitutions are independent, (2) evolution occurs only by substitution, and (3) a balance exists among specific classes of transversions and classes of transitions (Swofford and Olsen, 1990). When the assumptions of balance between specific classes of transitions and transversions is violated, Lake's method of invariants becomes inconsistent over a portion of the graph space.

Figure 5 also shows the performance of

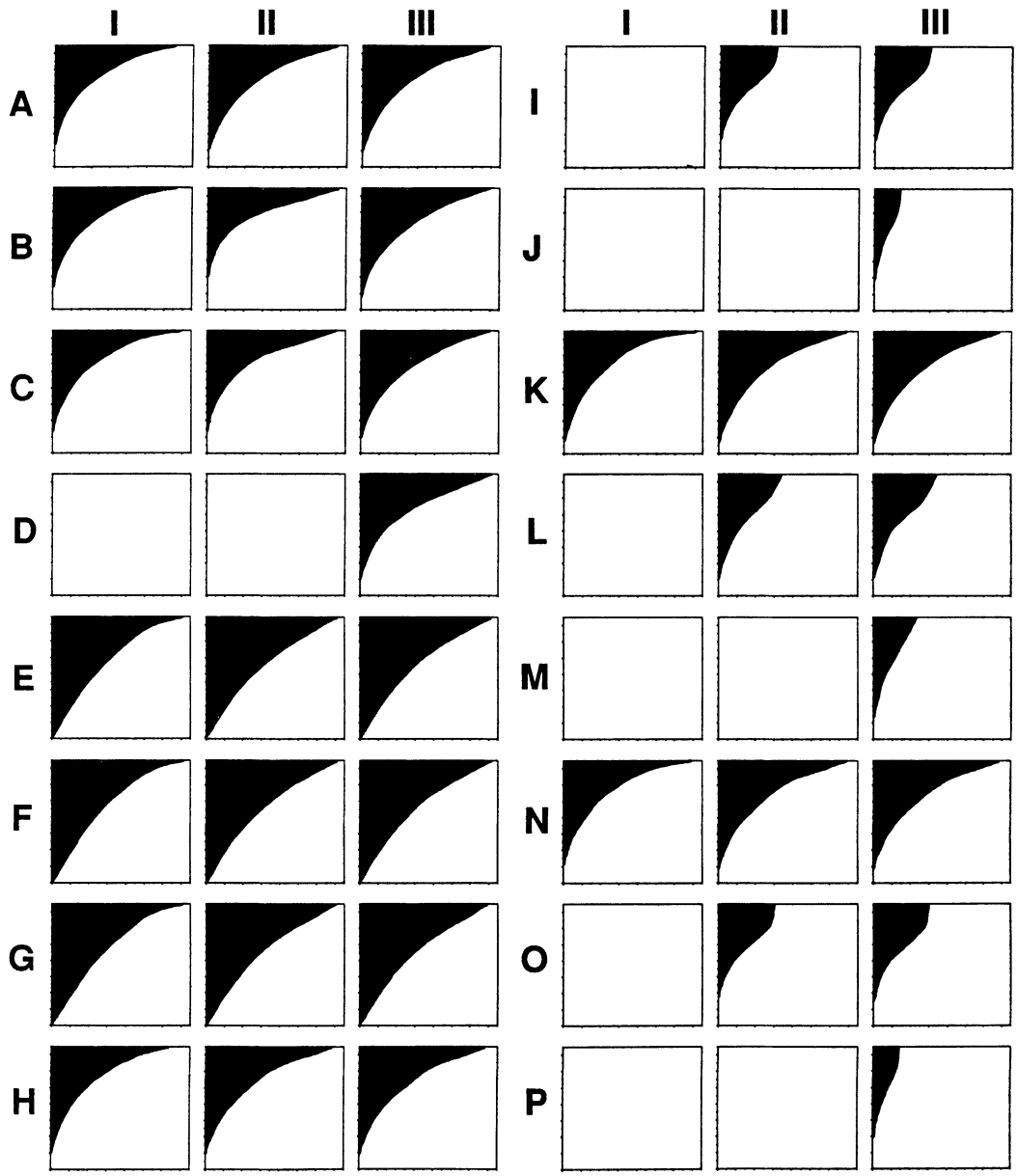


FIGURE 5. Results from the consistency study. White areas of the graph space represent areas of consistency (i.e., the true phylogeny is recovered), whereas black areas of the graph space represent areas of inconsistency (i.e., an incorrect phylogeny is recovered). The consistencies of 16 different phylogenetic methods (A–P) were examined under three models (I–III) of character change. A = parsimony; B = transversion parsimony; C = weighted parsimony; D = Lake's method of invariants; E = UPGMA with similarity distance; F = UPGMA with Jukes–Cantor distance; G = UPGMA with Kimura distance; H = neighbor joining with similarity distance; I = neighbor joining with Jukes–Cantor distance; J = neighbor joining with Kimura distance; K = weighted least squares with similarity distance; L = weighted least squares with Jukes–Cantor distance; M = weighted least squares with Kimura distance; N = unweighted least squares with simlarity distance; O = unweighted least squares with Jukes–Cantor distance; P = unweighted least squares with Kimura distance. I = equal probabilities of all nucleotide changes; II = transition : transversion ratio of 5:1; III = a mutation bias in which G ↔ C and A ↔ T changes are five times more probable than other changes.

