Syst. Biol. 47(1):3–8, 1998

Taxonomic Sampling, Phylogenetic Accuracy, and Investigator Bias

DAVID M. HILLIS

Department of Zoology and Institute of Cellular and Molecular Biology, University of Texas, Austin, Texas 78712, USA; E-mail: hillis@phylo.zo.utexas.edu

In this issue of *Systematic Biology*, a series of authors use several different approaches to examine the effects of taxonomic sampling on phylogenetic analysis. This topic is receiving increasing attention, in part because recent studies have reached a confusing diversity of conclusions about the effects of taxonomic sampling. For instance, contrast the conclusions reached in two recent papers on this topic:

> If the evolutionary question of interest does not require a large number of taxa, it seems best to use fewer taxa because larger trees are more likely to contain inconsistent branches. (Kim, 1996:372)

> Including large numbers of taxa in an analysis may be the best way to ensure phylogenetic accuracy. (Hillis, 1996:131)

These recommendations, taken at face value, appear to be in direct conflict with regard to advice on taxon sampling. The papers in this issue extend these studies and modify these recommendations on the basis of analyses of real data sets (Soltis et al., 1998; Poe, 1998), simulations (Graybeal, 1998), and theoretical considerations (Kim, 1998). One conclusion from reading these papers is that whether increased taxonomic sampling helps or hinders the process of accurate phylogenetic estimation depends to a great extent on how accuracy is evaluated and what is meant by ''taxonomic sampling.''

LOCAL VERSUS GLOBAL EFFECTS OF TAXONOMIC SAMPLING

Much of the apparent disagreement among authors on the effects of taxonomic sampling stems from the different evaluation criteria being evaluated. Kim (1998) discusses several of the criteria, and emphasizes the differences between evaluating efficiency versus consistency in phylogenetic analysis. Although this difference is important, a greater difference occurs depending upon whether investigators choose to evaluate the phylogenetic performance on a branch-by-branch basis or on a tree-by-tree basis. For every taxon added to an analysis, we are also attempting to estimate an additional internal branch. Thus, the problem gets more complex as we add taxa, and there are more places in the tree where problems with inconsistency may arise. This led Kim (1996) to his recommendation just quoted, and some authors (e.g., Graur et al., 1996) regularly heed this advice by reducing taxonomic problems to the simplest possible four-taxon trees. Under this strategy, a systematist samples all possible quartets of taxa that involve the internal branch of interest, and tabulates the number of times each of the three possible trees is supported. For instance, Graur et al. (1996) evaluated the relationships of rabbits by evaluating all possible quartets of taxa selected from rabbits, primates, other mammals, and an outgroup (a marsupial, monotreme, or reptile). None of the quartets supported the traditional group Glires (rabbits plus rodents), which they took as evidence that rabbits and rodents are not closely related.

One problem with reducing a phylogenetic analysis to its simplest possible form is that four-taxon trees can be very difficult to estimate correctly if rates of evolution are high (e.g., Hillis et al., 1994). Kim (1996:372) concluded ''that to be 95% confident of avoiding inconsistency problems, the expected number of changes over the entire tree for a given character must be less than one out of four.'' However, much higher rates of character evolution are acceptable (and even desirable) if the tree is densely sampled. To demonstrate this point, I modified the simulation of the 228-taxon tree from Hillis (1996) by increasing the expected amount of change along all the branches by 10-fold (Fig. 1). This tree

