Accuracy of Phylogenetic Analysis Including and Excluding Polymorphic Characters

John J. Wiens; Maria R. Servedio

Systematic Biology, Vol. 46, No. 2 (Jun., 1997), 332-345.

Stable URL:
http://links.jstor.org/sici?sici=1063-5157%28199706%2946%3A2%3C332%3AAOPAIA%3E2.0.CO%3B2-D

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Systematic Biology is published by Society of Systematic Biologists. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ssbiol.html.

Systematic Biology
©1997 Society of Systematic Biologists

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

http://www.jstor.org/
Thu Sep 18 10:38:51 2003

Syst. Biol. 46(2):332–345, 1997

ACCURACY OF PHYLOGENETIC ANALYSIS INCLUDING AND EXCLUDING POLYMORPHIC CHARACTERS

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 ubiquitous in systematic characters, yet systematists often do not deal with polymorphism explicity. For example, morphological systematists typically exclude characters in which any or "too much" polymorphism is observed, and molecular systematists often avoid intraspecific variation by sampling a single individual per species. Recent empirical studies have suggested that polymorphic characters contain significant phylogenetic information but are more homoplastic than fixed characters. Given these two observations, should including polymorphic characters increase or decrease accuracy? We addressed this question using simulated data sets that also show a strong relationship between homoplasy and intraspecific variability. Data sets were generated with eight species, two alleles per locus, and a variety of branch lengths, number of loci, and sample sizes (individuals sampled per species). The data sets were analyzed using eight parsimony coding methods (with and without a priori and successive weighting) and different variability thresholds for excluding polymorphic characters. Excluding polymorphic characters decreased accuracy under almost all conditions examined, even when only the more variable characters were excluded. Sampling a single individual per species also consistently decreased accuracy. Thus, two common approaches for dealing with intraspecific variation in morphological and molecular systematics can give relatively poor estimates of phylogeny. In contrast, the unweighted frequency method, including polymorphic characters and sampling a reasonable number of individuals per species ($n \geq 5$), can give accurate results under a variety of conditions. [Accuracy; coding methods; computer simulations; parsimony; polymorphic characters; sample size; weighting.]

Intraspecific variation, the raw material of evolutionary change, is abundant in all kinds of systematic characters, including morphology (e.g., Alberch, 1983), allozymes (e.g., Buth, 1984), and DNA sequences (e.g., Kreitman, 1983). Yet, except for studies using allozyme data, polymorphism is rarely dealt with explicitly in empirical or theoretical studies of interspecific phylogeny reconstruction. For example, morphologists tend to exclude characters in which any or "too much" polymorphism is observed (Campbell and Frost, 1993; Wiens, 1995), molecular systematists using DNA data often sample only a single individual from each species, and most recent simulation studies of the performance of phylogenetic methods have modeled evolutionary change only in terms of fixed mutations (e.g., Wheeler, 1992; Charleston et al., 1994; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995). In this paper, we explore the possible effects of excluding polymorphic characters on phylogenetic accuracy, the similarity of estimated trees to the true phylogeny.

Judging by the rarity with which intraspecific variation in morphological characters is reported in phylogenetic analyses, the practice of excluding polymorphic characters seems to be extremely common. This practice is difficult to document because conventions for dealing with intraspecific variation are rarely discussed in morphological studies (but see Campbell and Frost, 1993; Wiens, 1993; Domning, 1994; Reeder and Wiens, 1996). The exclusion of polymorphic characters presumably has its justification in the idea that characters that are more variable within species will be less reliable for inferring relationships between species (e.g., Darwin, 1859; Simpson, 1961; Farris, 1966; Kluge and Farris, 1969; Mayr, 1969). The idea that polymorphic characters may be less reliable for inferring interspecific phylogeny has been supported by two empirical studies of lizards (Campbell and Frost, 1993; Wiens, 1995). In these studies, polymorphic characters were more homoplastic than "fixed" characters (given that the absence of intraspecific variation may be an artifact of sample size; Campbell and Frost, 1993), and there was a significant positive relationship between levels of homoplasy and intraspecific variability (Wiens, 1995). However, the authors of both studies also concluded that polymorphic characters do appear to contain phylogenetic information. Computer simulations have indicated that increasing the number of characters in a parsimony analysis will generally increase accuracy (e.g., Huelsenbeck and Hillis, 1993; Hillis et al., 1994) but that accuracy may decrease if the added characters exhibit greater overall homoplasy (Bull et al., 1993). Given these facts alone, it is difficult to predict whether including polymorphic characters should increase or decrease accuracy. Furthermore, both Campbell and Frost (1993) and Wiens (1995) advocated character weighting as a means of including polymorphic characters while accommodating their higher levels of homoplasy (successive weighting [Farris, 1969] and downweighting based on intraspecific variability [Farris, 1966], respectively) but could not address the effectiveness of these weighting schemes with their empirical data sets.

Computer simulations are an important tool for understanding the performance and behavior of phylogenetic methods (see Hillis, 1995, for a recent review). Simulations allow one to determine how methods perform at estimating the true phylogeny, and their simplification allows the conditions that affect method performance to be known and altered systematically (Hillis, 1995). Thus, one cannot use simulations to generate completely realistic data sets, but one can test methods under a broad range of known conditions and determine when methods will perform well or poorly (Huelsenbeck, 1995). Because of their simplicity, simulations offer insights that could not be gained by examining the accuracy of methods with real data sets from known phylogenies.

