Syst. Biol. 47(4):702–710,1998

Taxon Sampling and the Accuracy of Large Phylogenies

BRUCE RANNALA,[1] JOHN P. HUELSENBECK,[2] ZIHENG YANG,[3] AND RASMUS NIELSEN[4]

[1]*Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245, USA; E-mail: rannala@life.bio.sunysb.edu*

[2]*Department of Biology, University of Rochester, Rochester, NY 14627-0211, USA*

[3]*Department of Biology, Galton Laboratory, University College London, 4 Stephenson Way, London NW1 2HE, England*

[4]*Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA*

With the advent of automated methods for rapid sequencing of DNA, and the availability of powerful microcomputers, many more attempts are being made to reconstruct large phylogenetic trees that may include hundreds of sequences and thousands of sites (Vigilant et al., 1991; Chase et al., 1993; Krings et al., 1997). This technological revolution has forced systematists to focus renewed attention on issues relating to the effects of sampling on the accuracy of reconstructed phylogenies. One avenue of research investigates the effect that character sampling has on phylogenetic estimation. That is, the number of taxa in the analysis is held constant but different samples of characters are drawn to investigate the accuracy of phylogenetic methods for samples comprising different numbers of sites and different genomic regions (Graybeal, 1994; Cummings et al., 1995). Another avenue of research has investigated the effect of taxon sampling on phylogenetic accuracy. For example, Hendy and Penny (1989) examined the consistency of the maximum parsimony (MP) method of inferring phylogeny for cases in which the molecular clock assumption (that substitution rates do not vary among lineages) is satisfied. They studied a simple (Poisson process) model of substitution, with only two possible states for each character, and focused on five and six taxon trees. They found that the longest branches were attracted to one another in the MP tree; the MP method can therefore be inconsistent (i.e., the estimated phylogeny will converge to an incorrect phylogeny as the number of independent characters in the analysis is increased), even in cases where the rates of substitution are equal among lineages and arbitrarily low. Hendy and Penny suggested that judicious addition of taxa can break up long branches and help the MP method to become consistent. Kim (1996) tested this prediction, using a combination of analytic theory and computer simulation, and argued that the problem of inconsistency becomes worse for the MP method as the number of taxa in the analysis increases. However, the manner in which the taxa were added to the analysis was somewhat unrealistic; instead of adding taxa within a monophyletic group, as most systematists attempt to do, Kim (1996) increased the age of the root while adding more taxa.

Most recently, Hillis (1996) has studied the effect of taxon sampling on phylogenetic accuracy more directly by attempting to evaluate the accuracy of a reconstructed phylogeny for a real taxonomic group, the angiosperms, which includes large numbers of taxa. The phylogeny of 228 species of angiosperms was first inferred from complete 18S ribosomal RNA genes by use of the MP method. Artificial data sets were then generated by computer simulation, based on the estimated phylogeny and a model of nucleotide substitution, after which the accuracy of phylogenies estimated from the artificial data sets, using either MP or neighbor-joining (NJ) methods, was determined. A remarkable result was that both procedures appear able to accurately reconstruct the phylogeny for this large group of taxa, using DNA sequences of only a few thousand sites. This is in sharp contrast with earlier simulation results, and empirical studies, which have suggested that much larger sequences may not provide sufficient information to accurately estimate phylogeny for as few as four taxa (Hillis et al., 1994).

In this note, we explore two aspects of the taxon sampling problem. First, we present a more general explanation for the increase in phylogenetic accuracy with an increase in the number of taxa sampled. We do this by considering the effect that taxon sampling has on the distribution of speciation times arising under a stochastic model of cladogenesis (the birth–death process; Kendall, 1948). Second, we consider several potentially misleading results that may arise in simulation studies aimed at evaluating the accuracy of phylogenies inferred for large numbers of taxa. We make several recommendations about how simulation studies should be conducted to reduce these potential biases.

Measuring Phylogenetic Accuracy By Using a Model of Cladogenesis

To study the effect on phylogenetic accuracy of increasing the number of taxa, we generated phylogenetic trees by computer simulation, using a birth–death process with taxon sampling to model cladogenesis (Nee et al., 1994; Yang and Rannala, 1997). The birth–death process allows both speciation and extinction events to occur (with rates λ and μ, respectively). The rates of speciation and extinction are assumed to be constant among lineages and over time, but this second assumption can be easily relaxed. Taxon sampling is modeled as a mass extinction event occurring at the present time; each extant taxon is sampled (i.e., survives the event) with probability ρ (Nee et al., 1994). Rannala (1997) noted that the joint probability distribution of the speciation times for s extant species generated by a birth–death process is equivalent to that of the order statistics (i.e., the rank-ordered observations) of $s - 1$ independent and identically distributed random variables, the common probability density of which may be determined analytically. Simulated speciation times for a sample of extant lineages generated by a birth–death process with taxon sampling can then be easily generated by simulating s–1 independent random variables from the kernel density and arranging them in rank order. The branching relationships among the sampled taxa are also easily simulated since all possible labeled trees are equally probable for a birth–death process. Details of the procedures we used to simulate phylogenies under the birth–death process may be found in Yang and Rannala (1997).

