Syst. Biol. 47(1):18–31, 1998

Sensitivity of Phylogeny Estimation to Taxonomic Sampling

STEVEN POE[1]

Department of Zoology and Texas Memorial Museum, University of Texas, Austin, Texas 78712-1064, USA; E-mail: stevepoe@mail.utexas.edu

Abstract.—Recent studies have shown that addition or deletion of taxa from a data matrix can change the estimate of phylogeny. I used 29 data sets from the literature to examine the effect of taxon sampling on phylogeny estimation within data sets. I then used multiple regression to assess the effect of number of taxa, number of characters, homoplasy, strength of support, and tree symmetry on the sensitivity of data sets to taxonomic sampling. Sensitivity to sampling was measured by mapping characters from a matrix of culled taxa onto optimal trees for that reduced matrix and onto the pruned optimal tree for the entire matrix, then comparing the length of the reduced tree to the length of the pruned complete tree. Within-data-set patterns can be described by a second-order equation relating fraction of taxa sampled to sensitivity to sampling. Multiple regression analyses found number of taxa to be a significant predictor of sensitivity to sampling; retention index, number of informative characters, total support index, and tree symmetry were nonsignificant predictors. I derived a predictive regression equation relating fraction of taxa sampled and number of taxa potentially sampled to sensitivity to taxonomic sampling and calculated values for this equation within the bounds of the variables examined. The length difference between the complete tree and a subsampled tree was generally small (average difference of 0–2.9 steps), indicating that subsampling taxa is probably not an important problem for most phylogenetic analyses using up to 20 taxa. [Multiple regression; modeling; phylogeny estimation; taxonomic sampling.]

It is now well established that the addition or deletion of taxa from a matrix can change the estimate of phylogenetic relationship for a subset of other taxa in that matrix (e.g., Doyle and Donoghue, 1987; Gauthier et al., 1988; Siddal, 1996). Various factors, such as extinction, unavailability of specimens, and the interests of the investigator, affect taxonomic sampling. These factors are not uncommon, and in fact very few studies sample all the species of a considered clade. Given that incomplete taxonomic sampling is the norm, that sampling can affect results, and that some authors have suggested the importance of greater sampling to improve phylogenetic accuracy (De Pinna, 1992; Wheeler, 1992; Lecointre et al., 1993; Hillis, 1996; but see Kim, 1996), it is surprising that relatively few studies have examined this issue systematically. The relationship of taxonomic sampling to accuracy remains unclear (contrast Kim, 1996, with Wheeler, 1992, and Hillis, 1996), especially with larger numbers of taxa (Hendy and Penny, 1989), and the factors that influence degree of sensitivity to taxonomic sampling have not been determined.

The effects of taxonomic sampling on phylogenetic inference most often have been assessed in the context of fossil taxa (see Donoghue et al., 1989; Huelsenbeck, 1991; Wheeler, 1992; Wiens and Reeder, 1995). Recent investigators have recognized that the fossil taxa problem is a subset of the general problem of taxonomic sampling in phylogenetic inference (de Pinna, 1992; Wheeler, 1992; Lecointre et al., 1993). The question of optimal taxonomic sampling in turn can be thought of as part of the more general issue of what combinations of number of taxa, number of characters, completeness, homoplasy, branch lengths, etc., are optimal for accurate estimation of phylogeny, and how relaxation of the optimality of any of these factors affects the estimate (as addressed in simulation studies; see Huelsenbeck, 1995).

Simulations, known phylogenies, statistical methods, and congruence studies, are

[1] Address until November 1, 1998: Division of Amphibians and Reptiles, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560.

complementary means by which to examine phylogenetic accuracy (Hillis, 1995). The effects of taxon sampling have been examined with simulations (Wheeler, 1992), a known phylogeny (Wiens and Reeder, 1995; Poe, in press), and data sets from the literature (Lecointre et al., 1993; Wiens and Reeder, 1995). Still, the great range of parameters to be altered (number of taxa, homoplasy, tree shape, branch lengths, etc.) and the diversity of questions to be asked (accuracy, consistency, sensitivity, etc.) leave the study of taxon sampling in phylogenetic inference in its infancy.

In this paper, I use data sets from the literature to address issues of sensitivity to taxonomic sampling. I ask: What is the magnitude of the effect of adding or deleting taxa and which factors influence this effect?

Sensitivity to Taxonomic Sampling

The sensitivity of a data set to sampling is the degree to which adding or removing taxa changes the estimate of phylogeny. Degree of sensitivity may vary within a given data set. For example, removing one or two taxa may not have a major effect on phylogeny, whereas removing 80% of taxa in a clade might drastically change the phylogenetic estimate for the remaining taxa. Sensitivity can also vary between data sets. For example, a more strongly supported tree may be more resistant to the effects of taxonomic sampling than a weakly supported tree.

The first aim of this paper is to assess the patterns of sensitivity to taxonomic sampling within data sets. I address this issue by performing taxon removal experiments within data sets and comparing the resulting tree to a pruned tree obtained with the complete matrix. Second, I examine the factors that influence the overall sensitivity of a data set to taxonomic sampling. In particular: What effect do homoplasy, number of informative characters, number of taxa, tree symmetry, and strength of support for a tree have on sensitivity to taxonomic sampling? For example, one might expect that more homoplastic or weakly supported trees would be more sensitive to the effects of taxonomic sampling.

Materials and Methods

Twenty-nine data sets (Table 1) from a variety of sources and covering several types of organisms were found that satisfy the following criteria: (1) At least five taxa were analyzed, so that taxon removal is not trivial. (2) A fully resolved, single most parsimonious tree was produced. (3) Only species level or higher phylogenies were examined (e.g., no intraspecific mtDNA phylogenies). (4) All or all except one of the known extant members of the studied (ingroup) clade are included in the analysis. These requirements were enacted in an attempt to standardize comparisons between data sets.

Operational Definitions

In order to evaluate sensitivity to sampling, one needs an operational definition of sensitivity and a ''baseline'' tree with which to compare trees from reduced sampling. The most desirable baseline tree would be the true tree. However, because this is not known, I use the most-parsimonious tree from the complete matrix as the tree for comparison (suitably pruned for adequate comparison). Although not as informative as the true tree, this complete tree is a convenient point of comparison for trees resulting from reduced sampling and is probably the most desirable tree for many workers (e.g., Lecointre et al., 1993).

