Syst. Biol. 47(3): 427–444, 1998

Testing Phylogenetic Methods with Tree Congruence: Phylogenetic Analysis of Polymorphic Morphological Characters in Phrynosomatid Lizards

JOHN J. WIENS

Section of Amphibians and Reptiles, Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213-4080, USA; E-mail: wiensj@ clpgh.org

Abstract.— Congruence between trees from separately analyzed data sets is a powerful approach for assessing the performance of phylogenetic methods but has been applied primarily to the analysis of molecular data. In this study, different methods for treating polymorphic characters were compared using morphological data from phrynosomatid lizards. Clades were identified that are both traditionally recognized and supported by recent molecular analyses, and species were sampled from these clades to make three "known" phylogenies of eight species each. The ability of different methods to estimate these "known" phylogenies with a finite sample of characters was tested. The phylogenetic methods included eight parsimony methods for coding polymorphism, three distance approaches (UPGMA, neighbor joining, and Fitch–Margoliash) applied to two genetic distance measures (Nei's and the modified Cavalli-Sforza and Edwards chord distance), and continuous maximum likelihood. The effects of excluding polymorphic characters and character weighting (a priori and successive) were also tested. Among the different parsimony approaches, the fixed-only method (excluding all polymorphic characters) performed relatively poorly, whereas the frequency method (including all polymorphic characters) performed relatively well. However, frequency-based distance methods consistently outperformed parsimony, especially with a small sample size (n = 1 individual per species). These results agree closely with those from recent simulation studies of polymorphic data and argue against the common practices of excluding polymorphic morphological characters, ignoring the frequencies of traits within species, and the exclusive use of parsimony to analyze morphological data. [Accuracy; congruence; distance; likelihood; morphology; parsimony; polymorphic characters.]

A typical phylogenetic analysis involves a bewildering array of decisions, including what type of data to sample (i.e., molecular, morphological), what phylogenetic method to apply (i.e., distance, likelihood, parsimony), whether or not to order or weight characters, and which taxa and characters to include or exclude. These decisions can be crucial, because different choices can lead to different trees, and at least some of these trees must be incorrect (because only one tree can be right). Research on phylogenetic methods can help determine which approaches are more likely to lead to correct or incorrect trees, and can help inform the methodological decisions that empirical systematists must make.

Hillis (1995) recently reviewed approaches for assessing the accuracy of phylogenetic methods, where accuracy is the ability of a method to estimate the true phylogeny. These approaches include (1) computer simulations, (2) known, laboratory-produced phylogenies, (3) congruence between trees from different data sets, and (4) statistical analyses. In general, the tree-congruence approach (hereafter, simply congruence) involves applying different analytical methods to one data set and determining which treatments consistently yield the tree that is well supported by other types of data (also called taxonomic congruence; Mickevich, 1978). Congruence is a particularly powerful tool for assessing accuracy because it uses real data, and avoids the necessary oversimplifications of simulations and laboratory-produced phylogenies (Miyamoto and Fitch, 1995). Despite these advantages, the congruence approach has not been widely used for testing phylogenetic methods, especially in comparison to computer simulations (Hillis, 1995). Furthermore, the questions addressed using congruence have mostly involved specific issues in the analysis of DNA sequence data, such as weighting of DNA characters (Allard and Miyamoto; 1992; Dixon and Hillis, 1993; Miyamoto et al., 1994; Cunningham, 1997) and DNA sequence alignment

(Wheeler, 1995; Titus and Frost, 1996). Few recent studies have applied this criterion to questions in morphological analysis (e.g., Slowinski, 1993; Thiele, 1993). Congruence was used extensively in the phenetics versus cladistics debate in the early 1980s (e.g., Colless, 1980; Schuh and Polhemus, 1980; Schuh and Farris, 1981; Rohlf et al., 1983; Sokal and Shao, 1985), although these studies used other criteria for comparing method performance besides accuracy per se (i.e., stability and predictivity of classifications).

This study applies the congruence approach to the phylogenetic analysis of morphological data, specifically, the relative accuracy of different ways of treating polymorphic morphological characters. Morphologists typically exclude characters in which any or "too much" polymorphism is observed (Campbell and Frost, 1993; Wiens, 1995), a practice seemingly rooted in the idea that characters that are highly variable within species will be less reliable for inferring the phylogeny between species. Empirical studies (Campbell and Frost, 1993; Wiens, 1995) have concluded that polymorphic morphological characters are more homoplastic (which supports their exclusion) but nevertheless contain significant signal (which supports their inclusion). Thus, the impact on accuracy of including or excluding polymorphic characters is difficult to predict. When polymorphic characters are included, different methods of coding, ordering, and weighting these data can give radically different trees for the same set of taxa and characters (Wiens, 1995). Campbell and Frost (1993) and Wiens (1995) advocated character weighting as a means to include polymorphic characters while accommodating their higher levels of homoplasy (successive weighting [Farris, 1969] and downweighting based on intraspecific variability [Farris, 1966], respectively), but did not address the accuracy of these weighting schemes.

Recent simulation studies have addressed the accuracy of different methods for treating polymorphic characters (Wiens and Servedio, 1997, 1998). These studies found that the most generally accurate parsimony approach is the unweighted frequency method, including all polymorphic characters, and that distance and likelihood methods may outperform parsimony, especially with small sample sizes. However, as with all simulation studies, these simulations relied on numerous simplifying assumptions, which may limit their relevance to real data in general and morphological data in particular.

The present study has two goals. First, to test the relative accuracy of different ways of treating polymorphic morphological characters with real data, namely, their exclusion versus inclusion, different weighting schemes (a priori and successive), and different phylogenetic methods (parsimony, distance, likelihood). The second goal is to determine how well results from congruence analysis of real data agree with comparable results from computer simulations.

MATERIALS AND METHODS

Phylogenies

The Phrynosomatidae comprise 10 genera and approximately 120 species of North American iguanian lizards. Phrynosomatid lizards have two important advantages for an analysis of this kind: (1) extensive molecular data have been collected for the family by T. Reeder (Reeder, 1995; Wiens and Reeder, 1997), and (2) the morphological data I gathered for these taxa did not exclude characters because of intraspecific variability (and therefore contain many polymorphic characters) and include information on the frequencies of traits within species.

Several criteria were applied to choose well-supported clades for this analysis. Clades were used that were supported by both molecular data (parsimony analysis of mitochondrial ribosomal DNA sequences; Reeder, 1995; Wiens and Reeder, 1997) and traditional taxonomy, and most clades were also supported by previous morphological cladistic analyses (which did not explicitly use polymorphic characters) or chromosomal and/or life-history characters. These clades are not truly "known" in the same sense that clades are known in studies of simulations and laboratory phylogenies (the main disadvantage of the congruence approach; Miyamoto and Fitch, 1995). However, phylogenetic history seems to be the best explanation for their strong support, given the fact that most are corroborated by both molecular and

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

Method	Summary
Any instance	0/1 or 1 = 1
Majority	If frequency of 1 is $\geq$50%, then 0/1 = 1, otherwise 0/1 = 0
Scaled	0 = 0, 0/1 = 1, 1 = 2; ordered 0 $\rightarrow$ 1 $\rightarrow$ 2, change from 0 $\rightarrow$ 2 is two steps
Unordered	Same as scaled, but unordered
Unscaled	Same as scaled (ordered), but change from 0 $\rightarrow$ 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

nonmolecular synapomorphies. Thus, I assume that the ability of methods to recover these clades indicates accuracy in the usual sense of the word.

The evidence for these clades is discussed below. The sand lizards constitute a distinctive clade of four genera (*Callisaurus*, *Cophosaurus*, *Holbrookia*, *Uma*), which is supported by molecular data (bootstrap = 67%; Reeder, 1995) and morphological characters (Etheridge and de Queiroz, 1988; de Queiroz, 1989; Frost and Etheridge, 1989). Monophyly of *Phrynosoma* is corroborated (weakly) by molecular data (Reeder, 1995) and is strongly supported by more than 30 morphological synapomorphies (Montanucci, 1987; Etheridge and de Queiroz, 1988; Frost and Etheridge, 1989). The clade consisting of sand lizards + *Phrynosoma* is strongly supported by molecular data (bootstrap = 91%; Reeder, 1995) and morphological characters (Etheridge and de Queiroz, 1988; Frost and Etheridge, 1989). The genera *Petrosaurus*, *Urosaurus*, and *Uta* are each strongly supported by molecular data (respective bootstrap values of 89%, 97%, and 100%; Reeder, 1995) as well as some morphological characters (Etheridge and de Queiroz, 1988; Frost and Etheridge, 1989).

