Syst. Biol. 51(4):664–671, 2002
DOI: 10.1080/10635150290102357

Increased Taxon Sampling Is Advantageous for Phylogenetic Inference

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

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

Until recently, it was believed that complex phylogenies might be extremely difficult to reconstruct due to the phenomenal rate of increase in the number of possible phylogenies as the number of taxa increases. However, Hillis (1996) showed through simulation that, for at least one complex phylogeny of angiosperms with 228 taxa, reconstruction was far more accurate than expected, even with relatively modest amounts of DNA sequence data. This led to a flurry of papers on the subject of taxon sampling and phylogenetic reconstruction, with focus quickly shifting from the question of whether complex

phylogenies can be reconstructed to whether and how much an existing phylogeny can be improved through increased taxon sampling (Hillis, 1998; Kim, 1998; Poe, 1998; Poe and Swofford, 1999; Pollock and Bruno, 2000; Rannala et al., 1998; Yang, 1998). Although a statistician might intuitively believe that it is generally better (or at least no worse) to increase the amount of data to resolve a question in statistical inference, the benefits of taxon addition for phylogenetic inference remain controversial. Some researchers have argued that taxon addition can decrease accuracy (Kim, 1996, 1998), while others believe that increased sampling improves accuracy (Graybeal, 1998; Hillis, 1996, 1998; Murphy et al., 2001; Poe, 1998; Pollock and Bruno, 2000; Pollock et al., 2000; Soltis et al., 1999). The reasons that different papers come to apparently contradictory conclusions deserve careful consideration.

An often cited factor affecting the benefits of taxon addition is the phenomenon of long-branch attraction (LBA). Some phylogenetic methods have a bias toward preferential clustering of long branches, leading to erroneous results when those long branches do not actually represent a monophyletic assemblage (Felsenstein, 1978; Hendy and Penny, 1989). This phenomenon has been cited in favor of increased taxon sampling, since sampling can be designed to break up long branches (Hillis, 1998). However, increased sampling has also been implicated as a potential cause of LBA because addition of a new long branch may wrongly attract a pre-existing long branch that had previously been inferred correctly (Poe and Swofford, 1999; Rannala et al., 1998). LBA may also explain some simulations that have found problems in phylogeny estimation when sampling outside the taxonomic group of interest (but see Pollock and Bruno [2000] for an alternative explanation). Outside sampling in these simulations tended to add long branches, which tended to attract the longest unbroken branch in the group of interest (Hillis, 1998; Rannala et al., 1998). The degree to which LBA is a problem depends greatly on the method of analysis, and LBA is much less of a problem for maximum likelihood (ML) than for parsimony or distance methods (Bruno and Halpern, 1999).

A recent paper on the subject of taxon addition (Rosenberg and Kumar, 2001) concludes that increased taxon sampling is of little benefit to phylogenetic inference when compared to increasing sequence length. We disagree with their interpretation and believe that their data support the importance of increased taxon sampling. In addition, some of their data were simulated under extreme conditions (i.e., substitution rates that were very high or low, or sequences that were unreasonably short). Large error values and nonlinear relationships at these extremes make it difficult to interpret effects for the majority of the range, and averaging across the entire range is inappropriate. Moreover, we do not believe that Rosenberg and Kumar (2001) used the most appropriate metric to measure the relative effect of taxon addition. Our reanalysis of their simulated data indicates that increased taxon sampling is highly beneficial for phylogenetic inference.

Reanalysis of Simulations on the Mammalian Phylogenetic Tree

Rosenberg and Kumar (2001) addressed the effects of partial taxon sampling on the error rate of phylogenetic estimation. Their main results are given in their Table 1, where each row represents the results of 100 simulations on a 66-taxon phylogenetic tree of eutherian mammals (Murphy et al., 2001). Sequences between 200 and 3,000 nucleotides in length (randomly chosen from the uniform distribution) were simulated under the Jukes-Cantor model of evolution (Jukes and Cantor, 1969) with substitution rates sampled from a gamma distribution with shape parameter equal to 1.0. The average rate of this distribution was not given directly, and can only be inferred visually from a scale bar on their tree, and hence is unclear. The error in the phylogenetic tree (E_G) determined from these simulated sequences was calculated as the fraction of internal branches at which the tree differed from the "true" tree used for the simulations (Robinson and Foulds, 1981). For each set of simulations, a subset of between 5 and 50 taxa was chosen, and the sub-tree relating this subset of taxa was determined in two ways (see Fig. 1): first, by using the subset of sequences (S), and second, by pruning the tree inferred from the complete set of sequences (P). The errors in these smaller phylogenetic trees (E_S and E_P) were calculated in a similar fashion by calculating the fraction of internal branches at which they differed from the corresponding