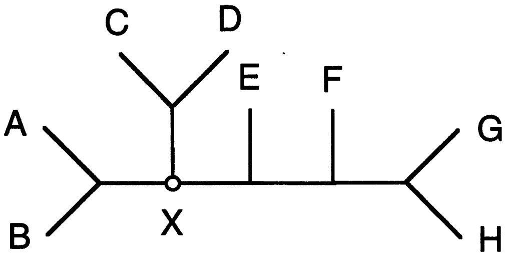

FIGURE 1. Topology of true tree used in simulations. X is the starting point used for determining initial frequencies.

In this study, we used simulations to test the effects of excluding versus including polymorphic characters on the accuracy of parsimony analysis. Because there are at least eight different parsimony methods for coding polymorphic characters (Wiens, 1995), we also compared the relative accuracy of these methods. We also addressed the effects of sample size (number of individuals per species) on accuracy and the success of the two weighting schemes recently advocated for dealing with polymorphic data. Although a plethora of genetic distance and maximum likelihood methods have also been developed for dealing with polymorphic data (see Swofford et al., 1996, for a recent review), they are not used on morphological data and are not particularly relevant to the practice of excluding polymorphic characters. The relative accuracy of these methods will be addressed elsewhere (Wiens and Servedio, in review).

MATERIALS AND METHODS

The simulated data sets for this study consist of allele frequencies for loci with two alleles per locus evolving by random genetic drift along a known phylogeny of eight species (Fig. 1). A character (locus) is considered polymorphic if it has two alleles within one or more terminal taxa. As in real data sets, a given locus in a given species may be fixed or polymorphic, and a given locus may be fixed in all species, polymorphic across all species, or fixed in some species and polymorphic in others.

For each locus, the data were generated as follows. First, the initial frequency of one of the alleles at the starting point (the internal node X, Fig. 1) was selected randomly (given that there are only two alleles, allele frequencies at a locus can be described based on the frequency of only one allele). The initial frequency was chosen randomly from a uniform distribution (i.e., any number from 0 to 1.000 inclusive had an equal probability of being selected). From this starting frequency, the frequency of the allele changed independently along each branch by the Wright–Fisher model of genetic drift (Fisher, 1930; Wright, 1931). To calculate the frequency at the end of a branch, the probabilities of fixation and loss were first calculated following Kimura (1955) based on the initial frequency, the population size, and the number of generations during which the lineage evolved. A random number was then used to determine if the allele would be fixed or lost or if the locus would remain polymorphic at the end of the branch, based on these probabilities. If the allele was lost or fixed, no further changes occurred. If the locus remained polymorphic, the terminal frequency of the allele for that lineage was then determined. The terminal frequency was calculated by first determining the probability density function for the allele frequency (given the initial frequency, population size, and number of generations), again following Kimura (1955). Then the rejection method for generating random deviates (Press et al., 1988) was used to pick the final frequency, based on this distribution. Lineages began their evolution with the terminal frequency of their ancestors. Mutations were assumed to be rare enough to be ignored, so loci that became fixed for an allele remained fixed. This assumption is justified because, given a two-allele model and the simulated population size (50,000 diploid individuals), mutation rates would have to be relatively high for mutation to predominate over drift in determining allele frequencies (Kimura and Crow, 1964). The branching pattern of the simulated phylogeny (Fig. 1) was chosen to be 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).

Three parameters varied in this study: number of loci (characters), sample size (individuals per species), and branch length. Branch length, defined here as the expected amount of character change for a lineage (following Huelsenbeck and Hillis, 1993), determined the rate of change in frequencies and the probabilities of alleles becoming fixed or lost. Thus, branch length determined the underlying levels of polymorphism in the simulated data sets. 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). Herein, we use this ratio synonymously with branch length. Branch length can be changed by changing the number of generations between splitting events while holding the effective population size constant. In our study, branch lengths ranged from a minimum of 0.2 (e.g., 10,000 generations with a population size of 50,000) to a maximum of 2.0 (100,000 generations). Alternatively, the differences in branch length can be viewed as differences in effective population size between lineages, with longer branch lengths being the result of, for example, founder effects or population bottlenecks. At the shortest branch lengths there is little probability of fixation or loss (exact values depend on the initial frequency) and generally only small changes in frequency, whereas at the longest branch lengths the chances of fixation and loss are very high and allele frequencies are effectively randomized between splitting events (see Kimura, 1955). Shorter or longer branch lengths could have been simulated, but these branch lengths (0.2–2.0) allowed for conditions where almost all characters are polymorphic and conditions where almost all characters are fixed. Branch lengths were either held constant across all the lineages to test the effects of a given branch

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 is ≥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 (0 ↔ 1 ↔ 2 ↔ 0)
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

length on accuracy or they were varied randomly across lineages. For runs with random branch lengths, the number of generations during which a lineage evolved (or the effective population size) was chosen randomly, but branch length was held constant across loci for that lineage. 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. One hundred replicated matrices were created for each combination of branch lengths and number of characters.

Sample size is also an important parameter because whether or not a character is considered polymorphic may depend on how many individuals are sampled (Campbell and Frost, 1993). For each set of 100 simulated data matrices, five sets of 100 matrices were created, one with perfect sampling (the original matrices) and one each with reduced sample sizes of 10, 5, 2, and 1 individual per species. Thus, each subsampled matrix corresponded to a single original array of frequencies. The allelic composition of each individual sampled from the simulated matrices (for each species and locus) was determined by randomly choosing a number and considering an allele present if the number was less than or equal to the frequency of the allele in that species (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); because the simulated species were diploids, this determination was made twice for each individual. The new matrices of frequencies were then calculated based on these sampled individuals.

