As expected, with more data (total nucleotides) we were able to reconstruct more accurate phylogenies. When the number of taxa sampled per clade was increased, the interrelationships of those clades could be inferred more accurately. However, sampling in experimental design is only relevant in the context of resource limitation; therefore, to compare apples to apples, we used the number of nucleotides per sequence (number of taxa × sequence length) as a control. In this case, trees are more accurately reconstructed when using more sites for fewer taxa than when using more taxa for fewer sites when the total number of nucleotides is held constant in a data set. This result is stronger for distance and likelihood methods of phylogeny reconstruction but less so for parsimony. We reconstructed most of the short internal branches with a reasonably degree of accuracy (Fig. 4) with an adequate amount of data (whether taxa or sites). This result is certainly encouraging for phylogenetic reconstruction in general.

The results presented here and by in Rosenberg and Kumar (2001) provide a useful framework for analyzing the effect of taxon sampling in phyloinformatic and phylogenomic studies.

ACKNOWLEDGMENTS

We thank B. Friedman, K. Tamura, J. Thorne, D. Hillis, D. Pollock, and two anonymous reviewers for comments on this manuscript. This research was supported by grants from the National Science Foundation (DBI-9983133), the National Institute of Health (HG-02096), and the Burroughs Wellcome Fund (BWI 1001311) to S.K.

REFERENCES

EIZRIK, E., W. J. MURPHY, AND S. J. O'BRIEN. 2001. Molecular dating and biogeography of the early placental mammal radiation. J. Hered. 92:212–219.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

HILLIS, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

KUMAR, S. 1996. A stepwise algorithm for finding minimum evolution trees. Mol. Biol. Evol. 13:584–593.

MURPHY, W. J., E. EIZIRIK, W. E. JOHNSON, Y. P. ZHANG, O. A. RYDER, AND S. J. O'BRIEN. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.

NEI, M., S. KUMAR, AND K. TAKAHASHI. 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc. Natl. Acad. Sci. USA 95:12390–12397.

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

POLLOCK, D. D., D. J. ZWICKL, J. A. MCGUIRE, AND D. M. HILLIS. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

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

ROSENBERG, M. S., AND S. KUMAR. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98:10751–10756.

SWOFFORD, D. L. 1998. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer, Sunderland, Massachusetts.

TAKAHASHI, K., AND M. NEI. 2000. Efficiencies of fast algorthims of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. 17:1251–1258.

TAMURA, K., AND M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

ZWICKL, D. J., AND D. M. HILLIS. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 8 July 2002; reviews returned 31 July 2002; final acceptance 23 September 2002
Associate Editor: Chris Simon

Syst. Biol. 52(1):124–126, 2003
DOI: 10.1080/10635150390132911

Is Sparse Taxon Sampling a Problem for Phylogenetic Inference?

DAVID M. HILLIS,[1] DAVID D. POLLOCK,[2] JIMMY A. MCGUIRE,[3] AND DERRICK J. ZWICKL[1]

[1]*Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, 1 University Station C0930, Austin, Texas 78712-0253, USA; E-mail: dhillis@mail.utexas.edu (D.M.H.), zwickl@mail.utexas.edu (D.J.Z.)*
[2]*Department of Biological Sciences and Biological Computation and Visualization Center, Louisiana State University, Baton Rouge, Louisiana 70803, USA; E-mail: dpollock@lsu.edu*
[3]*Museum of Natural Science and Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA; E-mail: jmcguire@lsu.edu*

Rosenberg and Kumar (2001) addressed the importance of taxon sampling in phylogenetic analysis and concluded that phylogenetic error is "largely independent of taxon sample size" (2001:10756) and that their "results do not provide evidence in favor of adding taxa to problematic phylogenies" (2001:10756). In response to these conclusions, Zwickl and Hillis (2002) and Pollock et al. (2002) conducted additional simulations and reanalyzed the data presented by Rosenberg and Kumar (2001). Zwickl and Hillis and Pollock et al. showed that these conclusions of Rosenberg and Kumar could not be supported either by analyses of their original data or by new simulations that corrected a number of deficiencies in Rosenberg and Kumar's original experimental design. Both Zwickl and Hillis and Pollock et al. found that increased taxon sampling resulted in greatly reduced phylogenetic estimation error, and Pollock et al. showed that the benefits of increased taxon sampling were similar to

adding an equivalent amount of sequence length for the same taxa (in the ranges simulated by Rosenberg and Kumar).