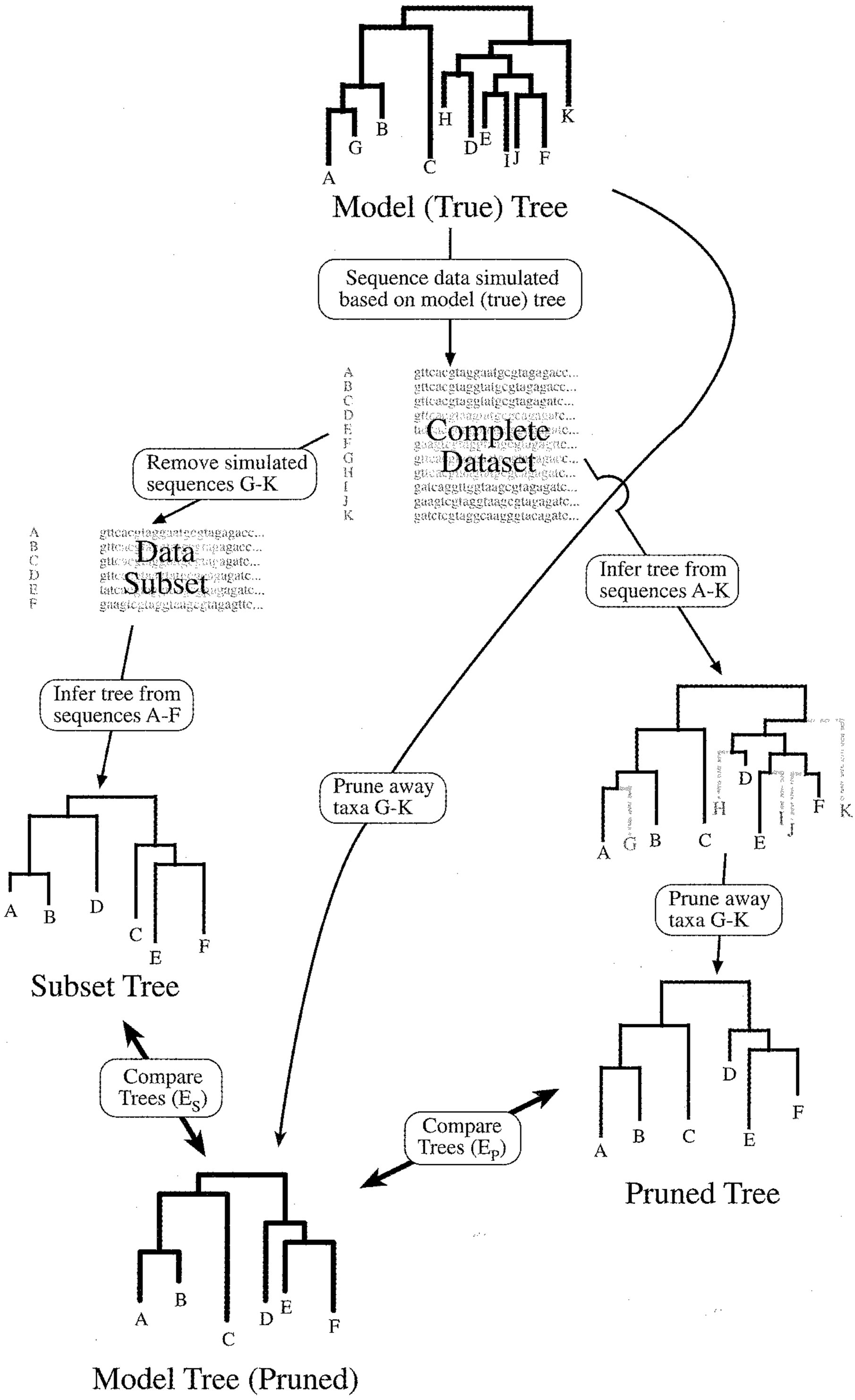

FIGURE 1. The distinction between subset and pruned tree inference from the Rosenberg and Kumar (2001) analysis. Sequence data is simulated based on a model (or "true") tree. For subset trees, the relationships among taxa in an example subset (A–F) are determined by inference of the tree using a corresponding subset of sequence data (A–F). For pruned trees, relationships are determined among all taxa using the complete sequence dataset (A–K). The extra taxa, G–K, are then pruned from the tree (removing the dotted lineages). The two inferred trees are compared to a pruned version of the model tree containing only taxa A–F (bold): E_S measures the error in the subset tree relative to the model tree; E_P measures the error in the pruned tree relative to the model tree. ΔE measures the proportion of error removed (or added, theoretically) by inferring the tree with the full set of sequences ($\Delta E = (E_S - E_P)/E_S$).