Two sets of phylogenetic analyses were performed. In the first set, the accuracy of excluding all polymorphic characters (the fixed-only approach of Campbell and Frost, 1993) was tested against the accuracy of eight different methods for including polymorphic characters. For this set of analyses, the fixed-only approach meant that characters were excluded if they exhibited any polymorphism in any species, thus reducing the overall number of characters. However, the determination of whether a character was polymorphic or fixed was based only on the individuals sampled (i.e., most loci are considered to be fixed with a sample size of one diploid individual regardless of the actual allele frequencies). The eight parsimony coding methods are summarized in Table 1 (reviewed/described in more detail by Wiens, 1995). For the frequency method, each taxon was given a unique character state, and the Manhattan distance (for a given character; Swofford and Berlocher, 1987) between each species was used to weight changes between these states in a step matrix (e.g., the method used for allozyme data sets by Wiens, 1995, suggested by D. Hillis). The performance of the methods was tested for all possible combinations of different branch lengths (random, 0.2, 0.8, 1.4, 2.0), number of loci (10, 25, 50, 75, 100), and sample size (all individuals, 10, 5, 2, 1).

For the second set of analyses, the effects of excluding characters with different levels of variability and the performance of a priori and a posteriori (successive) weighting were tested. Given that polymorphic characters contain significant phylogenetic information as a whole but contain increasing homoplasy with increasing variability, accuracy may be improved by excluding only the most polymorphic characters (as we believe is commonly done in real data sets). Levels of variability were calculated for each character using the mean intraspecific variability (MIV) as an index of variability (Wiens, 1995). The MIV is the sum of the frequencies of the rarer of two traits (alleles) for each species multiplied by 200 (to allow the index to range from 0 to 100) and then divided by the number of taxa. Thus, the MIV for a given character has a maximum of 100 when all the species are variable at a frequency of 50% and a minimum of 0 when there is no intraspecific variation in any of the species. The effects of excluding characters with MIV indices above 25, 50, and 75 were tested. Furthermore, the effects of excluding characters with high levels of variability relative to the other characters in the data set were tested. Analyses were performed excluding any characters with a MIV score greater than the mean MIV for the data set, greater than the mean MIV × 0.5, and greater than the mean MIV × 1.5. In simulated data sets with reduced sample sizes, levels of variability were based on the sampled allele frequencies, not on the actual allele frequencies. Because these six approaches for excluding polymorphic characters usually involved using at least some (less variable) polymorphic characters, the accuracy of each of the six exclusion criteria was tested using each of the eight polymorphism coding methods. The two weighting schemes were also applied to each of the coding methods. Because of the large number of approaches compared (64 per data matrix), the second set of analyses involved a more limited set of conditions. The number of loci and the sample size were held constant at intermediate values (50 loci, five individuals), and branch lengths were long (2.0) or short (0.2) or varied randomly.

Two weighting schemes were tested. Farris (1966) suggested weighting characters by the reciprocal of their intraspecific variability, and this method was also advocated by Wiens (1995). In this study, this scheme was implemented by weighting each character by 100 − MIV. Thus, characters with no intraspecific variation received a weight of 100, and loci having both alleles present at a frequency close to 50% in all taxa approached a weight of 0. Successive weighting can be implemented using a variety of measures of goodness of fit (e.g., consistency index [Kluge and Farris, 1969], retention index [Farris, 1989], and rescaled consistency index [Farris, 1989]) and ways for determining goodness-of-fit values from multiple equally parsimonious trees from the initial (unweighted) analysis (e.g., mean fit among shortest trees, highest fit among trees, lowest fit). Following the recommendations of Campbell and Frost (1993), the maximum value of the rescaled consistency index among the shortest trees from the unweighted analysis was used as the weighting function in this study. Our preliminary results suggested that the choice among the options listed does not greatly impact the results (Wiens and Servedio, unpubl. data). The frequency-bins method (Wiens, 1995) was used to code polymorphic data as frequencies for successive weighting because the Manhattan distance–step matrix approach does not allow goodness-of-fit statistics to be calculated.

For each set of conditions, the accuracy of methods was scored as the similarity between the true phylogeny (Fig. 1) and the strict consensus of the shortest estimated trees, averaged across the 100 replicated data sets. Similarity was measured using Colless's (1980) consensus fork index, the proportion of nodes in common between the true and estimated trees. An alternative approach to measuring performance would be to use the average of the mean similarity between the true tree and each of the shortest estimated trees for each matrix (e.g., counting the presence of the correct clade in one of multiple shortest trees as indicative of method success; Hillis et al., 1994). The consensus fork index may favor methods that give higher resolution (fewer polytomies), whereas the mean similarity measure may favor methods that give poorer resolution (it gives partial credit for polytomies). We consider the correct resolution of a clade to be the most relevant index of accuracy and therefore prefer the consensus fork index.

All phylogenetic analyses were implemented using PAUP 3.1.1 (Swofford, 1993) with the branch-and-bound search option. The protocol for generating frequency data was designed by M.R.S. and J.J.W. and programmed by M.R.S. in C. The programs for coding, screening, and subsampling these data were written in C by J.J.W.