the UPGMA, neighbor-joining, weighted least-squares, and unweighted least-squares methods using three different distances. The UPGMA method is inconsistent over a large region of the graph space, no matter which distance is used. The other three distance methods behave similarly to one another. In general, when the processes of evolution match the assumptions of the distances, neighbor-joining, weighted least-squares, and unweighted least-squares methods are consistent over all of the graph space. However, when the assumptions of the distances are violated, the methods are inconsistent over a portion of the graph space. For example, using similarity as a distance measure, the neighbor-joining method is inconsistent over a large region of the graph space for the Jukes–Cantor, Kimura, and modified Kimura models of evolution. When the Jukes–Cantor one-parameter distance correction is used, neighbor joining is consistent under the Jukes–Cantor model of evolution but inconsistent under the Kimura and modified Kimura models of evolution. Similarly, neighbor joining is consistent using the Kimura two-parameter correction when the model of evolution follows a Jukes–Cantor or Kimura model of evolution but is inconsistent when the model of evolution follows a modified Kimura model of evolution. The weighted and unweighted least-squares criteria behave the same way as the neighbor-joining algorithm, although the areas of inconsistency vary slightly in size (i.e., the general conditions under which the least-squares criteria are inconsistent are the same but the size of the areas of inconsistency differ).

Simulated Phylogenies

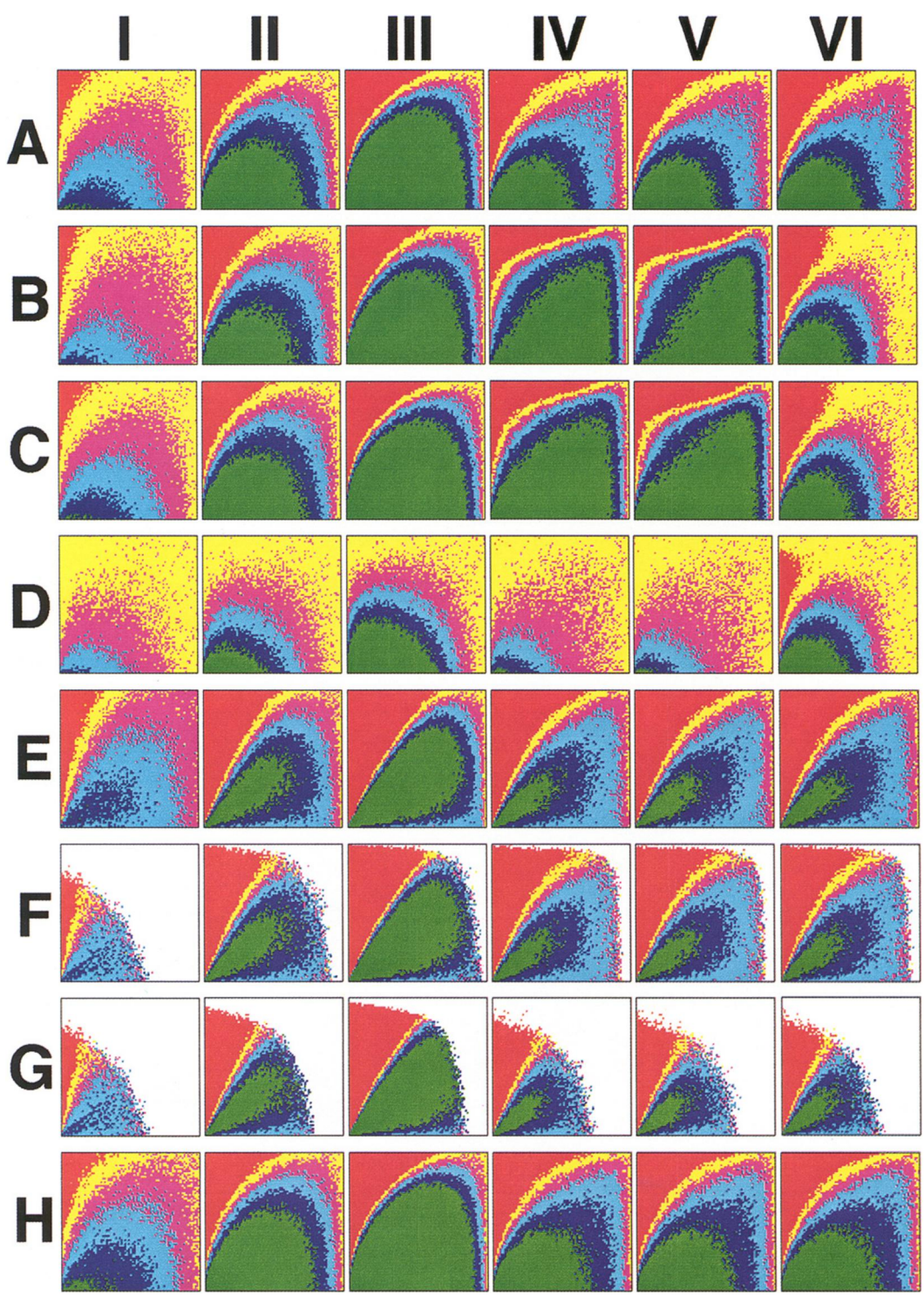
Figures 6 and 7 show the results from the computer simulations, with colors used to indicate relative performances of the different methods. White areas represent branch-length conditions under which the Jukes–Cantor and Kimura distance corrections are undefined in over 90% of the simulations. The performances of 16 different methods of phylogenetic inference (A–P) under six different models of evolution (I–VI) are depicted.

