Letter to the Editor

The Effect of Taxonomic Sampling on Accuracy of Phylogeny Estimation: Test Case of a Known Phylogeny

Steven Poe[1]

Department of Zoology and Texas Memorial Museum, University of Texas at Austin

Because of extinction, undiscovered species, and constraints of time, money and material, most molecular phylogenetic analyses are performed without extensive coverage of the group of interest. The degree of taxonomic coverage can affect hypotheses of tree topology (e.g., Gauthier, Kluge, and Rowe 1988), ancestral state assignment (Sanderson 1996), character evolution (Donoghue et al. 1989), and character correlation (Sillén-Tullberg 1993). Studies are beginning to examine the effect of taxon sampling on accuracy (e.g., Huelsenbeck 1991; Wheeler 1992; Wiens and Reeder 1995; Kim 1996). The favorability of including more taxa is now a working assumption for many (e.g., Baverstock and Moritz 1996, p. 26). This paper will examine the relationship of taxon sampling to accuracy using the known phylogeny of T7 bacteriophage of Hillis et al. (1992).

Hillis et al. (1992) created a known phylogeny of the T7 bacteriophage by serially propagating strains in the presence of a mutagen. For this first experimental phylogeny they enforced completely symmetrical branching and equal time between splitting events. This system has proven useful for questions of performance of phylogenetic methods (Hillis et al. 1992; Crandall 1994; Hillis and Huelsenbeck 1994; Wiens and Reeder 1995), molecular evolution (Bull et al. 1993; Cunningham et al. 1997), and confidence in phylogeny (Hillis and Bull 1993). Details of this phylogeny are available in Hillis et al. (1992), Bull et al. (1993), and Sober (1993). The advantages of the T7 system are that (unlike in studies of literature data sets) the phylogeny is known and (unlike in simulations) the tree is the result of biological evolution in a lineage of organisms (Hillis 1995). Disadvantages of this system have to do with applicability. For example, molecular evolution in T7 viruses in a controlled experiment may not be a useful model for evolution in many ''higher'' organisms under real-life conditions, and the single tree shape and set of branch lengths may not be representative of most phylogenies (although subsampling taxa and characters can ''create'' different phylogenetic conditions).

I perform subsampling experiments of taxa and characters from the T7 restriction site data to examine the effect of taxonomic sampling at three levels: (1) Does sampling more taxa improve or worsen accuracy for a given number of characters? (2) Does sampling more taxa improve or worsen accuracy for individual data sets? (3) Given a four-taxon tree, can adding a taxon affect accuracy for those four taxa?

To produce data sets with both large and small numbers of characters and taxa, I created 25 sets of 10, 25, 40, and 55 characters by choosing randomly without replacement (jackknifing) from the pool of 87 parsimony-informative characters (see Crandall 1994). These sets encompass the range of interesting variation: 10 characters is probably about the minimum for resolution, whereas at 55 characters, almost all character sets produce the correct tree. For each of these 100 data sets of subsampled characters, matrices for all informative combinations of ingroup taxon removal were compiled (the outgroup taxon was never removed). Because the T7 ingroup includes eight taxa, 219 ($n!/r![n - r]!$, $n = 8$, $r = 0, 1, \ldots, 5$) combinations are possible. Each of these 219 $\times$ 4 (numbers of characters) $\times$ 25 (sets per number of characters) = 21,900 data sets was analyzed with the branch-and-bound search with all characters unordered and equally weighted on PAUP 3.1 (Swofford 1993).

Comparing the relative accuracies of trees with different numbers of taxa is not simple; conclusions can depend on the measure of similarity used and the null expectations incorporated. For example, measuring success of cladogram estimation as to whether or not all clades are reconstructed correctly will bias results against trees with large numbers of taxa. In this paper, I use two measures of accuracy: The first measure, called ''raw accuracy,'' is simply the fraction of clades reconstructed correctly (in cases of multiple most-parsimonious trees, the average fraction was used). Although this measure is intuitively satisfying, results may be biased toward decreasing accuracy with the addition of taxa, because trees with low numbers of taxa will tend to have higher accuracy values by chance (e.g., a randomly chosen four-taxon tree has a ⅓ probability of 100% accuracy, whereas a random five-taxon tree has a $1/15$ probability of 100% accuracy). In an attempt to correct for this bias, I also included a measure that takes the expected values for randomly chosen trees into account. For each number of taxa, I scaled the range of raw accuracy values such that zero corresponds to the expected value for randomly chosen trees (values from Penny, Foulds, and Hendy 1982) of that number of taxa, with 1.0 still the maximum ([raw accuracy − expected value for random trees]/[1 − expected value for random trees]). This measure is called ''scaled accuracy.'' For example, in the four-taxon case, a raw accuracy of 0.67 corresponds to a scaled accuracy of 0.5, because 0.67 is 0.5 between the expected value for randomly chosen trees (0.33) and the maximum (1.0). Although the out-

[1] Present address: Division of Amphibians and Reptiles, National Museum of Natural History, Smithsonian Institute, Washington, D.C.

