Syst. Biol. 47(2): 228–253, 1998

Phylogenetic Analysis and Intraspecific Variation: Performance of Parsimony, Likelihood, and Distance Methods

JOHN J. WIENS[1] AND MARIA R. SERVEDIO[2]

[1]*Section of Amphibians and Reptiles, Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213-4080, USA; E-mail: wiensj@clpgh.org*
[2]*Department of Zoology, University of Texas, Austin, Texas 78712-1064, USA; E-mail: mservedio@mail.utexas.edu*

Abstract.—Intraspecific variation is abundant in all types of systematic characters but is rarely addressed in simulation studies of phylogenetic method performance. We compared the accuracy of 15 phylogenetic methods using simulations to (1) determine the most accurate method(s) for analyzing polymorphic data (under simplified conditions) and (2) test if generalizations about the performance of phylogenetic methods based on previous simulations of fixed (nonpolymorphic) characters are robust to a very different evolutionary model that explicitly includes intraspecific variation. Simulated data sets consisted of allele frequencies that evolved by genetic drift. The phylogenetic methods included eight parsimony coding methods, continuous maximum likelihood, and three distance methods (UPGMA, neighbor joining, and Fitch–Margoliash) applied to two genetic distance measures (Nei's and the modified Cavalli-Sforza and Edwards chord distance). Two sets of simulations were performed. The first examined the effects of different branch lengths, sample sizes (individuals sampled per species), numbers of characters, and numbers of alleles per locus in the eight-taxon case. The second examined more extensively the effects of branch length in the four-taxon, two-allele case. Overall, the most accurate methods were likelihood, the additive distance methods (neighbor joining and Fitch–Margoliash), and the frequency parsimony method. Despite the use of a very different evolutionary model in the present article, many of the results are similar to those from simulations of fixed characters. Similarities include the presence of the ''Felsenstein zone,'' where methods often fail, which suggests that long-branch attraction may occur among closely related species through genetic drift. Differences between the results of fixed and polymorphic data simulations include the following: (1) UPGMA is as accurate or more accurate than nonfrequency parsimony methods across nearly all combinations of branch lengths, and (2) likelihood and the additive distance methods are not positively misled under any combination of branch lengths tested (even when the assumptions of the methods are violated and few characters are sampled). We found that sample size is an important determinant of accuracy and affects the relative success of methods (i.e., distance and likelihood methods outperform parsimony at small sample sizes). Attempts to generalize about the behavior of phylogenetic methods should consider the extreme examples offered by fixed-mutation models of DNA sequence data and genetic-drift models of allele frequencies. [Accuracy; distance methods; maximum likelihood; parsimony; polymorphism; simulations.]

Intraspecific variation is ubiquitous in systematic characters, including morphology, allozymes, and DNA sequences (e.g., Alberch, 1983; Kreitman, 1983; Buth, 1984). Systematists deal with intraspecific variation in many different ways, including throwing out intraspecifically variable characters (often done with morphological data), sampling a single individual per species such that little or no variation is observed (with sequence data), treating each individual or haplotype as a separate terminal taxon (with sequence and restriction-site data), and using parsimony, distance, and likelihood methods designed for polymorphic data (see Wiens, 1995; Swofford et al., 1996). There has been considerable debate as to which of the methods for directly analyzing polymorphic data is superior (e.g., Farris, 1981; Felsenstein, 1981, 1985; Mickevich and Mitter, 1981, 1983; Rogers, 1986; Swofford and Berlocher, 1987; Crother, 1990; Swofford and Olsen, 1990; Campbell and Frost, 1993; Mabee and Humphries, 1993; Murphy, 1993; Wiens, 1995; Swofford et al., 1996; Wiens and Servedio, 1997). The choice among these methods is a serious concern because intraspecific variation is so widespread,

and because the application of different methods can give radically different trees for the same data set. Even subtle differences in how polymorphism is treated (e.g., coding a variable species as having the most common condition versus coding the derived state) can have a significant impact on tree topology (Wiens, 1995).

Results from computer simulations offer an important basis for choosing a phylogenetic method to apply to a given problem (Huelsenbeck, 1995). Because simulations use a known phylogeny, the criterion of accuracy (how well the method recovers the true phylogeny) can be used to choose among methods (Hillis, 1995). Furthermore, because simulations use simplified models, parameters affecting method performance can be varied systematically and thereby understood (e.g., Hillis, 1995; Huelsenbeck, 1995). Simulation of phylogenetic method performance has become an active area of research, yet virtually all recent simulation studies have implicitly assumed no variation within species (but see Kim and Burgman, 1988; Rohlf and Wooten, 1988). Wiens and Servedio (1997) compared a number of parsimony methods for excluding, including, coding, and weighting polymorphic characters using simulations of polymorphic data. They found that methods that include polymorphic characters and incorporate frequency information are generally the most accurate. However, they compared only parsimony methods and examined only a limited set of simulated conditions. This leaves open two questions: (1) What are the most accurate methods for phylogenetic analysis of polymorphic data? (2) Are generalizations about method performance based on simulations of fixed characters robust to an evolutionary model based on intraspecific variation? In this study we use simulations to test the accuracy of 15 methods for analyzing polymorphic data (eight parsimony coding methods, six genetic-distance methods, and continuous maximum likelihood).

MATERIALS AND METHODS

Simulated Data Sets

The simulated data sets consist of allele frequencies for unlinked loci with two or three neutral alleles per locus, evolving by genetic drift along a known phylogeny of four or eight species (Fig. 1). Geographic variation within species is not included in the evolutionary model. The simulated data represent generalized, discrete, heritable characters. They do not represent a specific kind of molecular or morphological data, so that many features that characterize specific data types have not been incorporated in the simulations. For example, the simulations do not include (1) differences between genotype and phenotype or the presence of multiple loci controlling a single trait as expected in morphological characters, (2) the presence of many alleles per locus as in allozyme data and many types of DNA data, and (3) nonindependence of character state frequencies among characters as in DNA sequence data. As in real data sets, a given locus may be fixed or polymorphic across all species or, most commonly, fixed in some species and polymorphic in others.

In the two-allele case, the data are generated using the following protocol for each locus. We follow the frequency of one of the two alleles throughout the description that follows. First, the initial frequency of one of the alleles at the starting point (the internal node X; Fig. 1) is selected randomly from a uniform distribution (i.e., any number from 0 to 1 inclusive has an equal probability of being selected). From this starting fre-

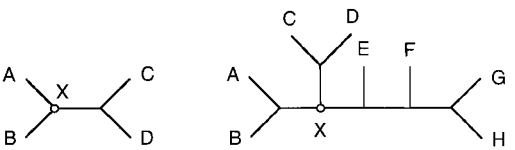


FIGURE 1. Topologies of the true four- and eight-taxon phylogenies used in simulations. X is the starting point used for determining initial frequencies.

quency, the allele frequency changes independently along each branch by the Wright–Fisher model of genetic drift (Fisher, 1930; Wright, 1931). The calculation of the frequency at the end of any branch occurs in two steps. (1) The probabilities of fixation and loss are calculated following Kimura (1955), using the initial frequency, the population size, and the number of generations for the lineage. A random number is then picked to determine if the allele will be fixed, lost, or if the locus will remain polymorphic at the end of the branch, based on these probabilities. If the allele is lost or fixed, no further changes occur. (2) If the locus remains polymorphic, the terminal frequency for the allele in that lineage is determined as follows. The probability density function for the allele frequency is calculated (given the initial frequency, population size, and number of generations), following Kimura (1955). The rejection method for generating random deviates (Press et al., 1988) is used to pick the final frequency, based on this distribution. A new lineage begins its evolution with the terminal frequency of its ancestors.

In parallel fashion, for the three-allele case, the initial frequencies of the alleles are first chosen randomly. As in step 1 already given, the alleles present at the ends of branches are determined using probabilities in Kimura (1955). If three alleles are present, their new frequencies are determined using a three-dimensional probability density function (Kimura, 1956). If only two alleles are present, their frequencies are determined using an approximation of the two-dimensional probability density function based on the initial frequencies of all three alleles (Kimura, 1955). In this case, the approximation can sometimes be inaccurate when branch lengths are short (ratio of population size to number of generations of 0.2 in the four-taxon case and 0.2, 0.5, and 0.8 in the eight-taxon case). To avoid this problem, we used the approximation method from the two-allele case at these lengths (to make the frequencies sum to 1.0 after the loss of an allele, the frequencies of the remaining two alleles were increased in proportion to their frequencies). As in the two-allele case, lineages begin with the terminal frequencies of their ancestors. The effects of more than three alleles per locus were not examined because the model of genetic drift becomes extremely complex (Kimura, 1956).

Mutations are assumed to be rare enough (over the limited time frame of the simulation) to be ignored, so loci that become fixed for an allele remain fixed. Given a two-allele model and the simulated population size (50,000 diploid individuals), mutation rates would have to be high for mutation to predominate over drift in determining allele frequencies (Kimura and Crow, 1964). Furthermore, we consider the most likely outcomes of mutation in the two-allele model to be (1) slowing the rate of change in allele frequencies at polymorphic loci (given a locus with two alleles, mutation and drift have opposing effects on allele frequencies) and (2) the rare generation of polymorphisms from fixed loci, since new mutations must realistically start at very low frequencies and are likely to be lost.

The position of the starting point (Fig. 1) of the simulations could influence the levels of polymorphism in different parts of the tree (i.e., less polymorphism further from the starting point). We therefore placed the starting point close to the ''center'' of the tree, to minimize this effect (Fig. 1). The branching pattern of the true phylogeny in the eight-taxon case (Fig. 1) is intermediate in its level of symmetry or balance, to avoid biasing the results by using a tree shape that is particularly easy or difficult to estimate correctly (e.g., Fiala and Sokal, 1985; Rohlf et al., 1990). A limited number of simulations was performed to test the effects of using maximally symmetric and asymmetric eight-taxon trees.

Two main sets of simulations were performed. The first set examined the effects of four parameters in the eight-taxon

case: number of loci (= characters), maximum number of alleles per locus, sample size (number of individuals sampled per species), and branch lengths. The second set of simulations examined the effects of branch lengths more extensively in the four-taxon, two-allele case.

Branch length, defined here as the expected amount of character change for a lineage, determined both the rate of change in frequencies and the probabilities of alleles becoming fixed or lost. Under the Wright–Fisher model of genetic drift, the change in the frequency of a neutral allele is a function of the number of generations during which evolution occurs divided by the effective population size (Kimura, 1955). Throughout the paper we use this ratio synonymously with branch length. Differences in branch length can be viewed as: (1) differences in the number of generations between splitting events, (2) differences in effective population size between lineages, with longer branch lengths being the result of (for example) founder effects or population bottlenecks, or (3) a combination of the two. In our calculations we used an effective population size of 50,000 diploid individuals. Branch lengths ranged from a minimum of 0.1 to a maximum of 2.0 (5,000 to 100,000 generations). At the shortest branch lengths there is a low probability of fixation or loss and generally only small changes in frequency. At the longest branch lengths the chances of fixation and loss are very high, and for those loci that remain polymorphic the allele frequencies are effectively randomized between splitting events (see Kimura, 1955). More extreme branch lengths could have been examined. However, at shorter branch lengths, calculations of frequencies become extremely time consuming. At longer branch lengths, fixation and loss result in no polymorphism. A limited set of analyses with branch lengths of 10.0 showed that the results are very similar to those using branch lengths of 2.0 (Wiens and Servedio, unpublished data). The branch lengths examined include a broad range of levels of variability. For example, given a sample of 50 loci for eight taxa and 100 replicated matrices, 95.1% of the loci are polymorphic within one or more species when all branches have a length of 0.2, 82.0% for a length of 0.8, 65.2% for 1.4, and 46.3% for 2.0.