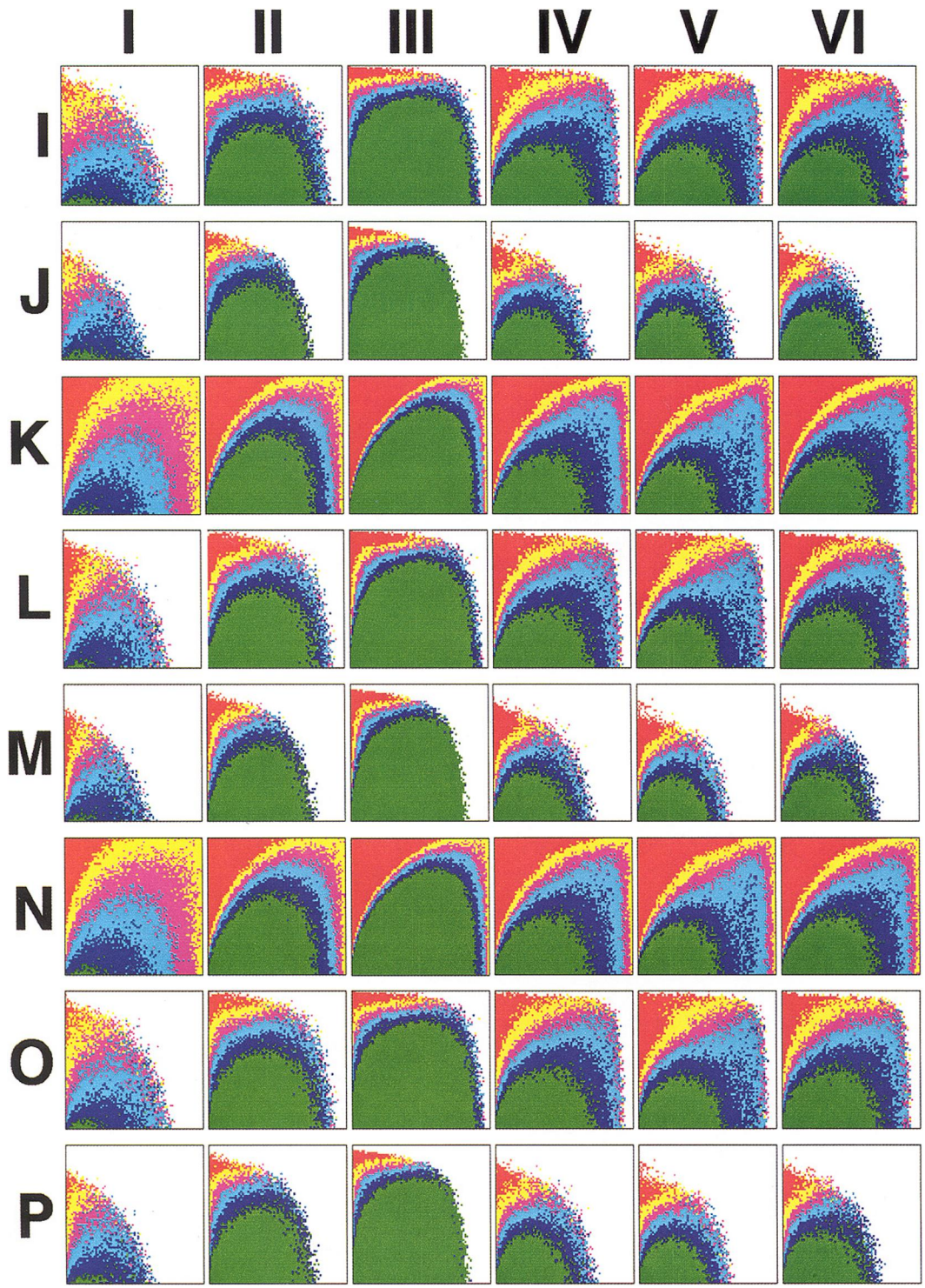
Analyses I, II, and III.—Analysis I, II, and III in Figures 6 and 7 illustrate the effect of addition of characters to the accuracy of phylogenetic analysis. Comparison among these columns shows that the probability of recovering the true phylogeny increases as more character data are added to the analysis. If the true phylogeny corresponds to an area of consistency, as more and more character data are added the methods estimate the true phylogeny with higher frequency over a larger portion of the graph space. If, however, the true phylogeny corresponds to an area of inconsistency, as more and more character data are added phylogeny estimation methods converge on an incorrect solution with higher and higher frequency. For example, the area in which parsimony estimates the true