Key words: accuracy, known phylogeny, T7 bacteriophage, taxonomic sampling, phylogeny estimation.

Address for correspondence and reprints: Steven Poe, Division of Amphibians and Reptiles, MRC 162, National Museum of Natural History, Smithsonian Institute, Washington, D.C. 20560. E-mail: poe.steve@nmnh.si.edu.

Mol. Biol. Evol. 15(8):1086–1090. 1998

group was never removed, comparisons were made using unrooted trees.

To assess whether accuracy increases with taxon addition, I calculated average accuracy for 10, 25, 40, and 55 characters for 4, 5, 6, 7, 8, and 9 taxa using both raw and scaled accuracies. Figure 1*A* and *B* shows the resulting patterns, including the expected values for randomly chosen trees. For the raw values, the 25- and 55-character data sets show very slight but significant (Spearman's rho, $P < 0.05$) increases in accuracy with taxon addition, and the 40-character data sets shows no significant increase or decrease. The 10-character data sets show a decrease in accuracy as taxa are added. For the scaled values, the 25-, 40-, and 55-character data sets show a slight but significant increase in accuracy as taxa are added. As with the raw values, the 10-character data sets show a significant decrease in accuracy as taxa are added.

Figure 1*C–F* shows the effect of taxon sampling on individual samples of jackknifed characters (rather than average performance for a given number of characters, as in fig. 1*A* and *B*). Only the raw accuracy results are presented, because scaled results are similar and would excessively clutter the graphs. For discussion, patterns are divided into three categories: 1) a positive relationship between number of taxa and accuracy, operationally defined as having increasing values of accuracy as the number of taxa increases and a significant ($P < 0.05$) Spearman signed-ranks correlation; (2) a negative relationship between number of taxa and accuracy, operationally defined as having decreasing values of accuracy as the number of taxa increases and a significant ($P < 0.05$) Spearman correlation; (3) no relationship between number of taxa and accuracy, defined as showing a nonsignificant Spearman correlation. I recognize the potential for multiple test problems here (i.e., Stevens 1992, p. 6). However, this phenomenon should result in spurious significance for both positive and negative relationships and thus should not bias the results in a particular direction.

For the 55-character data sets (fig. 1*F*), the correct tree is usually obtained regardless of how many taxa are sampled. Twenty-two of twenty-five 55-character data sets showed no relationship between number of taxa and accuracy. Some data sets of 40 characters (fig. 1*E*) show increasing accuracy with the addition of taxa, whereas some show decreasing accuracy and others show no effect with the addition of taxa. Overall, 5 of 25 data sets of 40 characters showed a positive relationship between number of taxa and accuracy, 4 showed a negative relationship, and 16 showed no relationship. The 25-character data sets (fig. 1*D*) produce the most ambiguous patterns. The relationships that were evident with the 40-character data sets are apparent here (e.g., decreasing accuracy with more taxa), but other patterns decrease and then increase in accuracy with the addition of taxa. Ten of twenty-five 25-character data sets showed a positive relationship between taxa and accuracy, and six showed a negative relationship. The remaining nine data sets showed various patterns, including insensitivity to sampling (as in the 55-character data sets) and curvilinear relationships. Most of the 10-character data sets (fig. 1*C*) decreased in accuracy as more taxa were added. Eighteen of twenty-five 10-character data sets showed a