For the simulated data sets to be relevant to the question of excluding versus including polymorphic characters, they should exhibit a significant correlation between levels of homoplasy and intraspecific variability, as in real data sets (Wiens, 1995). The general methodology used by Wiens (1995) to examine the relationship between homoplasy and variability was applied to the simulated data sets. Ten matrices were sampled from each of six simulated conditions. These conditions were random branch lengths, short branch lengths (0.2), and long branch lengths (2.0), with perfect sampling of individuals and two individuals sampled per species. For each of six coding methods (the polymorphic and unscaled methods give almost identical results to the missing and scaled methods, respectively), the homoplasy index (1 − consistency index) was calculated for each character. Each matrix contained 100 characters, and the trees were assumed to be correctly estimated. The relationship between homoplasy and variability (using the MIV) was analyzed using Spearman's rank correlation for the characters in each of the 10 data sets for the six sets of conditions using the six coding methods. For these analyses, the frequency-bins method instead of the Manhattan distance–step matrix approach was used to code polymorphic characters as frequencies because of the difficulty in calculating consistency indices for step matrix characters.

It may be argued that the assumptions of these simulations predispose the polymorphic characters to contain useful phylogenetic information. These assumptions include the fact that polymorphisms are shared because of common ancestry and that new lineages begin their evolution with the terminal frequencies of their ancestors. However, the parallel retention, fixation, or loss of a polymorphic allele in unrelated lineages can render the polymorphic characters positively misleading, and long branch lengths can effectively randomize allele frequencies between splitting events. Furthermore, the correlation between levels of homoplasy and variability in the simulated data sets (see below) is evidence that polymorphic characters are not necessarily predisposed to be informative in these simulations, but rather the opposite.

Results

Correlation between Homoplasy and Variability

The results of the analyses of the relationship between homoplasy and variability (Table 2) show that the strength of the correlation is sensitive to coding method, sample size, and branch lengths. In general, the results using variable branch lengths are similar to those from the comparably sized *Urosaurus* data set (nine species; Wiens, 1995), with the frequency and majority methods showing a strong positive correlation between homoplasy and variability (highly significant for the frequency method) and most methods showing a strong relationship with long and random branch lengths (depending on the sample size). These results suggest that the simulated polymorphic characters are similar to real polymorphic characters in this important aspect and thus can be used to test the effects of excluding, including, or downweighting polymorphic characters.

Accuracy of Including versus Excluding Polymorphic Characters

The results of the first set of analyses are shown in Figure 2. Under almost all conditions, the most accurate method is clearly the frequency method, whereas the least

TABLE 2. Spearman rank correlations of homoplasy and intraspecific variability for simulated data sets for three kinds of branch lengths (random, 0.2, 2.0), perfect (all) and limited ($n = 2$) sample sizes, and six coding methods. The mean results from 10 replicated matrices with 100 loci each are reported. Because of extensive polymorphism, there are no informative characters for the Missing method at low branch lengths (0.2).

	n = all		n = 2	
Method	Rho	P^a	Rho	P
Length = random				
Any-instance	0.301	0.112	0.383	0.005
Frequency	0.625	0.0001**	0.513	0.004
Majority	0.520	0.002	0.482	0.004
Missing	0.460	0.022	0.290	0.161
Scaled	0.341	0.069	0.461	0.002
Unordered	0.392	0.028	0.535	0.001
Length = 0.2				
Any-instance	−0.041	0.602	−0.024	0.597
Frequency	0.325	0.029	0.375	0.011
Majority	0.551	0.001*	0.402	0.011
Missing			0.462	0.082
Scaled	−0.071	0.578	0.012	0.515
Unordered	−0.071	0.578	−0.033	0.578
Length = 2.0				
Any-instance	0.433	0.007	0.417	0.018
Frequency	0.471	0.003	0.444	0.014
Majority	0.468	0.002	0.443	0.016
Missing	0.405	0.025	0.363	0.051
Scaled	0.430	0.008	0.427	0.014
Unordered	0.477	0.002	0.365	0.045

[a] Significance values are based on a sequential Bonferroni correction for multiple tests (Holm, 1979; Rice, 1989). * = correlation significant at 0.05 level; ** = correlation significant at 0.01 level.

accurate method is the exclusion of polymorphic characters. The accuracy of all the approaches becomes more similar when sample sizes are small ($n = 1$) and branches are long (2.0), i.e., when there is little or no polymorphism. As branch lengths decrease and sample sizes increase, the superiority of the frequency method and the inferiority of the fixed-only method become increasingly clear. These results also show that sampling a small number of individuals (e.g., $n = 1$ or 2) significantly decreases accuracy under almost all conditions, with the greatest decreases occurring when branch lengths are low and the number of characters is small. Although the impact of sampling only one or two individuals can be severe, the results also show that sample sizes greater than five individuals per species do not greatly increase accuracy.

Effects of Different Variability Thresholds and Weighting Schemes

The results of the second set of analyses are shown in Figure 3. For most combinations of variability thresholds, branch lengths, and coding methods, excluding polymorphic characters (Fig. 3, A–G) either decreased or did not affect accuracy. The only exceptions are (1) random branch lengths with the majority method excluding characters with MIV > 25, MIV > 50, and MIV > mean MIV × 1.5, (2) random branch lengths with the frequency method excluding characters with MIV > 50, (3) branch lengths of 0.2 with the majority method excluding characters with MIV > 50 and MIV > mean MIV × 1.5, and (4) branch lengths of 2.0 with the any-instance method and excluding characters with MIV > 25. All of these cases involve <4% increase in accuracy as the result of excluding polymorphic characters (the increase for the frequency method is only 0.4%). The use of mean MIVs as a criterion for excluding polymorphic characters tends to decrease accuracy more than does the use of absolute MIV values (25, 50, 75).