→

FIGURE 6. Performance of eight different tree-making methods: parsimony (A); transversion parsimony (B); weighted parsimony (C); Lake's method of invariants (D); UPGMA with similarity (E), Jukes–Cantor (F), and Kimura (G) distances; and neighbor joining with similarity distance (H). Simulations were of DNA characters for an unrooted four-taxon tree. Axes are the same as in Figure 4. Sequence strings were standardized based on the total number of variable characters, although invariant characters were recorded because of their effect on the various distance measures. Each graph is a 75 × 75 array in which each point represents 100 independent simulations. The precentage of simulated trees in which the correct tree was chosen is represented by different colors (white = undefined distances; red = 0–20%; yellow = 20–40%; pink = 40–60%; light blue = 60–80%; dark blue = 80–95%; green = 95–100%). For rate-corrected distance methods, the percentage of the time the correct tree was estimated out of at least 10 simulations in which all pairwise distances were defined is plotted. Analysis I = 10 characters, no mutational bias; analysis II = 100 characters, no mutational bias; analysis III = 500 characters, no mutational bias; analysis IV = 100 characters, 5:1 transition : transversion bias; analysis V = 100 characters, 10:1 transition : transversion bias; analysis VI = 100 characters, 5:1 G ↔ C or A ↔ T : G ↔ A, G ↔ T, C ↔ A, or C ↔ T bias.

FIGURE 7. The performance of eight different tree-making methods: neighbor joining with Jukes–Cantor (I) and Kimura (J) distances; weighted least squares with similarity (K), Jukes–Cantor (L), and Kimura (M) distances; and unweighted least squares with similarity (N), Jukes–Cantor (O), and Kimura (P) distances. Figure construction and analyses I–VI are the same as in Figure 6.





phylogeny with high frequency is quite limited if only 10 characters are variable, but the area is much larger when the analysis includes 100 or 500 variable positions. Similarly, parsimony converges on the incorrect phylogeny more strongly as more characters are added (the red area of the figures becomes larger).

Figures 6 and 7 illustrate differences in the effectiveness of methods of phylogenetic estimation.

1. The area in which transversion parsimony works well (i.e., estimates the true phylogeny >95% of the time) is slightly smaller than is the same region for parsimony or weighted parsimony. Weighted parsimony and parsimony are exactly equivalent in analyses I, II, and III because character changes are weighted equally in both cases. Transversion parsimony does not perform as well as parsimony or weighted parsimony under these conditions because it utilizes fewer characters.
2. Lake's method of invariants performs poorly over most of the graph space except in the Felsenstein zone, where it outperforms other methods of phylogenetic inference. Contrast this result with the results from the consistency analyses, which suggest that Lake's method of invariants would be a good choice because it is consistent over all of the graph space under some conditions of evolution (e.g., no mutation bias or transition : transversion mutation bias).
3. Not only is UPGMA inconsistent over a very large portion of the graph space (see Fig. 4), but UPGMA performs poorly over other areas of the graph space except along the diagonal, which represents equal branch lengths. Previous work suggested that UPGMA is sensitive to rate inequalities (e.g., Farris et al., 1970; Mickevich, 1978).
4. Methods that are relatively rate insensitive, such as parsimony, transversion parsimony, weighted parsimony, neighbor joining, and weighted and unweighted least squares, perform approximately equally well over most of the graph space when simple pairwise similarity is used.
5. The Jukes–Cantor one-parameter correction and Kimura's two-parameter correction make the Felsenstein zone smaller. However, these corrections for multiple substitution events do not completely eliminate the Felsenstein zone when the corrections match the processes of evolution perfectly, as the consistency analysis results would suggest. Because of underestimation of distances with finite data, a small area at the top-left corner of the graphs exists in which distance methods do not perform as well as random choice (i.e., the true phylogeny is chosen <33% of the time; the probability of choosing the correct tree at random).