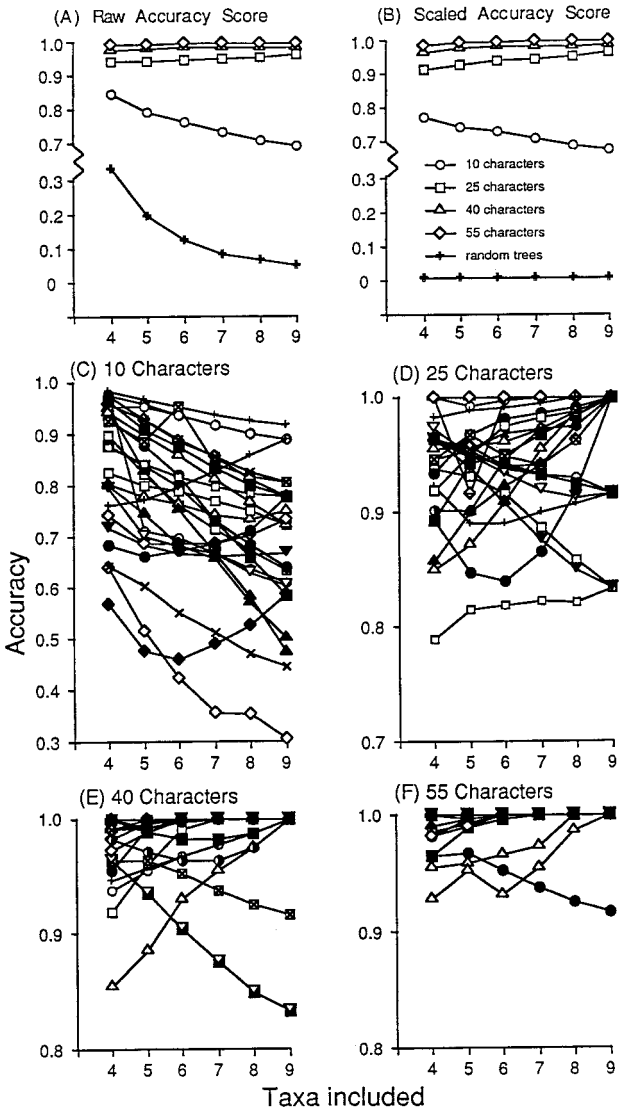

FIG. 1.—Effect of number of taxa on accuracy. *A,* Raw accuracy, measured as the fraction of clades shared by the estimated tree with the true tree. Each point is the mean accuracy value for 25 jackknifed sets of a particular number of characters and taxa. Lines connect values for each number of characters. *B,* Scaled accuracy, measured as in *A* but scaled such that the expected value for random trees corresponds to zero ([raw accuracy − expected value for randomly selected tree]/1 − expected value for random trees]). *C,* Effect of number of taxa on raw accuracy for each of 25 jackknifed subsamples of 10 characters. Each point is the mean raw accuracy value for a particular number of taxa in one set of 10 jackknifed characters. Lines connect values for each subsample. *D,* Effect of number of taxa on raw accuracy for each of 25 jackknifed subsamples of 25 characters. Each point is the mean raw accuracy value for a particular number of taxa in one set of 25 jackknifed characters. *E,* Effect of number of taxa on raw accuracy for each of 25 jackknifed subsamples of 40 characters. Each point is the mean raw accuracy value for a particular number of taxa in one set of 40 jackknifed characters. *F,* Effect of number of taxa on raw accuracy for each of 25 jackknifed subsamples of 55 characters. Each point is the mean raw accuracy value for a particular number of taxa in one set of 55 jackknifed characters. Note difference in scale of vertical axis between graphs.

Table 1
Results of Experiments Adding a Taxon to a Four-Taxon Tree

No. of Characters	No. of Data Sets	No. of Data Sets for which Adding a Taxon Affects Relationships	% Cases in which Adding a Taxon Improves Four-Taxon Relationships (% in which Improvement is to Single Correct Tree) ± SE	% Cases in which Adding a Taxon Worsens Four-Taxon Relationships (% in which Single Correct Tree is Worsened) ± SE
10	10	3	0.7 (0.04) ± 0.4	1.2 (0) ± 0.9
25	10	6	1.4 (0.3) ± 0.5	0 (0)
40	10	3	1.1 (0.4) ± 0.7	0 (0)
55	10	1	0.4 (0.07) ± 0.4	0.04 (0) ± 0.04
10, 25, 40, 55	40 (all)	13	0.9 (0.2) ± 0.3	0.3 (0) ± 0.2

NOTE.—Standard errors are between data sets (i.e., not individual trials). Data sets are not independent; standard errors are listed to describe the spre ad of the data.

negative relationship between number of taxa and accuracy; one showed a positive slope.

Figure 1 gives the results for general trends for adding or deleting taxa, but these trends do not address whether adding taxa improves or worsens a preexisting set of phylogenetic relationships. If a taxon is added in the wrong place on an otherwise correct tree and does not affect the relationships of that underlying tree, the accuracy score as measured in this paper will decrease. Likewise, adding a taxon in its correct place to an incorrectly resolved tree will cause an increase in accuracy score. In both of these cases, a change in accuracy is recorded, but in neither of them does adding the taxon affect relationships; that is, these cases do not argue for or against addition of taxa as a means to improve the estimate for a particular group of taxa.