In the first set of analyses (eight-taxon case), branch lengths were both (1) held constant across all the lineages in order to test the effects of a given branch length on accuracy, and (2) varied randomly across lineages, to generate conditions more like those presumably occurring in nature. In both cases, the length of each branch was held constant across loci for that lineage (i.e., there are no differences in number of generations or effective population size between loci in a given species). In the case of random branch lengths, the possible lengths included 0.2, 0.5, 0.8, 1.1, 1.4, 1.7, and 2.0, with an equal probability of each length being selected.

To examine the effects of sample size for a given set of conditions, five sets of 100 matrices each were created, one set with perfect sampling (every individual in the species sampled) and one each with reduced sample sizes of 10, 5, 2, and 1 diploid individual(s) per species. To create a matrix with a reduced sample size of x, x diploid individuals were sampled from the original matrix. This was done by randomly choosing two alleles per individual from the original matrix based on the allele frequencies. This sampling protocol is based on the idea that, given simplified conditions, the probability of sampling an allele is proportional to its frequency in the population (but see Rannala [1995] for a discussion of more complicated situations).

The first set of simulations addressed the performance of methods under all possible combinations of the following parameters: branch lengths (random, 0.2, 0.8, 1.4, and 2.0), sample size (perfect sampling, and $n = 10$, 5, 2, and 1 individ-

ual sampled per species), number of loci (10, 25, 50, 75, 100), and number of alleles per locus (2, 3). One hundred data matrices were generated for each set of conditions. This number was used because (1) there seems to be very little random variation in method performance for similar conditions (which suggests that 100 is adequate), and (2) simulations were extremely time-intensive at low branch lengths.

The second set of analyses explored the effects of branch lengths in more detail, using an unrooted four-taxon tree with the lengths of two sets of branches varied in increments (the general protocol used by Felsenstein [1978] and subsequent authors). One branch length was used for the internal branch and the terminal branches for species A and C, and the other was used for the terminal branches for species B and D (Fig. 1). Branch lengths were varied from a ratio of 0.1 to 1.9, in a total of 10 increments (0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9). These branch lengths were examined in the two-allele case for 10, 50, and 500 characters. Allele frequencies were assumed to be sampled without error. Results are presented in the graphical format used by Felsenstein (1978), Huelsenbeck and Hillis (1993), and others, to facilitate comparison of our results. A limited number of simulations was performed to compare the effects of extreme branch length variation in the two- and three-allele cases for 100 characters; the three-allele case was not used more extensively because three-allele simulations are extremely time-intensive for lower branch lengths. As in the first set of simulations, 100 matrices were generated for each set of conditions.

The protocol for generating frequency data was designed by Servedio and Wiens, and was programmed by Servedio in the C programming language. The implementation of Kimura's (1955) models in the simulations was checked using Mathematica (Wolfram, 1991). The programs for coding and subsampling the data were written in C by Wiens. The programs are available from the authors upon request.

Methods Examined

A total of 15 phylogeny-estimation methods were examined. These consisted of eight parsimony coding methods (Table 1), continuous maximum likelihood (Felsenstein, 1981), and six genetic distance methods (i.e., combination of tree-building method and genetic distance measure). The tree-building methods were UPGMA (Sokal and Michener, 1958), neighbor joining (Saitou and Nei, 1987), and Fitch–Margoliash (Fitch and Margoliash, 1967; or weighted least squares), applied to the genetic distances of Nei (1972) and the modified Cavalli–Sforza and Edwards (1967) chord distance (hereafter, chord distance). Many more tree-building methods and

TABLE 1. Summary of methods for coding polymorphic characters for parsimony analysis (from Wiens, 1995), where 0 is the primitive condition, 1 is derived, and 0/1 indicates polymorphism. Terminology largely from Campbell and Frost (1993).

Method	Summary
Any instance	0/1 or 1 = 1
Majority	if frequency of 1 ≥ 50%, then 0/1 = 1, otherwise 0/1 = 0
Scaled	0 = 0, 0/1 = 1, 1 = 2; ordered 0 → 1 → 2, change from 0 → 2 is two steps
Unordered	same as scaled, but unordered
Unscaled	same as scaled (ordered), but change from 0 → 2 is one step
Missing	0/1 = ?
Polymorphic	0/1 = (0, 1) either 0 or 1 depending on tree
Frequency	0/1 = weight based on frequency of trait 1

genetic distance measures could have been included in this analysis, but we restricted our analyses to these six because they are either widely used (e.g., UPGMA clustering of Nei's genetic distance) or because they are thought to perform well based on previous studies (e.g., neighbor joining and Fitch–Margoliash; Huelsenbeck, 1995; Cavalli–Sforza and Edwards modified chord distance; Felsenstein, 1985; Rogers, 1986). It should be understood that the performance of the tree-building methods found in this study may apply only to the distance measures we examined (e.g., UPGMA may not perform as well using ''overall similarity'' as a measure of distance). Continuous maximum likelihood and the chord distance explicitly assume no mutation and no fixation or loss of alleles, but do not assume equal rates of change among lineages (Felsenstein, 1985). Nei's distance assumes equal rates of change among lineages, but incorporates mutation (at equal rates among loci) and fixation and loss of alleles (Felsenstein, 1985).

The eight parsimony coding methods are summarized in Table 1, and are described in Wiens (1995). For the frequency method, each taxon was given a unique character state and the Manhattan distance between each pair of species for each character was used to weight changes between these states using a step matrix (following Wiens, 1995; suggested by D. Hillis). This method is an heuristic approximation of the FREQPARS approach of Swofford and Berlocher (1987) and is discussed by Berlocher and Swofford (1997). The scaled method was implemented in the three-allele case using the step-matrix approach of Mabee and Humphries (1993). For the majority (or ''modal'') method in the three-allele case, taxa were coded as unknown if only two alleles were present and occurred at equal frequencies.

Parsimony analyses were implemented using test versions of PAUP* (provided by D. L. Swofford), using the branch-and-bound search option. Distance and likelihood analyses were performed with PHYLIP 3.57c (Felsenstein, 1995) using the programs Gendist, Fitch, Neighbor, and Contml. UPGMA and neighbor joining do not have optimality criteria and therefore they always find the ''best'' neighbor joining and UPGMA tree (Swofford et al., 1996). For maximum likelihood and the Fitch–Margoliash method optimal trees were sought using the ''global rearrangements'' option. In the eight-taxon case, 10 different taxon-addition sequences were also used for each matrix. A set of analyses using 20 sequences per matrix showed little difference in the results, suggesting that 10 sequences should be sufficient to find the optimal tree. All trees were considered to be unrooted, and UPGMA was treated as estimating unrooted trees.

For each set of conditions, the accuracy of methods was scored as the similarity between the true phylogeny (Fig. 1) and the estimated tree (or the strict consensus of the shortest estimated trees for parsimony), averaged across the 100 replicated data sets. Similarity was measured using the consensus fork index of Colless (1980), the proportion of nodes in common between the true and estimated trees. We consider method success to be the proportion of correctly resolved nodes, and we did not give parsimony methods credit for having the correct clade among multiple shortest solutions for a given node (in empirical studies these would likely be rejected as being unsupported), although this convention may put methods that give poorly resolved estimates at a disadvantage. There are many different ways that accuracy can be measured in simulation studies, and most of these penalize methods to some degree for failing to produce a single shortest estimate of the true tree (Hillis et al., 1994).

Not every method could be applied to every data set. The any-instance and unscaled methods were not applied in the three-allele case, because they were described in the case where a character has a single derived condition (Campbell

and Frost, 1993). Felsenstein's (1981) Contml (continuous maximum likelihood) program crashes when there are two identical species in the matrix; such data matrices (usually occurring when there are long branches and small sample sizes) were excluded from maximum-likelihood analyses, and the results for this method for certain conditions are based on fewer than 100 data sets. This exclusion appears to have little impact on the results. The problem also occurred (to a lesser extent) using the Fitch–Margoliash method.

RESULTS

The Eight-Taxon Case

The first set of simulations (eight taxa) involved a large and varied parameter space (250 conditions). A limited sample of the results are shown in Figs. 2 and 3; the complete results are summarized graphically in the Appendix. Across all the conditions examined, the methods that are consistently most accurate are maximum likelihood, the six distance methods, and the frequency parsimony method. The performance of these methods is generally similar across the different conditions examined, but the distance and likelihood methods consistently outperform all the parsimony methods when sample sizes are small (one or two individuals sampled per species). We found few consistent differences in the accuracy of the distance methods using Nei's versus the Cavalli-Sforza and Edwards chord distance, despite the very different assumptions of these distance measures.

The frequency method was almost always the most accurate parsimony method. The scaled parsimony method performs as well as or slightly better than the frequency parsimony method under some conditions, particularly when branch lengths are equal. In general, the conditions where the scaled approach is most similar or superior to the frequency approach are also conditions where both methods perform very well (e.g., accuracy $>90\%$). The performance of the unscaled method is very similar to the scaled method, but is slightly less accurate under some conditions, particularly those in which there is little polymorphism (e.g., long branch lengths). The any-instance, unordered, majority, missing, and polymorphic parsimony methods are less accurate than the other methods under almost all conditions. The missing and polymorphic methods consistently performed very poorly.

Branch lengths, sample size, and number of characters and alleles all influenced the performance of methods, and sometimes interacted in complex ways. As expected, the number of characters was a crucial component of phylogenetic accuracy. All methods had difficulty in recovering the correct phylogeny with only 10 characters, but many methods could be quite accurate ($>90\%$) under certain conditions with only 25 (see Appendix). The impact of sample size depended largely on the branch lengths. With short branch lengths (length = 0.2), most methods had greatly reduced performance at small sample sizes, whereas the effect was less extreme (or negligible) at long branch lengths (length = 2.0). This is clearly related to levels of polymorphism, which are high with short branches and low with long branches. The effects of sample size were also influenced by the number of alleles per locus; methods were much less sensitive to small sample size with three alleles per locus rather than two (given the same branch length and number of characters). When methods were sensitive to sample size, generally only sample sizes of one or two decreased accuracy, and unless branches were very short (length = 0.2), sample sizes larger than five individuals per species did not greatly increase accuracy. Presumably, the extensive modification of allele frequencies on longer branches obviates the need for the precise estimates of frequencies obtained from larger sample sizes. Methods differed in their response to small sample