Analyses IV and V.—Analyses I, II, and III did not include a model of mutation bias in their construction. Mutational biases, especially transition–transversion biases, are often observed in analyses of DNA sequence data (e.g., Brown et al., 1982; Gojobori et al., 1982; Li et al., 1984). Analyses IV and V simulated 100 variable characters with a 5:1 and 10:1 transition : transversion bias, respectively. Some methods of phylogenetic inference would be expected to perform better with a transition : transversion bias because they were specifically designed to accommodate this type of bias (e.g., transversion parsimony, weighted parsimony, Lake's invariants, and Kimura corrected distances).

Comparison of analyses IV and V with analysis II (100 variable characters, no mutation bias) reveals several interesting aspects of the behavior of tree-making methods. For example, the Felsenstein zone for transversion parsimony is much smaller than the Felsenstein zone for parsimony. The Felsenstein zone becomes smaller with transversion parsimony because only the more slowly evolving characters are being used in phylogenetic analysis. Weighted parsimony represents a compromise between regular parsimony and transversion parsimony; the Felsenstein zone is larger with weighted parsimony than with transversion parsimony, but the area in which

weighted parsimony estimates the true phylogeny >95% of the time is also much greater. Weighted parsimony either outperforms or is nearly equivalent to unweighted parsimony over the entire graph space. Transversion parsimony and weighted parsimony outperform general parsimony under conditions of extreme rates. The Felsenstein zone of distance methods using the Jukes–Cantor one-parameter correction is much larger when a transition : transversion bias exists. This result is expected from the consistency analysis results, which show that the Jukes–Cantor corrected distance methods are inconsistent when the Jukes–Cantor assumptions are violated.

Analysis VI.—How sensitive are methods such as transversion parsimony, Lake's method of invariants, and distance methods using Kimura's two-parameter correction to violations of the assumed transition : transversion bias? Analysis VI incorporated a mutation model in which the probability of G ↔ C and A ↔ T changes were five times as likely as other mutations. With a 5:1 transition : transversion bias, however, G ↔ A and C ↔ T mutations are five times as likely as other mutations. Transversion parsimony, weighted parsimony, and especially Lake's method of invariants are extremely sensitive to violations of the assumed transition : transversion bias. Transversion and weighted parsimony do not perform as well as parsimony, whereas Lake's method of invariants does not perform as well as random tree choice in situations of extreme rate inequality.

Discussion

The Relationship between Evolutionary Process and Success of Methods

The performance of methods for phylogenetic inference can be viewed as a problem of how well the model of evolution assumed by the estimation method fits the actual processes of evolution. Invariably, because any model of evolution used in phylogeny estimation is a simplification of actual processes, the model of evolution assumed by the estimation method cannot exactly match these processes. The question of the relative performances of different tree-making methods boils down to the robustness of the methods to violations of their underlying assumptions and the degree to which these assumptions are violated in the real world. In many cases, the differences between model and process cause inconsistent results in phylogenetic analysis (e.g., Felsenstein, 1978).

In this study, we examined the performance of tree-making methods under best-case (all assumptions are realized) and worse-case (at least one assumption is violated) situations. It is important to have an idea of how methods of phylogenetic inference behave when one or more of their assumptions are violated because this gives an idea of how the method can be expected to perform in the real world. Over the past decade, numerous studies have shown that some of the basic assumptions of most phylogenetic methods are violated with sequence data. For example, Gojobori et al. (1982) and Li et al. (1984) have shown that the assumption of symmetry of nucleotide substitutions is violated for actual sequence data. Similarly, compensatory mutations in stem regions of ribosomal DNA show that the assumption of character independence is often violated with sequence data (Wheeler and Honeycutt, 1988). The approach advocated in this study provides information on the robustness of phylogenetic methods over a wide range of conditions.

Performance in the Four-Taxon Case

The consistency analyses of this study revealed the conditions under which estimation methods fail with infinite sample size. In general, when the assumptions of a estimation method closely match the processes of evolution, the method is consistent over all of the graph space. Conversely, if the assumptions of the method are violated, the method is typically inconsistent over portions of the graph space. If one were to pick a method based only on the consistency analysis, one might choose methods that are consistent over all of the graph space under at least some conditions.

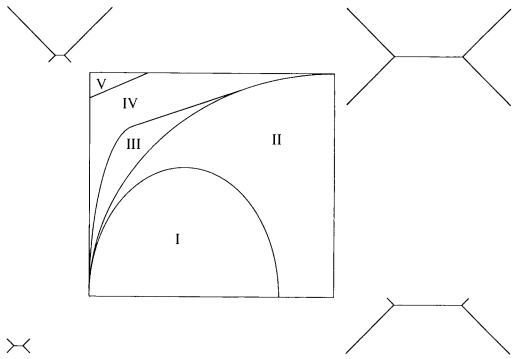


FIGURE 8. Regions of the graph space in which different phylogenetic methods perform best. The *x*-axis represents the three-branch length and the *y*-axis represents the two-branch length.