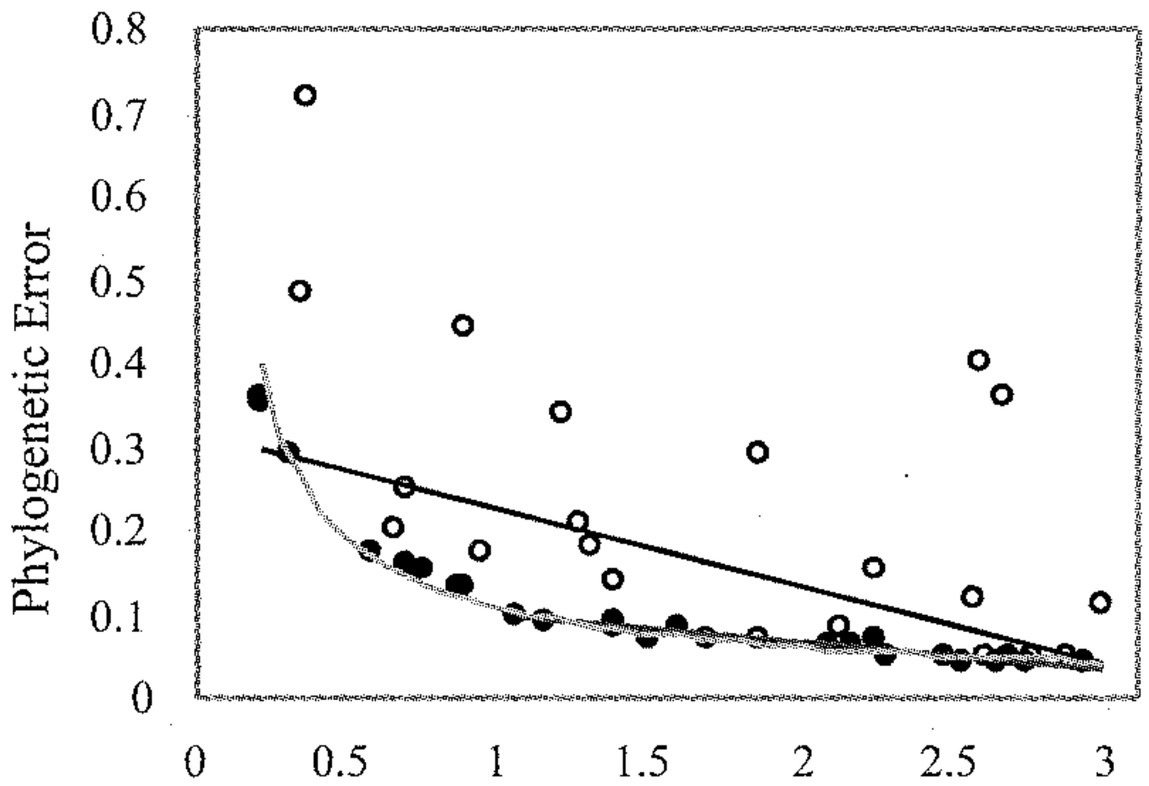

FIGURE 2. The relationship between phylogenetic error and sequence length for the data of Rosenberg and Kumar (2001). The closed circles represent a subset of simulations with evolution rates set between 0.7 and 4.5; open circles represent simulations with rates outside this range. The least squares linear regression lines are plotted for the entire data set (solid line) and for points greater than 1 kb in the subset (dark gray line). A least squares power regression line for the complete subset is also plotted (light gray line), and this line very closely overlaps the linear dark gray line for points greater than 1 kb.

"true" sub-tree. Trees were obtained using the minimum evolution method in PAUP* (Swofford, 2000). Results from other methods referred to in the text (parsimony and maximum likelihood) were not given in the paper and cannot be evaluated here.