Although monophyly of *Sceloporus* is not unambiguous (e.g., Reeder, 1995), several well-supported clades within this speciose genus (approximately 80 species; Sites et al., 1992) were used in a separate set of analyses. The *variabilis* group (Smith, 1939), represented in this study by *S. couchii* and *S. variabilis*, is strongly supported by molecular data (bootstrap = 92%; Wiens and Reeder, 1997). The clade of *Sceloporus* above the basal *variabilis* group is supported by mtDNA sequence data (bootstrap = 62%; Wiens and Reeder, 1997), and the location of the ribosomal gene on the long arm of chromosome pair 2 (rather than on a single pair of microchromosomes; Porter et al., 1994). The *jalapae* group (*S. jalapae* and *S. ochoterenae*; Thomas and Dixon, 1976) is strongly supported by molecular data (bootstrap = 94%; Wiens and Reeder, 1997). The clade of *Sceloporus* above the *jalapae* and *variabilis* groups corresponds to the large-scaled, large-bodied radiation of Smith (1939) and Hall (1973), and is supported by molecular data (bootstrap = 74%; Wiens and Reeder, 1997). Within this clade, two groups were used. One of these is the *clarkii* group of Hall (1973; *S. clarkii* and *S. melanorhinus*), which is weakly supported by molecular data but strongly supported by chromosomal characters, including the Em9 mutation and centric fissions of macrochromosome pairs 1, 3, 4, and 5 (Hall, 1973; Sites et al., 1992). The other clade is the *formosus* group (Smith, 1939; Hall, 1973), from which two species were sampled (*S. smaragdinus* and *S. taeniocnemis*). This species pair is supported by molecular data (bootstrap = 83%; Wiens and Reeder, 1997), and these species also share viviparous reproduction (Guillette et al., 1980), X and Y sex chromosomes that are indistinct from each other, and a highly reduced number of microchromosomes (Hall, 1973; Sites et al., 1992; Goyenechea Mayer-Goyenechea and Mendoza Quijano, 1993)—although whether all three characters are synapomorphies depends on the position of the *formosus* group within *Sceloporus*. It should be noted that the evidence that supports the "true" clades used in this study does not come from characters that are known to be polymorphic.

Species were sampled from each of these clades to make three unrooted "known" phylogenies of eight species each (Fig. 1). The two trees for phrynosomatid genera differ in their selection of taxa. Despite some uncertainties in the phylogeny of phrynosomatid genera and with-in *Sceloporus* (Reeder and Wiens, 1996;

Wiens and Reeder, 1997), the relationships depicted in these three trees are uncontroversial. Analyses were restricted to eight species to ensure that all parts of the phylogeny would be unambiguous and that the results would be directly comparable to simulation results in the eight-taxon case (Wiens and Servedio, 1997, 1998). All three trees have the same symmetrical shape. This tree shape was determined largely by necessity rather than choice but is similar to the eight-taxon model tree used extensively in simulation studies of polymorphic characters (except that the simulated tree is slightly more unbalanced/asymmetric; Wiens and Servedio, 1997, 1998). Simulations suggest that, in general, tree shape has a relatively minor impact on the performance of these methods in the eight-taxon case (Wiens and Servedio, 1998). A limited set of analyses was also performed using 40 taxa to test the robustness of some of the results to tree shape and taxon sampling (see below); these analyses compared the ability of different methods to resolve the monophyly of the six generic-level clades (*Phrynosoma* + sand lizards, *Phrynosoma*, sand lizards, *Petrosaurus*, *Urosaurus*, *Uta*).

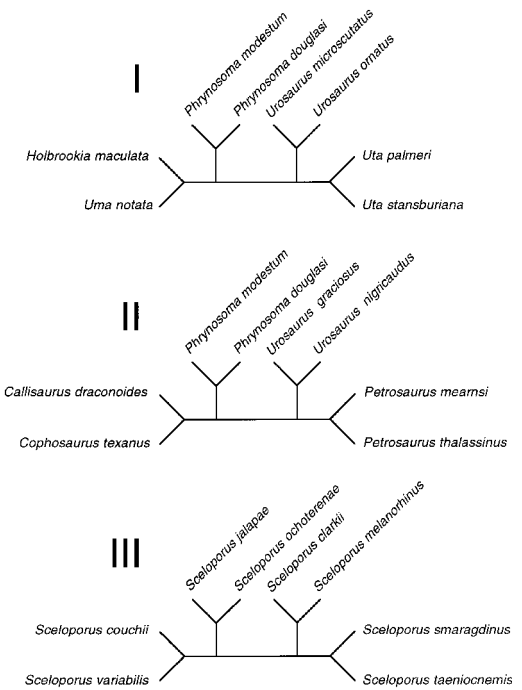

FIGURE 1. The three well-supported phylogenies of phrynosomatid lizards used in this study. Trees are unrooted.

Character Data

The raw data for these analyses consist of the frequencies of qualitative binary morphological character states within species. These characters describe variation in osteology (cranial and postcranial), scalation, and coloration. The characters range from discrete to those that are more-or-less continuous but described in a qualitative manner (Stevens, 1991; Thiele, 1993). For the analyses of phrynosomatid genera (trees I and II; Fig. 1), the morphological data consist of 80 characters from Reeder and Wiens (1996). The data for *Sceloporus* (tree III; Fig. 1) consist of 131 characters from Wiens and Reeder (1997). For tree I, 32.5% of the characters vary within one or more species, for tree II 30.0% are polymorphic, and for tree III (within *Sceloporus*) 92.4% are polymorphic. Descriptions of the characters, specimens examined, and other details are provided in the respective papers; the raw matrices used for resampling are available at http://www.utexas.edu/depts/systbiol.

Several of the characters used by Reeder and Wiens (1996) and Wiens and Reeder (1997) were excluded from the present study. Because the program used to implement distance and likelihood methods (PHYLIP; Felsenstein, 1995) does not allow missing frequency data, characters were excluded if they contained any missing data in the taxa sampled. Most of the characters used by Reeder and Wiens (1996) and Wiens and Reeder (1997) were binary (two conditions), and the few multistate characters were excluded for simplicity (for example, the any-instance and unscaled parsimony coding methods are difficult to apply to multistate characters; Wiens, 1995). Characters based on observations from the literature (i.e., myological characters) were also excluded.

For each of the three raw data sets, new matrices were created by randomly subsampling characters. Sampling was done without replacement, so that no character was represented more than once in any single data matrix. For each of the three trees, new matrices were made with 10, 25, 50, and 75 characters

TABLE 2. Excluding polymorphic characters significantly decreases phylogenetic accuracy, relative to the unweighted frequency parsimony method including all characters. Means and standard errors for each method (based on 100 matrices) are presented. *P*-values are based on paired *t*-tests. ch = number of characters.

Data set	Frequency	Fixed-only	*P*
Tree I, 10 ch	0.476 ± 0.016	0.340 ± 0.014	< 0.0001
Tree II, 10 ch	0.538 ± 0.018	0.362 ± 0.017	< 0.0001
Tree III, 10 ch	0.524 ± 0.023	0.086 ± 0.011	< 0.0001
Tree I, 25 ch	0.624 ± 0.017	0.460 ± 0.016	< 0.0001
Tree II, 25 ch	0.786 ± 0.015	0.616 ± 0.017	< 0.0001
Tree III, 25 ch	0.758 ± 0.018	0.184 ± 0.014	< 0.0001
Tree I, 50 ch	0.752 ± 0.014	0.628 ± 0.013	< 0.0001
Tree II, 50 ch	0.880 ± 0.013	0.772 ± 0.016	< 0.0001
Tree III, 50 ch	0.846 ± 0.012	0.292 ± 0.013	< 0.0001
Tree I, 75 ch	0.780 ± 0.007	0.766 ± 0.008	0.0189
Tree II, 75 ch	0.992 ± 0.004	0.948 ± 0.009	< 0.0001
Tree III, 75 ch	0.888 ± 0.011	0.362 ± 0.008	< 0.0001

each. One hundred pseudoreplicates (replicates hereafter), each with a different selection of characters, were made for each tree and set of conditions (number of characters). The mean accuracies for a limited set of analyses (Table 2) show very small standard errors, suggesting that 100 replicates are adequate.