Methods that were consistent over all of the graph space included neighbor joining, weighted least squares, and unweighted least squares, when the assumptions of the distance corrections are met, and Lake's method of invariants. When the assumptions of these methods are violated, Lake's invariants, neighbor joining, and the two least-squares methods become inconsistent. However, the area of inconsistency is small (compared with the area of inconsistency of parsimony) for the distance methods when just one assumption is violated.

The simulations show the performance of different phylogenetic methods under the condition of limited numbers of character data. The simulation analyses, in many cases, gave results that were quantitatively different from those of the consistency study. For example, the consistency study showed that Lake's method of invariants is a consistent estimator of phylogeny under some conditions. However, the simulation study showed that although Lake's method of invariants may be a consistent estimator of phylogeny (i.e., there is no Felsenstein zone), it is also a very poor estimator of phylogeny given finite data. Lake's method of invariants sacrifices performance in other areas of the graph space to obtain even limited performance in the top-left corner of the graph space examined in this study. Lake's method of invariants was a poor estimator of phylogeny even when trees with 500 variable characters were simulated. Similarly, the neighbor-joining method and weighted and unweighted least-squares criteria were consistent estimators of phylogeny when certain distance corrections were used. However, the simulations also show that under conditions with limited numbers of character data there is a small area in the top-left corner of the graph space (=Felsenstein zone) in which the performance is worse than would be expected from choosing a tree at random. It is important to examine tree-making methods using both analytical and simulation techniques to obtain an accurate picture of performance.

This study also indicates which methods perform well under different branch-length conditions. In Figure 8, the graph space examined in this study is subdivided into five different regions, which varied among methods. These subdivisions are meant as a qualitative assessment of the results of this study. Most methods of phylogenetic inference (except UPGMA and Lake's method of invariants) estimate the true phylogeny with high frequency in region I. This is true for numbers of variable sites that are characteristic of many molecular studies (about 50–100 variable sites). The performance of tree-making methods falls off in region II, an area in which branch lengths are very long. To achieve high performance in region II, many variable sites (>500) must be included in the analysis. Alternatively, the more slowly evolving character-state changes can be weighted more heavily. Many methods are positively misleading in regions III, IV, and V of Figure 8. Weighting the more slowly evolving character substitutions (e.g., as is done with transversion or weighted parsimony) is one way of making this area of inconsistency smaller. Transversion parsimony and weighted parsimony perform relatively well in region III of the graph space. Rate-insensitive distance methods with corrections for multiple substitution events also perform well in the Felsenstein zone; neighbor joining and the least-squares methods with corrected distances perform well in regions III and IV. Lake's

method of invariants was superior to the other tree-making methods examined in region V. However, to achieve even moderate performance in region V, Lake's method of invariants sacrifices performance in other parts of the graph space.

How do the results from this study compare with those of previous simulation analyses of methods of phylogenetic inference? Simulations that have examined the four-taxon case have taken a few model trees along a short transect from equal branch lengths to Felsenstein-like branch lengths (Li et al., 1987; Jin and Nei, 1990; Nei, 1991; Sidow and Wilson, 1991). In general, the conclusions from these simulations are in close agreement with those from the simulations of this study, i.e., Lake's method of invariants performs poorly over most of the graph space and is outperformed by distance methods with corrections for multiple substitution events under many Felsenstein branch-length conditions. Furthermore, these simulations show that Lake's method of invariants is sensitive to violations of its assumptions and that distance methods, when the assumptions of the distance corrections are met, outperform parsimony in the Felsenstein zone. However, when rates of change are high along all branches, methods that give higher weights to the more slowly evolving characters (e.g., transversion parsimony, weighted parsimony) significantly outperform the corrected distance methods. Previous simulations also indicate that the results from this study may hold under some conditions for larger numbers of taxa (Nei, 1991).

Other Considerations

Although the ability of different methods to correctly identify the correct tree under a wide variety of evolutionary conditions is certainly an important criterion by which to judge methods, performance is by no means the only attribute a tree-making method should possess. For example, a method may give the correct branching order but provide poor estimates of branch lengths (Hillis et al., 1992). If branch lengths are the primary consideration, then the performance criterion used herein is inappropriate. Furthermore, ease of calculation and the nature of the method (i.e., whether the method incorporates an optimality criterion or is simply a clustering algorithm; Swofford and Olsen, 1990) are also important criteria to keep in mind when choosing among tree-making methods. The neighbor-joining method estimates four-taxon phylogenies with high frequency under a wide variety of conditions. It can also be a consistent estimator of phylogeny when the assumptions of its distance correction are met. However, Nei (1991) noted that the neighbor-joining method is merely a clustering algorithm for estimating trees under the minimum-evolution criterion (see also Rzhetsky and Nei, 1992). Although clustering algorithms are very fast, they only provide a point estimate of the phylogeny of the group. Suboptimal trees cannot be examined using the UPGMA or neighbor-joining methods but can be examined using methods that include optimality criteria (e.g., parsimony, Lake's method of invariants, or the least-squares criteria). Moreover, as with any heuristic tree-estimation method, the neighbor-joining algorithm does not guarantee an optimal solution under the minimum-evolution criterion.