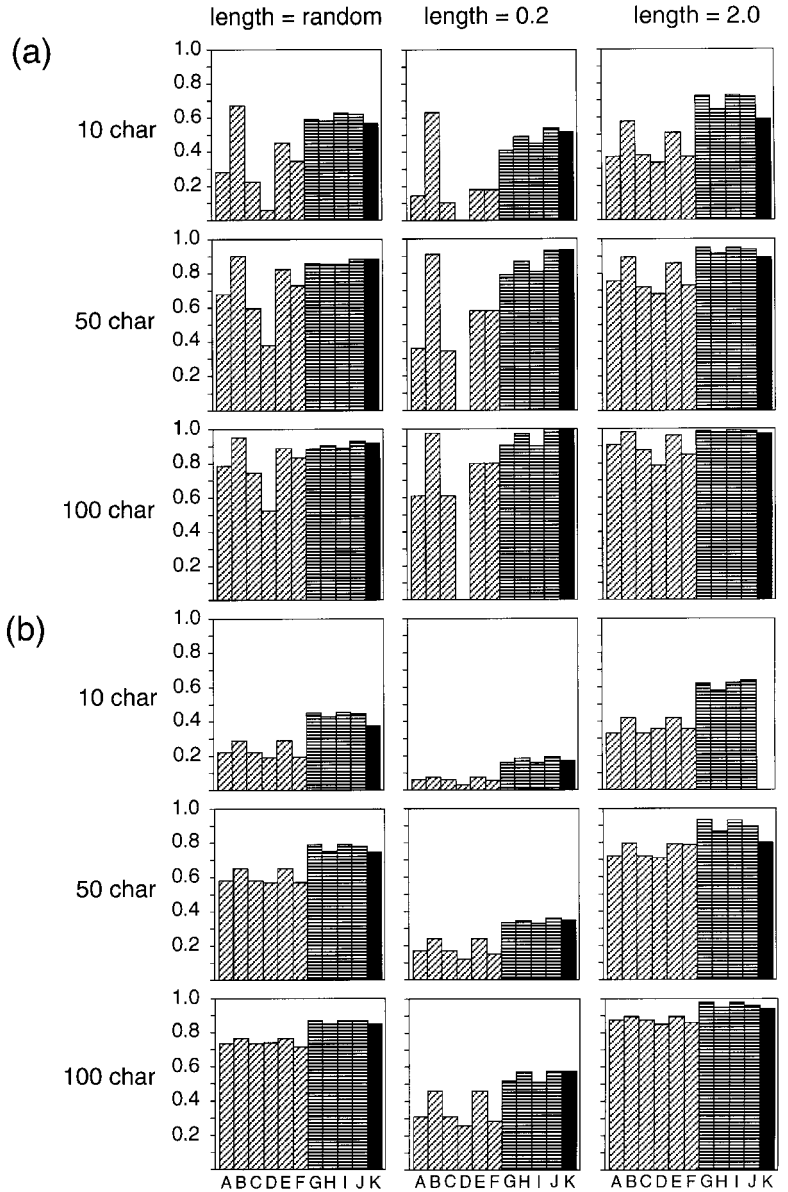


FIGURE 2. Sample of results from the first set of simulations, showing the effects of branch length, sample size, and number of characters in the eight-taxon, two-allele case on accuracy of phylogenetic methods. Methods are parsimony (▨), distance (▤), and likelihood (■): A = any-instance; B = frequency; C = majority; D = missing and polymorphic; E = scaled and unscaled; F = unordered; G = UPGMA, Nei's distance; H = neighbor joining and Fitch–Margoliash, Nei's distance; I = UPGMA, modified Cavalli-Sforza and Edwards chord distance; J = neighbor joining and Fitch–Margoliash, modified Cavalli–Sforza and Edwards chord distance; and K = continuous maximum likelihood. The missing and polymorphic methods have an accuracy of 0 when branch length is 0.2 and $n = 10$ individuals. Results for maximum likelihood with 10 characters, length = 2.0, and $n = 1$ could not be determined because of the large number of identical taxa. (a) Ten individuals per species. (b) One individual per species.

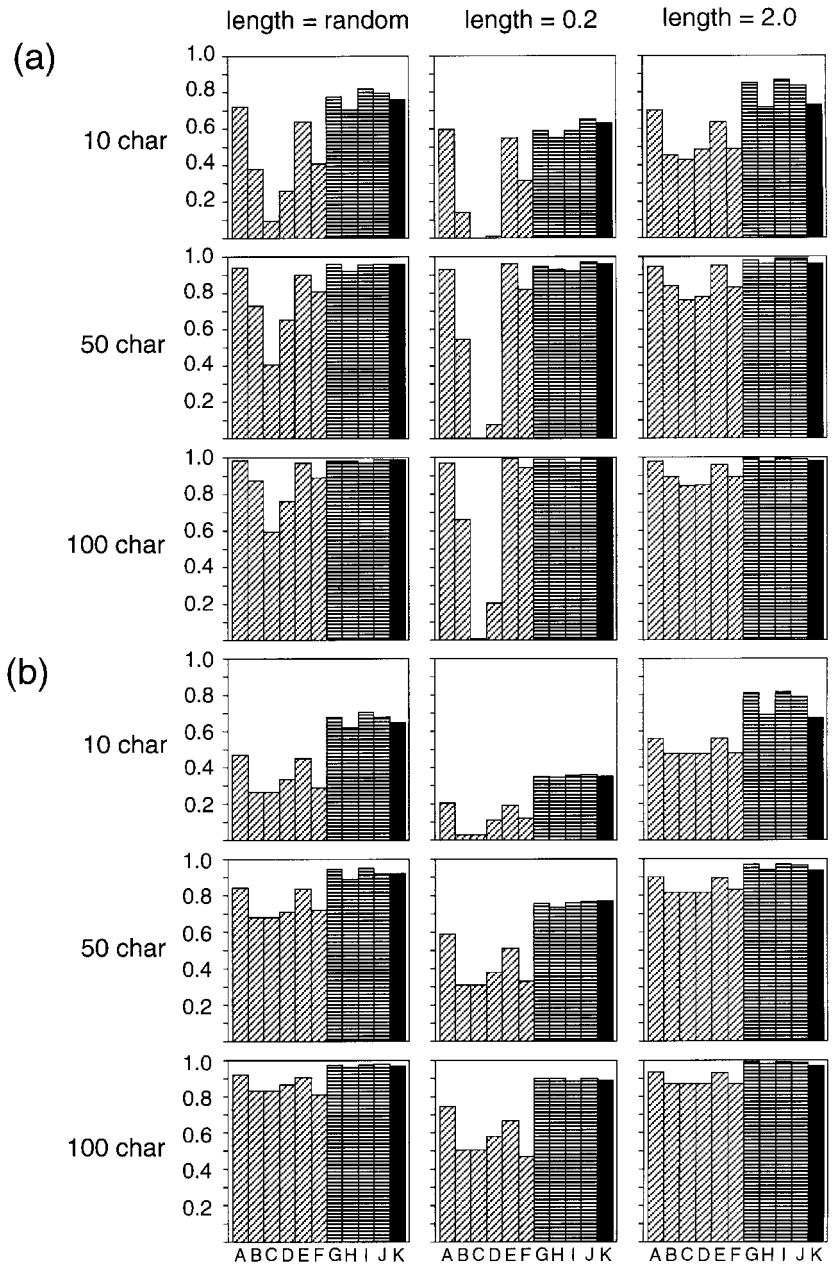


FIGURE 3. Sample of results from the first set of simulations, showing the effects of branch length, sample size, and number of characters in the eight-taxon, three-allele case on accuracy of phylogenetic methods. Methods are parsimony (▨) distance (▤) and likelihood (■): A = frequency; B = majority; C = missing; D = polymorphic; E = scaled; F = unordered; G = UPGMA, Nei's distance; H = neighbor-joining and Fitch–Margoliash, Nei's distance; I = UPGMA, modified Cavalli–Sforza and Edwards chord distance; J = neighbor-joining and Fitch–Margoliash, modified Cavalli-Sforza and Edwards chord distance; and K = continuous maximum likelihood. (a) Ten individuals per species. (b) One individual per species.

sizes. In particular, distance and likelihood methods were far more robust to small sample sizes than most parsimony methods. Two parsimony methods (missing and polymorphic) actually showed increasing accuracy with smaller sample sizes, but parsimony methods that were not greatly affected by sample size were those that performed poorly with both large and small sample sizes. Methods were generally more efficient at high branch lengths than at low branch lengths (i.e., more accurate given the same number of characters).

The number of alleles per locus also influenced the performance of methods. In general, methods could achieve more accurate results in the three-allele case than in the two-allele case with the same sample size and number of loci. The number of alleles also influenced the relative accuracy of parsimony methods. For example, with short branches (length = 0.2), the scaled method was highly accurate in the three-allele case but performed poorly in the two-allele case. Many of the differences related to the number of alleles may be caused by the higher frequency of loss of alleles in the three-allele case (Kimura, 1956); these losses and fixations can be informative for all methods and do not contribute to the errors caused by small sample size.

A limited set of simulations examined the effects of tree shape on method performance (Fig. 4), holding constant the number of characters, alleles, and sample size. In general, tree shape did not have a large impact on relative or absolute method performance. There was a trend for many methods to perform better with increasing symmetry or balance, as seen

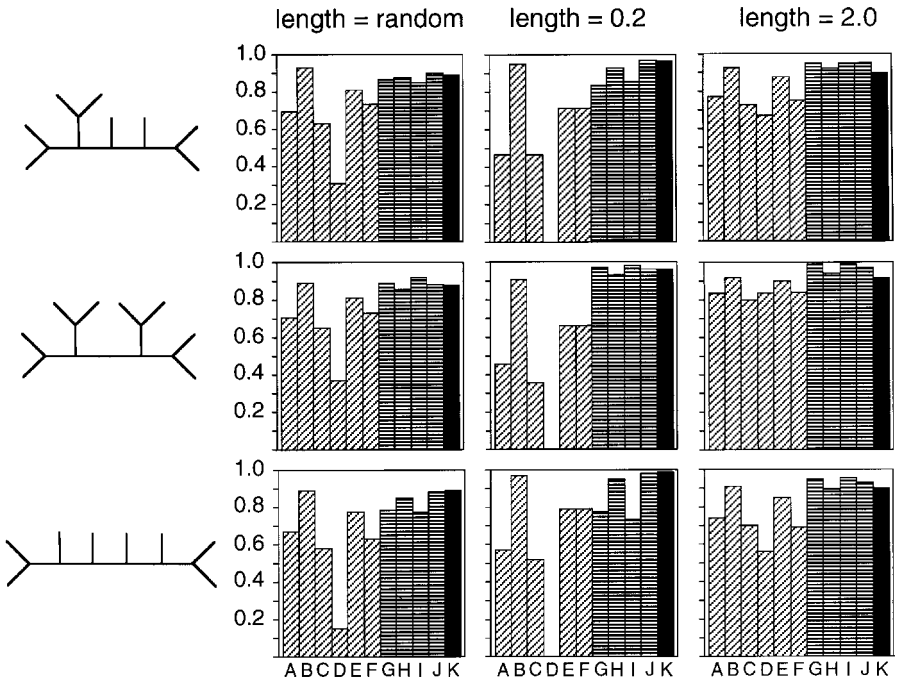


FIGURE 4. The effects of tree shape on the performance of methods in the eight-taxon, two-allele case, with 50 characters and complete sampling of allele frequencies within species. Methods are parsimony (▨), distance (▤), and likelihood (■): A = any-instance; B = frequency; C = majority; D = missing and polymorphic; E = scaled and unscaled; F = unordered; G = UPGMA, Nei's distance; H = neighbor joining and Fitch–Margoliash, Nei's distance; I = UPGMA, modified Cavalli–Sforza and Edwards chord distance; J = neighbor joining and Fitch–Margoliash, modified Cavalli–Sforza and Edwards chord distance; and K = continuous maximum likelihood.