Since our goal is to resolve the effects of the various factors in these simulations (sequence length, substitution rate, and taxon addition), we first considered the errors in the complete tree, which are independent of the subsequent sub-tree analyses. Considering the reduction in error created by increasing sequence length (see Fig. 2), there is an apparent clustering of points along a curved line at the bottom of the distribution. Removal of the points with extremely high and low substitution rates (< 0.7 and > 4.5; see below for justification of these cutoff values) completely removes the points that deviate from this line, and we see the classic relationship between sequence length and phylogenetic error (Hillis et al., 1994), showing a sharp decrease in error as data are added to the shorter sequences, followed by a much shallower decrease for sequences above 1 kb (Fig. 2). Linear regression of both the complete data set and sequences greater than 1 kb in the thinned data set showed that for the complete data set, the slope is large (slope = -0.09), but the correlation is weak ($r^2 = 0.31$), while for the thinned data set the magnitude of the slope is much smaller (slope = -0.03), and the correlation much stronger ($r^2 = 0.92$), despite a reduction in the number of points considered from 50 to 22. We also found that a power curve (phylogenetic error = 31.9* sequence length$^{-0.8256}$) gave an extremely good fit ($r^2 = 0.98$) to the entire thinned data set, and closely matches the linear correlation for points greater than 1 kb. This suggests that if 1 kb or more of sequence has already been obtained (for a problem similar to the one modeled), as is standard in most modern phylogenetic analyses, the amount of error reduction with increasing sequence length is considerably smaller than indicated by Rosenberg and Kumar based on the overall average.

The relationship between substitution rate and phylogenetic error is only slightly more complex than that between sequence length and error. To infer the existence of a branch, the substitution rate must be large enough that there is a reasonable probability that a substitution occurred along the branch. Moreover, the substitution rate should not be so large that the ability to infer any substitutions that did occur is obscured by multiple subsequent substitutions at the same site. This results in a distorted U-shaped relationship between substitution rate and error (or inversely, dome-shaped relationship when considering accuracy; e.g., Goldstein and Pollock, 1994; Pollock, 1998), in which error rates initially fall rapidly with increasing substitution rates, and then slowly rise as substitution rates increase further. Looking at average effects, Rosenberg and Kumar attributed a general reduction in error to an increase in substitution rate. The noisy complete data set reduces to a clear U-shaped curve when sequences below 1 kb are removed (see Fig. 3). Although linear regression of the complete data set shows a weak negative correlation between error and substitution rate (slope = -0.037; $r^2 = 0.10$), if one considers only the linear portion of the thinned data curve beyond a substitution rate of 0.7, the correlation becomes stronger and positive (slope = 0.017; $r^2 = 0.43$). For a broad range of substitution rates between 0.7 and 4.5, it is not clear that there is any important effect of substitution rate on phylogenetic error.