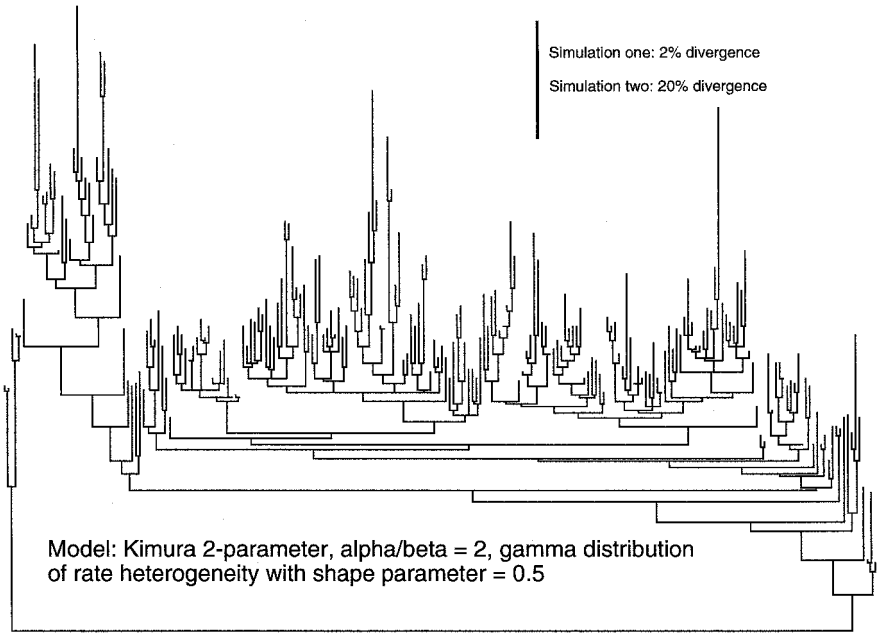

FIGURE 1. A model tree based on the phylogenetic analysis of angiosperm diversity by Soltis et al. (1997). In the original simulation (Hillis, 1996), rates of divergence were based on the observed rates among the angiosperms; in this case (simulation one), the scale bar represents 2% divergence. In the present paper, the simulation was repeated (simulation two), but evolutionary rates were increased so that the expected divergence was 10-fold greater (the scale bar represents 20% divergence in this case).

is based on a phylogenetic estimate of angiosperm relationships (Soltis et al., 1997), and as such it represents an approximation of the topology of the kind of tree that systematists are actually attempting to estimate. The characters are evolving according to a Kimura two-parameter model of evolution, with a 2:1 transition:transversion ratio, and rate heterogeneity among sites (modeled with a gamma distribution with the shape parameter $\alpha = 0.5$). Under these conditions, the average character is changing 23.6 times across the tree, and because of the rate heterogeneity among sites, some characters change many more times. At these high rates of evolution, many of the terminal sequences are so dissimilar that no biologist would recognize them as homologous. Nonetheless, the tree is accurately reconstructed with just a few thousand nucleotides, and many of the branches require fewer data to reconstruct than at lower rates of evolution (Fig. 2).

Suppose we are interested in a particular internal branch in the tree (marked with an arrow in Fig. 3). This branch is correctly estimated in the full tree if all the taxa are included. If we sample a quartet of taxa to examine this same branch (e.g., as in Fig. 3), then the branch will be inconsistently estimated for almost every possible quartet. For the quartet of taxa shown in Figure 3, the probability that a single nucleotide will be misinformative about the relationships of the four taxa under the parsimony criterion is approximately 0.4. The probability that a single nucleotide will be informative about the relationships of the four taxa under the parsimony criterion is approximately 0.006. Thus, one would expect to converge on the wrong solution for these four taxa with great speed under these conditions; only a few nucleotides would need to be sequenced to guarantee finding the wrong solution. In contrast, if all the taxa are included in the analysis, then the branch is correctly reconstructed with a few thousand nucleo-

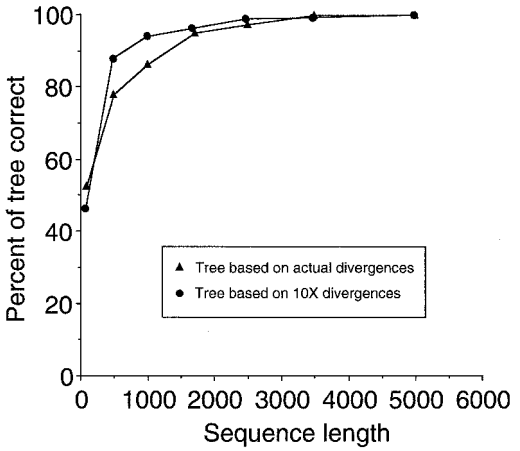

FIGURE 2. Performance of parsimony in estimating the 228-taxon tree shown in Figure 1. "Percent of tree correct" is based on the partition metric (Robinson and Foulds, 1981; Penny and Hendy, 1985). All internal branches in the tree are correctly estimated with 5,000 nucleotides, for either simulation.

tides. Clearly, at least some phylogenetic problems require intensive and extensive taxonomic sampling.