It must be noted that due to the existence of undiscovered species, ancestors, and countless incipient and failed lineages, the true ''completeness'' of a clade is probably unknowable and is perhaps even a meaningless concept. However, it is doubtful that the real-life number of species (or lineages or populations) in a clade has any effect on the conclusions of this paper, which are based on knowable factors such as the number of possible trees (discussed later).

Operationally, I consider sensitivity to sampling to be the length difference between (a) the most parsimonious trees from a subsampled matrix (one in which

TABLE 1. Statistics and results for the data sets analyzed in this study. RI = retention index (Farris, 1989); Total support index = Bremer's (1994) measure of total clade support; Tree symmetry = Colless's (1982) *I*. $\bar{s}$ = number of steps difference between a tree from reduced sampling and the pruned complete tree, averaged for all trials and all numbers of taxa removed; b = regression parameter for the parabola that describes the relationship of fraction of taxa sampled t to s; R^2 = Coefficient of determination for the relationship of t to s, describing the degree of fit for the second-order regression model. s and b are measures of composite sensitivity of a data set to taxonomic sampling.

Reference	Organism	Number of taxa	Number of inf. characters	RI	Total support index	Tree symmetry	$\bar{s}$	b	R^2
Winterbottom (1990)	Bony fish	5	37	0.90	0.57	0.33	0.00	0.00	NA
Wiens and Titus (1991)	Amphibians	5	12	0.83	0.13	1.00	0.00	0.00	NA
Mayden et al. (1991)	Bony fish	5	27	0.68	0.25	1.00	0.00	0.00	NA
Trueb and Cannatella (1986)	Amphibians	6	23	0.73	0.40	0.50	0.08	0.41	0.82
Sang et al. (1995)	Angiosperms	6	16	0.95	0.74	0.60	0.00	0.00	NA
Gatesy et al. (1993)	Crocodilians	7	52	0.69	0.24	0.33	0.00	0.00	NA
Cannatella et al. (unpubl.)	Amphibians	7	199	0.53	0.10	0.13	0.24	1.14	0.93
Hillis and de Sá (1988)	Amphibians	8	15	0.77	0.50	0.24	0.02	0.11	0.54
Gardner (1991)	Mammals	8	21	0.74	0.12	0.29	0.38	1.75	0.71
Stark (1993)	Dipterans	8	15	0.70	0.37	0.43	0.20	0.98	0.87
Sytsma and Gottlieb (1986)	Angiosperms	9	55	0.93	0.82	0.11	0.00	0.00	NA
Arnold (1989)	Lizards	9	51	0.78	0.51	0.36	0.58	3.13	0.68
Cox and Urbatsch (1990)	Angiosperms	9	23	0.66	0.29	0.21	0.40	1.85	0.58
Ranker (1990)	Angiosperms	9	40	0.73	0.25	0.36	0.00	0.00	NA
Wiens (1993)	Lizards	9	23	0.70	0.41	0.43	0.12	0.63	0.94
Norell and de Queiroz (1991)	Lizards	10	41	0.62	0.17	0.47	0.11	0.53	0.86
Schultz (1990)	Arachnids	11	60	0.64	0.25	0.33	0.19	0.99	0.85
Geraads (1992)	Mammals	11	50	0.60	0.16	0.53	0.62	3.23	0.85
Rosenberg (1996)	Gastropods	11	29	0.85	0.42	0.47	0.20	1.02	0.75
Vrba et al. (1994)	Mammals	12	33	0.74	0.25	0.49	0.38	1.86	0.69
Futuyma and McCafferty (1990)	Coleopterans	13	66	0.64	0.11	0.42	1.46	7.28	0.78
Weller et al. (1995)	Angiosperms	14	20	0.77	0.34	0.26	0.18	0.91	0.63
Krajewski and Fetzner (1994)	Birds	15	166	0.54	0.23	0.54	2.93	15.90	0.94
Shaffer et al. (1991)	Amphibians	16	111	0.52	0.16	0.19	1.38	6.94	0.95
Salles (1992)	Mammals	17	21	0.77	0.34	0.42	0.42	2.27	0.94
Wild (1995)	Amphibians	17	14	0.61	0.32	0.73	1.08	5.83	0.93
Carpenter (1990)	Bony fish	20	79	0.72	0.28	0.64	1.23	6.78	0.88
Iverson (1991)	Turtles	20	34	0.74	0.37	0.57	0.68	3.73	0.87
Morrone (1994)	Coleopterans	20	37	0.47	0.15	0.40	2.27	12.23	0.94

Note. Although every effort was made to recreate results exactly, in some cases trees and tree statistics were different from those presented in the original papers. If reanalysis of published matrices yielded different results than those presented, the tree and tree statistics obtained from reanalysis were still used if the requirements outlined in the Methods section (e.g., a single optimal tree) were met.

one or more taxa has been culled from the matrix before analysis), and (b) the length of the subsampled matrix mapped on the most parsimonious tree from the complete matrix, pruned of taxa that were deleted from the matrix to obtain the reduced most parsimonious tree in (a). The procedure is as follows: First, a most parsimonous tree is obtained from the complete matrix including all taxa. Call this tree $mpt_{complete}$. Next, taxa are culled from the complete matrix to get a reduced matrix, and these same taxa are pruned from $mpt_{complete}$ to get mpt_{pruned}. A most parsimonious tree is then obtained from this reduced matrix. Call this tree $mpt_{reduced}$. The reduced matrix is then mapped onto the pruned and the reduced trees, and the sensitivity of a data set to sampling for a particular trial is the length of mpt_{pruned} minus the length of $mpt_{reduced}$. I call this value s (s_i for these individual trials, s for the average for a particular number of taxa in a particular data set, $\bar{s}$ for the average for a data set; see below). If mpt_{pruned} and $mpt_{reduced}$ are the same, then s_i is zero and the data set can be said to be insensitive to that particular removal of taxa. If the lengths are different, then s_i is the additional number of steps to fit the reduced matrix to mpt_{pruned}. Figure 1 depicts the procedure undertaken to obtain this measure for a single instance of removal of taxa. Phylogenetic analyses were performed with PAUP version 3.1 (Swofford, 1993). Statistical analyses were performed with StatView (Abacus Concepts, 1992).