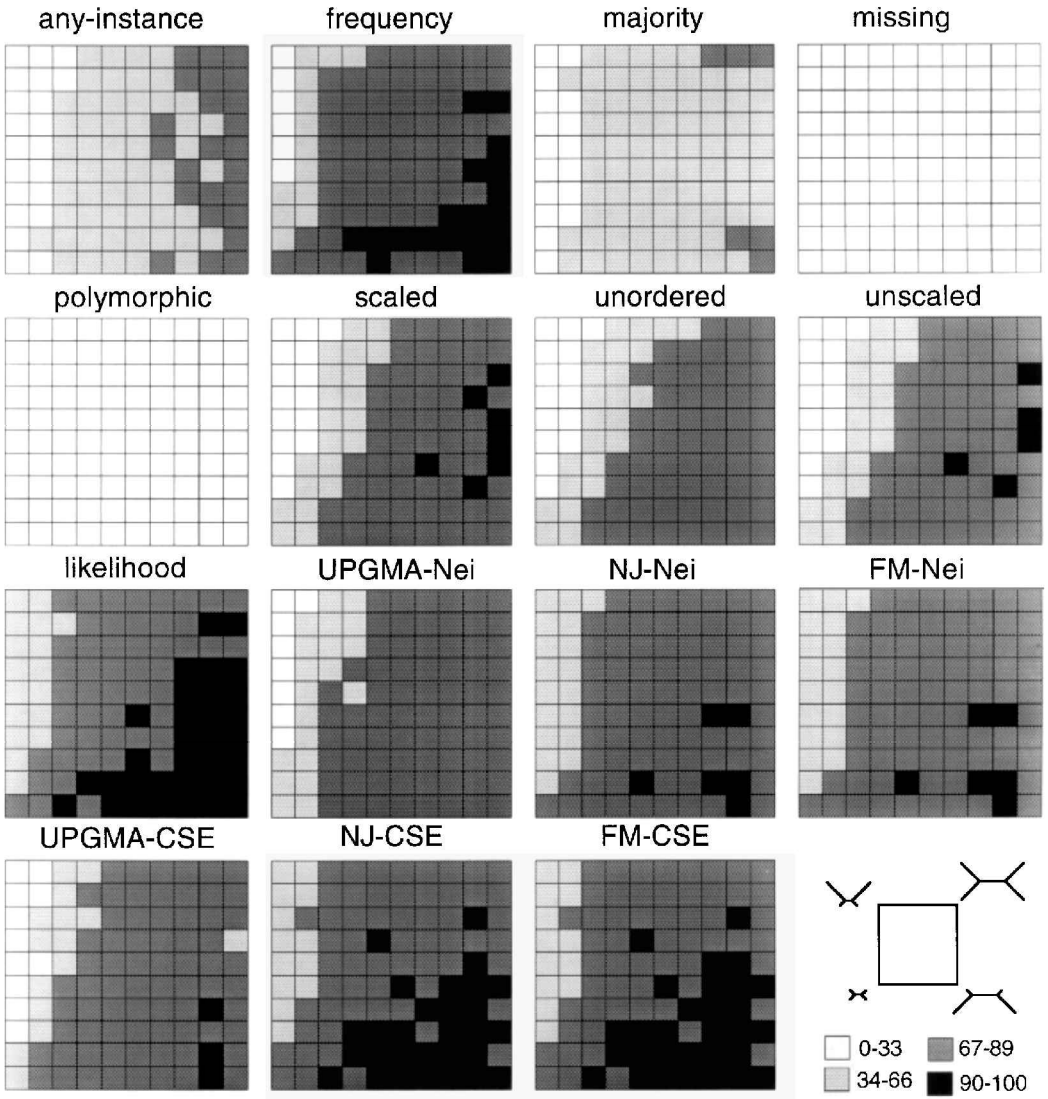


FIGURE 5. Effects of branch length on the accuracy of phylogenetic methods in the four-taxon, two-allele case with 10 characters and perfect sampling of allele frequencies. Different shadings indicate the proportion of 100 matrices in which the correct tree is estimated. The x-axis is the length of two terminal branches and the internal branch, and the y-axis is the length of the other two terminal branches. Branch lengths vary from 0.1 to 1.9. CSE = Cavalli-Sforza and Edwards modified chord distance; FM = Fitch–Margoliash; NJ = neighbor joining.

in other simulation studies (e.g., Rohlf et al., 1990). However, there were many exceptions, and with short branch lengths many methods performed better on more asymmetric trees. Tree shape affected the accuracy of most methods only slightly, but UPGMA was particularly sensitive. When branch lengths were short or variable, UPGMA (both distances) was the most accurate method on symmetric trees but relatively inaccurate on asymmetric trees.

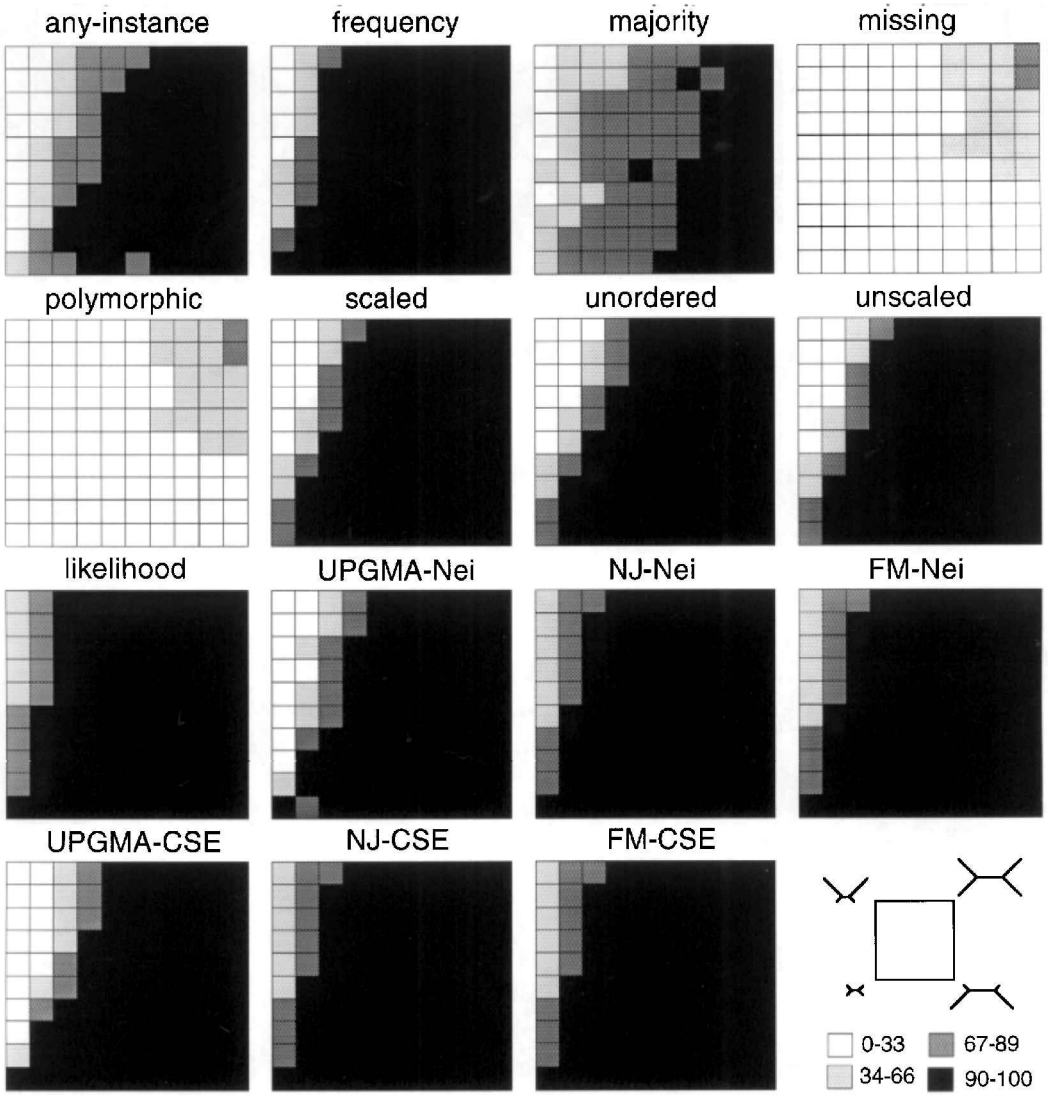


FIGURE 6. Effects of branch length on the accuracy of phylogenetic methods in the four-taxon, two-allele case with 50 characters and perfect sampling of allele frequencies. Different shadings indicate the proportion of 100 matrices in which the correct tree is estimated. The x-axis is the length of two terminal branches and the internal branch, and the y-axis is the length of the other two terminal branches. Branch lengths vary from 0.1 to 1.9. CSE = Cavalli-Sforza and Edwards modified chord distance; FM = Fitch–Margoliash; NJ = neighbor joining.

Examination of Branch Lengths

The effects of branch length variation in the two-allele, four-taxon case are shown in Figs. 5–7. All methods have difficulty in reconstructing the true phylogeny in the ''Felsenstein zone'' (Huelsenbeck and Hillis, 1993), the upper left-hand corner of the graph. Under these conditions, all parsimony methods and UPGMA are positively misled (accuracy $< 33\%$; the probability of estimating the correct tree is less than picking one of the three trees randomly), regardless of the number of characters included. Likelihood and the additive

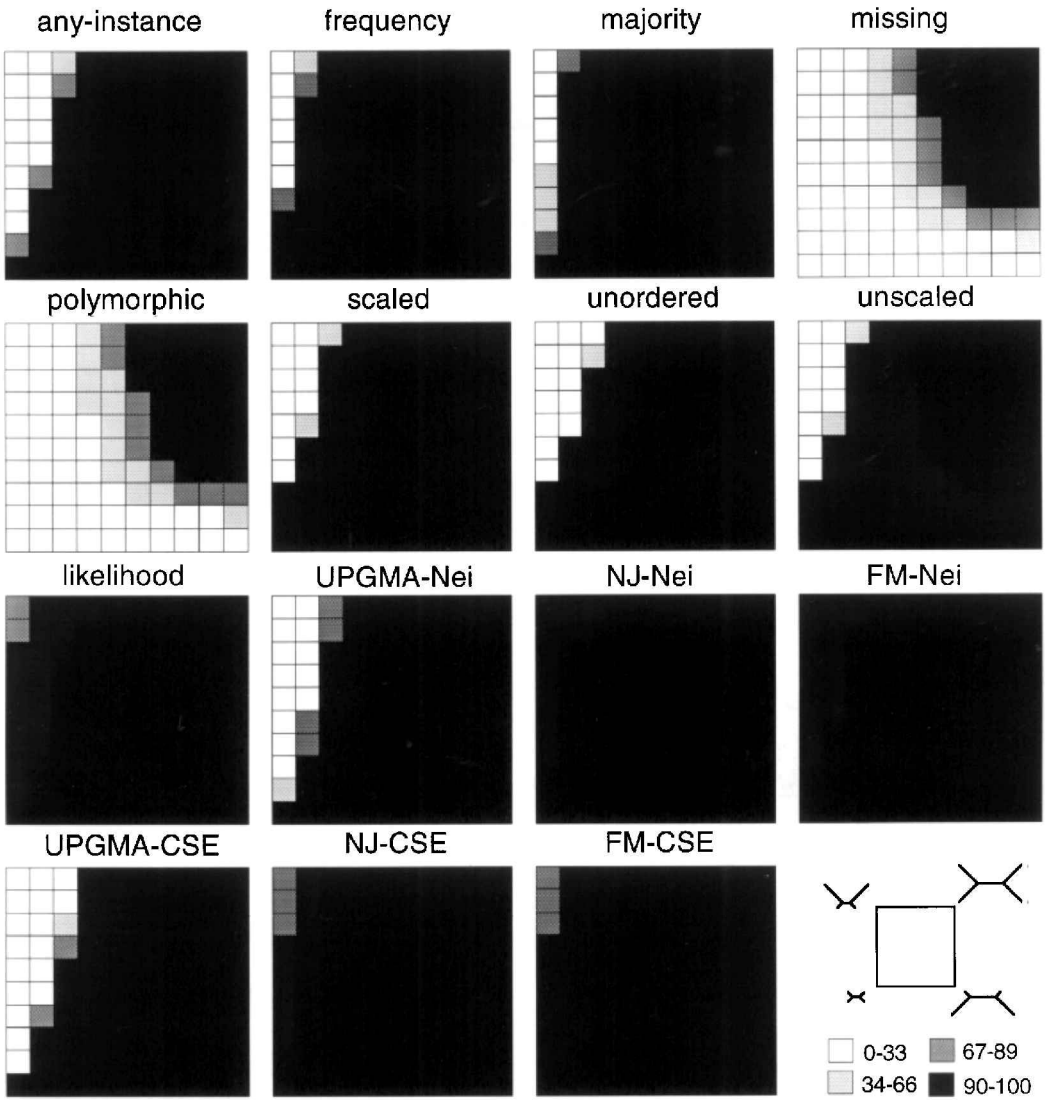


FIGURE 7. Effects of branch lengths on the accuracy of phylogenetic methods in the four-taxon, two-allele case with 500 characters and perfect sampling of allele frequencies. Different shadings indicate the proportion of 100 matrices in which the correct tree is estimated. The x-axis is the length of two terminal branches and the internal branch, and the y-axis is the length of the other two terminal branches. Branch lengths vary from 0.1 to 1.9. CSE = Cavalli–Sforza and Edwards modified chord distance; FM = Fitch–Margoliash; NJ = neighbor joining.

distance methods (neighbor joining and Fitch–Margoliash) also show reduced accuracy under these conditions, but are not positively misled (accuracy $>33\%$), even with only 10 characters. Given enough characters, these methods give highly accurate results in the Felsenstein zone (Fig. 7).