The effects of small sample size ($n = 1$ individual per species) were also tested. The original data were based on a mean of 3.9 individuals per species for osteological characters and 11.1 individuals per species for external characters (although for certain characters the sample sizes were smaller in some or all species, such as sex-specific color patterns). For each set of 100 matrices, 100 new matrices were made with $n = 1$. For a given character in a given taxon, small sample size was simulated by randomly choosing a number (from 0 to 1.0) and considering a trait present in the individual if the number was less than or equal to the frequency of the trait in that species. This procedure assumes (among other things) that (1) any combination of character states is possible in an individual (e.g., no genetic linkage) and (2) each individual has only one of the two possible traits (e.g., heterozygotes are not detectable as such).

Phylogenetic Methods

Three sets of analyses were performed. The first examined the effects of including, excluding, and weighting polymorphic characters on the accuracy of parsimony analysis. The second compared the performance of various parsimony, distance, and likelihood methods with different numbers of characters and sample sizes. The third tested the robustness of some of the conclusions from the eight-taxon analyses in the 40-taxon case.

In the first set of analyses, the accuracy of excluding polymorphic characters was tested against the accuracy of eight different methods for including polymorphic characters. The eight parsimony coding methods (Table 1) were reviewed previously (Wiens, 1995). For the frequency method, each taxon was given a different character state, and the Manhattan distance (for a given character) between each species was used to weight changes between these states in a step matrix. This method was used by Wiens (1995; suggested by D. Hillis) and is a heuristic approximation of the FREQPARS method (Swofford and Berlocher, 1987; see also Ber-locher and Swofford, 1997). All eight polymorphism coding methods give identical results when there is no polymorphism (in this study, when $n = 1$), and the missing and polymorphic coding methods give identical results for binary characters (polymorphic data cells treated as "unknown" by both methods).

Exclusion criteria.— Several different criteria for excluding polymorphic characters were tested. The most extreme is the fixed-only approach (Campbell and Frost, 1993), in which characters are excluded if they exhibit any polymorphism in any species. However, given that polymorphic characters can contain

significant phylogenetic information as a whole but exhibit increasing homoplasy with increasing intraspecific variability (Wiens, 1995), accuracy may be more likely to be improved by excluding only the most polymorphic characters. The effects of excluding characters with different levels of polymorphism were tested. Levels of variability were calculated for each character by 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 vary from 0 to 100) and 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, as were the effects of excluding characters with high levels of variability relative to the other characters in the data set. Analyses were performed excluding: (1) any characters with a MIV score greater than the mean MIV for the data set, (2) a score greater than the mean MIV times 0.5, and (3) a score greater than the mean MIV times 1.5. The determinations of whether a character was polymorphic or fixed and its MIV score were based only on those taxa and characters sampled for a given replicate. Because these six approaches for excluding polymorphic characters usually involved using at least some (less variable) polymorphic characters, they were tested using each of the eight polymorphism coding methods.

Weighting schemes.— Two weighting schemes also were tested using these eight coding methods. Farris (1966) suggested weighting characters by the reciprocal of their intraspecific variability, a method I have also advocated (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 characters with both traits 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, and 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. Limited simulation results suggest that the choice among the options listed does not greatly impact the results (J. Wiens and M. Servedio, unpubl. data). The frequency-bins method (Wiens, 1995) was used to code polymorphic data as frequencies for successive weighting because of the difficulty of calculating goodness-of-fit statistics with step matrix-coded characters.

Comparison of tree-building methods.— In the second set of analyses, 15 phylogenetic methods were examined. These consisted of the eight coding methods used with parsimony (Table 1), continuous maximum likelihood (Felsenstein, 1981), and six genetic distance methods. The distance methods were UPGMA (Sokal and Michener, 1958), neighbor-joining (Saitou and Nei, 1987), and Fitch–Margoliash ([FM]; Fitch and Margoliash, 1967; or weighted least squares), applied to the genetic distance of Nei (1972) and the modified Cavalli-Sforza and Edwards (1967 [CSE]) chord distance. For distance and likelihood methods, the frequencies of qualitative traits among individuals within species were treated the same as allele frequencies at an allozyme locus. The frequency parsimony method and these distance and likelihood methods are similar in that they all make direct use of frequency information. Distance and likelihood methods are applied only rarely to morphological data (Felsenstein, 1988; Lynch, 1989), and the choice of these particular distance and likelihood methods was based on previous recommendations and simulation studies (Felsenstein, 1988; Wiens and Servedio, 1998 [and references therein]).

40-Taxon case.— To test the sensitivity of the results to tree shape and subsampling of

taxa, I performed a limited set of analyses that included all 40 species of phrynosomatids used in the molecular analysis of Reeder (1995). The ability of the eight parsimony coding methods and the fixed-only method (excluding all polymorphic characters) to recover the six well-supported clades at the intergeneric level was assessed (*Phrynosoma* + sand lizards, *Phrynosoma*, sand lizards, *Petrosaurus*, *Urosaurus*, *Uta*). For each replicate, accuracy was scored as the proportion of these six clades correctly resolved as monophyletic. As is typical for morphological data, many characters had missing or inapplicable data in one or more taxa, and distance and likelihood analyses were therefore not attempted in the 40-taxon case. Given that only parsimony approaches were compared, characters with missing data could be included, and a total of 105 characters were randomly subsampled to make new matrices with 25, 50, 75, and 100 characters each. The large number of taxa necessitated using the heuristic search option, and 20 addition sequence replicates (with TBR branch swapping) per data matrix were used to estimate the most-parsimonious tree(s). Because of the time-intensive nature of these searches, only 20 matrices were analyzed for each set of conditions in the 40-taxon case, and the frequency-bins method was used (instead of step matrices) to code polymorphic characters.

Programs used.— Parsimony analyses were implemented using a test version of PAUP* (provided by David Swofford; 4.0d52), with the branch and bound search option (in the eight-taxon case). Distance and likelihood analyses were implemented using PHYLIP 3.57c (Felsenstein, 1995). The programs for subsampling, coding, and scoring the data were written in C by me.

UPGMA and neighbor-joining are clustering algorithms (Swofford and Olsen, 1990) and do not have optimality criteria (they always find the "best" UPGMA and neighbor-joining tree). For maximum likelihood and the FM method, optimal trees were searched for by using the "global rearrangements" option with 10 different taxon-addition sequences per 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. 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 few characters and little polymorphism) were excluded from maximum likelihood analyses, and the results for this method are, for certain conditions, based on fewer than 100 data sets. This problem also occurs with the FM method.

Measuring accuracy.— For each analysis in the eight-taxon case, accuracy was scored as the similarity between the well-supported 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 by using the consensus fork index of Colless (1980), the proportion of nodes in common between the "known" and estimated trees. Because I consider method success to be the correct resolution of a clade, I did not treat having the correct tree among one of multiple shortest trees in a parsimony analysis as contributing to the accuracy of a method (but see Hillis et al., 1994). The same procedure for scoring method success was used by Wiens and Servedio (1997, 1998). Comparisons with an alternative method for scoring accuracy are presented in the Results. The approach for scoring accuracy is effectively the same in the eight- and 40-taxon cases. The eight-taxon case compares the ability of methods to correctly resolve five clades, whereas in the 40-taxon case there are six clades to be resolved as monophyletic and there are many more species both inside and outside these clades.