In order to address the effects of adding taxa to a given set of relationships, I examined in detail four- and five- (including outgroup) taxon trees from 10 jackknifed samples of 10, 25, 40, and 55 characters. For each data set, I compared each of the 56 possible four-taxon trees with the 5 possible five-taxon trees that can result from addition of a taxon ($56 \times 5 \times 10$ data sets $\times$ 4 numbers of characters = 11,200 comparisons). I recorded cases in which (1) the true four-taxon tree was among the optimal trees before taxon addition but not after, taking special note of cases in which a single correct tree was made incorrect by addition of taxa, and (2) the true four-taxon tree was not among the optimal trees before taxon addition but was among them after, taking special note of cases in which a single correct tree was obtained with the addition of a taxon.

Table 1 shows the results of the four-taxon experiments. Adding a taxon usually does not affect preexisting relationships (0.6% of trials), but when relationships are affected, they are usually improved. Improvement to a single correct tree from one or more incorrect trees occurred relatively often (42% of improvements). In no case did addition of a taxon cause a correct single most-parsimonious tree to be incorrectly estimated. When four-taxon relationships were damaged by the addition of a taxon, change was always from multiple most-parsimonious trees, one of which was correct, to one or more incorrectly estimated trees.

The three sets of results in this paper (fig. 1 and table 1) offer different levels of insight into the effects of taxonomic sampling on accuracy. Figure 1*A* and *B* show that overall accuracy decreases with a low number of characters but increases with larger numbers of characters. This increase is negligible, though, in raw accuracy (e.g., from 0.9756 accuracy for four taxa to 0.9833 for nine taxa at 40 characters) and slight in scaled accuracy (e.g., from 0.9634 accuracy for four taxa to 0.9824 for nine taxa at 40 characters). In the case of this T7 data set, adding taxa does not help much, and in one case (10 characters), it is decidedly detrimental. Also, adding characters is better for accuracy than adding taxa for this data set and this range of characters and taxa (two-way ANOVA with the independent variables no. of taxa, no. of characters, and dependent variable arcsin transformed raw accuracy; no. of taxa: df = 5, $F = 0.5$, $P = 0.7736$; no. of characters: df = 3, $F = 388$, $P < 0.0001$).

Figure 1*C–F* show that adding taxa consistently decreases accuracy in 10-character data sets, has little effect on accuracy for 55-character data sets, and does not consistently increase or decrease accuracy for sets of 25 or 40 characters. This result suggests that adding taxa may have predictable effects when character numbers are low enough that resolution may be a problem or when support is strong, but that effects may be harder to predict in other cases. Certainly it is safe to conclude that the tree with the most taxa does not necessarily have the highest percentage of correct clades. Further, it is clear that for increased sampling to be helpful, enough characters must be present to resolve relationships (fig. 1*C–F*).

Wiens and Reeder (1995) examined the effect of adding incompletely scored taxa to parsimony analyses of the T7 data in a set of experiments similar to those described in figure 1. The aspects of that study that overlapped with this one were congruent. For example, Wiens and Reeder (1995) found that when larger numbers of more completely scored characters are used taxonomic sampling has little effect on results, as is shown from this study in figure 1*F*.

In the four-taxon experiments, over 99% of taxon additions had no effect on the original four-taxon relationships. Perhaps this result is not surprising given that this phylogeny is especially amenable to correct reconstruction (Sober 1993). However, one might expect that random variation would create approximately equal

numbers of cases in which adding taxa improves original relationships and in which adding taxa worsens relationships. But this pattern did not emerge; adding taxa seldom worsened original relationships and never caused a correct single most-parsimonious tree to be incorrectly estimated. This result suggests the possibility that natural conditions under which adding taxa damages original relationships may be rare. More concretely, these experiments showed that adding taxa can help resolve a set of relationships correctly under conditions of real evolution.

The lack of change in most preexisting relationships in the four-taxon experiments suggests that increases in accuracy with taxon addition shown by some data sets in figure 1*C–F* may generally be due to the placement of the added taxon in its correct position (rather than improvement of preexisting relationships). Decreases in accuracy with addition of taxa appear to be due mainly to incorrect placement (as opposed to upsetting of existing relationships) caused by an insufficient number of characters. Without added characters as taxa are added, accuracy would be expected to decrease eventually because more characters are needed to resolve a greater number of branches. The 10-character data sets exemplify this phenomenon. Virtually all of them decrease in accuracy (fig. 1*C*), but in none of them did adding a taxon worsen preexisting four-taxon relationships (table 1).