TAXONOMIC SAMPLING SCHEMES

There are many different ways that systematists might select taxa for analysis. In many cases, the taxa selected will be based on availability. In other cases, it might be possible to select taxa according to a sampling strategy. Let us assume that a sampling strategy is possible, and imagine that a systematist is interested in analyzing the phylogeny of a large and diverse group, such as angiosperms. We will also assume that preliminary data are available for 20 species. The systematist now has time and money to add 200 more species to the analysis, so some strategy for taxonomic sampling is necessary. Consider five of the many possible strategies:

1. Add the 200 additional taxa randomly from living organisms (e.g., the systematist would sample randomly from the tree of life).
2. Choose taxa randomly within the monophyletic group of interest (in this example, the systematist would randomly sample 200 additional angiosperms).
3. Select taxa within the monophyletic group of interest that will represent the overall diversity of the group. For example, the systematist might select two divergent representatives from each of 100 different families of angiosperms, purposefully chosen to best represent angiosperm diversity.
4. Select taxa within the monophyletic group of interest that are expected (based on current taxonomy or previous phylogenetic studies) to subdivide long branches in the initial tree.
5. Add (and delete) taxa until the a priori biases of the systematist are supported. I call this last strategy the Theriot Effect after the tongue-in-cheek practices of Theriot et al. (1995:4): "We added or discarded characters [taxa] until we achieved the results we believed, then stopped."

Although this range of options may seem extreme, they reflect the range of studies that have been conducted on the topic of "taxonomic sampling." I expect few practicing systematists would purposefully choose sampling strategies 1, 2, or 5. The first strategy would ensure the inclusion of very long branches in the tree, and genes that were evolving at an appropriate rate for elucidating the phylogenetic relationships among angiosperms would likely be saturated for changes among the other taxa. Adding additional taxa would not reduce the branch lengths in the tree, and the additional taxa would be highly unlikely to help resolve angiosperm phylogeny. The second strategy might seem more likely, but I doubt any systematist would choose this approach either. If he or she did, a large percentage of the added taxa would be composites and orchids, and most of the families of angiosperms would be unrepresented. The dangers of the Theriot Effect (strategy 5) should be clear, and hopefully this strategy would not be selected. I would expect the typical plant systematist to choose something similar to the third sampling strategy, or, if he or she was ex-

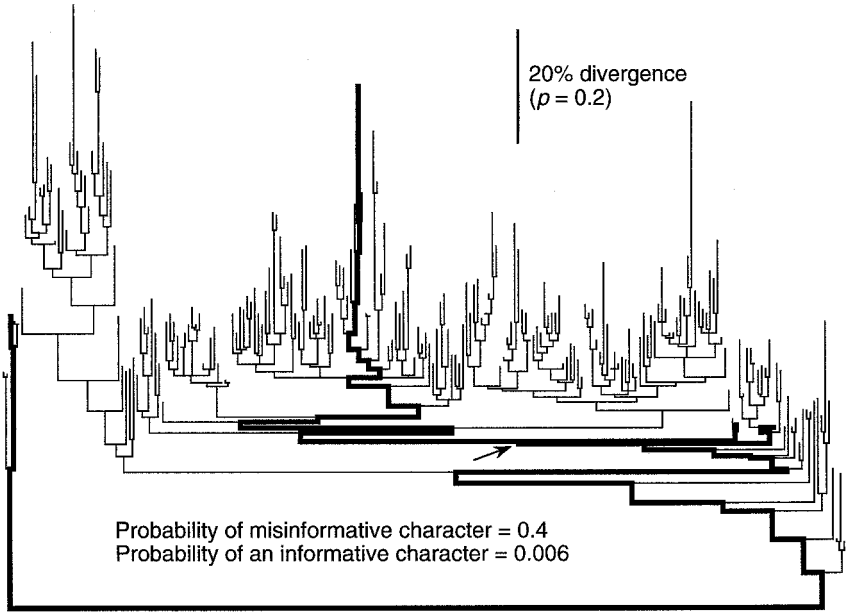

FIGURE 3. Correct phylogenetic estimation of a small internal branch (indicated by the arrow) is strongly dependent on taxonomic sampling. Under the conditions simulated (10 times the observed rates of evolution for angiosperms), if only the four taxa highlighted in bold were sampled, then a misinformative character (one that would support one of the two wrong trees for four taxa) is approximately 67 times more likely than an informative character (one that supports the correct tree). The wrong tree would be estimated with virtual certainty if more than a few characters were collected. However, the branch in question is reconstructed correctly in the analysis of all the taxa. The vast majority of other quartets of taxa defined by this (and many other) internal branches show the same effect.

plicitly adding taxa to reduce problems with long-branch attraction, the systematist might choose the fourth option. It is likely that he or she would choose some combination of strategies 3 and 4.