RESULTS

Including, Excluding, and Weighting Polymorphic Characters

The results of the analyses testing the effects of excluding and weighting polymorphic characters are shown in Figures 2–4. The practice of excluding all polymorphic characters (the fixed-only approach; black bars in Figs. 2–4) performs relatively poorly, and significantly

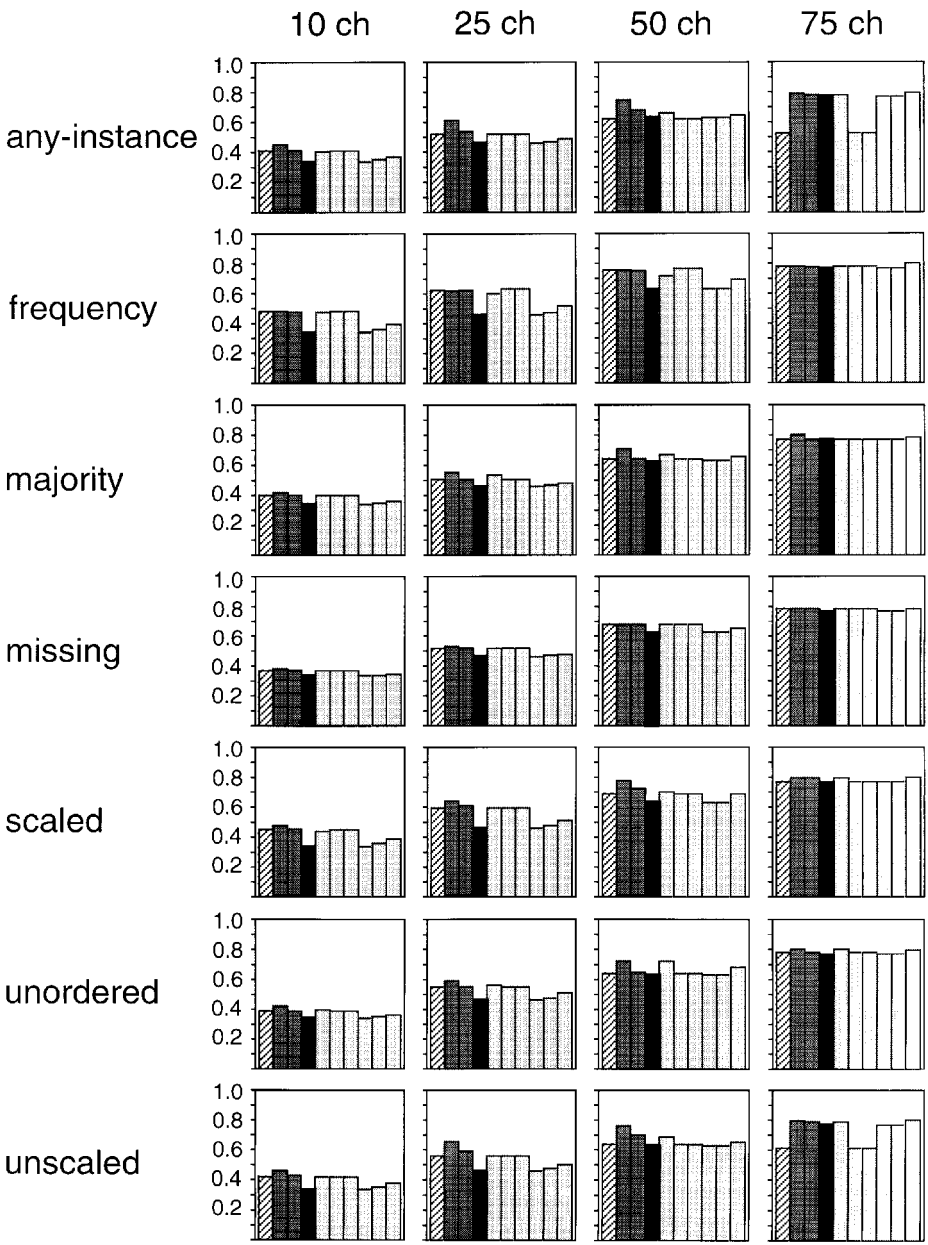

FIGURE 2. Accuracy of parsimony analysis using different coding methods, weighting schemes, and different variability thresholds for excluding polymorphic characters. Results are based on tree I (Fig. 1) with different numbers of characters (ch), and each bar represents the mean results from 100 replicated matrices. Methods (left to right) are: (1) unweighted analysis including all characters (diagonal lines); (2) including all characters and with variable characters downweighted (dark gray); (3) including all characters and with successive weighting (dark gray); (4) fixed-only (black); the last six methods (light gray) employ different variability thresholds for excluding polymorphic characters: (5) excluding characters with MIV > 25, (6) MIV > 50, (7) MIV > 75, (8) MIV > (mean MIV of data set)(0.5), (9) MIV > mean MIV, and (10) MIV > (mean MIV) (1.5). The missing and polymorphic coding methods give identical results for these data.

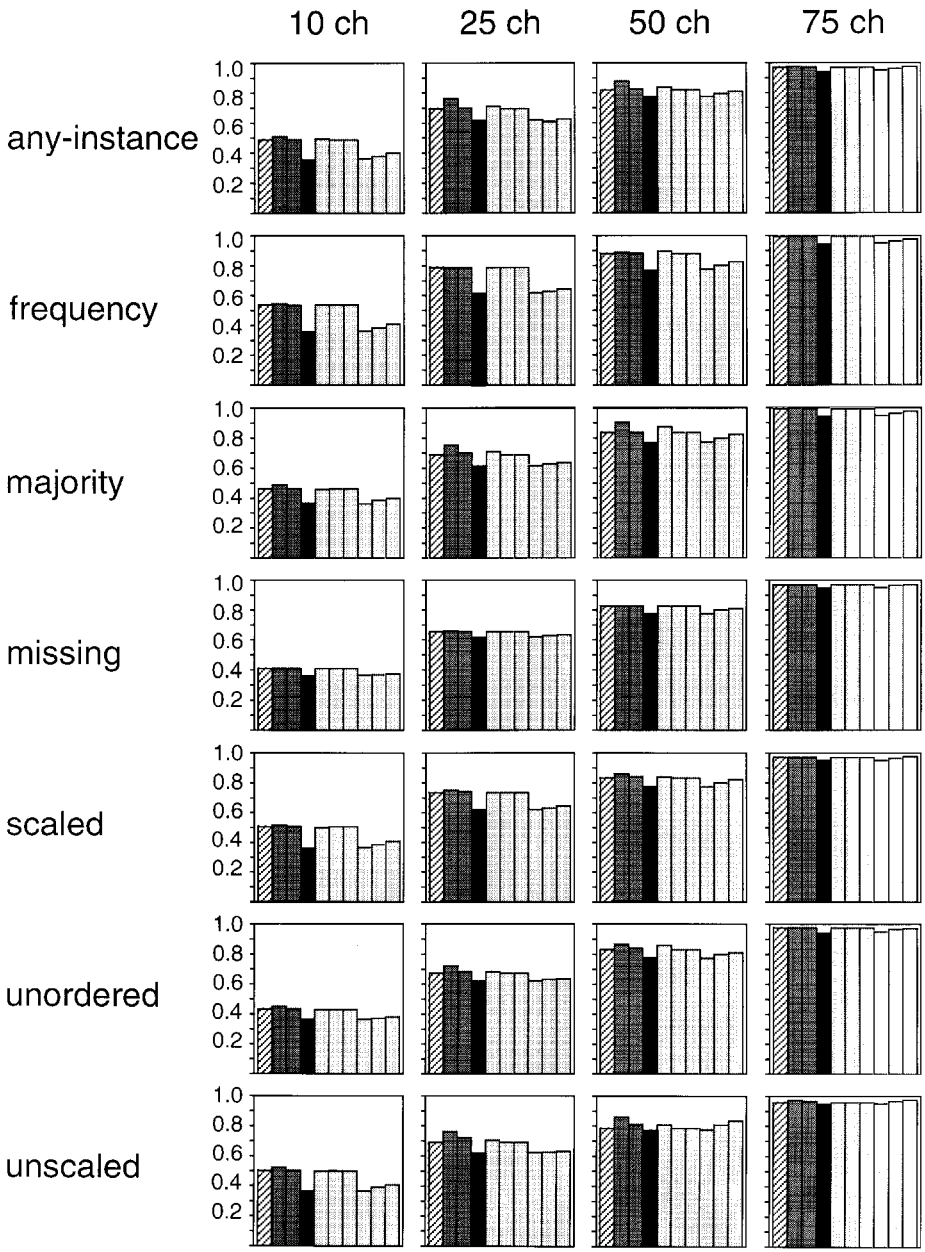

FIGURE 3. Accuracy of parsimony analysis using different coding methods, weighting schemes, and different variability thresholds for excluding polymorphic characters. Results are based on tree II (Fig. 1) with different numbers of characters (ch), and each bar represents the mean results from 100 replicated matrices. Methods (left to right) are: (1) unweighted analysis including all characters (diagonal lines); (2) including all characters and with variable characters downweighted (dark gray); (3) including all characters and with successive weighting (dark gray); (4) fixed-only (black); the last six methods (light gray) employ different variability thresholds for excluding polymorphic characters: (5) excluding characters with MIV > 25, (6) MIV > 50, (7) MIV > 75, (8) MIV > (mean MIV of data set)(0.5), (9) MIV > mean MIV, and (10) MIV > (mean MIV)(1.5). The missing and polymorphic coding methods give identical results for these data.