Sensitivity Within Data Sets

For a data set of n ingroup taxa, the procedure in Figure 1 was repeated to get s_i for 20 removals of random samples of $n - 3$ taxa, $n - 4$ taxa, . . . , 1 taxon. Taxa were selected for removal using a random number generator written with Mathematica (Wolfram, 1991). Taxon removals that only allowed a single tree to be constructed with the unremoved taxa were ignored, and the outgroup was never removed. In cases where multiple outgroups were used in the original study, a composite outgroup was created by reconstructing ancestral states following Maddison et al. (1984), with topology determined by the states in the outgroup. This composite outgroup was used instead of pruning all but a single outgroup taxon in an attempt to recreate results from the original paper (e.g., using just one of several outgroups could change ingroup relationships and/or result in more than one most parsimonious trees). Thus, 20 s_i values were obtained for each removal of a particular number of taxa. I obtained an average for each of these sets of 20 values, and I call this average for a particular number of taxa removed s. The number of ingroup taxa included was graphed versus s to look for patterns within and between data sets and potential models for regression. Using the average rather than the raw values for this purpose amounts to the assumption that the sample mean is equal to the population mean (= the mean sensitivity to sampling for all possible removals of n – k taxa). Although this assumption is probably not met exactly, this practice has a number of beneficial effects. First, using the averages allows a much better fit of the within-data-set model (a second-order relationship—discussed later) to the data. Second, the regression assumptions of normality and equality of errors are much better satisfied by using the averages (data not shown, available on request). The main drawback is that the error obtained from this model will be underestimated because variance about the mean is not taken into account. Although this practice precludes estimation of confidence intervals for resulting equations, I consider this a necessary trade-off.

Sensitivity Between Data Sets

In order to compare the relative sensitivity of data sets to taxonomic sampling, a measure of composite sensitivity for a data set is needed. One possible measure is simply the average s for a data set ($\bar{s}$). I calculated $\bar{s}$ for each data set, and it is a useful heuristic tool for intuitively examining relative sensitivity. However, because the values of s follow a clear parabolic pattern (discussed later), a simple mean is

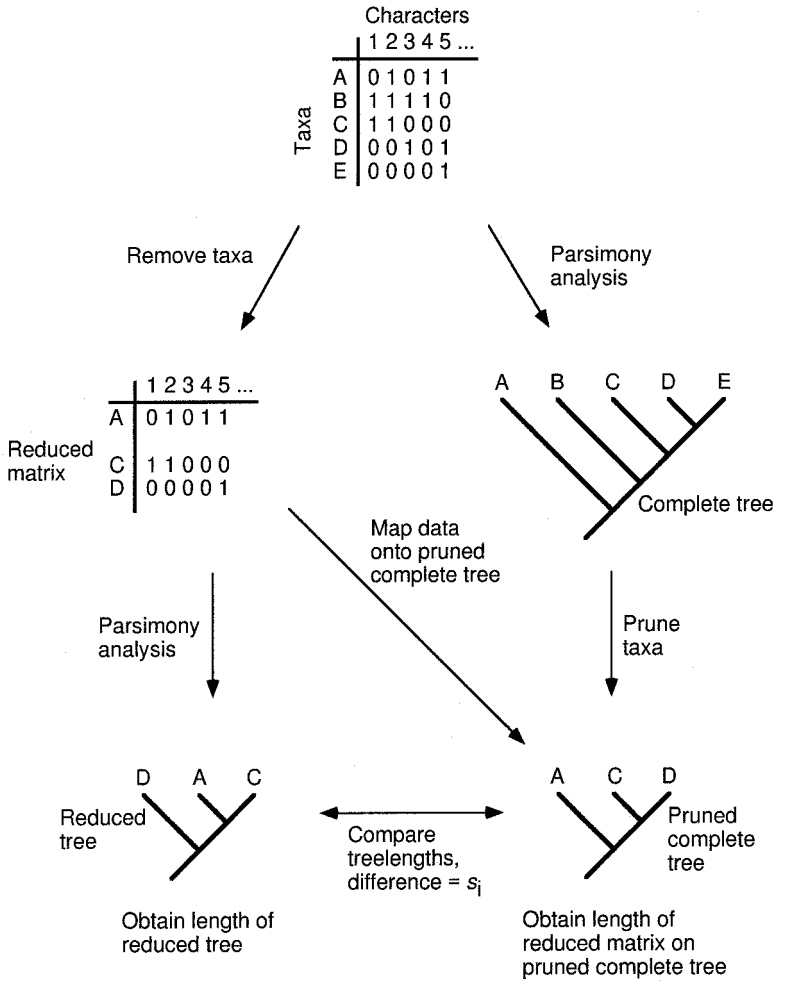

FIGURE 1. Procedure for measuring sensitivity of a data set to taxonomic sampling.

probably not the best descriptor of composite sensitivity. Furthermore, using $\bar{s}$ does not allow for a predictive equation to take into account the effect of the fraction of ingroup sampled on s_i for a particular case. These problems can be overcome by using the parameter(s) that describes the within-data-set patterns as a composite measure of sensitivity to sampling. The regression equation for within-data-set patterns of sensitivity can be described by a single variable regression parameter (discussed later). I used multiple regression to discern the behavior of this regression parameter relative to the following independent variables: (1) homoplasy, as measured by the ensemble retention index RI (Farris, 1989); (2) number of taxa; (3) number of informative characters; (4) strength of support for a tree, as measured by Bremer's (1994) total support index; and (5) tree symmetry, as measured by Colless's (1982) *I*. These variables were evaluated with both stepwise (forward and backward) and multiple regression (including all variables in the model). I then derived a regression equation to predict the sensitivity of a data set, given information on the significant variables.