In their response, Rosenberg and Kumar (2002) focused on a slightly different conclusion from that in their original paper, which was that "longer sequences, rather than extensive sampling, will better improve the accuracy of phylogenetic inference" (2001:10751). In 2001, Rosenberg and Kumar argued that the beneficial effect of increasing taxa was 10-fold lower than the beneficial effect of increasing sequence length and that the effects of increased taxon sampling for the same genes were negligible ("largely independently" of phylogenetic error). Rosenberg and Kumar (2002) have now concluded that the beneficial effect of increasing taxon sample size is not small, but they suggested that the benefit comes simply from the overall increase in size of the data matrix (the total number of characters × taxa). Furthermore, they maintained that there is a greater benefit to increasing the total sequence length for few taxa than can be obtained by increasing taxon sampling for the same genes. Here, we discuss the two sets of conclusions reached by Rosenberg and Kumar (2001, 2002).

Is Phylogenetic Error Independent of Taxon Sample Size?

> The use of different sample sizes (number of taxa) may lead to different phylogenetic inferences; however, the error associated with these estimates is largely independent of the sample size (Rosenberg and Kumar, 2001:10756).

> [I]ncreased sampling of taxa is one of the most important ways to increase overall phylogenetic accuracy (Zwickl and Hillis, 2002:588).

> A directed strategy of adding taxa to a phylogenetic analysis will often be one of the most profitable uses of time and resources (Pollock et al., 2002:670).

The conclusion reached by Rosenberg and Kumar (2001) is clearly in conflict with the other two quoted conclusions, and all three cannot be correct. Either the number of taxa in a phylogenetic analysis is largely independent of the error in the phylogenetic estimates, or it is not. Rosenberg and Kumar (2001) concluded that there was a large effect of adding more sequence data per taxon examined but that there was only a minimal effect (10-fold lower) of adding more taxa for the same genes. Rosenberg and Kumar (2002) then stated that their "results indicate that increasing the number of taxa can dramatically increase the accuracy of the relationships among the sampled clades" (2002:122). Therefore, we all now appear to agree that phylogenetic error is strongly and negatively correlated with taxon sample size and that phylogenetic error is strongly and negatively correlated with character sample size (number of characters examined per taxon).

Although Rosenberg and Kumar (2002) argued that the absolute increase in accuracy that resulted from increased taxon sampling in their original study was small, this is clearly because the total error in these simulations was relatively low to begin with. One point presented by Pollock et al. (2002) was that a measure of percentage of error removed rather than a measure of total error removed gives a clearer picture of the effects of taxon sampling. When there is only one incorrect branch in an analysis and it is corrected by taxon sampling, then taxon sampling can hardly be faulted for not having a larger effect. In the case of small taxon samples, there are very few branches being estimated to begin with; correcting one incorrect branch in a four-taxon tree by adding additional taxa can hardly be said to be insignificant. In cases where overall error rates are higher (e.g., Rosenberg and Kumar, 2002: fig. 3), the absolute and the relative benefits of increased taxon sampling are substantial and similar to the effects of increased sequence length. Thus, the argument of Rosenberg and Kumar (2002) that the absolute effects of taxon sampling are small is specious. The absolute effects of taxon sampling are dependent on the amount of error present in the particular problem, but in general taxon sampling represents an excellent means of reducing or eliminating whatever phylogenetic error may exist.

Rosenberg and Kumar (2002) also emphasized a second conclusion: if resources are limited, it is better to collect more characters for fewer taxa than to collect fewer characters for more taxa. Therefore, we now turn to that question.

With Finite Resources, Is It Better to Add More Taxa or More Characters?

> When resources are limited, one would appear to do better by sequencing more sites/genes per taxon than by increasing the number of taxa for shorter sequences (Rosenberg and Kumar, 2002:000).

> [U]nder the conditions of Rosenberg and Kumar's [2001] simulations, error reduction can be achieved equally well by taxon addition or by increasing sequencing length (Pollock et al., 2002:669).

> Accuracy improved dramatically with the addition of taxa and much more slowly with the addition of characters. If taxa can be added to break up long branches, it is much more preferable to add taxa than characters (Graybeal, 1998:9).

> 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 (Hillis, 1998:7).

Rosenberg and Kumar (2002) concluded that the beneficial effects of adding taxa to a phylogenetic analysis are simply an effect of adding more total data and that one would actually do better by adding more characters and holding the number of taxa sampled constant. Our position is that the answer to this question (which is better: more taxa or more characters?) depends entirely on the starting point and conditions of the study. If many characters have already been obtained for few taxa, it is often better to add more taxa than to add additional characters for the same taxa (as was nicely demonstrated by Graybeal, 1998). However, if relatively few characters have been obtained for the taxa analyzed to date, one would often do better by adding more characters per taxon. Moreover, the effects of adding taxa to an analysis are not simply explained by the increase in overall data.