The eight parsimony methods and UPGMA differ in their sensitivity to long-branch attraction. UPGMA and the qualitative parsimony methods (any-instance, scaled, unscaled, unordered) are misled over a larger area of the graph than the

frequency parsimony method (Figs. 5–7). However, the effects of the Felsenstein zone on UPGMA are more dependent on the number of characters; with only 10 characters UPGMA is misled over a smaller area than the qualitative parsimony methods, a roughly equal area with 50 characters, and a slightly larger area with 500 characters. It is difficult to judge whether or not the missing and polymorphic methods are misled in the Felsenstein zone. These methods are generally unable to resolve (correctly or incorrectly) any parts of the tree when there is extensive polymorphism, since they effectively treat polymorphic data cells as ''unknown.'' They perform well only with a very large number of characters and when all the branches are long (Fig. 7).

The results based on only 10 characters (Fig. 5) highlight the relative efficiencies of the methods. With 10 characters maximum likelihood, frequency parsimony, and neighbor joining and Fitch–Margoliash with the chord distance give highly accurate results (>90%) over more of the graph space than the other methods. In contrast, the any-instance, majority, missing, polymorphic, and unordered methods and UPGMA are conspicuously inefficient under these conditions.

A limited set of simulations was performed to examine the effects of extreme branch length differences among lineages in the three-allele case (Fig. 8). The results are virtually identical between the two and three-allele cases for the lower right-hand corner of the graph space (long internal branch, two long terminal branches and two short terminal branches); all methods except for the missing and polymorphic parsimony methods perform very well. In the Felsenstein zone, however, the presence of three alleles greatly reduces the sensitivity of several methods to long branch attraction, relative to the two-allele case. The accuracy of likelihood, neighbor joining, and Fitch–Margoliash increases considerably, and UPGMA with Nei's distance and the frequency parsimony method are no longer positively misled. UPGMA with the chord distance and the qualitative parsimony methods remain positively misled under these conditions in the three-allele case. The improved performance of methods in the three-allele case in the Felsenstein zone may be caused by a decrease in the number of alleles shared between the two long, unrelated branches, relative to the two-allele case.

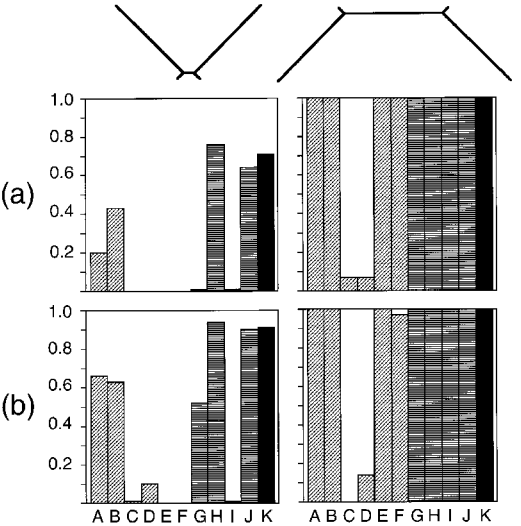


FIGURE 8. Effects of number of alleles on accuracy of phylogenetic methods when branch lengths are highly unequal. The long branches have a length of 2.0, and the short branches have a length of 0.2 (smaller lengths are extremely difficult to calculate for the three-allele case); there are 100 characters. Methods are parsimony (▨), distance (▤), and likelihood (■): A = frequency; B = majority; C = missing; D = polymorphic; E = scaled; F = unordered; G = UPGMA, Nei's distance; H = neighbor joining and Fitch–Margoliash, Nei's distance; I = UPGMA, modified Cavalli-Sforza and Edwards chord distance; J = neighbor joining and Fitch–Margoliash, modified Cavalli–Sforza and Edwards chord distance; and K = continuous maximum likelihood. (a) Two-allele case. (b) Three-allele case.

DISCUSSION

Relative Accuracy of Methods

No single method has the highest accuracy under all the simulated conditions.

However, under most conditions in the eight-taxon case, maximum likelihood, the distance methods, and the frequency parsimony method are superior to the remaining methods and perform similarly to each other. Among these methods, UPGMA is far more sensitive to the extreme differences in branch lengths in the four-taxon case (Figs. 5–7). Frequency parsimony is less accurate than likelihood and the distance methods with small sample sizes and is also positively misled in the Felsenstein zone (Figs. 5–7). Thus, our results suggest that methods that make direct use of frequency information are generally the most accurate (be they parsimony, distance, or likelihood). Among these methods, continuous maximum likelihood and the additive distance methods may have additional advantages under certain conditions (small sample sizes and the Felsenstein zone).

Maximum likelihood and the additive distance methods performed surprisingly well given that their assumptions were often violated. Continuous maximum likelihood and the chord distance assume that there is no fixation or loss of alleles (Felsenstein, 1985), yet performed well under conditions when there was extensive fixation and loss (e.g., length = 2.0). Nei's genetic distance assumes that branch lengths are equal among lineages, yet the additive distance methods (neighbor joining and Fitch–Margoliash) were not positively misled in the Felsenstein zone using this distance measure. In contrast to the conclusions of Huelsenbeck and Hillis (1993) and Huelsenbeck (1995) based on simulations of DNA sequence data, we found that the model-based methods (distance and likelihood) were consistent for all branch length combinations examined (including the Felsenstein zone), even when the assumptions of their models were violated.

While these results might be taken as a strong endorsement for widely applying these distance and likelihood methods to polymorphic data, some qualifications should be mentioned. An additional assumption of continuous maximum likelihood and the chord distance is that there is no mutation; the effects of violating this assumption were not tested in the present study, although it may be that these methods are robust to violations of this assumption as well. Furthermore, current implementations of maximum-likelihood and distance methods for polymorphic characters make it difficult to incorporate characters with missing data entries (Felsenstein, 1995). It may also be difficult to effectively search for optimal trees (maximum likelihood) and to evaluate ''suboptimal'' trees (neighbor joining) using these methods (Huelsenbeck, 1995), and to simultaneously analyze diverse kinds of data.

A number of simplifying assumptions are made in generating the simulated data sets in our study. Because these assumptions are violated in real data sets, the results should be taken with some caution (i.e., they may not apply). However, just because these assumptions are violated in nature does not necessarily mean that the conclusions will not apply to many empirical data sets. For example, congruence studies of method performance show that polymorphic morphological characters give results very similar (both qualitatively and quantitatively) to our simulation results for the eight-taxon, two-allele case (Wiens, 1998). Many qualitatively similar results are also obtained in simulations of polymorphism in DNA sequence data, which incorporate mutation, linkage, and geographic variation (Wiens, M. Servedio, and R. Servedio, unpublished data). Despite the many simplifying assumptions of the simulations of this study (e.g., no mutation, selection, linkage, geographic variation, differences between genotype and phenotype), it appears that many of the conclusions may nevertheless apply to data sets where these assumptions are clearly violated.

Polymorphism and the Performance of Phylogenetic Methods

Most recent studies of method performance have been based on simulated data sets that include only fixed mutations, and do not explicitly include intraspecific variation. How do our results compare to those based on fixed-mutation models? In the following discussion, we do not claim that the results of fixed-mutation studies are necessarily inapplicable when there is variation within species. For example, the results of those studies are relevant to the ability of methods to estimate the true gene tree, even if sampling different individuals within the same species yields different gene trees. Instead, we evaluate whether their conclusions are truly general and apply to a very different model of evolution, which explicitly incorporates intraspecific variation. In our study, the data consist of multiple unlinked genes, and the phylogeny being estimated is the species tree.

A major result of recent simulation studies based on fixed mutations is that all methods have reduced accuracy when there are long terminal branches separated by a short internal branch (the Felsenstein zone; Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995). Our results show that the Felsenstein zone can be problematic for polymorphic data as well (see also Kim and Burgman, 1988), despite the fact that one model assumes change due only to new mutations whereas the other assumes no new mutations.

These results are somewhat disturbing in that they suggest a mechanism by which the Felsenstein zone effect could occur among closely related species. Most empirical examples of long-branch attraction involve only distantly related species (e.g., major groups of mammals [Allard and Miyamoto, 1992], tetrapods [Huelsenbeck and Hillis, 1993], and insects [Carmean and Crespi, 1995; Huelsenbeck, 1997]). Conventional wisdom is that such situations can be ameliorated by breaking up long branches by including additional taxa (e.g., Swofford et al., 1996). Unfortunately, our simulations with polymorphic characters suggest that a Felsenstein zone problem could occur also among closely related species—for example, when population sizes are generally large (short branch lengths under our model) but two unrelated species in the same group have very small effective population sizes (long branch lengths resulting from peripheral isolation, founder effects, population bottlenecks, etc.). Many phylogenetic methods seem to be systematically misled by the fixation and loss of polymorphic traits in the unrelated lineages with small population sizes. In such a situation, there is no possibility of subdividing these long branches with the inclusion of additional taxa, and the only solution may be to use methods that are consistent under these conditions (e.g., continuous maximum likelihood, Fitch–Margoliash, neighbor joining).

This Felsenstein zone effect for polymorphic data might be seen as an example of the problem of lineage sorting of ancestral polymorphisms. Several authors have discussed how the problem of lineage sorting might mislead phylogenetic inferences based on DNA sequences from a single locus when divergence times are short and population sizes are large, but that the problem can be overcome by using several unlinked loci and/or sampling multiple individuals from each species (Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991). However, our results suggest that this effect may occur with many unlinked loci. Sampling many loci and/or individuals therefore may not help resolve the species tree correctly, because many gene trees seem to systematically converge on the same incorrect answer. The unrelated species with small population sizes may tend be clustered, even with perfect sampling of individuals within species.

Using simulations of fixed DNA mutations, Huelsenbeck and Hillis (1993) and

Huelsenbeck (1995) examined thoroughly the effects of branch length variation on method performance in the four-taxon case. They found that parsimony and UPGMA are positively misled in the Felsenstein zone, whereas maximum likelihood, neighbor joining, and the Fitch–Margoliash method can be accurate given enough characters. These findings are consistent with our results. However, Huelsenbeck and Hillis (1993) and Huelsenbeck (1995) also found several patterns that we did not find, including (1) parsimony generally performs well outside the Felsenstein zone, given sufficient characters; (2) UPGMA performs well only when branch lengths are similar among lineages; (3) maximum likelihood, neighbor joining, and the Fitch–Margoliash method are positively misled in the Felsenstein zone without a very large numbers of characters and a close fit between the models used in simulating the data and estimating the tree; and (4) most methods have reduced accuracy when branch lengths are very long. We relate these differences in method performance to differences in the evolutionary models used as follows.

We found that the accuracy of parsimony depends largely on how polymorphic characters are coded. For example, the frequency method is quite accurate for most of the graph space outside the Felsenstein zone, whereas the missing and polymorphic methods are hopelessly inefficient under all but a limited set of conditions (long branches and a large numbers of characters; Figs. 5–7). We also found that parsimony methods differ in their sensitivity to long-branch attraction (the Felsenstein zone, Figs. 5–7). As found in fixed-mutation studies (e.g., Huelsenbeck and Hillis, 1993), parsimony methods that are modified to reflect relevant details of the simulated evolutionary process seem to have a smaller region of inconsistency than those that do not (i.e., area of the graph where they give positively misleading results). For example, Kimura's (1955) model assumes that it is more likely for an allele present at a high frequency to go to fixation over time rather than be lost. Thus, those parsimony methods that assume that a change in an allele's frequency from 99% to fixation (100%) is ''easier'' than a change from 1% to 100% (the frequency and majority methods) are consistent over more of the graph space than the qualitative parsimony methods that would give equal weight to these allele frequency changes (any-instance, scaled, unscaled, unordered). We believe that these qualitative parsimony methods are more easily misled when there is a high probability of allele fixation and loss in two unrelated terminal lineages (as in the Felsenstein zone) because these methods only consider the fixation and loss of alleles to be informative, and ignore any changes in frequency along the internal branch. Thus, these methods are misled under branch length combinations that are not misleading for parsimony methods that incorporate even crude frequency information (e.g., majority, frequency).