Various transformations of the within-data-set descriptive parameter were tried to best satisfy the assumptions of multiple re-

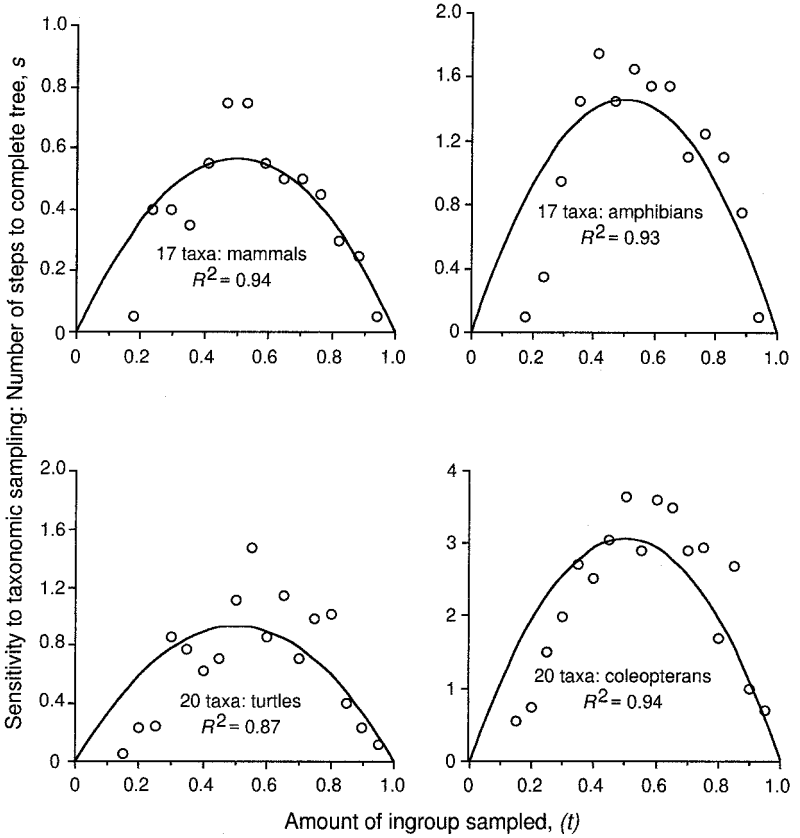

FIGURE 2. Examples of the second-order relationship between number of taxa sampled relative to the number of taxa in the complete clade *(t)*, and sensitivity to sampling *(s)*.

gression. Cook and Weisberg's (1983) diagnostic for nonconstant variance was applied to ensure that the assumption of constant variance was not significantly violated.

RESULTS

Sensitivity to Taxonomic Sampling Within Data Sets

Table 1 provides results and statistics for the 29 data sets. Small data sets are generally resistant to the effects of taxonomic sampling; seven data sets, all with nine or fewer taxa, had $\bar{s} = 0$, indicating that the same tree was obtained no matter which taxa were sampled. For all data sets, the (pruned) complete tree is on average within 0.52 steps (range = 0–2.9) of the tree obtained from reduced sampling.

It became apparent in preliminary investigations that the relationship of percentage of taxa sampled to sensitivity to sampling could be described very neatly by a second-order equation in almost all cases where taxonomic sampling had an effect (mean $R^2 = 0.82$). Table 1 lists the correlation coefficients (R^2), and Figure 2 shows graphical examples of this relationship. Assuming the second-order model, the equation

$$s = at^2 + bt + c$$

then describes the relationship of fraction of ingroup taxa sampled t to sensitivity to sampling (number of extra steps to the complete tree) s, with unknown parameters a, b, and c. Assuming that $s = 0$ when $t = 1$ (the complete tree is within 0 steps of itself) and that $s = 0$ when $t = 0$ (a tree

with no ingroup taxa is within 0 steps of the complete tree pruned of all ingroup taxa), $c = 0$ and $a = -b$, and the equation reduces to one parameter. The variable s is also 0 when $t = 1/T$ and when $t = 2/T$, where T is the total number of taxa in the clade, so one could also force the parabola through zero at either of these points. However, to do so would greatly complicate the simple parabolic relationship (increasing the number of parameters to be estimated) without greatly improving the fit of the curve. Modeling by the equation that follows allows a simple and very precise (see R^2 values in Table 1) depiction of the t to s relationship.

From the preceding arguments, the equation reduces to

$$s = bt(1 - t)$$

Thus the sensitivity to taxonomic sampling of each data set can be summarized by a single parameter b, which varies among data sets; that is, the size of the parabola varies depending on the data set. Values for b (under the assumptions just given) for each data set are listed in Table 1.

Factors Affecting Sensitivity to Taxonomic Sampling

The transformation $\ln(b + 1)$ was found to reduce correlation of variance with the magnitude of the dependent variable (e.g., Weisberg, 1985). Under this transformation, none of the independent variables were found to significantly violate the assumption of equal variance according to the test of Cook and Weisberg (1983). Although this result does not necessarily mean that the assumption of equal variance is satisfied, it does suggest that this assumption was not grossly violated.

The results of multiple regression with dependent variable $\ln(b + 1)$ (composite sensitivity to taxonomic sampling) and independent variables number of taxa, number of characters, RI, total support index, and tree symmetry are summarized in Table 2. Only number of taxa is a significant predictor of sensitivity to taxonomic sampling. Under stepwise regression (forward and backward, F to remove/include = 4), the final model included both number of taxa and RI. Figure 3 shows the relationship of number of taxa to $\ln(b + 1)$.

TABLE 2. Regression coefficients and P values for five independent variables and dependent variable $\ln(1 + b)$, where b is a measure of composite sensitivity of a data set to taxonomic sampling, model using all variables. See Table 1 and text for descriptions of independent variables.