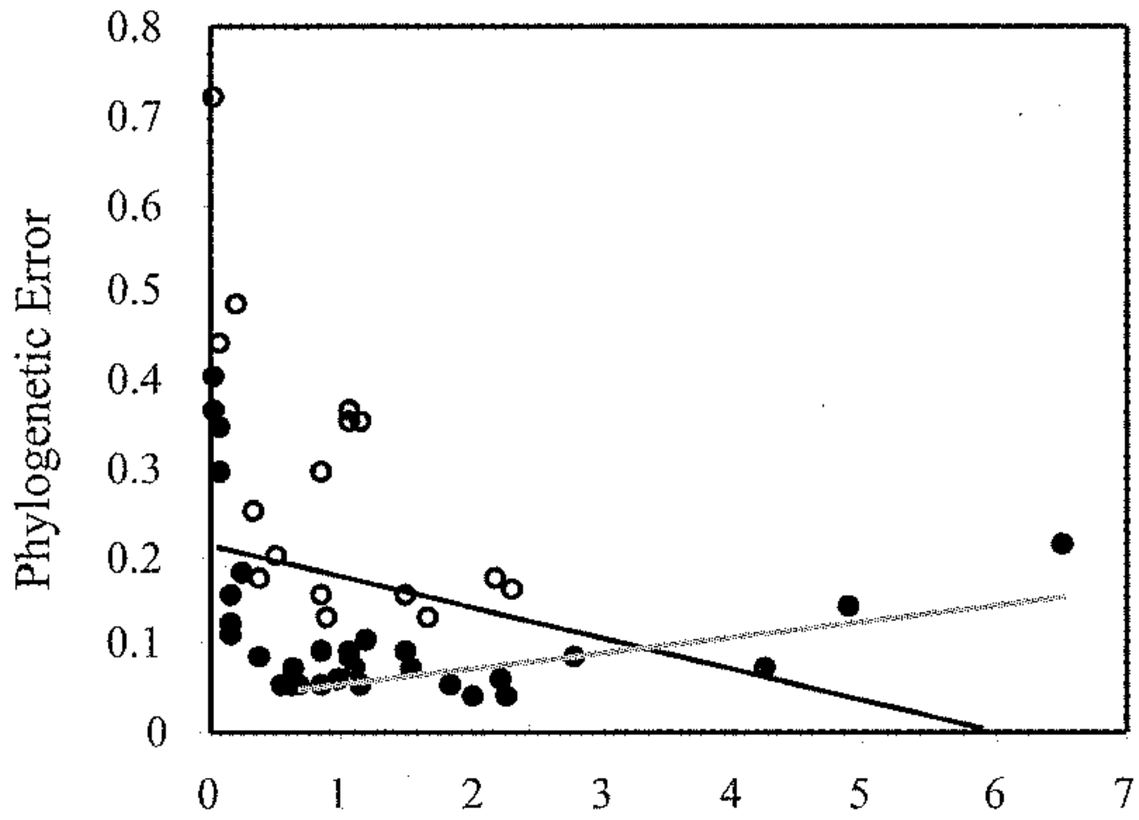

FIGURE 3. The relationship between phylogenetic error and substitution rate for the data of Rosenberg and Kumar (2001). The closed circles represent a subset of simulations with sequence lengths of at least 1 kb; open circles represent simulations with shorter sequence lengths. Linear regression lines are plotted for the entire dataset (solid line) and for points in the subset with rates greater than 0.7 (dark gray line).

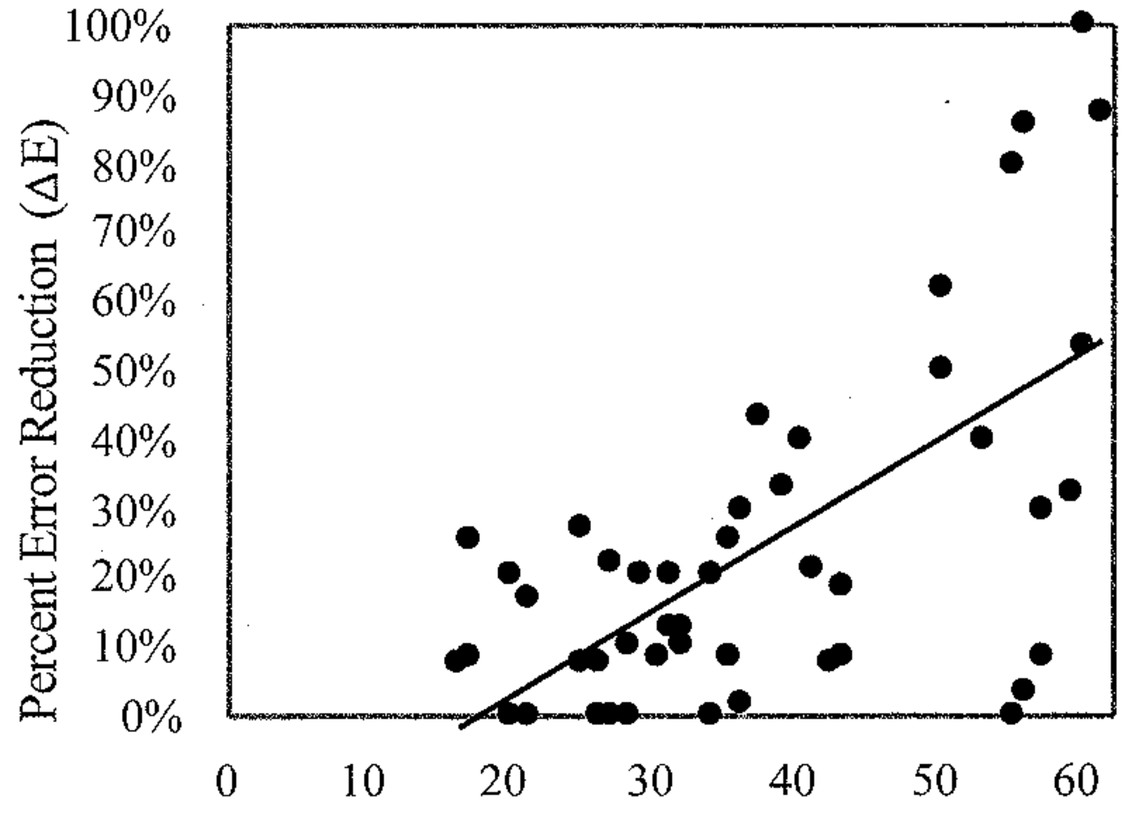

FIGURE 4. Error reduction with increased taxon sampling. The percent reduction in phylogenetic reconstruction error (ΔE) plotted versus the increase in taxon number from the subsample to the full sample in the simulation of Rosenberg and Kumar (2001). The least-squares regression line is shown.