Both a priori and successive weighting increased accuracy using most methods (Fig. 3, H and I), but usually by <10%. A priori weighting generally increases accuracy more than does successive weighting. Unfortunately, successive weighting decreased the accuracy of the frequency method under all three branch lengths (by almost 10% with random branch lengths and branch lengths of 0.2), and a priori weighting reduced accuracy with branch lengths of 0.2 and 2.0 (the method increases accuracy by only 0.4% with random branch lengths). Despite the slight improvements sometimes produced by character weighting or excluding characters, the most generally accurate method for dealing with intraspecific variation appears to be an unweighted frequency analysis including all polymorphic characters.

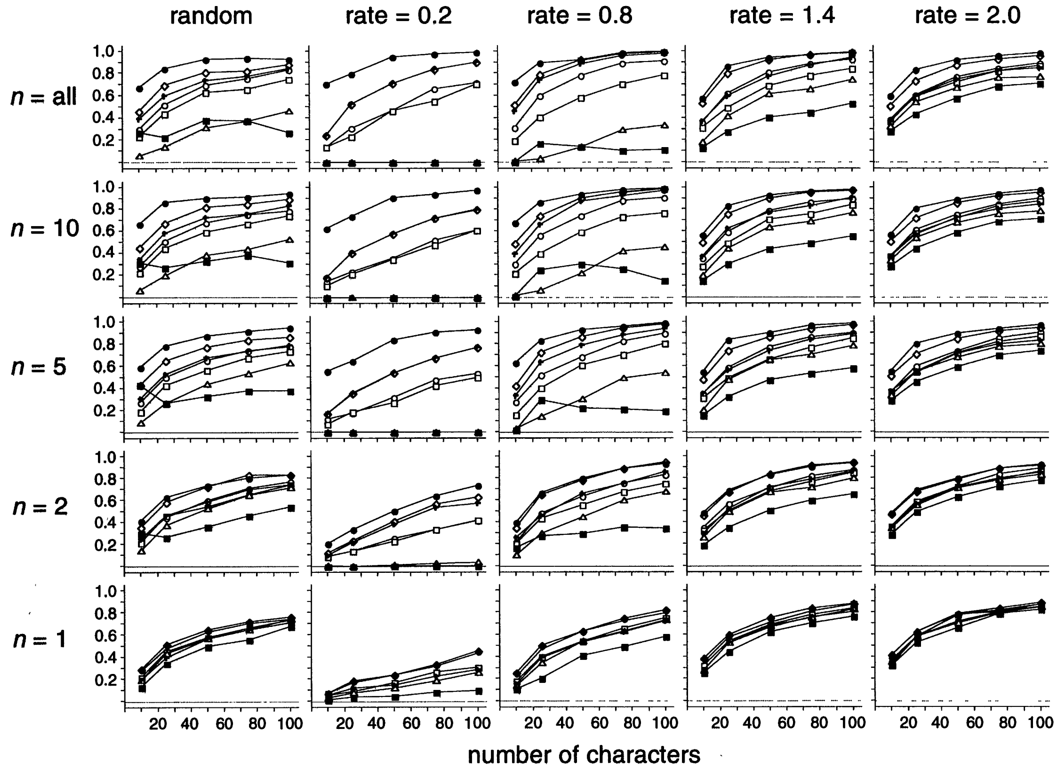

FIGURE 2. Accuracy of phylogenetic methods for including and excluding polymorphic characters at different sample sizes (n), branch lengths (rate), and numbers of characters. ● = frequency; ◇ = scaled and unscaled (almost identical for all conditions); ✚ = unordered; ○ = any-instance; □ = majority; △ = missing and polymorphic; ■ = fixed-only (excluding all polymorphic characters). Polymorphic coding methods are described in Table 1 (and by Wiens, 1995). Each point is the mean accuracy for 100 replicated matrices.

DISCUSSION

Performance of Methods

The results of this study strongly suggest that the costs of excluding informative polymorphic characters outweigh the potential benefits of excluding characters with greater average homoplasy. This statement is generally true even when only the most variable (and potentially most homoplastic) characters are excluded. The analysis using different variability levels as thresholds for exclusion was only performed under a limited set of conditions (50 characters, n = 5 individuals). Under other conditions (e.g., more characters), the different exclusion criteria may have had a better relative performance. But even if this were the case, Figure 2 shows that, given >50 characters (for five or more individuals), the unweighted frequency method has an accuracy of ≥95% for all branch lengths examined; it would therefore be impossible for excluding polymorphic characters to improve on this accuracy level very much. Furthermore, the fixed-only approach seems to become less accurate with larger sample sizes while the frequency method becomes more accurate (Fig. 2), suggesting that testing these exclusion approaches with larger sample sizes would not support excluding polymorphic characters, regardless of the variability threshold. An analysis using data sets with a stronger relationship between homoplasy and variability might show a benefit for excluding polymorphic characters. However, morphological data for phrynosomatid lizards (Reeder and Wiens,