UPGMA performed much better in our study than expected based on studies of fixed mutations. We found that UPGMA was quite efficient in the lower right corner of the graph space (Figs. 5–8), conditions where studies of fixed characters have found its accuracy to be considerably reduced (Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995), However, it is possible that the branch length differences under our model are not extreme enough to mislead UPGMA in this part of the graph space. Previous studies have characterized UPGMA as generally less accurate than parsimony (Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995), but in our study UPGMA was (1) consistent over roughly the same area as six of the eight parsimony methods (all but the frequency and majority methods), (2) more efficient over much of the graph space than four of the eight parsimony methods (any-instance, majority, missing, polymorphic), and (3) about as efficient as three others (scaled, unordered, unscaled). Under some conditions (Figs.

5, 8), the qualitative parsimony methods were positively misled by unequal branch lengths whereas UPGMA was not. We consider the best explanation for these surprising results to be that UPGMA (as applied in this study) uses fine-grained information on allele frequencies that are ignored by all but one of the parsimony methods. Thus, UPGMA, at least when used with appropriate genetic distances, may offer a better fit to the evolutionary model than do the nonfrequency parsimony methods.

Huelsenbeck and Hillis (1993) and Huelsenbeck (1995) found that maximum likelihood, neighbor joining, and the Fitch–Margoliash method are positively misled in the Felsenstein zone (with four taxa) unless there is a very large numbers of characters ($\geq$1,000) and the model of evolution assumed by the methods matches the simulated evolutionary model. In contrast we found that, while the accuracy of these methods was often highly reduced in the Felsenstein zone, they were never positively misled. This was true even with only 10 characters and when the assumptions of the methods were violated. One possible explanation for this difference is that the branch lengths we examined were not extreme enough to mislead these methods (although parsimony and UPGMA were easily misled). To test this hypothesis, we performed a limited set of simulations in the Felsenstein zone with the length of the long, unrelated terminal branches increased from 2 to 10 (branch lengths higher than 10 would not affect the results, because alleles were always fixed or lost along these long branches). With 50 characters and two alleles, maximum-likelihood and the additive-distance methods were not positively misled, despite the hundredfold difference in branch lengths between lineages. We do not doubt that there are potential branch lengths combinations where these methods could be misled, yet there seems to be a clear difference in how the methods perform under these models. Other possible explanations for the greater success of these methods with polymorphic data include (1) that allele frequencies are more informative per unit character than the presence or absence of nucleotides (at least in these simulations) and/or (2) that fixation and loss of alleles effectively limit the length of the longest branches in the genetic drift model (i.e., because alleles frequently evolve to loss or fixation on a branch of length 2, a length of 10 is not really five times longer).

We also found that methods did not suffer from reduced accuracy at very high branch lengths outside the Felsenstein zone, in contrast to results based on fixed characters (e.g., Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995). Under the fixed-mutation model, DNA sequences of simulated taxa are effectively randomized at high branch lengths and contain little or no phylogenetic information. However, when the internal branch is long under the drift model, alleles have a high probability of being fixed or lost along this branch, even though allele frequencies of nonfixed loci are effectively randomized. These fixations and losses of alleles provide phylogenetic information for all the methods, and allow methods to perform relatively well with long branch lengths under this model.

Finally, our results suggest that sample size (individuals per species) is an important and underappreciated component of phylogenetic accuracy. Under many conditions, small sample sizes can have a devastating effect on method performance. For example, with 50 characters, two alleles, and eight taxa, the average accuracy of the frequency parsimony method with low branch lengths (length = 0.2) is 91% with 10 individuals per species and 24% with a single individual per species. Sample size affects the relative performance of the methods as well. For example, for the same conditions described earlier with random branch lengths, the frequency parsimony method is more accurate than UPGMA (with the chord distance) when 10 indi-

viduals are sampled per species (frequency = 90%, UPGMA = 85%), but with a sample size of 1 individual per species, UPGMA is more accurate (UPGMA = 79%, frequency = 65%). The reason for the greater accuracy of distance and likelihood methods at small sample sizes seems to be a combination of (1) the tendency of these methods to give fully resolved estimates despite coarse frequency information and (2) a superior ability to deal with random noise (Wiens, in press).

Several previous authors have addressed the effects of sample size using empirical resampling analyses (e.g., Gorman and Renzi, 1979; Hillis, 1987; Archie et al., 1989; Smouse et al., 1991). The true phylogeny in these studies was not known, so the effect of subsampling on phylogenetic accuracy could not be assessed. Several of these studies predicted a greater need for larger sample sizes when levels of polymorphism are relatively high (e.g., Hillis, 1987; Archie et al., 1989). These predictions are supported by the results of this study. Thus, in our study, small sample sizes have relatively little impact on the accuracy of most methods when there are relatively few polymorphic loci (i.e., length = 2.0) and much greater impact when most loci are polymorphic (i.e., length = 0.2).

Simulations are an important tool because they allow testing of method behavior under known but simplified conditions. The simulations in this study are very simplistic, and do not include many realistic complexities or features specific to certain kinds of data. They represent an extreme model, where genetic drift is the only evolutionary force. Similarly, other simulation studies, most of which are designed to emulate DNA sequence data, also represent an extreme model, in which mutation is the only evolutionary force. Attempts to generalize about the performance of phylogenetic methods should consider the extreme examples offered by simulations of DNA sequences evolving by mutation and allele frequencies evolving by genetic drift. It is encouraging that many common patterns are evident.

Previous Studies of Method Performance Using Polymorphic Data

Some previous simulation studies have addressed the performance of phylogenetic methods using data with intraspecific variation. Two studies (Kim and Burgman, 1988; Rohlf and Wooten, 1988) compared the performance of continuous maximum likelihood, a parsimony-based distance method (WAGPROC; Swofford, 1983), and UPGMA (with various distances) using simulated data sets with allele frequencies evolved by random genetic drift. Kim and Burgman (1988) examined the four-taxon, two-allele case with various combinations of branch lengths, and showed the apparent inconsistency of parsimony and UPGMA and the superiority of maximum likelihood in the Felsenstein zone. Rohlf and Wooten (1988) examined data sets with 20 taxa and equal branch lengths among lineages and concluded that UPGMA and likelihood generally gave the most accurate results. These results are both consistent with each other and with the results of this study. However these authors did not examine the performance of additive distance methods (neighbor joining and Fitch–Margoliash) or the eight parsimony coding methods, and did not incorporate the effects of sample size or the fixation and loss of alleles. Furthermore, Rohlf and Wooten (1988) only simulated equal branch lengths among lineages, whereas Kim and Burgman (1988) examined only a limited set of branch length combinations. Wiens and Servedio (1997) addressed the accuracy of different parsimony methods for including, excluding, weighting, and coding polymorphic characters, using a subset of the data analyzed here.

Wiens (1995) compared the eight parsimony coding methods with statistical analyses of seven morphological and molecular data sets, and found (in agreement with the results of our simulations)

that the frequency method generally performed best for all the performance criteria used. Wiens (1995) also found that the scaled method never estimated the same tree as the frequency method for any of the data sets examined; in our simulations these methods are most likely to yield consistently different trees when the scaled method is being misled by the Felsenstein zone effect and the frequency method is not (e.g., Fig. 8b). Although other explanations are certainly possible for the differences in the estimated trees, these results suggest the possibility of branch-length effects in real polymorphic data and the importance of further testing of these methods and hypotheses with empirical and simulated data.

ACKNOWLEDGMENTS

We are grateful to R. Servedio for invaluable mathematical and computational advice. We thank D. Swofford for allowing us to use test versions of his PAUP* software package for this research. D. Cannatella, P. Chu, B. Livezey, C. Simon, and an anonymous reviewer provided useful comments on the manuscript. M.R.S. acknowledges the support of a National Science Foundation Graduate Fellowship.

REFERENCES

ALBERCH, P. 1983. Morphological variation in the neotropical salamander genus *Bolitoglossa*. Evolution 37:906–919.

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.

ARCHIE, J. W., C. SIMON, AND A. MARTIN. 1989. Small sample size does decrease the stability of dendrograms calculated from allozyme-frequency data. Evolution 43:678–683.

BERLOCHER, S. H., AND D. L. SWOFFORD. 1997. Searching for phylogenetic trees under the frequency parsimony criterion: An approximation using generalized parsimony. Syst. Biol. 46:211–215.

BUTH, D. G. 1984. The application of electrophoretic data in systematic studies. Annu. Rev. Ecol. Syst. 15:501–522.

CAMPBELL, J. A., AND D. R. FROST. 1993. Anguid lizards of the genus *Abronia*: Revisionary notes, description of four new species, a phylogenetic analysis, and key. Bull. Am. Mus. Nat. Hist. 216:1–121.

CARMEAN, D., AND B. CRESPI. 1995. Do long branches attract flies? Nature 373:666.

CAVALLI-SFORZA, L. L., AND A. W. F. EDWARDS. 1967. Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet. 19:233257.

COLLESS, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: A reappraisal. Syst. Zool. 29:288–299.

CROTHER, B. I. 1990. Is ''some better than none'' or do allele frequencies contain phylogenetically useful information? Cladistics 6:277–281.

FARRIS, J. S. 1981. Distance data in phylogenetic analysis. Pages 3–23 *in* Advances in cladistics, Volume 1. Proceeding of the first meeting of the Willi Hennig Society (V. A. Funk and D. R. Brooks, eds.). New York Botanical Garden, New York.

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 gene frequencies and quantitative characters: Finding maximum likelihood estimates. Evolution 35:1229–1242.

FELSENSTEIN, J. 1985. Phylogenies from gene frequencies: A statistical problem. Syst. Zool. 34:300–311.

FELSENSTEIN, J. 1988. Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst. 19:445–471.

FELSENSTEIN, J. 1995. PHYLIP, version 3.57c. Department of Genetics, Univ. Washington, Seattle.

FIALA, K. L., AND R. R. SOKAL. 1985. Factors determining the accuracy of cladogram estimation: Evaluation using computer simulation. Evolution 39:609–622.

FISHER, R. A. 1930. The genetical theory of natural selection, 1st edition. Clarendon, Oxford, England.

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

GORMAN, G., AND J. RENZI, JR. 1979. Genetic distance and heterozygosity estimates in electrophoretic studies: Effects of sample size. Copeia 1979:242-249.

HILLIS, D. M. 1987. Molecular versus morphological approaches to systematics. Annu. Rev. Ecol. Syst. 18:23–42.

HILLIS, D. M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44:3–16.