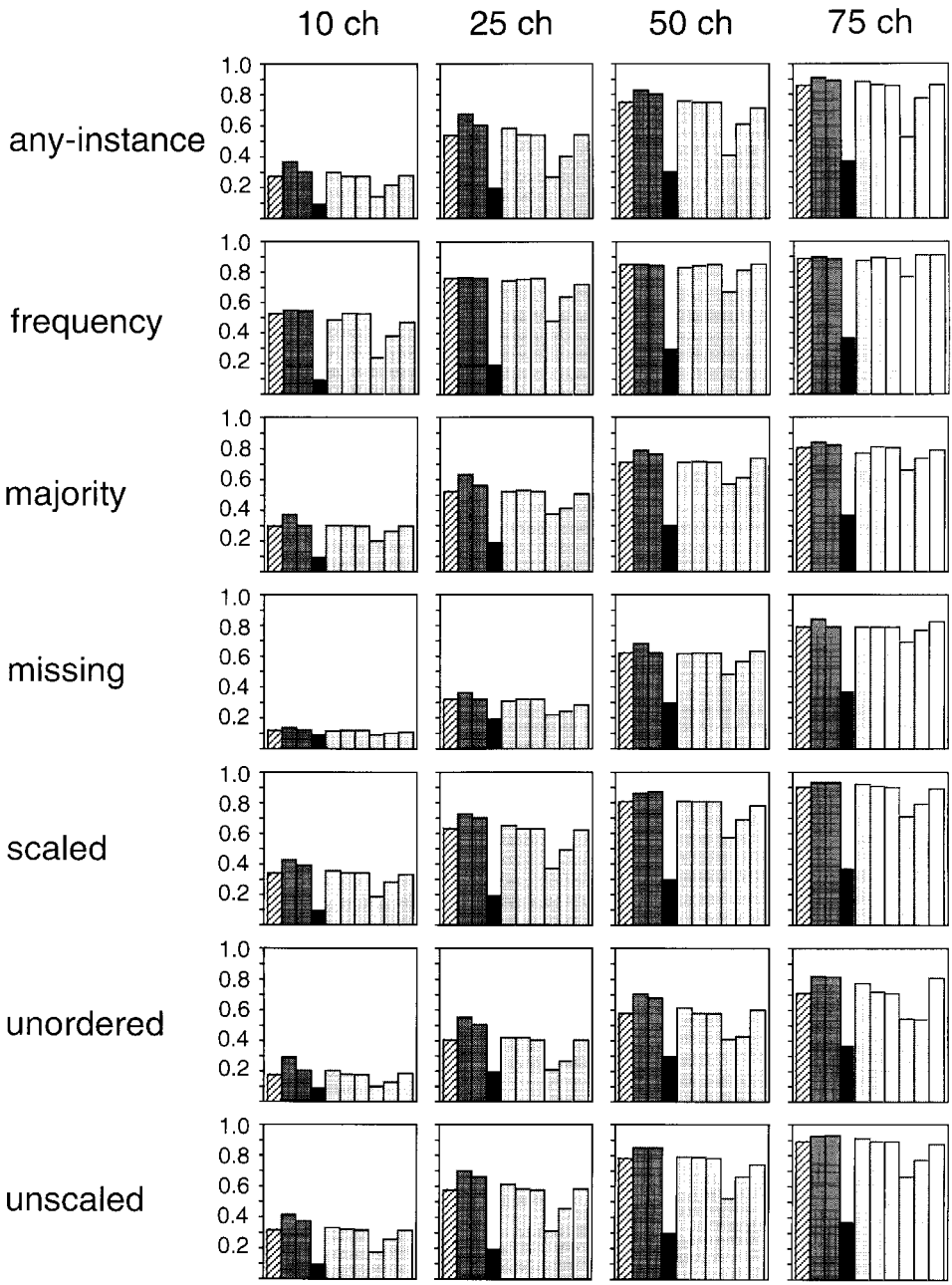

FIGURE 4. Accuracy of parsimony analysis using different coding methods, weighting schemes, and different variability thresholds for excluding polymorphic characters. Results are based on tree III (Fig. 1) with different numbers of characters (ch), and each bar represents the mean results from 100 replicated matrices. Methods (left to right) are: (1) unweighted analysis including all characters (diagonal lines); (2) including all characters and with variable characters downweighted (dark gray); (3) including all characters and with successive weighting (dark gray); (4) fixed-only (black); the last six methods (light gray) employ different variability thresholds for excluding polymorphic characters: (5) excluding characters with MIV > 25, (6) MIV > 50, (7) MIV > 75, (8) MIV > (mean MIV of data set)(0.5), (9) MIV > mean MIV, and (10) MIV > (mean MIV) (1.5). The missing and polymorphic coding methods give identical results for these data.

decreases accuracy under many conditions (Table 2). Even for the data sets with relatively little polymorphism (trees I and II), the accuracy of the fixed-only method was at least 10% lower than the most accurate method for including them (unless the overall number of characters was very large), and for the *Sceloporus* data set (tree III), accuracy was consistently lower by at least 50%. Excluding only the more polymorphic characters (light gray bars; Figs. 2–4) may have no effect on accuracy (relative to including all polymorphic characters), increase accuracy slightly, or decrease accuracy, depending upon the data set analyzed, the number of characters sampled, the variability threshold used for exclusion, and the method used to code the included polymorphic characters. There is no exclusion criterion that increases accuracy under most conditions (relative to an analysis including all the polymorphic characters), or even one that does not decrease accuracy under some conditions.

A priori and successive weighting improved the accuracy of most methods under most conditions examined, and a priori weighting (downweighting characters based on their levels of intraspecific variability) performed better than successive weighting. However, a priori and successive weighting generally had little effect on the accuracy of the frequency method, and sometimes slightly decreased its accuracy. With or without character weighting, the frequency coding method is generally more accurate than the other coding methods. Under certain conditions, a priori weighting allows the majority, scaled, or unordered methods to have slightly greater accuracy than the frequency method (particularly with a larger number of characters). In summary, no single approach to coding, weighting, and excluding polymorphic characters is the most accurate under every set of conditions examined, but the frequency method (either unweighted or with a priori weighting), coupled with including all polymorphic characters, seems to give consistently accurate results.

Accuracy of Parsimony, Distance, and Likelihood Methods

The relative performance of the 15 parsimony, distance, and likelihood methods using different numbers of characters and sample sizes is shown in Figure 5. Under most conditions, UPGMA (regardless of the distance measure used) gives the most accurate results. The next most accurate methods are generally neighbor-joining and FM (which perform similarly), but only when these methods are used with the CSE chord distance. On data sets with the full sample sizes, the next most accurate method is generally the frequency parsimony method, followed by maximum likelihood and neighbor-joining and FM used with Nei's genetic distance. The least accurate methods are generally the non–frequency parsimony methods. Among these methods, the scaled approach tends to perform best, generally followed by the unscaled method. The relative performance of the remaining five parsimony methods (any-instance, majority, missing, polymorphic, unordered) varies considerably among data sets and conditions. In general, the methods that perform best are those that make direct use of frequency information, whether they be parsimony, distance, or likelihood (although there are clearly differences in performance among these methods that are unrelated to the use of frequency information). Part of the higher success of frequency-based methods is that they give well-resolved estimates (few or no polytomies), but the frequency-based methods still appear to be generally superior when all methods are standardized to the same level of resolution (Table 3).