Kim's (1996) study is most relevant to sampling strategy 1, or adding increasingly distantly-related taxa to the analysis. In his principal simulation, Kim (1996) evaluated a sampling scheme in which taxa are added without reducing the average branch length of taxa in the tree. Namely, he randomly selected a tree relating t taxa, to which he randomly assigned branch lengths from an exponential distribution. To examine the effects of taxonomic sampling, he held the average branch length in the tree constant while changing the number of taxa included in the analysis. In the real world, this sampling scheme could only be approximated by adding successively more distantly related taxa (i.e., outside the original group of interest), so that the addition of taxa did not reduce the length of the average branch in the tree. Kim's (1996) simulation suggests that systematists are correct to avoid this strategy.

Kim (1998) conducted new simulations to evaluate strategy 2, of randomly adding taxa from the group of interest to the analysis. Under these conditions, he found that addition of taxa can either increase or decrease the difference in parsimony scores between the model tree and its nearest-neighbor trees. Although this measure does not directly assess the accuracy of phylogenetic estimates, it does suggest that it is better to add some taxa than others. Which are the best taxa to add? Not surprisingly, the taxa that break up long branches (and thereby make the trees less star-like) are the best ones to add. This adds support for strategy 4, or the pur-

poseful division of long branches in the tree.

Yang and Goldman (1997) also recently reported a set of simulations in which taxa were randomly selected for analysis from the group of interest (see also Purvis and Quicke, 1997). They found that the percentage of taxa sampled from a clade had a greater effect on phylogenetic accuracy than did the absolute number of taxa sampled. This is expected under random sampling of taxa if the speciation and extinction rates are held constant through time in the modeled tree. Under these conditions, the estimated tree for 20 taxa sampled from a model tree of 1,000 taxa will be nearly star-like (very small internal branches with long peripheral branches), whereas the estimated tree for 20 taxa sampled from a model tree of 20 taxa will have many relatively large internal branches. Once again, this suggests that investigator control of the addition of taxa can have a highly beneficial effect on phylogenetic analyses.

Graybeal (1998) evaluated strategy 4, namely, purposefully breaking up long branches in the tree by judicious addition of taxa. This follows the recommendations of most recent authors on the subject of taxon sampling (e.g., Hendy and Penny, 1989; Swofford et al., 1996). She found that addition of such taxa is not only strongly beneficial, but under many conditions accuracy of the phylogenetic estimate improves with the addition of taxa *even if the total number of characters examined remains unchanged.* In other words, given a limited amount of time and money for phylogenetic analysis, one can sometimes improve the accuracy of the phylogenetic estimate by collecting fewer data for more taxa. Obviously, there are limits to this effect, but Graybeal's (1998) results highlight just how beneficial judicious taxon sampling can be.

What are the effects of taxon sampling as practiced by real systematists? Obviously, this will vary from case to case, but the studies by Soltis et al. (1998) and Poe (1998) provide some insight. The apparent tractability of the real angiosperm tree sampled by Soltis et al. (1998) indicates that systematists have chosen taxa well. The study of empirical data sets by Poe (1998) indicates that for clades with small numbers of taxa, incomplete sampling is not likely to be a serious problem. This reinforces the idea that the percentage of included taxa in a clade may be a more important consideration than the total number of included taxa. However, more empirical studies are needed to examine the effects of sampling few taxa from a clade of many taxa; the angiosperm data set of Soltis et al. (1998) appears to be ideal for this purpose.