HILLIS, D. M., J. P. HUELSENBECK, AND C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

HUELSENBECK, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

HUELSENBECK, J. P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

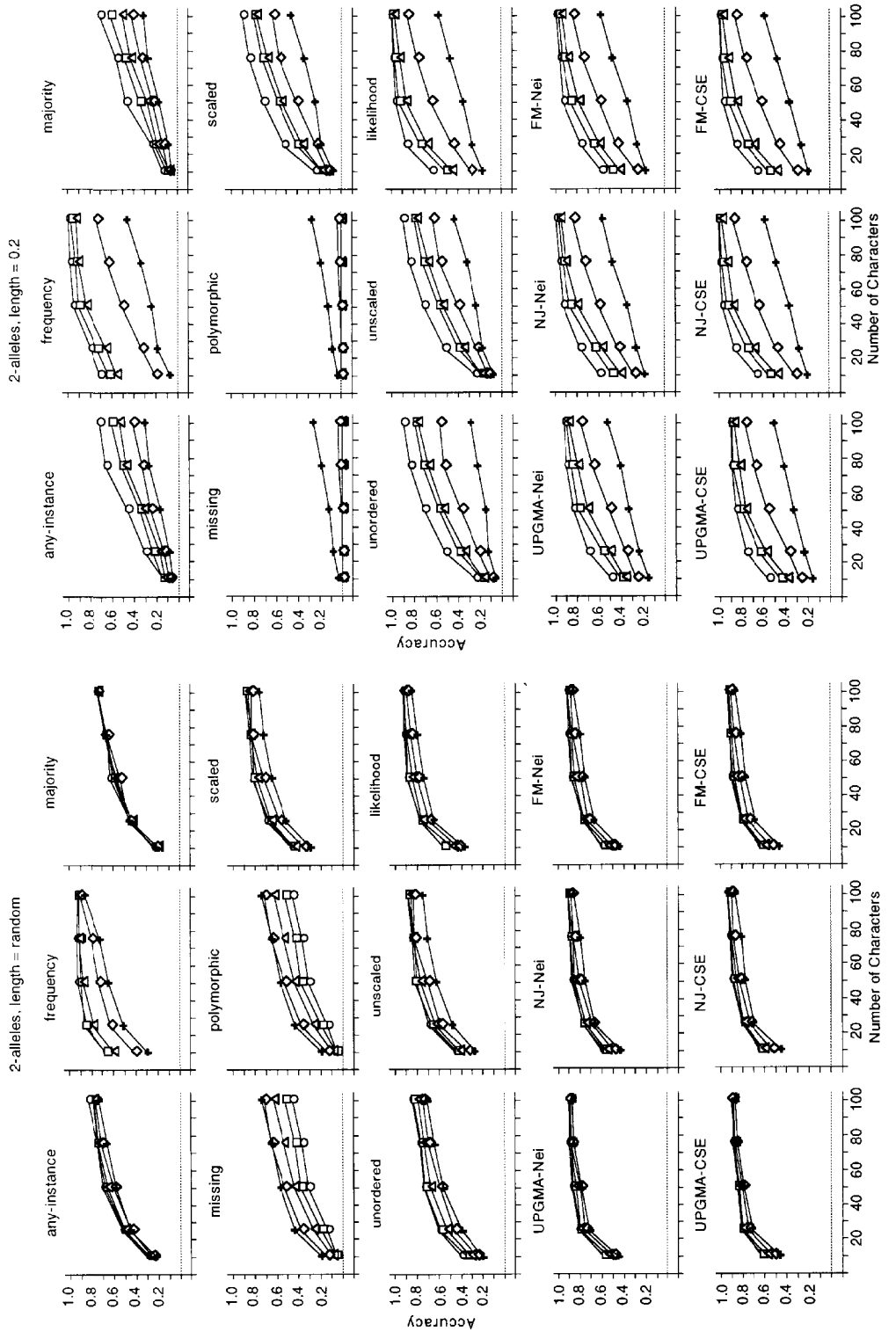
KIM, J., AND M. A. BURGMAN. 1988. Accuracy of phylogenetic-estimation methods under unequal evolutionary rates. Evolution 42:596–602.

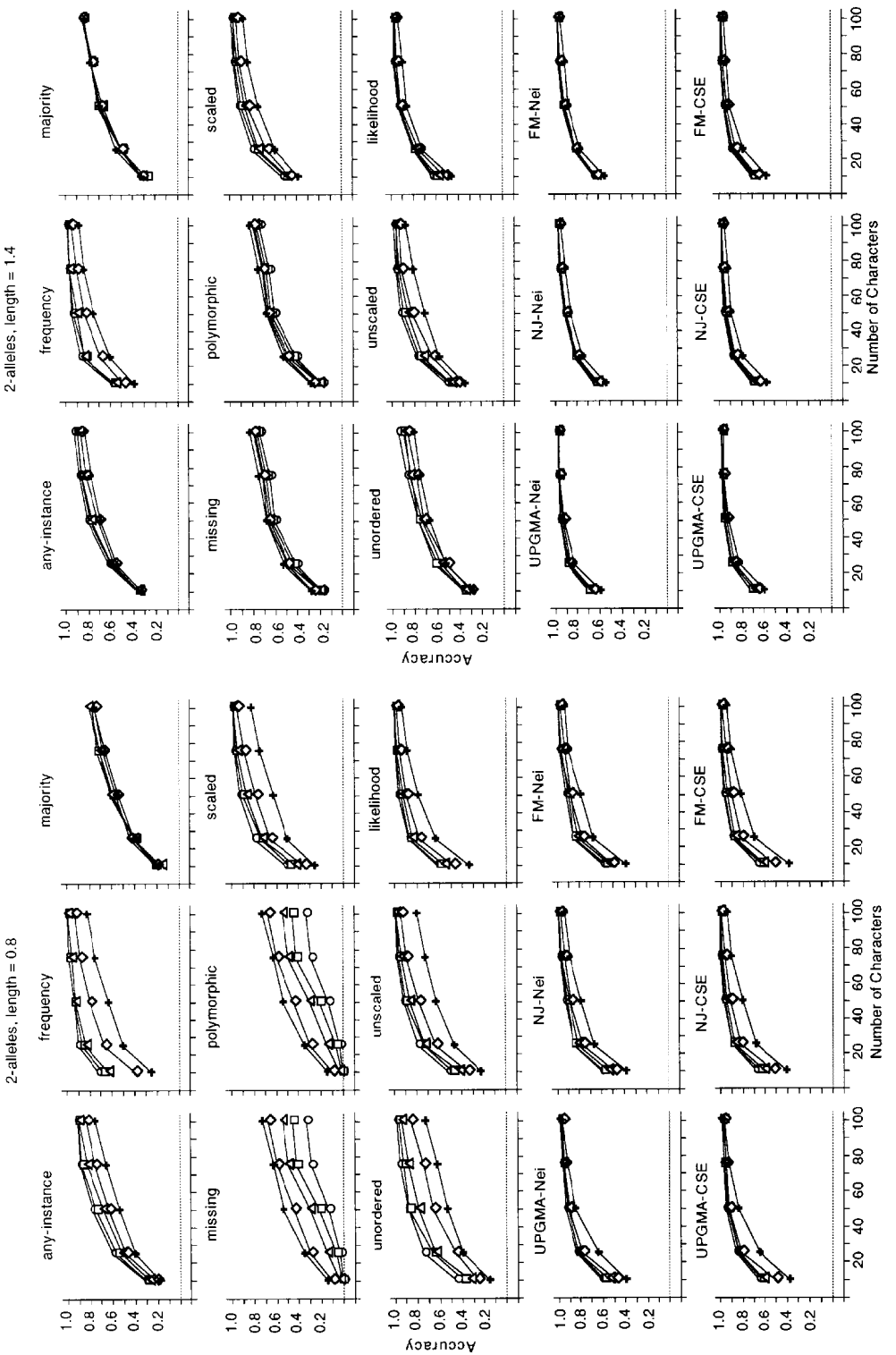
KIMURA, M. 1955. Random genetic drift in multiallelic locus. Evolution 9:419–435.
KIMURA, M. 1956. Random genetic drift in a triallelic locus; Exact solution with a continuous model. Biometrics 12:57–66.
KIMURA, M., AND J. F. CROW. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.
KREITMAN, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature 304:412–417.
MABEE, P. M., AND J. HUMPHRIES. 1993. Coding polymorphic data: Examples from allozymes and ontogeny. Syst. Biol. 42:166–181.
MICKEVICH, M. F., AND C. MITTER. 1981. Treating polymorphic characters in systematics: A phylogenetic treatment of electrophoretic data. Pages 45–58 *in* Advances in cladistics, Volume 1. Proceeding of the first meeting of the Willi Hennig Society (V. A. Funk and D. R. Brooks, eds.). New York Botanical Garden, New York.
MICKEVICH, M. F., AND C. MITTER. 1983. Evolutionary patterns in allozyme data: A systematic approach. Pages 169–176 *in* Advances in cladistics, Volume 2. Proceedings of the second meeting of the Willi Hennig Society (V. A. Funk and D. R. Brooks, eds.). Columbia Univ. Press, New York.
MURPHY, R. W. 1993. The phylogenetic analysis of allozyme data: Invalidity of coding alleles by presence/absence and recommended procedures. Biochem. Syst. Ecol. 21:25–38.
NEI, M. 1972. Genetic distance between populations. Am. Nat. 106:238–292.
PAMILO, P., AND M. NEI. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING. 1988. Numerical recipes in C. Cambridge Univ. Press, New York.
RANNALA, B. 1995. Polymorphic characters and phylogenetic analysis: A statistical perspective. Syst. Biol. 44:421–429.
ROGERS, J. S. 1986. Deriving phylogenetic trees from allele frequencies: A comparison of nine genetic distances. Syst. Zool. 35:297–310.
ROHLF, F. J., AND M. C. WOOTEN. 1988. Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. Evolution 42:581–595.
ROHLF, F. J, W. S. CHANG, R. R. SOKAL, AND J. KIM. 1990. Accuracy of estimated phylogenies: Effects of tree topology and evolutionary model. Evolution 44:1671–1684.
SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.
SMOUSE, P. E., T. E. DOWLING, J. A. TWOREK, W. R. HOEH, AND W. M. BROWN. 1991. Effects of intraspecific variation on phylogenetic inference: A likelihood analysis of mtDNA restriction site data in cyprinid fishes. Syst. Zool. 40:393–409.
SOKAL, R. R., AND C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 28:1409–1438.
SWOFFORD, D. L. 1983. WAGPROC, Version 4.2. Wagner Procedure Program. Illinois Natural History Survey, Champaign.
SWOFFORD, D. L., AND S. H. BERLOCHER. 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. Syst. Zool. 36:293–325.
SWOFFORD, D. L., AND G. J. OLSEN. 1990. Phylogeny reconstruction. Pages 411–501 *in* Molecular systematics, 1st edition (D. M. Hillis and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.
SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogeny reconstruction. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.
TAKAHATA, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.
WIENS, J. J. 1995. Polymorphic characters in phylogenetic systematics. Syst. Biol. 44:482–500.
WIENS, J. J. 1998. Testing phylogenetic methods with tree-congruence: Phylogenetic analysis of polymorphic morphological characters in phrynosomatid lizards. Syst. Biol. 47:(in press).
WIENS, J. J., AND M. R. SERVEDIO. 1997. Accuracy of phylogenetic analysis including and excluding polymorphic characters. Syst. Biol. 46:332–345.
WOLFRAM, S. 1991. Mathematica: A system for doing mathematics by computer. Addison-Wesley, Reading, Massachussetts.
WRIGHT, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.
WU, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435.