Small sample size ($n = 1$) generally has a strong negative impact on the performance of methods, although the extent of this impact varies among data sets, conditions, and methods. For example, in the *Sceloporus* data set with a small number of characters (10 or 25), the success of the most accurate parsimony method is effectively cut in half by small sample size; this is clearly related to the abundant polymorphism in this data set. Parsimony appears to be more sensitive to small sample size than do the distance and likelihood methods. The tendency of distance and likelihood methods to give well-resolved trees (even with coarse frequency information) appears to contribute to their success relative to parsimony but accounts for only part of their greater accuracy (Table 4). Presumably, the superior performance of these methods with small sample

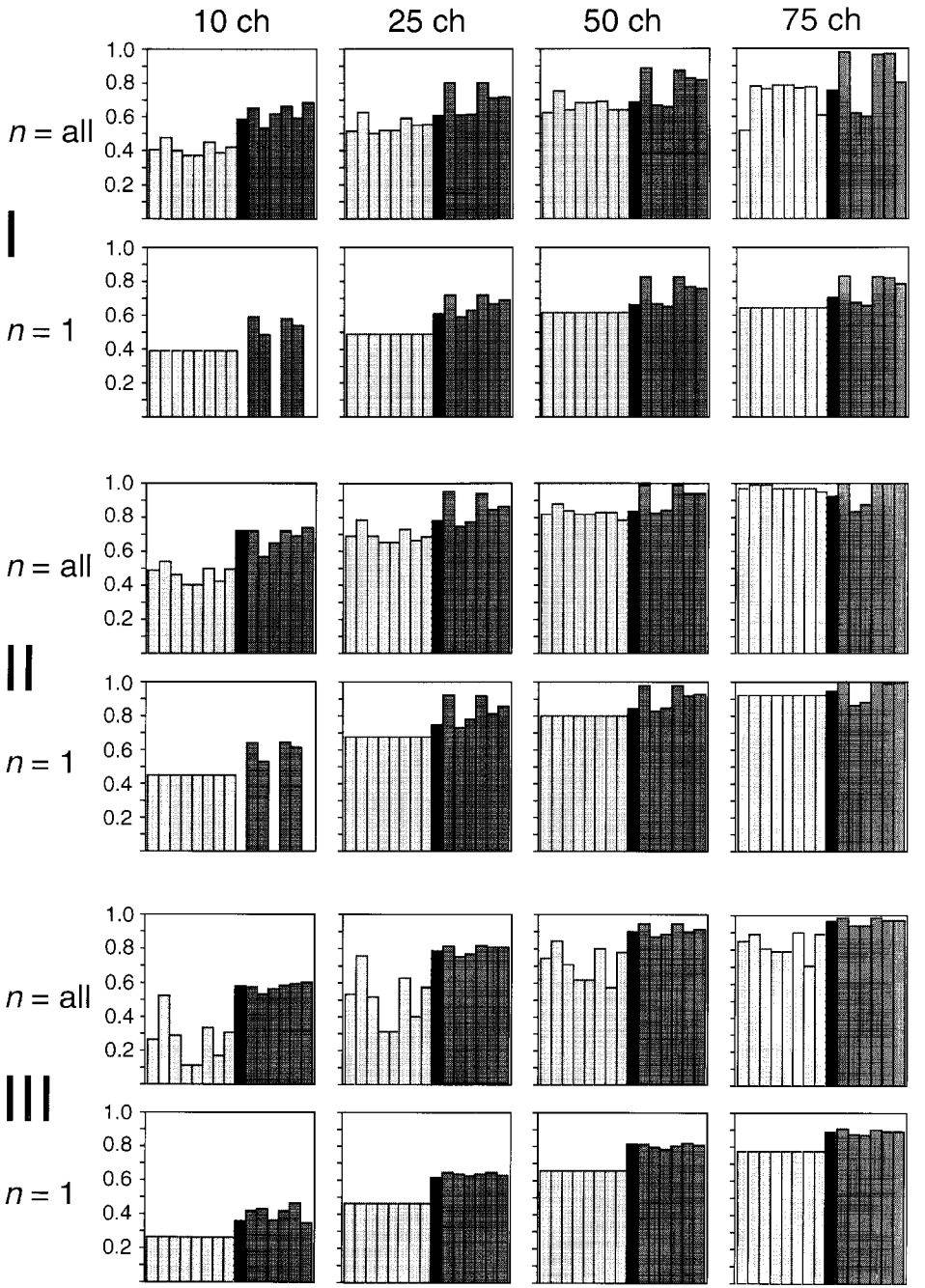

FIGURE 5. Accuracy of parsimony (light gray), likelihood (black), and distance (dark gray) methods for trees I–III (Fig. 1) with different numbers of characters (ch) and sample sizes (the original data or one individual per species). Methods are (left) parsimony (light gray): (1) any-instance, (2) frequency, (3) majority, (4) missing, (5) polymorphic, (6) scaled, (7) unordered, 8) unscaled; (center) likelihood (black): continuous maximum likelihood; (right) distance (dark gray): (1) UPGMA with Nei's distance, (2) neighbor-joining with Nei's distance, (3) FM with Nei's distance, (4) UPGMA with CSE distance, (5) neighbor-joining with CSE distance, and (6) FM with CSE distance. Because many taxa are identical with $n = 1$ and 10 characters, a few cases have no values for likelihood and FM.

TABLE 3. A comparison of two measures of accuracy. When there are multiple shortest trees from a parsimony analysis, the accuracy of the method is based on either the number of correctly resolved clades (the measure used throughout the paper) or the number of clades shared between the true and a randomly selected, fully resolved shortest tree (so that all methods have the same level of resolution). Results are for tree III (Fig. 1; *Sceloporus*) with 25 characters, conditions where the relative accuracy of methods is well-differentiated. Selected nonparsimony methods are included for comparison. Each value is the mean from 100 data matrices.

	Accuracy	
Method	Correctly resolved clades	Single shortest tree
Fixed-only	0.184	0.230
Any-instance	0.534	0.644
Frequency	0.758	0.762
Majority	0.520	0.640
Missing/polymorphic	0.316	0.496
Scaled	0.628	0.720
Unordered	0.404	0.514
Unscaled	0.574	0.674
Maximum likelihood	0.786	0.786
UPGMA–Nei	0.814	0.814
FM–CSE	0.808	0.808

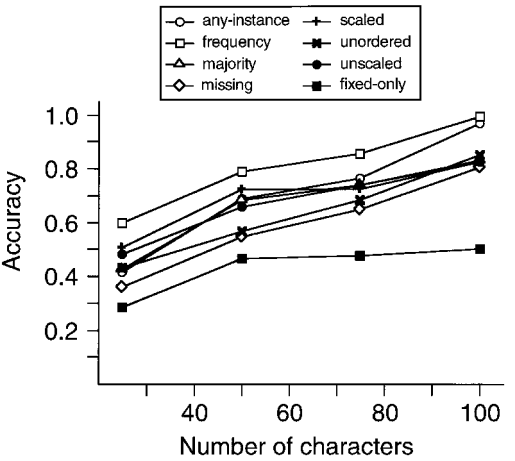

FIGURE 6. Accuracy of seven parsimony coding methods and the fixed-only approach (excluding all polymorphic characters) in the 40-taxon case. Accuracy is the proportion of six well-supported clades of phrynosomatid lizards that are resolved correctly. The missing and polymorphic coding methods give identical results for these data.

sizes is a combination of greater resolution and insensitivity to random noise.

Forty-Taxon Case

Results from the 40-taxon case (Fig. 6) are similar to those obtained in the eight-taxon case, although methods are generally less accurate given the same number of characters in the 40-taxon case. Among the parsimony methods tested, the frequency method always gives the most accurate results and the fixed-only approach (excluding all polymorphic characters) always gives the least accurate results. Under all the conditions examined, the accuracy of the fixed-only approach is about half the accuracy of the frequency method. The per-

TABLE 4. Accuracy of parsimony and distance methods with a sample size of one individual per species, showing that the superior accuracy of distance methods at small sample sizes (relative to parsimony methods) is based partly on their tendency to estimate fully resolved trees. The accuracy of parsimony was based both on correctly resolved clades and (to standardize resolution) on a randomly selected, fully resolved tree from among the shortest trees from a given analysis. Each accuracy value is an average based on 100 replicated data sets. ch = number of characters.

	Parsimony		Distance (CSE)	
Data set	Correctly resolved clades	Randomly selected tree	UPGMA	Neighbor-joining
Tree I, 10 ch	0.388	0.506	0.580	0.540
Tree I, 25 ch	0.492	0.668	0.720	0.664
Tree I, 50 ch	0.614	0.712	0.824	0.764
Tree I, 75 ch	0.646	0.732	0.826	0.818
Tree II, 10 ch	0.448	0.518	0.646	0.614
Tree II, 25 ch	0.674	0.764	0.914	0.810
Tree II, 50 ch	0.800	0.870	0.974	0.914
Tree II, 75 ch	0.918	0.954	1.000	0.990
Tree III, 10 ch	0.266	0.414	0.418	0.464
Tree III, 25 ch	0.464	0.612	0.634	0.646
Tree III, 50 ch	0.660	0.756	0.806	0.822
Tree III, 75 ch	0.774	0.824	0.902	0.890

formance of the other parsimony methods is generally similar to that of the eight-taxon case, although the any-instance method does surprisingly well with 75 and 100 characters.