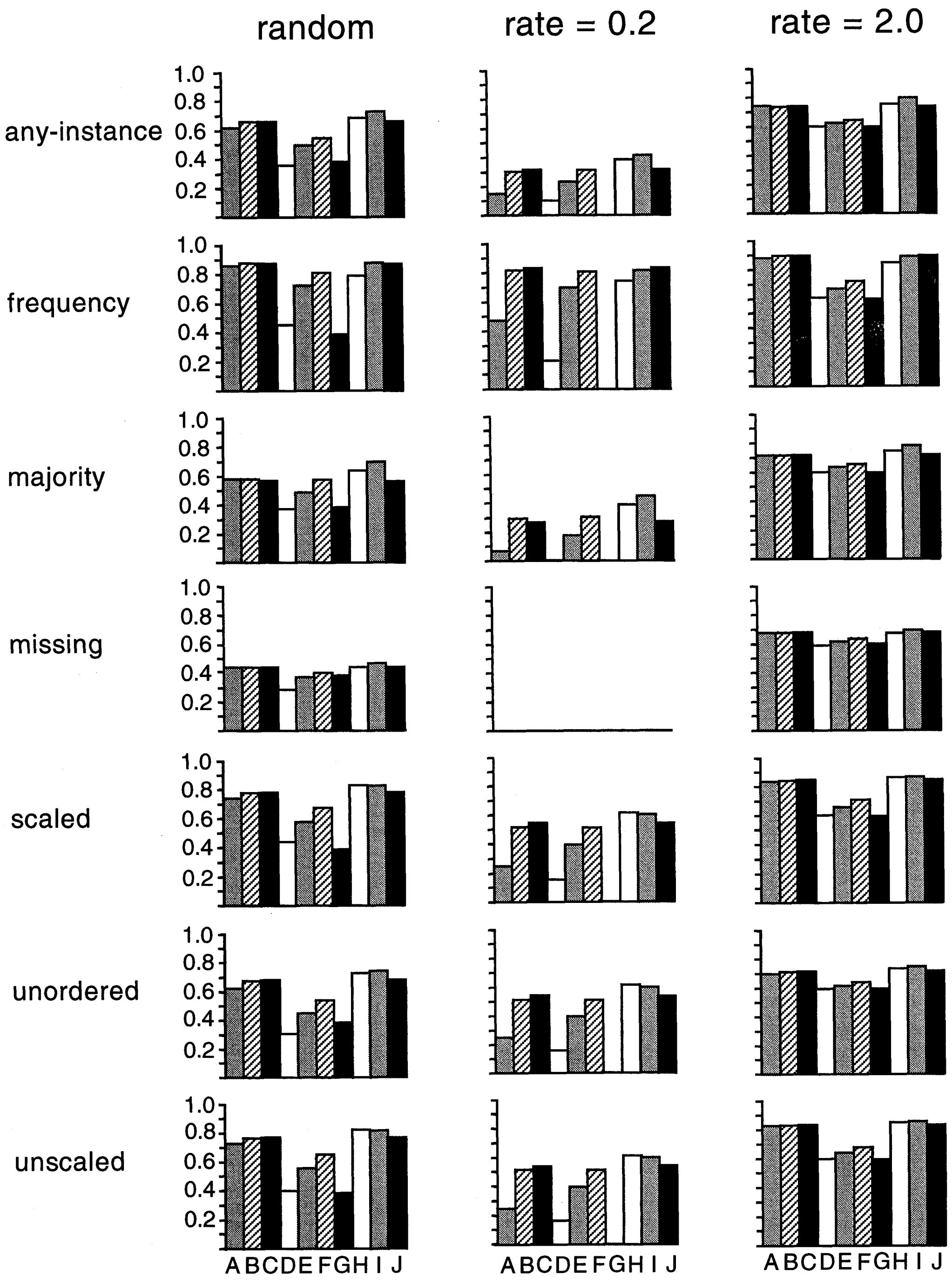

FIGURE 3. Accuracy of phylogenetic methods using different variability thresholds for excluding polymorphic characters, different weighting schemes, and different coding methods. Data sets have 50 characters and five individuals sampled per species. A = excluding characters with MIV > 25; B = with MIV > 50; C = with MIV > 75; D = with MIV > (mean MIV of data set)(0.5); E = with MIV > mean MIV; F = with MIV > (mean MIV)(1.5); G = excluding all polymorphic characters (fixed-only); H = including all characters and with successive weighting; I = including all characters and with variable characters downweighted; J = unweighted analysis including all characters. Each bar represents the mean results from 100 replicated matrices.

1996) exhibit a particularly strong relationship between homoplasy and variability (Wiens, 1995), yet excluding polymorphic morphological characters decreased the probability of recovering clades supported by separately analyzed molecular and morphological data sets (Wiens, in review).

The relative performance of the nine methods (the eight coding methods and the fixed-only approach) appears to be determined largely by the amount of phylogenetic information that each can utilize. The frequency method uses the most fine-grained information possible and outperforms the other methods under almost all conditions. The only exceptions involve a few cases where branches are long and equal and the number of characters is relatively large. The scaled and/or unscaled methods may be slightly superior under these conditions (and then only at some sample sizes), but the frequency method generally is very accurate under these conditions as well (>90%), which strongly suggests the frequency method is the safest method to apply to real data sets where the specific evolutionary conditions are unknown. The scaled, unscaled, and unordered methods recognize three character states (a trait is absent, polymorphic, or fixed) and generally were exceeded in accuracy only by the frequency method. The two ordered methods (absent to polymorphic to fixed; scaled and unscaled) generally outperformed the unordered method, and the scaled method was sometimes slightly more accurate than the unscaled method (not shown in Fig. 2). The any-instance and majority approaches recognize only two possible states and consistently performed worse than the frequency, scaled, unscaled, and unordered methods. The missing and polymorphic methods (which behave identically given only two traits/alleles; Wiens, 1995) treat all variable species for a polymorphic character as unknown and performed far worse than the preceding methods. The fixed-only approach throws out the polymorphic characters altogether and generally performed the worst of all. The missing, polymorphic, and fixed-only methods, probably the most widely used in morphological systematics, performed increasingly worse with increasing sample size, while the accuracy of the other methods improved. The explanation for this behavior is clear; as more individuals are sampled, more polymorphism is detected and more characters are either thrown out (fixed-only method) or rendered uninformative (missing and polymorphic methods). When underlying levels of polymorphism are high, these three methods are usually unable to resolve any parts of the tree correctly.