Branch Lengths Encountered in Real Character Data

How do the conditions examined in this study relate to real systematics problems? Figure 9 shows the branch lengths of two examples from the literature: a study of tetrapod phylogeny based on 18S ribosomal RNA (rRNA) sequence data (Hedges et al., 1990) (Fig. 9a) and a study of lipotyphan mammals based on 12S rRNA gene sequence data (Allard and Miyamoto, 1992) (Fig. 9b). Branch lengths for both data sets were estimated using likelihood (Felsenstein, 1981), assuming a Jukes–Cantor model of sequence evolution and equal frequencies of nucleotides. Only four of the 27 taxa examined in the Hedges et al. (1990) data set were analyzed (mouse: *Mus musculus*; bird: *Turdus migratorius*; lizard: *Sceloporus undulatus*; alligator; *Alligator mississippiensis*), and an unrooted tree consistent

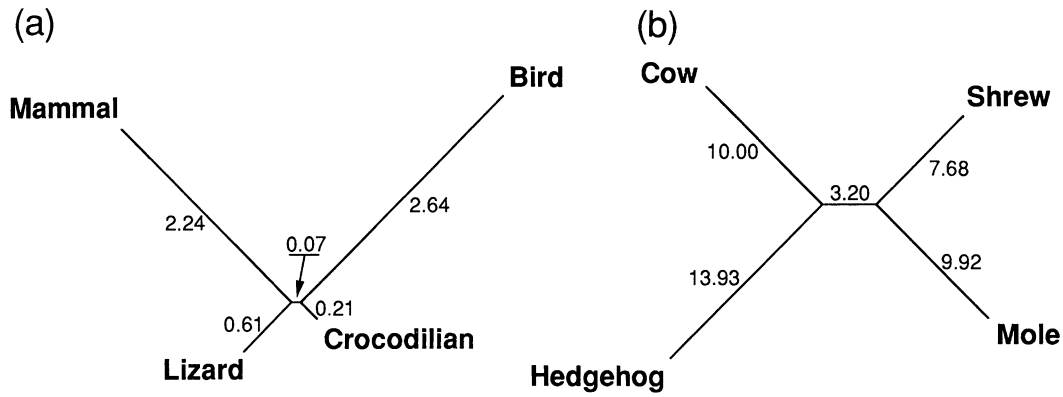


FIGURE 9. Likelihood estimates of the relative branch lengths of (a) tetrapods and (b) lipotyphan mammals.

with the "traditional" view of tetrapod phylogeny was adopted (Gauthier et al., 1988). The same tree assumed to be correct in the Allard and Miyamoto (1992) study was used in this analysis.

Figure 9 shows that the lipotyphan mammal data set probably falls in an area of consistency. The branch lengths, although relatively long (about 10% expected internodal change between each node), fall in an area in which most phylogenetic methods reliably infer the correct phylogenetic tree. In contrast, the phylogeny based on the Hedges et al. (1990) data falls on or near the boundary of consistency/inconsistency. However, it is clear that the character data analyzed in the tetrapod and mammal examples do not closely match the Jukes–Cantor model of evolution used in estimating branch lengths. An important assumption that is violated in these data sets is the assumption of rate homogeneity among sites. This assumption is probably violated in 18S rRNA and 12S rRNA gene sequence data (Hillis and Dixon, 1991) because many sites are invariant across life (or nearly so). The inclusion of invariant sites in a likelihood estimation of branch lengths would cause an underestimation of the actual branch lengths. A parsimony analysis of the bird, mammal, alligator, and lizard sequences using PAUP (Swofford, 1992) results in a tree consistent with a bird–mammal grouping ((bird, mammal)(alligator, lizard)). This tree is very strongly supported: in a bootstrap analysis of these data, the bird–mammal grouping was found in 98.5% of the bootstrap replicates. Two interpretations of these results are possible: (1) the bird–mammal grouping represents the true phylogeny (i.e., the traditional view of tetrapod relationships is incorrect) or (2) the actual tetrapod phylogeny has two very long opposing peripheral branches that attract one another. If interpretation 2 is correct, then the Hedges et al. (1990) data provide an interesting example of a tree evolving under conditions that result in an inconsistent analysis under the parsimony criterion.