Variable	Regression coefficient	t Value	P
Intercept	0.79	0.85	0.4014
Number of taxa	0.12	6.03	<0.0001
RI	−2.20	−1.70	0.1018
Total support	0.20	0.26	0.7959
Tree symmetry	0.36	0.85	0.4065
Number of inf. characters	0.004	1.59	0.1247

Predicting Sensitivity to Taxonomic Sampling

Given that sensitivity to taxonomic sampling is dependent on at least one of the variables studied here, a regression equation can be derived that predicts the sensitivity of a data set to taxonomic sampling. This equation is derived here for number of taxa in a clade T and fraction of taxa sampled t. Homoplasy (RI) was included with number of taxa in the stepwise regression results, but it is not incorporated into this model because number of taxa is a much more significant predictor of sensitivity to taxonomic sampling than is RI (Table 2), and the model using number of taxa alone is almost as good a fit as the model using both RI and number of taxa (removing RI from the model only lowers R from 0.87 to 0.80). Given that fraction of taxa t is related to sensitivity s by

$$s = bt(1 - t),$$

and b can be described by a linear relationship with number of taxa T such that

$$\ln(b + 1) = mT + k$$

where m is the slope and k the intercept of the regression line in Figure 3, the relationship of T and t to s can be summarized by

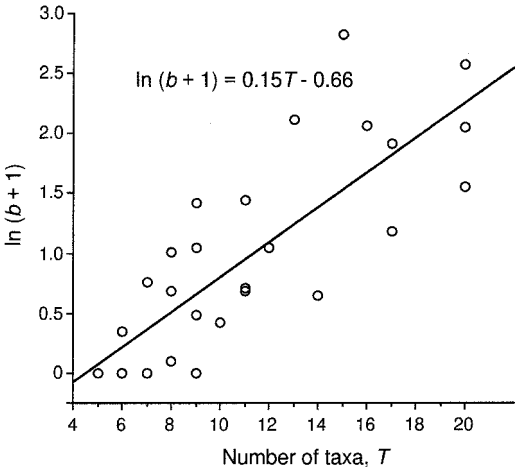

FIGURE 3. Regression relating independent variable number of taxa in a clade *(T)* to composite sensitivity of a data set to sampling *(b)*.

$$s = (e^{0.15T-0.66} - 1)(t - t^2)$$
$$3/T \leq t \leq 1.0 \qquad 5 \leq T \leq 20$$

To reiterate, T is the number of taxa potentially sampled, t is the fraction of T actually sampled, 0.66 is the intercept derived from the regression analyses, and s is the predicted number of extra steps for the hypothesized tree relative to the tree obtained following reduced sampling.

DISCUSSION

Factors Affecting Sensitivity to Taxonomic Sampling

All data sets that were sensitive to the effects of taxonomic sampling showed a pattern that was well described by a second-order equation (Fig. 2), and the composite sensitivity of data sets appears to increase predictably with greater numbers of taxa (Fig. 3). Both the within-data-sets parabolic pattern and the between-data-sets increase of s with more taxa are most easily explained as simple consequences of the number of trees compared to a single tree (the complete tree). The parabolic pattern can be explained as a consequence of few taxa having fewer solutions different from the complete tree (with four taxa there is a 1/3 probability of matching the complete tree by chance; when t is small, s is near 0), with larger numbers of taxa having solutions similar to the tree for all the taxa (using more of the same taxa results in a similar tree; when t is large, s is near 0), and intermediate numbers of taxa having neither of these ''advantages'' (when t is near 0.5, s is maximized). The symmetry of this pattern remains unexplained, and may be a fortuitous consequence of the small numbers of taxa analyzed. The relationship of increasing total number of taxa T to increasing s can be explained by a similar argument. Clades with more taxa have more potential solutions and thus more solutions with increases in tree length (s increases with T).

Only number of taxa was found to be a significant predictor of sensitivity to taxonomic sampling. Retention index, number of informative characters, tree symmetry, and, most surprisingly, total support index are not significant predictors. The nonsignificance of total support index and RI indicates that strongly supported or consistent trees are not guarantees that results will be stable to differing taxonomic coverages. Although nonsignificant in this study, retention index and number of informative characters are both nearly significant predictors. When a wider range of values is included, these variables could become significant. One might expect number of informative characters to be significant because the measure of sensitivity in this paper incorporates tree length, which obviously is a function of the number of characters. And experiments with the data from this paper suggest that both retention index and number of characters may eventually turn out to be significant: Analyzing the residuals from the T versus $\ln(b + 1)$ relationship as the dependent variable in separate simple regressions with the independent variables RI and number of informative characters produces a significant slope in each case (data not shown), indicating a nonrandom relationship. These results and experiments removing correlated variables (see later discussion) coupled with the nonsignificance of these variables in the multiple regression may suggest that although there is an (apparently weak) relationship between RI

and sensitivity to sampling and between character numbers and sensitivity to sampling, these relationships are overwhelmed by the dominant effect of number of taxa.

The principal limitation to the generality of these multiple regression conclusions is the narrow range of variable values examined. Conclusions from a regression should not be extended beyond the bounds of the variables used, so the preceding generalities and the predictive equation are limited by the values of Table 1 for number of taxa (20 or less), number of informative characters (12–199), RI (0.47–0.93), total support index (0.10–0.82), and tree symmetry (0.11–1.0). This limitation is not serious for the more completely sampled ranges (total support index and tree symmetry), but it may be a problem for RI and number of characters (which, as discussed earlier, may affect results), and it is a serious limitation for number of taxa. The most profitable extension of this work would probably be the inclusion of more data sets with greater numbers of taxa. However, the requirements set out in the beginning of this paper become more difficult to satisfy with greater numbers of taxa. These requirements hinder the procurement of a wider range of values for the other variables as well. For example, the use of only data sets that produce single most parsimonious trees may (or may not) ensure a high and narrow range of values for the retention index. Simulations are likely the only means by which these studies can be greatly extended.

A further potential difficulty in interpreting the multiple regression results is multicollinearity, the interdependence of predictor variables in a multiple regression. Homoplasy has been suggested to be dependent on number of taxa (Sanderson and Donoghue, 1989), and some of the other variables are likely to be correlated as well. Among pairwise comparisons of the independent variables used here, total support and RI were most highly correlated in the overall correlation matrix (not shown), with a correlation coefficient of 0.76. No other coefficient was greater than 0.52 (except that between taxa and $\ln(b + 1)$: 0.80), and number of taxa and RI had a coefficient of 0.42. Although none of these values reaches the rule-of-thumb problem threshold of 0.8 suggested by many authors (e.g., Licht, 1995), the RI/total support value of 0.76 does suggest the potential for problems (as well as for future research) with these two variables. One way to surmount multicollinearity problems is to remove one of the correlated variables (e.g., Dillon and Goldstein, 1984). Removing total support from the multiple regression analysis results in RI becoming a significant variable ($t = -2.21$; $P = 0.04$) in addition to number of taxa ($t = 6.93$; $P < 0.0001$). Removing total support from the analysis results in number of informative characters becoming significant ($t = 2.27$; $P = 0.03$) in addition to number of taxa ($t = 6.24$; $P < 0.0001$). Although the significance values for these variables were increased slightly in this procedure, the model using only number of taxa is still preferred because adding either RI or number of characters to the predictive equation relating T and b does not greatly increase the fit of the data to the model (R goes from 0.80 to 0.87 when RI is added and from 0.80 to 0.85 when number of characters is added).