If the substitution rates are constant among lineages, then data sets generated under the birth–death process will satisfy a molecular clock. The rate of substitution in our analyses was assumed to be constant across lineages and was parameterized as the tree height (m), which is the expected number of substitutions per site for a single lineage that persists from the root of the tree to the tip. In our simulations, we used $m = 0.15$, which is approximately the value used by Hillis (1996). For each random tree (a single realization of the birth–death process), we simulated sequence data under the K80 + Γ model (Kimura 1980; Yang, 1993) of nucleotide substitution. This model of substitution allows for a bias in the rate of transitions (κ = transition rate/transversion rate; $\kappa = 2$ in our simulations) and a substitution rate that varies across sites according to a gamma distribution (with shape parameter α; $\alpha = 0.5$ in our simulations, so the variance of substitution rates across sites was $\mathrm{Var}[r] = 2$). This model-based approach allowed us to separately examine the effects of taxon sampling and sequence length on the accuracy of inferred phylogenies.

The fraction of the extant taxa that are sampled (sharing a most recent common ancestor, MRCA) is known to have an important effect on the tree shape (Nee et al., 1994; Yang and Rannala, 1997), such that the average length of the terminal branches is increased with a decrease in this fraction. This result is illustrated in Figures 1 and 2. Figure 1a shows 10 phylogenetic trees typical of those generated from a birth–death process when all of the extant taxa of a monophyletic group ($s = 20$, $\rho = 1.0$) are sampled ($\lambda = 6.09$, $\mu = 3.04$, $\mathrm{E}[s] = 20$). $\mathrm{E}[s]$ denotes the expected number of species generated under a birth–death process with sampling for the given parameter values (i.e., the average number of taxa observed in a sample from this process over many realizations). Fig-