Computer Simulation in Phylogenetic Analysis

The testing of tree-making procedures with reference to known phylogenies and processes is an important step in improving methods of phylogenetic inference. However, it is important that the testing be performed in a manner that shows those conditions under which methods perform well and those conditions under which methods perform poorly. The exhaustive approach taken in this study is an attempt to accurately and fairly portray the performance of a large number of tree-making methods for the four-taxon case. This study also suggests several avenues of research that future simulations of tree-making methods could take. We did not examine a large number of tree-making methods and distance corrections (e.g., maximum

likelihood, the minimum-evolution criterion, and the three-, four-, and six-parameter distance corrections). We also did not examine the performance of different methods for a larger number of taxa under a wide variety of branch-length conditions. Future simulations could examine other tree-making methods under a wider variety of conditions.

ACKNOWLEDGMENTS

This paper was improved through the comments of Jim Bull, David Cannatella, Paul Chippindale, Clifford Cunningham, Michael Miyamoto, David Swofford, and two anonymous reviewers. We especially thank Jim Bull and Clifford Cunningham for their advice on simulation and methodological matters. Computer programs for this study were written using a Macintosh computer supplied by Project Quest. This work was supported by NSF grants DEB-9106746 and DEB-9221052.

REFERENCES

ALLARD, M. W., AND M. M. MIYAMOTO. 1992. Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. Mol. Biol. Evol. 9:778–786.

BLANKEN, R. L., L. C. KLOTZ, AND A. G. HINNEBUSCH. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J. Mol. Evol. 19:9–19.

BROWN, W. M., E. M. PRAGER, A. WANG, AND A. C. WILSON. 1982. Mitochondrial DNA sequences in primates: Tempo and mode of evolution. J. Mol. Evol. 18:225–239.

CAMIN, J. H., AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution 19:311–326.

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phlogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet. 19:233–257.

DEBRY, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol. Biol. Evol. 9:537–551.

FARRIS, J. S., A. G. KLUGE, AND M. J. ECKARDT. 1970. A numerical approach to phylogenetic systematics. Syst. Zool. 19:172–191.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

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

FITCH, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20:406–416.

FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 155:279–284.

GAUTHIER, J., A. G. KLUGE, AND T. ROWE. 1988. Amniote phylogeny and the importance of fossils. Cladistics 4:105–209.

GOJOBORI, T., W.-H. LI, AND D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360–369.

HEDGES, S. B., K. D. MOBERG, AND L. R. MAXSON. 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

HILLIS, D. M., M. W. ALLARD, AND M. M. MIYAMOTO. 1993. Analysis of DNA sequence data: Phylogenetic inference. Methods Enzymol. 224:456–487.

HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT, AND I. J. MOLINEUX. 1992. Experimental phylogenetics: Generation of a known phylogeny. Science 255:589–592.

HILLIS, D. M., AND M. T. DIXON. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. Q. Rev. Biol. 66:411–453.

JIN, L., AND M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7:82–102.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. Munro, ed.). Academic Press, New York.

KIDD, K. K., AND L. A. SGARAMELLA-ZONTA. 1971. Phylogenetic analysis: Concepts and methods. Am. J. Hum. Genet. 23:235–252.

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

LAKE, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. Mol. Biol. Evol. 4:167–191.

LI, W.-H., K. H. WOLFE, J. SOURDIS, AND P. M. SHARP. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harbor Symp. Quant. Biol. 52:847–856.

LI, W.-H., C.-I. WU, AND C.-C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. 21:58–71.

MICKEVICH, M. F. 1978. Taxonomic congruence. Syst. Zool. 27:143–158.

NEI, M. 1991. Relative efficiences of different tree-making methods for molecular data. Pages 90–128 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, Oxford, England.

OLSEN, G. J. 1988. Phylogenetic analysis using ribosomal RNA. Methods Enzymol. 164:793–838.

PEACOCK, D., AND D. BOULTER. 1975. Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J. Mol. Biol. 95:513–527.

RZHETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.

SAITOU, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. 27:261–273.

SAITOU, N., AND M. NEI. 1987. The neighbor-joining

method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

SANKOFF, D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28:35–42.

SIDOW, A., AND A. C. WILSON. 1991. Compositional statistics evaluated by computer simulations. Pages 90–128 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, Oxford, England.

SOKAL, R. R., AND C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 28:1409–1438.

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

SWOFFORD, D. L. 1992. PAUP: Phylogenetic analysis using parsimony, version 3.0s. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L., AND G. J. OLSEN. 1990. Phylogeny reconstruction. Pages 411–501 *in* Molecular systematics (D. M. Hillis and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.

TATENO, Y., M. NEI, AND F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. 18:387–404.

WHEELER, W. C., AND R. L. HONEYCUTT. 1988. Paired sequence difference in ribosomal RNAs: Evolutionary and phylogenetic implications. Mol. Biol. Evol. 5:90–96.

Received 23 October 1992; accepted 2 March 1993