Some may find it surprising that polymorphic characters and the frequency method perform well under conditions where frequencies are effectively randomized between speciation events (e.g., by long branch lengths). For example, many authors have argued that frequency data should not be used in phylogenetic analysis because of their potential variability in space and time (e.g., Crother, 1990). However, when the frequency of a neutral allele is modified extensively there is a high probability that the allele will be lost or fixed (Kimura, 1955), and allele fixations and losses are phylogenetically informative for most methods. Through the synapomorphic loss and fixation of alleles, polymorphic characters seem to retain considerable phylogenetic information, despite the high rates of change. The frequency method also makes use of the fact that, even under high rates of change, an allele with a very high frequency is more likely to be fixed than lost and vice versa (e.g., a change from 95% to 100% receives a smaller weight than a change from 5% to 100%), whereas this information is ignored by most qualitative methods (e.g., any-instance, scaled, unscaled, unordered).

The results of this study also suggest that the qualitative parsimony methods (e.g., any-instance, scaled, unordered, unscaled) appear to be sensitive to differences in branch lengths among lineages. Specifically, the average for the randomly varied branch lengths was 1.1, but the qualitative methods were less accurate (and less accurate relative to the frequency method) for the variable lengths than for lengths of 0.8 and 1.4 (Fig. 2). A more detailed study of the effects of branch length on the performance of these methods (using the same evolutionary model with a four-taxon tree) suggested that the qualitative methods are particularly sensitive to the problem of long branch attraction (Felsenstein, 1978) and become positively misled more easily than do the methods that make use of frequency information (Wiens and Servedio, in review). Apparently these qualitative methods are misled by the fixation and loss of alleles in unrelated terminal lineages (with long branches) while ignoring informative changes in frequency along internal branches. Previous authors have asserted the superiority of the frequency approach based on the sensitivity of qualitative methods to finite sample sizes (Swofford and Berlocher, 1987; see Wiens, 1995, for an empirical demonstration), yet these qualitative methods appear to have additional problems that are unrelated to sample size.

Overall, the performances of successive and a priori weighting were somewhat disappointing. These methods improve the accuracy of the nonfrequency methods probably in part because they increase their resolving power (their ability to discriminate among possible trees) and because they allow these methods to give less weight to small changes in frequency (as would be common in more variable characters, which are downweighted). However, the unweighted frequency method produces highly resolved trees and gives less weight to small changes in frequency, leaving these weighting schemes little on which to improve.

Comparison with Empirical Results

We used simulated genetic data to address the accuracy of a common practice in morphological systematics, the exclusion of polymorphic characters. Although we made a number of simplifying assumptions in generating our data (e.g., one locus per character; two alleles per locus; phenotype is equivalent to genotype; no selection, mutation, or geographic variation), there is broad agreement between our results and those from a recent study of polymorphic characters using morphological and allozyme data sets (Wiens, 1995). As would be expected given the results of this study, Wiens (1995) found that (1) there is a significant correlation between levels of homoplasy and intraspecific variability using the majority and frequency methods, (2) frequency-coded polymorphic characters contain significant phylogenetic information, whereas other methods yield random levels of noise for some data sets, and (3) among the eight coding methods, the frequency method generally performs best for five criteria that are likely to be correlated with accuracy (number of shortest trees, number of informative characters, phylogenetic signal, nodes supported by bootstrapping, and sensitivity of tree topology to variation in sample size). Some might argue that the extensive levels of polymorphism simulated in this study would not be found in real morphological data sets. Few morphological studies have addressed the levels of intraspecific variability in discrete phylogenetic characters with large sample sizes, yet Wiens (1993) found that 23 of 24 informative morphological characters in *Urosaurus* were variable within one or more species. Some might also question whether a genetic drift model is appropriate for simulating the evolution of morphological characters. Felsenstein (1988) argued that seemingly discrete, polymorphic morphological characters may evolve by random genetic drift, although he discussed a somewhat more complex model than simulated here. Adding complexity (e.g., multiple loci per character, environmental effects) to our model might impact the results, but there is no a priori reason to expect increasing complexity to favor discarding information from trait frequencies and polymorphic characters.

Effects of Sampling Single Individuals

As with the exclusion of polymorphic morphological characters, the practice of sampling a single individual per species is common (especially in DNA studies) but rarely justified. Our results suggest that

TABLE 3. Accuracy of the frequency method with different branch lengths, demonstrating the effects of increasing sample size versus increasing number of characters.