How does adding taxa help or hurt preexisting relationships? Adding taxa can ''break up'' branches that have a concentration of homoplasy that results in spuriously attracted lineages, thus nullifying that attraction and causing an incorrect estimate to be improved (Hendy and Penny 1989). Similarly, addition of a taxon may create an imbalance in homoplasy and a worsening of relationships (Kim 1996). Lineages that previously were not spuriously attracted to each other could become ''long'' in a relative sense by virtue of the shortening of another branch on which the added taxa connect. Both of these circumstances may occur in the T7 data set, with the former much more common. Relative to the parsimony method used in this paper, phylogenetic methods such as maximum likelihood that allow for flexibility in model choice may better reconstruct phylogenies in cases of suboptimal taxonomic sampling. Greater accuracy does not necessarily follow from use of likelihood, but improvement is likely if the model used is a better description of evolutionary change for the phylogeny of interest than is the parsimony model (e.g., Felsenstein 1978).

Overall conclusions are that: (1) adding taxa tends to cause a slight increase in accuracy if enough characters are used (fig. 1*A* and *B*); (2) adding taxa can cause a decrease in accuracy (fig. 1*C*—*F*), but this decrease does not generally involve preexisting relationships (table 1); and (3) adding taxa can improve preexisting relationships (table 1). The utility of these findings for an investigator is, of course, directly related to one's willingness to generalize results from the T7 bacteriophage system. In assessing properties of tree estimation, the role of known phylogenies is to act as an experimental test, or reality check, for generalizations or predictions made from simulation studies or from theory (Hillis 1995). For example, theory, simulations (e.g., Hendy and Penny 1989; Kim 1996), and now, in this paper, a known phylogeny have demonstrated cases in which adding a taxon improves accuracy for a subset of taxa. Conditions also exist under which adding taxa can worsen relationships (Kim 1996), but these conditions were found to be rare for the T7 data set. Future work should be directed toward determining not only what conditions can produce inaccuracy in phylogeny reconstruction, but also whether these conditions are achievable or common in nature.

Acknowledgments

I thank David Cannatella for helpful discussions and for reading the manuscript, David Hillis for reading an early version of the manuscript and providing the T7 data on disk, Craig Moritz and anonymous reviewers for comments, and members of the Hillis/Bull labs for discussions on this paper. Financial support was provided by fellowships from the University of Texas and the NSF.

LITERATURE CITED

BAVERSTOCK, P., and C. MORITZ. 1996. Project design. Pp. 17–28 *in* D. M. HILLIS, C. MORITZ and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

BULL, J. J., C. W. CUNNINGHAM, I. J. MOLINEAUX, M. R. BADGETT, and D. M. HILLIS. 1993. Experimental molecular evolution of bateriophage T7. Evolution **47**:993–1007.

CRANDALL, K. 1994. Intraspecific cladogram estimation: accuracy at higher levels of divergence. Syst. Biol. **43**:222–235.

CUNNINGHAM, C. W., K. JENG, J. HUSTI, M. BADGETT, I. J. MOLINEAUX, D. M. HILLIS, and J. J. BULL. 1997. Parallel molecular evolution of deletions and nonsense mutations in bateriophage T7. Mol. Biol. Evol. **14**:113–116.

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.

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

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

HENDY, M. D., 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., and J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–192.

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

HILLIS, D. M., and J. P. HUELSENBECK. 1984. To tree the truth: biological and numerical simulations of phylogeny. Pp. 55–67 *in* D. M. FAMBROUGH, ed. Molecular evolution of physiological processes. Rockefeller University Press, New York.

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

KIM, J. 1996. General inconsistency conditions for parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. **45**:363–373.

PENNY, D., L. R. FOULDS, and M. D. HENDY. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature **297**:197–200.

SANDERSON, M. 1996. How many taxa must be sampled to identify the root node of a large clade? Syst. Biol. **45**:168–173.

SILLÉN-TULLBERG, B. 1993. The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. Evolution **47**:1182–1191.

SOBER, E. 1993. Experimental tests of phylogenetic inference methods. Syst. Biol. **42**:85–89.

STEVENS, J. 1992. Applied multivariate statistics for the social sciences. 2nd edition. Lawrence Erlbaum Associates, Hillsdale, N.J.

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

WHEELER, W. 1992. Extinction, sampling, and molecular phylogenetics. Pp. 205–215 *in* M. J. NOVACEK and Q. D. WHEELER, eds. Extinction and phylogeny. Columbia University Press, New York.

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

CRAIG MORITZ, reviewing editor

Accepted April 2, 1998