As a final caveat, the P values in Table 2 are statistically limited to the population of matrices from which the present sample of 29 matrices were drawn: that is, matrices that produce a single optimal tree and include all or all but one known members of the clade sampled. However, there is no a priori reason to expect that either of these factors specifically affects the relative importance of the independent variables, beyond the potential corollary effects discussed earlier (e.g., broader range of RIs).

Despite these limitations, the regression results should be applicable to a wide variety of situations. Simulations may identify conditions wherein the general conclusions of this paper do not hold (e.g., data sets with a total support index of 1.0 may be insensitive to sampling), but these conditions may not be achievable in real-life evolution (Huelsenbeck, 1995) or feasible with real-life funding (e.g., hundreds of

thousands of characters). This paper has demonstrated which factors are important in predicting sensitivity to taxonomic sampling under a diversity of real-life conditions.

The Utility of $s = (e^{0.15T-0.66} - 1)(t - t^2)$

Although this equation should be informative for examining the effects of taxonomic sampling, it is not very useful for rigorous hypothesis testing. Because the s_i versus t regression does not meet the assumptions necessary for realistic error calculation, the overall error for this equation cannot be calculated with accuracy. Without knowledge of the error involved, confidence intervals for hypothesis testing cannot be calculated. An example of a hypothesis test with this equation is as a ''sampling correction'' for results from studies using incomplete ingroup coverage. For example, if one is sampling 9 of 15 species in a clade ($T = 15$, $t = 9/15 = 0.6$), the complete tree will on average be within 0.9 steps of the tree obtained. One could then examine all trees within this extra length (e.g., with a strict consensus tree) to correct for reduced sampling. If the hypothesis under question (of monophyly, character correlation, etc.) still holds in trees s steps longer than the most parsimonious trees, this result could be considered to be robust to taxonomic sampling. Tests such as this one may be possible when an equation has been derived from more extensive sampling of data sets in simulation. Table 3 lists values for the preceding equation within the limits examined in this study.

Perhaps the most useful way to think about the equation for s is to consider T to be the optimal number of taxa needed for accurate estimation of phylogeny (David Baum, pers. comm.). In this case, T has nothing to do with the completeness of the clade but rather is concerned with the number of taxa one would sample under ideal conditions

If one considers T to represent the ''complete'' clade (in some sense) rather than the optimal clade, the accuracy of this equation relative to the true tree depends on the suitability of the complete tree as a baseline comparison for reduced trees (see Fig. 1). Next I consider three possibilities: (1) The best estimate of phylogeny is the complete tree, (2) a combination of trees from reduced sampling is the best estimate of phylogeny, and (3) a tree completely different from the complete tree and the subsampled trees is the best estimate of phylogeny.

If the complete tree is the best estimate of phylogeny, then the equation gives an unbiased sampling correction of the expected result if more taxa are included, relative to the best estimate of phylogeny. But although the complete tree is probably the most desirable tree for most workers (e.g., Baverstock and Moritz, 1996), it may be no more likely to be the true tree than trees from reduced sampling (Poe, in press).

If the true tree is more similar to trees from reduced sampling than to the complete tree, then the predictive equation will overestimate the number of steps s to the true tree (while correctly estimating the number of steps s to the complete tree), because the true tree will be ''closer'' (have lower s) to the trees from reduced sampling than will be the tree to which the reduced sampling trees are being compared. This scenario is in fact very likely in cases where the true tree is not the com-

TABLE 3. Values for the predictive regression equation $s = (e^{0.15T-0.66} - 1)(t - t^2)$.

Number of taxa in clade, T	Number of taxa sampled (t)	Average number of steps to pruned complete tree, s
10	3 (0.30)	0.3
10	5 (0.50)	0.3
10	7 (0.70)	0.3
10	9 (0.90)	0.1
15	3 (0.20)	0.6
15	6 (0.40)	0.9
15	9 (0.60)	0.9
15	12 (0.80)	0.6
15	14 (0.93)	0.2
20	3 (0.15)	1.2
20	7 (0.35)	2.1
20	10 (0.5)	2.3
20	13 (0.65)	2.1
20	17 (0.85)	1.2
20	19 (0.95)	0.4

plete tree. In subsampling experiments with the known T7 phylogeny of Hillis et al. (1992), Poe (in press) found that when the complete tree (the tree with all taxa) is not the true tree, trees from reduced sampling are better estimates of phylogeny than are trees with greater sampling. In these cases the predictive equation would give a conservative estimate of number of steps to the true tree.

The third possibility is that the best tree is completely different from the complete tree and from the subsampled trees; that is, the reduced trees and the complete tree are more similar to each other than either is to the true tree. In this case, the predictive equation will give an underestimate of *s* and, potentially, a false sense of security. Although this situation is certainly undesirable, its possibility may not be too damaging to the utility of the equation. Under positively misleading conditions such as these, correcting for incomplete taxonomic sampling is probably of limited concern, as the optimal and ''corrected'' estimates will result in differently resolved trees but perhaps approximately equally bad estimates of phylogeny. Applying the equation provides no advantages, but the potential costs are trivial.

Thus it appears that the predictive equation for *s* gives either a conservative or an unbiased estimate of the number of steps to the true tree, if the true tree is obtainable from that data set.

As with the significance results for the multiple regression, an additional factor to consider in evaluating the utility of the predictive regression equation is the restricted population of matrices from which it is based—those that include all or all except one known members of the clade sampled and produce a single optimal tree. However, extrapolation of these results to other types of matrices is possible because relaxation of these criteria has predictable effects. The factors that affect the preceding equation are related to the number of trees potentially compared (see earlier discussion). Because the ''completeness'' of the clade sampled has nothing to do with the number of trees compared (beyond *T* and *t*, which are already taken into account by the equation), there is no reason to expect that relaxation of this criterion will bias the equation; that is, setting *T* at a specific number should get the same results regardless of whether *T* is the total number of taxa in the clade or some other number of taxa. Thus, *T* in the predictive equation can be thought of as the number of taxa potentially sampled or, as discussed earlier, the number needed to increase accuracy rather than as necessarily some absolute number of taxa in the clade.