DISCUSSION

Comparison to Simulation Results

How do the conclusions based on congruence compare to results from simulations of polymorphic data? For comparable conditions (number of taxa, characters, states per character, sample sizes), the major conclusions of simulations (Wiens and Servedio, 1997, 1998) and congruence analysis are nearly identical: (1) excluding all polymorphic characters (the fixed-only method) decreases phylogenetic accuracy; (2) excluding polymorphic characters (using the same six variability thresholds) and the two weighting schemes (a priori and successive) do not consistently increase accuracy when compared to the unweighted frequency coding method using all polymorphic characters; (3) the frequency coding method appears to be the most generally accurate parsimony method, followed by the scaled method; (4) small sample sizes decrease accuracy considerably under many conditions (particularly when levels of polymorphism are high); and (5) distance and likelihood methods are less sensitive to small sample size than parsimony. Although the simulations of Wiens and Servedio (1997, 1998) made a number of unrealistic assumptions (i.e., phenotype equals genotype and no selection, mutation, or geographic variation), the agreement between the congruence and simulation results is striking. An obvious candidate explanation for the agreement between the congruence and simulation results is that the simple genetic drift model (Fisher, 1930; Wright, 1931; Kimura, 1955) used in the simulations may provide a reasonable approximation for the evolution of at least some morphological characters (see also Felsenstein, 1988; Lynch, 1989), at least for the purposes of comparing phylogenetic methods.

The strong performance of UPGMA on these real data sets is surprising, but not unprecedented. In many of the simulated conditions examined by Wiens and Servedio (1998), UPGMA gave more accurate results than any other method tested, especially when any of the following was true: (1) branch lengths were long (length = 1.4 and 2.0), (2) sample sizes were small (n = 1 or 2), and/or (3) the model tree was fully symmetric/balanced. The symmetry of the model trees used in this study probably contributes to the success of UPGMA, but simulations suggest that UPGMA can outperform all other methods on fully asymmetric trees with relatively long branches (Wiens and Servedio, 1998). These simulations (Wiens and Servedio, 1998) also suggest that UPGMA may be less sensitive to unequal branch lengths than previously thought, at least for the conditions examined. UPGMA also performed surprisingly well in some simulation studies of DNA sequence data. For example, Huelsenbeck and Kirkpatrick (1996, Fig. 4) found that UPGMA outperformed parsimony, neighbor joining, and likelihood at high rates of change in the eight-taxon case (tree shape varied randomly), and the graphs of Huelsenbeck and Hillis (1993, Fig. 6E) show that UPGMA outperforms parsimony and neighbor-joining over many combinations of branch lengths in the four-taxon case when using a limited number of characters. The relative merits of phenetic methods (such as UPGMA) and parsimony were debated extensively in the early 1980s using congruence analyses (Colless, 1980; Schuh and Polhemus, 1980; Schuh and Farris, 1981; Rohlf et al., 1983; Sokal and Shao, 1985). The results of the present study and others suggest that this debate may still be surprisingly unresolved, at least in terms of which approach gives the most accurate estimate of a known phylogeny.

An important caveat that should be made about the results of this study is that trees estimated by UPGMA were treated as unrooted. This is not the traditional usage of UPGMA (although this is how it was treated in the simulation studies cited above), and probably contributes to its surprising success. Clearly, the results should not be taken as an endorsement for using UPGMA to root trees. It is also possible that UPGMA and the other distance methods might perform worse if other distance measures were used, such as overall similarity.

Recommendations for Empirical Studies

These conclusions lead to several recommendations regarding the phylogenetic analy-

sis of polymorphic morphological characters. The results suggest that polymorphic characters should not simply be excluded, as seems to be common practice in morphological analyses. This conclusion is even more compelling when one considers that the morphological data used in this study show a very strong positive relationship between levels of homoplasy and intraspecific variability (Wiens, 1995), a potential "worst-case scenario" for including polymorphic characters. The results also suggest that information on the frequencies of traits within species should be collected and utilized. The results of this study, simulations (Wiens and Servedio, 1997, 1998), and statistical analyses of real data sets (Wiens, 1995) all suggest that methods that ignore frequencies tend to perform poorly relative to methods that incorporate these data. This study also shows that sampling a reasonable number of individuals per species may greatly increase accuracy (relative to sampling a single individual), especially when using parsimony.

Distance methods frequently outperformed parsimony with real data in this study, but certain practical limitations of the distance methods should be mentioned. For example, it is difficult to include characters with missing data when using distance methods designed for polymorphic characters (at least in current versions of PHYLIP). It may also be difficult to combine characters evolving under different models of evolution (i.e., trait frequencies, DNA sequences) without resorting to distance measures that ignore the differences in these models (e.g., overall similarity). Several other disadvantages of distance methods unrelated to phylogenetic accuracy have also been noted, such as the absence of a clear relationship between character state changes and clades (e.g., Wiley, 1981).

An important issue in the use of distance methods that was not addressed in this study concerns the effects of branch lengths. Simulation studies of both fixed and polymorphic characters suggest that UPGMA is more sensitive than most parsimony methods to certain misleading combinations of branch lengths (i.e., the Felsenstein Zone; Huelsenbeck and Hillis, 1993), whereas neighbor-joining and FM (and likelihood) are less sensitive to this problem than parsimony is (e.g., Huelsenbeck, 1995; Wiens and Servedio, 1998). The results of the present study, which show that distance and likelihood methods perform relatively well on the morphological data sets examined, suggest promise for the use of these methods with morphological data on more difficult phylogenetic problems where parsimony is likely to fail (e.g., Felsenstein, 1978).

Finally, one should not construe from this paper that I am arguing that distance and likelihood methods should supplant the use of parsimony for analyzing morphological data. Rather, I suggest that the range of methods that are effective with morphological characters may be greater than parsimony alone, and that non-parsimony methods may be advantageous in at least some situations.

Pros and Cons of the Congruence Approach

This study exemplifies many of the strengths and weaknesses of the congruence approach. The most obvious strength is that it tests the performance of methods with real data, while avoiding many of the simplifications and unrealistic assumptions of simulations (Miyamoto and Fitch, 1995). Morphological characters may be particularly difficult to model realistically because their genetic basis is usually unknown, they may have complex ontogenies, and they may pass through a poorly defined process of selection and delimitation by the investigator (Stevens, 1991). Polymorphic characters in general may also be difficult to simulate, given the variety of processes that may act in concert on the frequencies of traits within species (selection, mutation, drift, migration).

The congruence approach also has serious limitations. The most obvious of these is that the phylogenies are never truly known (Miyamoto and Fitch, 1995). Another problem is that well-supported clades are unlikely to represent a random or even "natural" sample of the clades within a group, and some common types of phylogenetic problems (such as short internodes caused by rapid speciation) are less likely to be represented. In this study, extensive subsampling of taxa was required to make fully "known" phylogenies, and the data matrices used are (to a certain extent) a product of this artificial sampling. On the positive side, a limited set of analyses suggests that at least some

of the major results of this study are not particularly sensitive to taxon sampling (Fig. 6). Perhaps the most serious limitation of congruence studies is that, in contrast to simulations, the parameters that affect method performance are difficult to vary and understand. Although some parameters can be varied systematically in congruence studies (number of characters, sample size), others are largely unknown or out of the investigators' control (branch lengths, model of evolution, tree shape). Thus, both congruence and computer simulations have distinct advantages and disadvantages for testing methods, and the best inferences may be gained through a combination of these approaches (e.g., Allard and Miyamoto, 1992).

A major weakness of this particular congruence analysis is that all the results come from a single family of lizards. Although the major conclusions are also supported by simulations, the generality of these results should be verified by conducting similar congruence analyses with polymorphic morphological characters in other groups of organisms. The limiting factor for such analyses may not be the absence of well-supported molecular phylogenies, but rather the absence of morphological data sets that include polymorphic characters and information on the frequencies of traits within species. The importance of collecting these data in future morphological studies is clear.