Branch length	Sample size	No. characters 25	50	100
Random	1	0.512	0.646	0.762
	2	0.630	0.736	0.834
0.2	1	0.184	0.238	0.456
	2	0.330	0.504	0.738
0.8	1	0.502	0.630	0.748
	2	0.666	0.800	0.888
1.4	1	0.598	0.750	0.882
	2	0.686	0.832	0.948
2.0	1	0.624	0.792	0.892
	2	0.696	0.798	0.926

sampling single individuals can greatly decrease accuracy under many conditions, especially when the number of characters is low or when levels of polymorphism are high (e.g., short branch lengths). For example, with short branch lengths (ratio = 0.2) and 50 characters, the frequency method has an accuracy of 91% with 10 individuals per species and only 24% with one individual per species. Conventional wisdom and some empirical resampling studies have suggested that it is more effective to sample more characters rather than to sample more individuals (e.g., Hillis, 1987; Kesner, 1994). However, the results of this study suggest that the opposite is true for shorter branch lengths; for example, more accurate results would be obtained by sampling twice as many individuals as characters (Table 3). For random and long branch lengths, more accurate results would indeed be obtained by sampling twice as many characters rather than by sampling twice as many individuals (Table 3). Although the results for longer branch lengths would seem to support the emphasis on sampling characters over individuals, there is an underlying assumption inherent in this discussion that is highly questionable: that it is just as easy to double the number of characters as it is to double the number of individuals. The limited number of genes routinely used in phylogenetic analyses of DNA sequence data and the limited number of characters and loci sampled in morphological and allozyme studies suggest that it is much more difficult to sample more characters for a given phylogenetic problem than to sample more individuals (although rare species may make the reverse true). Sampling more individuals may often be a more efficient way to improve phylogenetic results, and the problem of intraspecific variation should be addressed routinely by all systematists, including those working with DNA data (see also Smouse et al., 1991).

CONCLUSIONS

Recent empirical studies have suggested that polymorphic characters contain significant phylogenetic information but are more homoplastic than fixed characters. Thus, empirical data are ambiguous as to whether polymorphic characters should be included in phylogenetic analyses. Our simulated data sets also show a general correlation between homoplasy and variability but are unambiguous about the effects of excluding polymorphic characters. Excluding polymorphic characters decreases accuracy under almost all conditions examined, even when only the most variable characters are excluded. Sampling a single individual per species, a common practice in DNA studies, also consistently decreases accuracy. Thus, the most common approaches for dealing with intraspecific variation in morphological and molecular systematics can give relatively poor estimates of phylogeny. In contrast, the unweighted frequency method, including polymorphic characters and sampling a reasonable number of individuals per species, can give accurate results under a variety of conditions. These results are in broad agreement with recent empirical studies that show that polymorphic characters (both molecular and morphological) can be highly informative when treated as frequencies. The inclusion of polymorphic characters and information on the frequency of traits within species (based on multiple individuals) may greatly improve the accuracy and rigor of both molecular and morphological systematics.

ACKNOWLEDGMENTS

We are grateful to R. Servedio for invaluable mathematical and computational consultations and to J. Bull and J. Huelsenbeck for useful suggestions and advice. M.R.S. acknowledges the support of an NSF Graduate Fellowship. We thank D. Cannatella, P. Chu, D. Faith, B. Livezey, S. Poe, and two anonymous reviewers for comments on the manuscript.

REFERENCES

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

BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD, AND P. J. WADDELL. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42:384–397.

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.

CHARLESTON, M. A., M. D. HENDY, AND D. PENNY. 1994. The effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. J. Comput. Biol. 1:133–151.

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.

DARWIN, C. 1859. On the origin of species. Harvard Univ. Press, Cambridge, Massachusetts.

DOMNING, D. P. 1994. A phylogenetic analysis of the Sirenia. Proc. San Diego Soc. Nat. Hist. 29:177–189.

FARRIS, J. S. 1966. Estimation of conservatism of characters by constancy within biological populations. Evolution 20:587–591.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18:374–385.

FARRIS, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 5:417–419.

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

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

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 Press, Oxford, England.

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.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6:65–70.

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

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

KESNER, M. H. 1994. The impact of morphological variants on a cladistic hypothesis with an example from a myological data set. Syst. Biol. 43:41–57.

KIMURA, M. 1955. Random genetic drift in multi-allelic locus. Evolution 9:419–435.

KIMURA, M., AND J. F. CROW. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725–738.

KLUGE, A. G., AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.

KREITMAN, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature 304:412–417.

KUHNER, M. K., AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

MAYR, E. 1969. Principles of systematic zoology. McGraw-Hill, New York.

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.

REEDER, T. W., AND J. J. WIENS. 1996. Evolution of the lizard family Phrynosomatidae as inferred from diverse types of data. Herpetol. Monogr. 10:43–84.

RICE, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223–225.

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.

SIMPSON, G. G. 1961. Principles of animal taxonomy. Columbia Univ. Press, New York.

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.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, version 3.1. 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., 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.

WHEELER, W. C. 1992. Extinction, sampling and molecular phylogenetics. Pages 205–215 *in* Extinction and phylogeny (M. Novacek and Q. Wheeler, eds.). Columbia Univ. Press, New York.

WIENS, J. J. 1993. Phylogenetic systematics of the tree lizards (genus *Urosaurus*). Herpetologica 44:399–420.

WIENS, J. J. 1995. Polymorphic characters in phylogenetic systematics. Syst. Biol. 44:482–500.

WRIGHT, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Received 1 April 1996; accepted 1 September 1996
Associate Editor: Daniel Faith