Relaxation of the requirement of a single most parsimonious tree also has predictable effects that are related to the number of possible trees. With more than one most parsimonious tree, the proportion of potential ''wrong'' matches is reduced because more than one tree is acceptable as a match. The practical effect of multiple most parsimonious trees is a decrease in actual *s* due to the greater number of times *s* will be 0 (or near 0), and a concomitant overestimate of *s* by the predictive equation. The magnitude of this effect is dependent at least on the number of most parsimonious trees or, more generally, on the distribution of tree lengths. A greater number of optimal or near-optimal trees leads to more *s* values approaching zero and a lower *s* than would be predicted by the equation. As the number of optimal trees decreases (causing the proportion of trees with low *s* values to decrease), actual *s* should approach the value of *s* predicted by the equation. In sum, if this equation is applied to situations with greater than one optimal tree, it should give conservative estimates of *s*.

Additional Conclusions and Future Directions

Taxonomic sampling appears not to be a problem for many data sets in the range sampled here; the complete tree is on average within 0.52 steps of the obtained tree, and several small data sets are completely resistant to taxonomic sampling (Table 1). I take this result as an encouraging indication that taxonomic sampling is not causing large errors in phylogeny reconstruction in small clades. The size of

the clade from which taxa have been sampled can provide a useful guideline for assessing the probable effects of taxonomic sampling (Table 2, Fig. 3), and the predictive regression equation gives quantitative values of sampling effects for clades of up to 20 taxa (Table 3). However, a strongly supported tree is no guarantee of resistance to taxonomic sampling (Table 2), and the effects of taxonomic sampling on larger data sets (>20 taxa) remain to be determined.

The regression line of Figure 3 is not likely to be useful for much larger numbers of taxa for two main reasons. First, there may be an upper limit to s dependent on the number of taxa and informative characters involved (as well as other factors). It is possible that wider sampling of data sets will produce an exponential curve that levels off with greater numbers of taxa, and that number of informative characters will turn out to be an important variable. Second, because s increases exponentially, it is impossible that the equation could describe this relationship for much larger numbers of taxa (e.g., with 100 taxa, s values around 1 million are returned). Figure 3, which indicates greater scatter in the data as T increases, may suggest that the limits of this equation may be reached relatively close to 20 taxa. Greater sampling of data sets is needed to assess the limits of this equation, the significance of other variables outside of the sampled ranges, and the patterns for greater numbers of taxa. Wider sampling would come most efficiently not from more real data sets but rather from simulations. Simulation studies are currently underway testing the effect of taxonomic sampling on tree length, in the hope that the predictive regression equation of this paper can be derived for greater numbers of taxa and, more importantly, relative to a true tree rather than to a completely sampled tree. Alternatively, it may be found that, in general, increasing taxonomic sampling does not help accuracy at all (Kim, 1996).

This paper has assessed the average effects of taxonomic sampling, without regard for where in the tree those taxa are placed. The results of this study should be most pertinent when taxa have been subsampled from a clade for which little or nothing is known of the phylogeny. In these cases, information on strategic sampling—where in the tree to add taxa (e.g., Felsenstein, 1978; Hendy and Penny, 1989; Huelsenbeck, 1991)—is less helpful than information on the average effects of taxonomic sampling because the placement of taxa cannot be estimated a priori. The investigation of such general trends and the effects of particular placements of taxa are both fruitful areas for future research.

ACKNOWLEDGMENTS

I thank David Baum, Jim Bull, David Cannatella, David Hillis, Jim McGuire, and two anonymous reviewers for useful reviews of the manuscript. I thank David Cannatella, Matt Brauer, and the Systematics Discussion Group at the University of Texas at Austin for helpful discussion. Financial support was provided by a University of Texas Fellowship and a National Science Foundation Graduate Fellowship.

REFERENCES

ABACUS CONCEPTS, INC. 1992. StatView 4.0. Berkeley, CA.

ANSCOMBE, F. J. 1973. Graphs in statistical analysis. Am. Statist. 27:17–21.

ARNOLD, E. N. 1989. Towards a phylogeny and biogeography of the Lacertidae: Relationships within an Old-World family of lizards derived from morphology. Bull. Br. Mus. Nat. Hist. (Zool.) 55:209–257.

BAVERSTOCK, P., AND C. MORITZ. 1996. Project Design. Pages 17–27 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

BREMER, K. 1994. Branch support and tree stability. Cladistics 10:295–304.

CARPENTER, K. E. 1990. A phylogenetic analysis of the Caesionidae (Perciformes: Lutjanoidea). Copeia 1990:692–717.

COLLESS, D. H. 1982. Review of *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*, by E. O. Wiley. Syst. Zool. 31:100–104.

COOK, R. D., AND S. WEISBERG. 1983. Diagnostics for heteroscedasticity in regression. Biometrika 70:1–10.

COX, P. B., AND L. E. URBATSCH. 1990. A phylogenetic analysis of the coneflower genera (Asteraceae: Heliantheae). Syst. Bot. 15:394–402.

DE PINNA, M. C. 1992. A new subfamily of Trichomycteridae (Teleostei, Siluriformes), lower loricarioid relationships and a discussion of the impact of additional taxa for phylogenetic analysis. Zool. J. Linn. Soc. 106:175–229.

DILLON, W. R., AND M. GOLDSTEIN. 1984. Multivariate analysis: Methods and applications. John Wiley and Sons, New York.

DONOGHUE, M. J., J. DOYLE, J. GAUTHIER, A. G. KLUGE, AND T. ROWE. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20: 431–460.

DOYLE, J., AND M. DONOGHUE. 1987. The importance of fossils in elucidating seed plant phylogeny and macroevolution. Rev. Palaeobot. Palynol. 50:63–95.

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

FUTUYMA, D. J., AND S. S. MCCAFFERTY. 1990. Phylogeny and the evolution of host plant associations in the leaf beetle *Ophraella* (Coleoptera, Chrysomelidae). Evolution 44:1885–1913.