The papers in this issue are useful for identifying the range of outcomes of taxonomic sampling schemes, from the very bad (e.g., strategy 1: randomly adding taxa from the tree of life) to the very good (e.g., strategy 4: adding taxa to break up long branches). Random sampling of taxa from a group of interest (strategy 2) can be effective or not, depending on the details of the true tree. However, it is obviously not the best strategy, nor is it the strategy likely to be used by most systematists. Careful addition of taxa to ensure coverage of the group of interest and to purposefully break up long branches (a combination of strategies 3 and 4) seems to be optimal. In some cases, deletion of problematic taxa (e.g., taxa with abnormally high rates of evolution) may also be warranted. Unfortunately, purposeful addition and deletion of taxa allows the possibility of consciously or unconsciously biasing the results (the dreaded Theriot Effect). This problem would be easy to overcome through use of a simple method, namely, the blinding of taxon names during analysis. If taxa are to be selected for inclusion or exclusion after an initial analysis, this should be done without the a priori knowledge of the investigator of the effects on the analysis of the additions or deletions. Thus, all decisions about inclusion or exclusion of taxa would be based only on information about the tree itself, thus avoiding the possibility of an investigator selecting taxa on the basis of how closely the results match his or her preconceived notions of relationship.

Blinding of taxon names should be a standard feature of programs for phylogenetic analysis.

Although there is still much disagreement about the expected effects of taxonomic sampling in phylogenetic analysis, there are a few conclusions that seem to be uncontroversial. First, at least some large, very complex trees are far easier to estimate than most systematists would have guessed. Second, some small trees (e.g., quartets) are among the hardest possible phylogenetic trees to estimate correctly. Third, inclusion of many taxa in a densely sampled tree permits more effective use of rapidly evolving characters than in a poorly sampled tree. Fourth, judicious addition of taxa can move some phylogenetic problems from the virtually impossible to the tractable. Fifth, addition of taxa does not always make problems easier; adding highly divergent taxa, for instance, can make phylogenetic estimation harder. Sixth, taxonomic sampling, as practiced by systematists, typically does not involve random sampling of taxa, nor is this expected to be a particularly effective strategy. Finally, given the role of a systematist in selecting taxa for inclusion or exclusion in an analysis, and given the possibility of thereby biasing the results of the analysis, systematists should use blinding of taxon names during the decision-making process.

It is clear that taxonomic sampling can have important consequences for phylogenetic analysis. Therefore, systematists should give careful consideration to how they decide which taxa to add to an analysis, and should describe their sampling strategy. Theorists should evaluate competing sampling strategies, and emphasize realistic sampling strategies rather than invent new sampling strategies that no systematist could or would use. Perhaps then we can begin to formulate more practical advice on the subject of how to best sample taxa to estimate relationships within the tree of life.

REFERENCES

GRAUR, D., L. DURET, AND M. GOUY. 1996. Phylogenetic position of the order Lagomorpha (rabbits, hares, and allies). Nature 379:333–335.

GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.

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

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

HILLIS, D. M., J. P. HUELSENBECK, AND D. L. SWOFFORD. 1994. Hobgoblin of phylogenetics? Nature 369:363–364.

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

KIM, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47:43–60.

PENNY, D., AND M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool. 34:75–82.

POE, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47:18–31.

PURVIS, A., AND D. L. J. QUICKE. 1997. Are big trees indeed easy? Reply from A. Purvis and D. L. J. Quicke. TREE 12:357–358.

ROBINSON, D. F., AND L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

SOLTIS, D. E., P. S. SOLTIS, M. E. MORT, M. W. CHASE, V. SAVOLAINEN, S. B. HOOT, AND C. M. MORTON. 1998. Inferring complex phylogenies using parsimony: An empirical approach using three large DNA data sets for angiosperms. Syst. Biol. 47:32–42.

SOLTIS, D. E., P. S. SOLTIS, D. L. NICKRENT, L. A. JOHNSON, W. J. HAHN, S. B. HOOT, J. A. SWEERE, R. K. KUZOFF, K. A. KRON, M. W. CHASE, S. M. SWENSEN, E. A. ZIMMER, S.-M. CHAW, L. J. GILLESPIE, W. J. KRESS, AND K. J. SYTSMA. 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Missouri Bot. Gard. 84:1–49.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd. edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

THERIOT, E. C, A. E. BOGAN, AND E. E. SPAMER. 1995. The taxonomy of Barney: Evidence of convergence in hominid evolution. Ann. Improb. Res. 1:3–7.

YANG, Z., AND N. GOLDMAN. 1997. Are big trees indeed easy? TREE 12:357.

Received 10 November 1997; accepted 20 November 1997
Associate Editor: D. Cannatella