Given these results, it is appropriate to avoid the shorter sequences (less than 1 kb) and the extreme substitution rates (less than 0.7 and greater than 4.5) in the analysis of these simulations. It is reasonable to conclude that if all sites are evolving extremely slowly, and if, for example, only 200 bp have been collected, it is certainly advantageous to collect longer sequences. For the vast majority of datasets, however, these conditions will not hold, and reconstruction properties in the parameter range we are considering will be more applicable.

Although Rosenberg and Kumar stated that the average branch was slightly better resolved in more complex trees (i.e., those with more taxa), the support for this conclusion is extremely weak ($r^2 = 0.003$ between E_S and taxon sample size). We removed short sequences and low rates from the dataset in an attempt to remove the greatest sources of noise from this analysis, and the support for correlation was improved, but not dramatically ($r^2 = 0.203$). There are many possible explanations for noisiness of the error statistics, including a large variance in the difficulty of estimating branches in different trees. The full benefit of increased taxon sampling can be better calculated by taking the difference between the errors in the subsample tree and the pruned tree, $E_S - E_P$. We can then determine the proportion or percentage error removed through increased sampling by taking this as a fraction over E_S, namely $\Delta E = (E_S - E_P)/E_S$. We note that this improvement metric is different than Rosenberg and Kumar's D_E, which is the fraction of branches that are different between the subsample and pruned sample trees. ΔE provides a straightforward measure of the change in error between a subsampled tree and the same tree pruned from the full dataset. Also, rather than considering the number of taxa in the subsample, it is clearer to view the benefits of increased taxon sampling in terms of the increase in taxon number going from the subsample to the full sample. A graph of ΔE versus increase in taxon number (see Fig. 4) indicates a strong and positive correlation between error reduction and increase in taxon number (slope = 1.2% per taxon; $r^2 = 0.41$). We note that one point in Kumar and Rosenberg's data had $E_S = 0$ (also, $E_P = 0$ for this point). Since ΔE is undefined for this point, it was excluded from all ΔE analyses.

In all of the simulation conditions examined by Rosenberg and Kumar, $\Delta E \geq 0$, and the reduction in error ranged from 0 to 100% (Fig. 4). If increased taxon sampling on average has no effect on phylogenetic accuracy, we would expect the average ΔE to be 0, and we would expect as many negative values as positive values for ΔE. The fact that increased taxon sampling never reduced (and usually greatly increased) phylogenetic accuracy under the conditions examined by

TABLE 1. Effects test: Multiple regression analysis of number of taxa, length of sequence, and substitution rate as predictors of ΔE.

Source	DF	Sum of squares	F ratio	Probability
Number of taxa	1	13097.0	35.5	<0.0001
Length of sequence	1	1658.7	4.49	0.0397
Substitution rate	1	1982.6	5.37	0.0251

Rosenberg and Kumar is strong evidence for the benefits of increased taxon sampling. Such a result is to be hoped for, but is not necessarily certain, with any robust statistical method.

One problem with comparing the percentage of error removed due to taxon addition for the full data set is that the simulations are potentially confounded by variation in the number of sites and rate of evolution. A multiple regression analysis of all three variables (Table 1) indicates that taxon sample size is by far the strongest predictor of ΔE ($P < 0.0001$). The independent contributions of the other two variables are significant, however, and our earlier analysis suggests that inclusion of short sequences and low mutation rates contributes the most noise. A graph of ΔE versus increase in taxon number for longer sequences (> 500 bp, the stated lower limit in Rosenberg and Kumar's materials and methods) and rates greater than 0.7 (see Fig. 5) shows a much clearer and stronger positive correlation between error reduction and increase in taxon number (slope = 1.8% per taxon; $r^2 = 0.76$).