The Need for Investigating Method Success with Morphological Data

Most recent studies of phylogenetic method performance have concentrated on the analysis of molecular data, particularly DNA sequences (note the dearth of morphological studies in the review by Hillis, 1995). It is important that the accuracy of methods for analyzing morphological data be tested as well (e.g., Sokal, 1983; Lamboy, 1994). Despite the increasing use of DNA sequence data, morphology remains the most widely used type of phylogenetic evidence (Sanderson et al., 1993). Insights about method performance gained from real or simulated molecular data may also be relevant to morphologists (Wiens and Hillis, 1996), but this is rarely addressed. There are also many questions that are largely specific to morphological data, such as the treatment of continuous variation, shape differences, fossils, and ontogeny. Finally, many of the common practices of morphological phylogenetics seem to owe their widespread use to historical inertia rather than quantitative investigation, and their choice is rarely even discussed. The results of this study suggest that at least three standard practices in morphological studies— excluding polymorphic characters, ignoring the frequencies of traits within species, and using only parsimony— may lead to relatively poor phylogenetic estimates, at least in some cases. Congruence studies are a useful tool for assessing the performance of methods with morphological data, and offer a way for the growth of molecular systematics to contribute to the development and rigor of morphological phylogenetics.

ACKNOWLEDGMENTS

I am indebted to Tod "Gigabase" Reeder for collecting the extensive DNA sequence data sets for phrynosomatid lizards, which made this analysis possible. Thanks are due also to the many museum curators who allowed me to examine phrynosomatid lizard specimens in their care (curators and institutions are thanked individually in the papers by Reeder and Wiens, 1996, and Wiens and Reeder, 1997). I am grateful to David Swofford for graciously allowing me to publish results based on a test version of his PAUP* program. I thank David Cannatella, Kevin de Queiroz, Junhyong Kim, Brad Livezey, Maria Servedio, and an anonymous reviewer for many useful comments on the manuscript.

REFERENCES

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

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.

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. Natl. Hist. 216:1–121.

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

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

CUNNINGHAM, C. W. 1997. Is congruence between data partitions a reliable predictor of phylogenetic accu-

racy? Empirically testing an iterative procedure for choosing among phylogenetic methods. Syst. Biol. 46:464–478.

DE QUEIROZ, K. 1989. Morphological and biochemical evolution in the sand lizards. Ph.D. Dissertation, Univ. of California, Berkeley.

DIXON, M. T., AND D. M. HILLIS. 1993. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. 10:256–267.

ETHERIDGE, R., AND K. DE QUEIROZ. 1988. A phylogeny of Iguanidae. Pages 283–368 *in* Phylogenetic relationships of the lizard families: Essays commemorating Charles L. Camp (R. Estes and G. K. Pregill, eds.). Stanford Univ. Press, Stanford, California.

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. 1981. Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. Evolution 35:1229–1242.

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

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

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

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

FROST, D. R., AND R. ETHERIDGE. 1989. A phylogenetic analysis and taxonomy of iguanian lizards (Reptilia: Squamata). Misc. Publ. Mus. Natl. Hist. Univ. Kans. 81:1–65.

GOYENECHEA MAYER-GOYENECHEA, I., AND F. MENDOZA QUIJANO. 1993. Descripcion del cariotipo de *Sceloporus smaragdinus* (Sauria: Phrynosomatidae). Bol. Soc. Herpetol. Mex. 5:62–64.

GUILLETTE, L. J., JR., R. E. JONES, K. T. FITZGERALD, AND H. M. SMITH. 1980. Evolution of viviparity in the lizard genus *Sceloporus*. Herpetologica 36:201–215.

HALL, W. P. 1973. Comparative population cytogenetics, speciation, and evolution of the crevice-using species of *Sceloporus*. Ph.D. Dissertation, Harvard Univ., Cambridge, Massachusetts.

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

HUELSENBECK, J. P., AND M. KIRKPATRICK. 1996. Do phylogenetic methods produce trees with biased shapes? Evolution 50:1418–1424.

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

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

LAMBOY, W. F. 1994. The accuracy of the maximum parsimony method for phylogeny reconstruction with morphological characters. Syst. Bot. 19:489–505.

LYNCH, M. 1989. Phylogenetic hypotheses under the assumption of neutral quantitative-genetic variation. Evolution 43:1–17.

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

MIYAMOTO, M. M., M. W. ALLARD, R. M. ADKINS, L. L. JANECEK, AND R. L. HONEYCUTT. 1994. A congruence test of reliability using linked mitochondrial DNA sequences. Syst. Biol. 43:236–249.

MIYAMOTO, M. M., AND W. M. FITCH. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44:64–76.

MONTANUCCI, R. R. 1987. A phylogenetic study of the horned lizards, genus *Phrynosoma*, based on skeletal and external morphology. Contrib. Sci. Natl. Hist. Mus. Los Angel. Cty. 113:1–26.

NEI, M. 1972. Genetic distance between populations. Am. Nat. 106:238–292.

PORTER, C. A., M. W. HADUIK, AND K. DE QUEIROZ. 1994. Evolution and phylogenetic significance of ribosomal gene location in chromosomes of squamate reptiles. Copeia 1994:302–313.

REEDER, T. W. 1995. Phylogenetic relationships among phrynosomatid lizards as inferred from mitochondrial ribosomal DNA sequences: Substitutional bias and information content of transitions relative to transversions. Mol. Phylogenet. Evol. 4:203–222.

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.

ROHLF, F. J., D. H. COLLESS, AND G. HART. 1983. Taxonomic congruence re-examined. Syst. Zool. 32:144–158.

SAITOU, N., AND M. NEI. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

SANDERSON, M. J., B. G. BALDWIN, G. BHARATHAN, C. S. CAMPBELL, C. VON DOHLEN, D. FERGUSON, J. M. PORTER, M. F. WOJCIECHOWSKI, AND M. J. DONOGHUE. 1993. The growth of phylogenetic information and the need for a phylogenetic data base. Syst. Biol. 42:562–568.

SCHUH, R. T., AND J. T. POLHEMUS. 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera). Syst. Zool. 29:1–26.

SCHUH, R. T., AND J. S. FARRIS. 1981. Methods for investigating taxonomic congruence and their application to the Leptopodomorpha. Syst. Zool. 30:331–351.

SITES, J. W., JR., J. W. ARCHIE, C. J. COLE, AND O. F. VILLELA. 1992. A review of phylogenetic hypotheses for the lizard genus *Sceloporus* (Phrynosomatidae): Im-

plications for ecological and evolutionary studies. Bull. Am. Mus. Nat. Hist. 213:1–110.

SLOWINSKI, J. B. 1993. "Unordered" versus "ordered" characters. Syst. Biol. 42:155–165.

SMITH, H. M. 1939. The Mexican and Central American lizards of the genus *Sceloporus*. Field Mus. Nat. Hist. Zool. Ser. 26:1–397.

SOKAL, R. R. 1983. A phylogenetic analysis of the Caminacules. II. Estimating the true cladogram. Syst. Zool. 32:185–201.

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

SOKAL, R. R., AND K. SHAO. 1985. Character stability in 39 data sets. Syst. Zool. 34:83–89.

STEVENS, P. F. 1991. Character states, morphological variation, and phylogenetic analysis: A review. Syst. Bot. 16:553–583.

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.

THOMAS, R. A., AND J. R. DIXON. 1976. A re-evaluation of the *Sceloporus scalaris* group (Sauria: Iguanidae). Southwest. Nat. 20:523–536.

TITUS, T. A., AND D. R. FROST. 1996. Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). Mol. Phylogenet. Evol. 6:49–62.

WHEELER, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44:321–331.

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

WIENS, J. J., AND D. M. HILLIS. 1996. Accuracy of parsimony analysis using morphological data: A reappraisal. Syst. Bot. 21:237–243.

WIENS, J. J., AND T. W. REEDER. 1997. Phylogeny of the spiny lizards (*Sceloporus*) based on molecular and morphological evidence. Herpetol. Monogr. 11:1–101.

WIENS, J. J., AND M. R. SERVEDIO. 1997. Accuracy of phylogenetic analysis including and excluding polymorphic characters. Syst. Biol. 46:332–345.

WIENS, J. J., AND M. R. SERVEDIO. 1998. Phylogenetic analysis and intraspecific variation: Performance of parsimony, distance, and likelihood methods. Syst. Biol. 47:228–253.

WILEY, E. O. 1981. Phylogenetics. The theory and practice of phylogenetic systematics. John Wiley and Sons, New York.

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

Received 15 January 1997; accepted 10 February 1998
Associate Editor: K. de Queiroz