GARDNER, S. L. 1991. Phyletic coevolution between subterranean rodents of the genus *Ctenomys* (Rodentia: Hystricognathi) and nematodes of the genus *Paraspidodera* (Heterakoidea: Aspidoderidae) in the Neotropics: Temporal and evolutionary implications. Zool. J. Linn. Soc. 102:169–201.

GATESY, J., R. DESALLE, AND W. C. WHEELER. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. 2: 152–157.

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

GERAADS, D. 1992. Phylogenetic analysis of the tribe Bovini (Mammalia: Artiodactyla). Zool. J. Linn. Soc. 104:193–207.

HENDY, M. C., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

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

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.

HILLIS, D. M., AND R. DE SÁ. 1988. Phylogeny and taxonomy of the *Rana palmipes* group (Salientia: Ranidae). Herp. Monog. 2:1–26.

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

HUELSENBECK, J. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. 40: 458–469.

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

IVERSON, J. B. 1991. Phylogenetic hypotheses for the evolution of modern kinosternine turtles. Herp. Monog. 5:1–27.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

KRAJEWSKI, C., AND J. W. FETZNER, JR. 1994. Phylogeny of cranes (Gruiformes: Gruidae) based on cytochrome-B DNA sequences. Auk 111:351–365.

LECOINTRE, G., H. PHILIPPE, H. L. VAN LE, H. LE GUYADER. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. 2: 205–224.

MAYDEN, R. L., W. J. RAINBOTH, AND D. G. BUTH. 1991. Phylogenetic systematics of the cyprinid genera *Mylopharodon* and *Ptychocheilus*: Comparative morphometry. Copeia 1991:819–834.

MORRONE, J. M. 1994. Systematics, cladistics, and biogeography of the Andean weevil genera *Macrostyphlus, Adioristidius, Puranius,* and *Amathynetoides,* new genus (Coleoptera: Curculionidae). Amer. Mus. Nov. 3104:1–63.

NETER, J., W. WASSERMAN, AND M. KUTNER. 1983. Applied linear regression models. Richard D. Irwin, Homewood, Illinois.

NORELL, M., AND K. DE QUEIROZ. 1991. The earliest iguanine lizard (Reptilia: Squamata) and its bearing on iguanine phylogeny. Am. Mus. Nov. 2997:1–16.

PENNY, D., AND M. D. HENDY. 1985. Testing methods of evolutionary tree reconstruction. Cladistics 1:266–272.

POE, S. In press. The effect of taxonomic sampling on phylogeny estimation: Test case of a known phylogeny. Mol. Biol. Evol.

RANKER, T. A. 1990. Phylogenetic systematics of neotropical *Hemionitis* and *Bommeria* (Adiantaceae) based on morphology, allozymes, and flavonoids. Syst. Bot. 15:442–453.

ROSENBERG, G. 1996. Independent evolution of terrestriality in Atlantic truncatellid gastropods. Evolution 50:682–693.

SALLES, L. O. 1992. Felid phylogenetics: Extant taxa and skull morphology (Felidae, Aeluroidea). Amer. Mus. Nov. 3047:1–67.

SANG, T., D. J. CRAWFORD, AND T. F. STUESSY. 1995. ITS sequences and the phylogeny of the genus *Robinsonia* (Asteraceae). Syst. Bot. 20:55–64.

SCHULTZ, J. W. 1990. Evolutionary morphology and phylogeny of Arachnida. Cladistics 6:1–38.

SHAFFFER, J. B., J. M. CLARK, AND F. KRAUS. 1991. When molecules and morphology clash: A phylogenetic analysis of North American ambystomatid salamanders (Caudata: Ambystomatidae). Syst. Zool. 40:284–303.

SIDDALL, M. 1996. Another monophyly index: Revisiting the jackknife. Cladistics 11:33–56.

SOKAL, R. R., AND F. J. ROHLF. 1981. Biometry, 2nd edition. W. H. Freeman, San Francisco.

STARK, J. 1993. A revision of the neotropical genus *Dagus* Cresson (Diptera: Ephydridae). Am. Mus. Nov. 3080:1–21.

SYTSMA, K. J., AND L. D. GOTTLIEB. 1986. Chloroplast DNA evolution and phylogenetic relationships in *Clarkia* sect. Peripetasma (Onagraceae). Evolution 40:1248–1261.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, version 3.1. Illinois Natural History Survey, Champaign.

TRUEB, L., AND D. C. CANNATELLA. 1986. Systematics, morphology, and phylogeny of genus *Pipa* (Anura: Pipidae). Herpetologica 42:412–449.

VRBA, E. S., J. R. VAISNYS, J. E. GATESY, R. DESALLE, AND K. Y. WEI. 1994. Analysis of paedomorphosis using allometric characters: The example of Reduncini antelopes (Bovidae, Mammalia). Syst. Biol. 43: 92–116.

WEISBERG, S. 1985. Applied linear regression, 2nd edition. John Wiley and Sons, New York.

WELLER, S. G., W. L. WAGNER, AND A. K. SAKAI. 1995. A phylogenetic analysis of *Schiedea* and *Alsinidendron* (Caryophyllaceae: Alsinoideae): Implications for the evolution of breeding systems. Syst. Bot. 20: 315–337.

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

WIENS, J. J. 1993. Phylogenetic systematics of the tree lizards (genus *Urosaurus*). Herpetologica 49:399–420.

WIENS, J. J., AND T. W. REEDER. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. 44:548–558.

WIENS, J. J., AND T. TITUS. 1991. A phylogenetic analysis of *Spea* (Anura: Pelobatidae). Herpetologica 47: 21–28.

WILD, E. R. 1994. New genus of Amazonian microhylid frog with a phylogenetic analysis of New World genera. Copeia 1994:837–849.

WINTERBOTTOM, R. 1990. The *Trimmatom nanus* species complex (Actinopterygii, Goviidae): Phylogeny and progenetic heterochrony. Syst. Zool. 39:253–265.

WOLFRAM, S. 1991. Mathematica, 2nd edition. Addison-Wesley, Reading, Massachusetts.

Received 17 September 1996; accepted 10 December 1996
Associate Editor: D. Baum