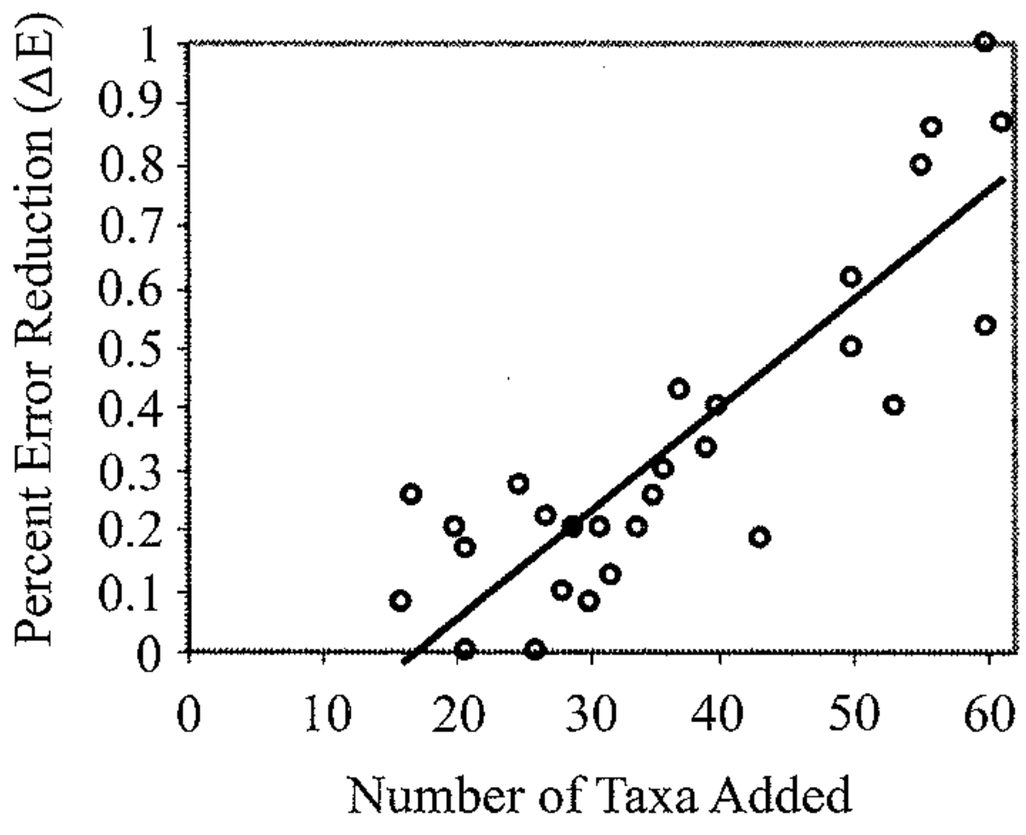

FIGURE 5. Error reduction with increased taxon sampling. This is the same plot shown in Figure 4, except that data points from simulations in which sequences were less than 500 bp and/or substitution rates were less than 0.7 have been excluded.

A simple means of comparing the effects of increasing taxon number versus sequence length is to consider the effect of doubling or tripling the total amount of sequence. If the sequence length is held constant at 1 kb, and the number of sequences obtained is doubled by moving from 33 to 66 taxa, the expected error reduction for branches in the original 33-taxon tree will be 0.03, while if the number of sequences is tripled by moving from 22 to 66 taxa, the expected error reduction in the 22-taxon tree will be 0.06. By comparison, if the number of taxa is held constant at 22 and the amount of sequence is doubled or tripled by increasing the sequence length to 2 kb or 3 kb, the expected amount of error reduction also is 0.03 and 0.06, respectively. In other words, under the conditions of Rosenberg and Kumar's simulations, error reduction can be achieved equally well by taxon addition or by increasing sequencing length.

DISCUSSION

Our main conclusion is that for most realistic situations, phylogenetic reconstruction will be negligibly affected by substitution rate, and that taxon addition will have an effect at least equal to increasing sequence length. Substitution rate will primarily be an important factor in phylogenetic analyses if very slow or very fast evolving sequences are selected—a point well understood by most practicing systematists. Assuming sequences that evolve at appropriate rates of evolution have been selected for analysis, systematists should focus on both increasing numbers of taxa as well as increasing sequence length (or other phylogenetically informative characters) to increase the accuracy of their phylogenetic estimate. This contradicts Rosenberg and Kumar's interpretation, as they attributed a large effect to increasing sequence length, a moderate effect to substitution rate, and a trivial effect (only one-tenth the magnitude of the effect of sequence length) to taxon addition. By removing noise and separating out the individual effects of rate, sequence length, and taxon addition, it can be determined from Rosenberg and Kumar's data that the benefits of doubling or tripling the sequence length are approximately equal to the benefits of doubling or tripling the number of taxa while holding sequence length constant. This result is likely somewhat

dependent on the particulars of the Murphy et al. (2001) tree used in the Rosenberg and Kumar (2001) study, but we have no reason to believe that this tree is not representative of the kinds of trees that are commonly examined in phylogenetic studies.