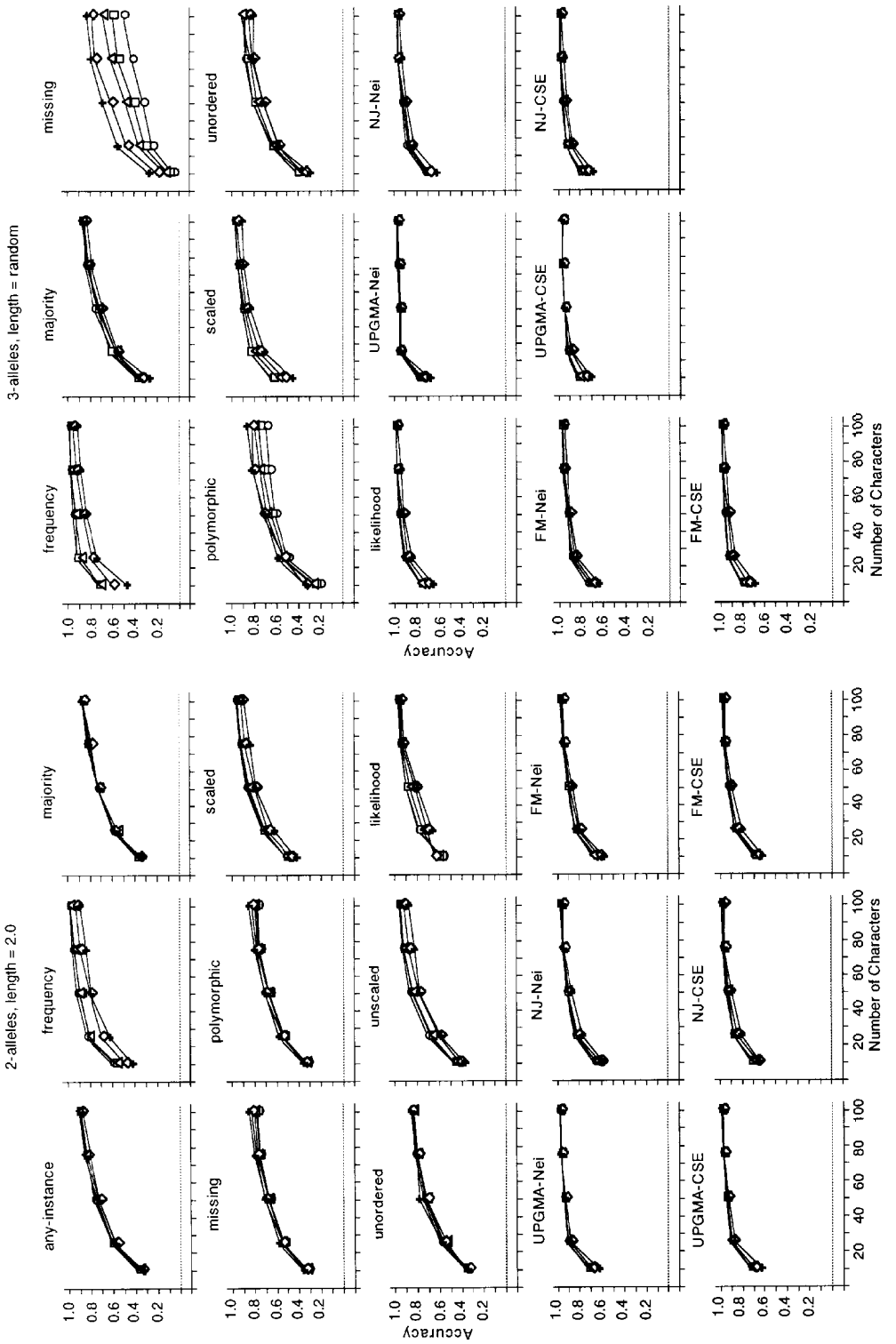
Received 25 June 1996; accepted 24 May 1997
Associate Editor: C. Simon

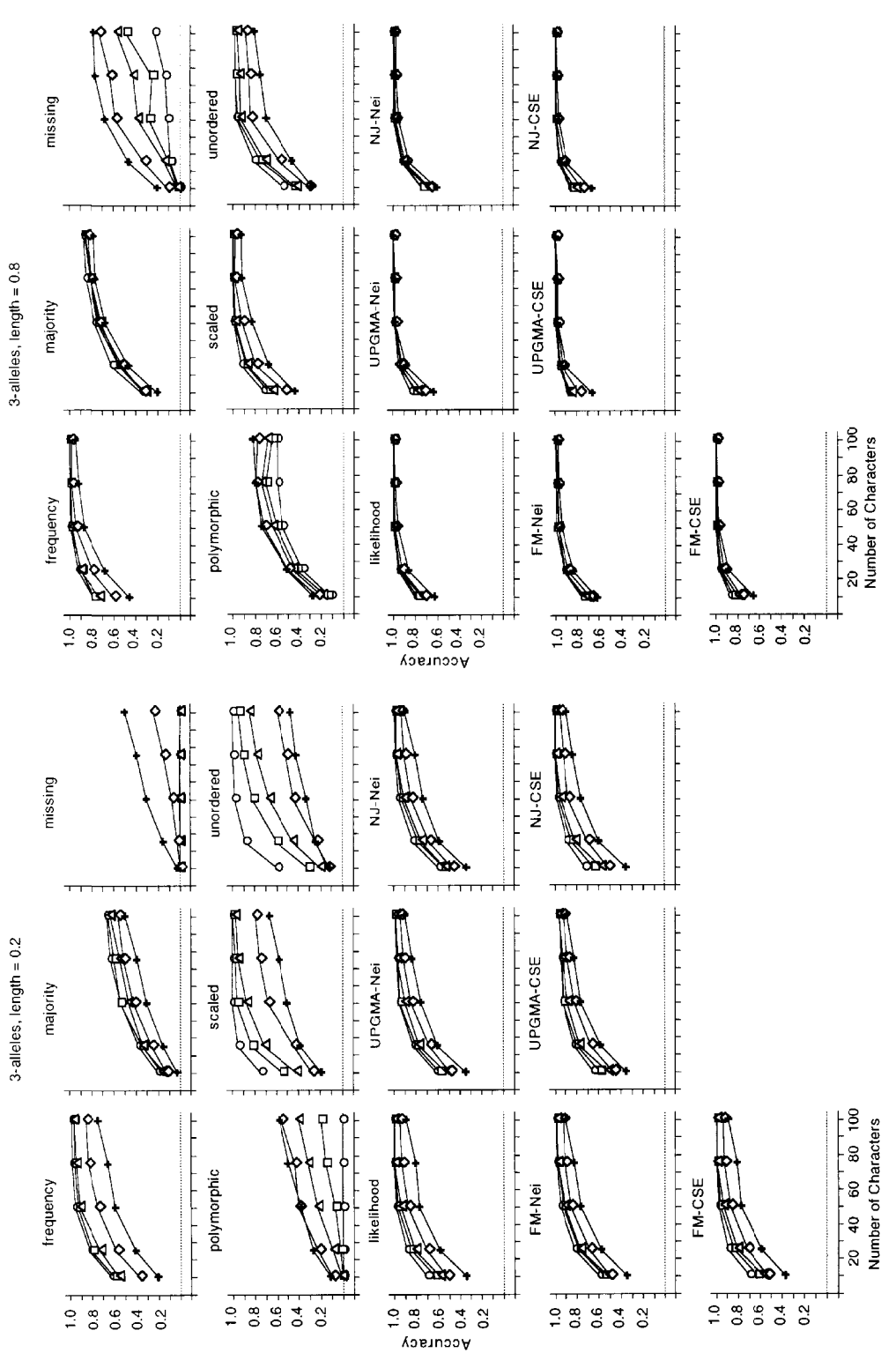
APPENDIX

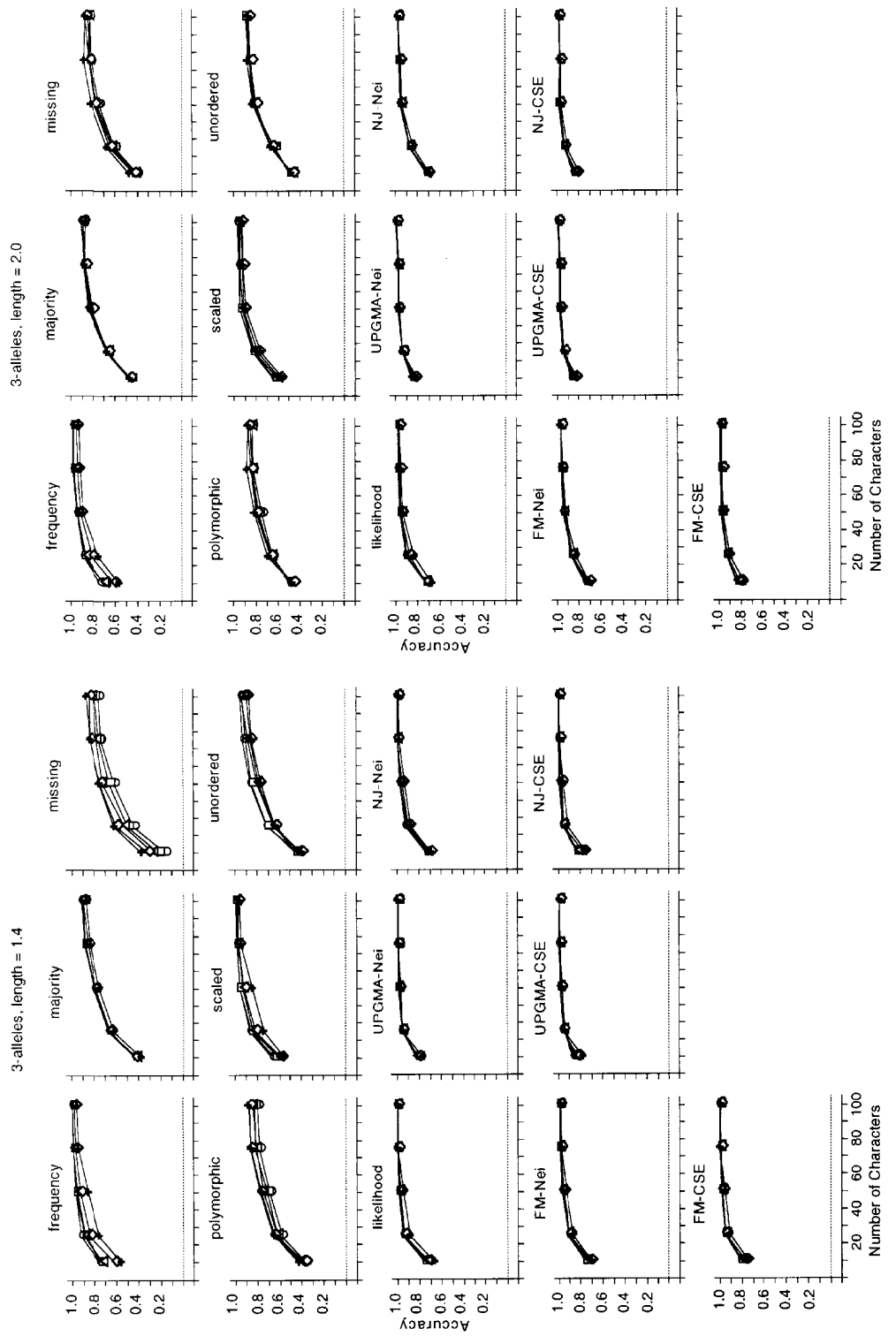
Accuracy of phylogenetic methods in the eight-taxon case, with different number of alleles per locus, number of characters, sample sizes, and branch lengths. Symbols represent accuracy at different sample sizes (individuals sampled per species): —○— = perfect sampling; —□— = 10; —△— = five; —◇— = two; —+— = one. Each point is the mean accuracy for 100 replicated matrices. CSE = Cavalli–Sforza and Edwards modified chord distance; FM = Fitch–Margoliash; NJ = neighbor joining.


2-alleles, length = random
2-alleles, length = 0.2
any-instance
missing
UPGMA-Nei
UPGMA-CSE
frequency
polymorphic
unscaled
NJ-Nei
NJ-CSE
majority
scaled
likelihood
FM-Nei
FM-CSE
Accuracy
Number of Characters


2-alleles, length = 0.8
2-alleles, length = 1.4
any-instance
frequency
majority
missing
polymorphic
scaled
unordered
unscaled
likelihood
UPGMA-Nei
NJ-Nei
FM-Nei
UPGMA-CSE
NJ-CSE
FM-CSE
Accuracy
Number of Characters


2-alleles, length = 2.0
3-alleles, length = random
any-instance
frequency
majority
missing
polymorphic
scaled
unordered
unscaled
likelihood
UPGMA-Nei
NJ-Nei
FM-Nei
UPGMA-CSE
NJ-CSE
FM-CSE
Accuracy
Number of Characters


3-alleles, length = 0.2
3-alleles, length = 0.8
frequency
majority
missing
polymorphic
scaled
unordered
likelihood
UPGMA-Nei
NJ-Nei
FM-Nei
UPGMA-CSE
NJ-CSE
FM-CSE
Accuracy
Number of Characters


3-alleles, length = 1.4
3-alleles, length = 2.0
frequency
majority
missing
polymorphic
scaled
unordered
likelihood
UPGMA-Nei
NJ-Nei
FM-Nei
UPGMA-CSE
NJ-CSE
FM-CSE
Accuracy
Number of Characters
1.0
0.8
0.6
0.4
0.2
20
40
60
80
100