Numerous simulation studies (see Hillis et al., 1994, for examples) have shown that adding characters in a phylogenetic analysis typically leads to rapid convergence on a particular solution (the exact rate of convergence is dependent on the details of the underlying tree and model of evolution). However, most methods of analysis are inconsistent for certain small-taxon problems unless the underlying processes of evolution have been modeled perfectly (e.g., the well-studied Felsenstein zone problem; Felsenstein, 1978; Huelsenbeck and Hillis, 1993). Because complete knowledge of evolutionary processes is unobtainable in most realistic situations, it is important to add enough taxa to make the phylogenetic problem tractable. When the data set includes only a few taxa that diverged a long time ago (i.e., the taxa are separated by long branches), virtually any method of phylogenetic analysis is likely to be inaccurate across almost any real sample of characters (especially because the accuracy of estimation of the parameters of any evolutionary model are also dependent on taxon sample size). Therefore, addition of data for these same few taxa will likely lead to convergence on an incorrect solution (i.e., the estimation methods will be inconsistent). In this case, it is clearly better to add taxa to the analysis, thereby making the problem tractable. It is also biologically unrealistic to assume that all sites in a sequence behave identically. When the underlying evolutionary processes are different across different sites, evolutionary inferences can be improved dramatically by adding taxa to the analysis, whereas adding longer sequences is not similarly beneficial (Pollock and Bruno, 2000). However, when the divergence times among taxa are short and the parameters of the evolutionary models are relatively easy to estimate, then addition of characters for the same number of taxa will lead to quick convergence on a correct solution.

Rosenberg and Kumar (2002) based their conclusion that more characters are better than more taxa on an analysis that compared sampling 15–45 taxa for 500–2,000 nucleotide positions. However, as shown by Pollock et al. (2002: fig. 2) there is a very rapid decrease in phylogenetic error under the conditions simulated by Rosenberg and Kumar (2001) as one increases from 500 to 1,000 nucleotides, with comparatively little benefit gained by adding additional sequence length beyond 1,000 nucleotides. In contrast, the decrease in phylogenetic error that results from increased taxon sampling appears to be close to linear across the entire range of taxon sample sizes examined by Rosenberg and Kumar (2001), as shown in Pollock et al.'s (2002) Figure 4. Thus, it is not surprising that the benefit of adding characters is somewhat greater than the benefit of adding taxa in this limited range of parameter space; the examined conditions are on the part of the curve where adding characters results in the greatest reduction of error. Thus, one indeed may be better off sequencing an additional 500 nucleotides for 30 taxa than randomly adding another 30 taxa with the same 500 nucleotides. In contrast, however, if one already has 5,000 nucleotides sequenced across 30 taxa (under the conditions simulated by Rosenberg and Kumar), then it would be much better to collect data on another 30 taxa than to collect data on another 5,000 nucleotides for the same taxa. In this range, addition of more characters makes little difference, but addition of more taxa is still greatly beneficial. If the taxa can be added purposefully (e.g., to break up long branches), then the benefits of increased taxon sampling would be even greater.

There is no simple answer to the question posed in the heading of this section; the answer will depend on the particular situation being examined (the scope of the problem, the number of taxa already sequenced, the number of characters already collected, and the quantity and availability of additional relevant taxa to include). We disagree with the assertion of Rosenberg and Kumar (2002) that more characters per taxon is necessarily a better strategy than more taxa for the same characters. Rosenberg and Kumar (2002) put their argument in terms of the current genome sequencing studies, in which many genes (or complete genomes) are examined from very few taxa. Rosenberg and Kumar (2002) argued that their conclusions "mesh well" with this scattered genome approach. In contrast, we propose that this approach will likely result in poorly estimated evolutionary models, poorly estimated phylogenetic trees, and a poor overall view of evolutionary history. If one is interested in inferring the evolutionary history of life, a much broader sample of taxa (perhaps sequenced for far less than full genomes) will result in a much more accurate estimate of phylogeny than will complete genomes of only a small sample of taxa.

REFERENCES

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

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

HILLIS, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.

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

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

POLLOCK, D. D., AND W. J. BRUNO. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. 17:1854–1858.

POLLOCK, D. D., D. J. ZWICKL, J. A. MCGUIRE, AND D. M. HILLIS. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

ROSENBERG, M. S., AND S. KUMAR. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98:10751–10756.

ROSENBERG, M. S., AND S. KUMAR. 2002. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

ZWICKL, D. J., AND D. M. HILLIS. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 26 September 2002; reviews returned 1 October 2002; final acceptance 8 October 2002
Associate Editor: Jeffrey Thorne