Consideration of other factors leads to the conclusion that taxon addition will provide benefits above and beyond those that were evaluated in the present simulations. For instance, taxa were subsampled randomly in these simulations. It is generally believed that adding taxa for the purpose of breaking up long branches (rather than adding taxa haphazardly) improves accuracy (e.g., Hillis, 1998, but see Poe and Swofford, 1999). Targeted taxon addition is possible with real data, as Goldman (Goldman, 1998; Massingham and Goldman, 2000) developed a methodology based on information theory that identifies branches that would benefit most from bisection under simple models of evolution. With approximately 4,000 mammals to choose from, there will be considerable room for intelligent direction of taxon addition in investigations of higher-level mammalian phylogeny. In general, historical sampling of mammalian taxa has had more to do with an anthrocentric viewpoint and an interest in sequencing representatives of the more divergent groups. It is not clear that either of those sampling priorities approximates the optimal sampling design.

Another concern is that Rosenberg and Kumar simulated their data under a Jukes-Cantor model with no among-site rate variation, and analyzed them using the minimum evolution criterion. Real data will require more complicated models and stand to benefit from analytical approaches that better utilize model information, such as maximum likelihood (ML) or posterior probability (Bayesian) approaches. Although ML has been proven to be consistent given the correct model and unlimited data (Rogers, 1997), optimization of the model and its parameter estimates is an important aspect of maximizing the accuracy of estimated trees (e.g., see Cunningham et al., 1998; Posada and Crandall, 2001). Both of these tasks are better served by taxon addition than by increasing sequence lengths for a fixed taxon sample (Pollock and Bruno, 2000). Pollock and Bruno (2000) also showed that when the model varies among sites, a dramatic increase in accuracy can be achieved when the rate at individual sites can be determined. This increase in accuracy can be achieved only by adding taxa, not by increasing sequence length.

Our results provide good evidence in favor of adding taxa (when feasible) to difficult phylogenetic problems as a means of reducing overall phylogenetic error. Further support of this conclusion is provided by additional simulations using the Rosenberg and Kumar model tree by Zwickl and Hillis (2002). Because there are a number of additional benefits associated with taxon addition, our results and conclusions are encouraging for the phylogenetic analysis of large datasets. A directed strategy of adding taxa to a phylogenetic analysis will often be one of the most profitable uses of time and resources.

ACKNOWLEDGMENTS

We thank Michael Hellberg and Mohamed Noor for constructive comments on the manuscript. D.D.P. was supported by a Research Competitiveness Subprogram grant LEQSF(2001-04)-RD-A-08 from the State of Louisiana, and by the State of Louisiana's Millennium Research Program: Biological Computation and Visualization Center. D.J.Z. was supported by an NSF IGERT fellowship in Computational Phylogenetics and Applications to Biology (DGE-0114387). J.A.M. was supported by a National Science Foundation grant (DEB-010855), and D.M.H. was supported by a National Science Foundation ITR grant (EIA-0121680).

REFERENCES

BRUNO, W. J., AND A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564–566.

CUNNINGHAM, C. W., H. ZHU, AND D. M. HILLIS. 1998. Best-fit maximum likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52:978–987.

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

GOLDMAN, N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. London B Biol. Sci. 265:1779–1786.

GOLDSTEIN, D. B., AND D. D. POLLOCK. 1994. Least squares estimation of molecular distance—Noise abatement in phylogenetic reconstruction. Theor. Popul. Biol. 45:219–226.

GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problems? 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 383:130–131.

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.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

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.

MASSINGHAM, T., AND N. GOLDMAN. 2000. EDIBLE: Experimental design and information calculations in phylogenetics. Bioinformatics 16:294–295.

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.

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

POE, S., AND D. L. SWOFFORD. 1999. Taxon sampling revisited. Nature 398:299–300.

POLLOCK, D. D. 1998. Increased accuracy in analytical molecular distance estimation. Theor. Popul. Biol. 54:78–90.

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., J. A. EISEN, N. A. DOGGETT, AND M. P. CUMMINGS. 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. Mol. Biol. Evol. 17:1776–1788.

POSADA, D., AND K. A. CRANDALL. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

RANNALA, B., J. P. HUELSENBECK, Z. YANG, AND R. NIELSEN. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

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

ROGERS, J. S. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Syst. Biol. 46:354–357.

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.

SOLTIS, P. S., D. E. SOLTIS, P. G. WOLF, D. L. NICKRENT, S. M. CHAW, AND R. L. CHAPMAN. 1999. The phylogeny of land plants inferred from 18S rDNA sequences: Pushing the limits of rDNA signal? Mol. Biol. Evol. 16:1774–1784.

SWOFFORD, D. L. 2000. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer, Sunderland, Massachusetts.

YANG, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

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

First submitted 27 December 2001; reviews returned 5 April 2002; final acceptance 2 May 2002
Associate Editor: Keith Crandall