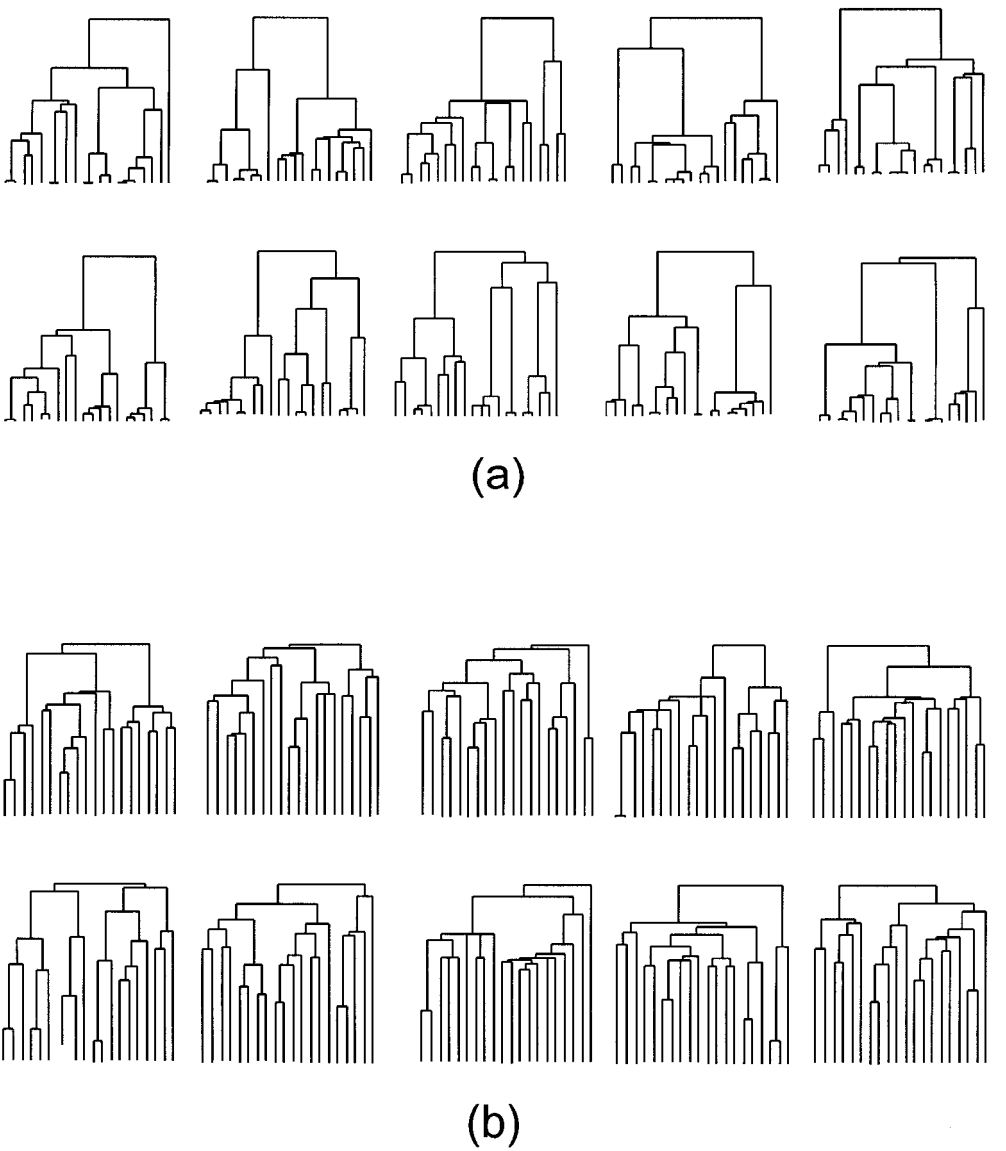

FIGURE 1. Twenty simulated realizations of a birth–death process with sampling used to model cladogenesis. The speciation and extinction rates were chosen so that the expected number of sampled taxa was 20. The ratio of the speciation rate (λ) to the extinction rate (μ) was arbitrarily set to be $\lambda/\mu = 2$. (a) Phylogenetic trees generated in 10 realizations of a birth–death process with complete taxon sampling ($\rho = 1.0$). (b) Phylogenetic trees generated in 10 realizations of a birth–death process with 0.1% taxon sampling ($\rho = 0.001$).

ure 1b shows 10 phylogenetic trees typical of those generated from a birth–death process when only 1 in 1,000 ($s = 20{,}000$, $\rho = 0.001$) of the extant taxa of a monophyletic group is sampled ($\lambda = 19.70$, $\mu = 9.85$, $E[s] = 20$). For both sampling regimes, values of λ and μ were chosen so that the expected number of sampled taxa was $E[s] = 20$ and the arbitrary constraint $\lambda/\mu = 2$ was satisfied. Furthermore, the branch lengths were generated under the condition that the root of the tree was maintained when sampling additional

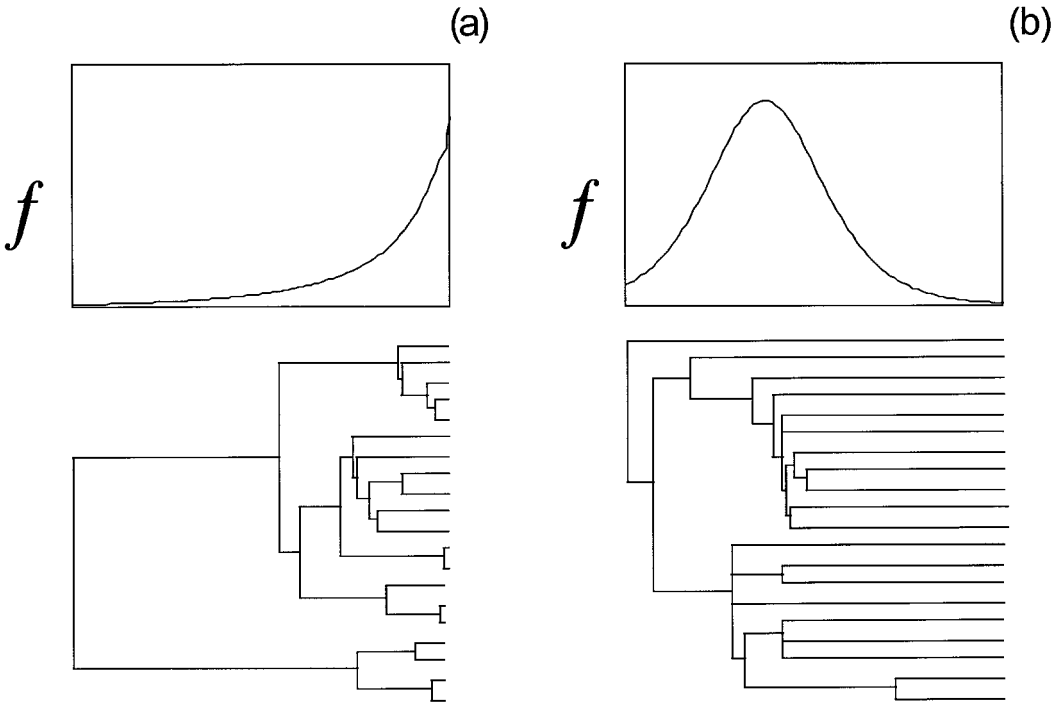

FIGURE 2. Increased taxon sampling decreases (on average) the times at which speciation events occur in a phylogenetic tree. (a) The distribution of speciation times for $s = 20$ taxa with complete taxon sampling ($\rho = 1.0$). (b) The distribution of speciation times for $s = 20$ taxa with 0.1% taxon sampling ($\rho = 0.001$). The speciation and extinction rates used in constructing these distributions were identical to those used in generating the simulated trees shown in Figure 1.

taxa (i.e., by conditioning on the age of an MRCA). This is roughly equivalent to sampling additional taxa from a well-defined group sharing an MRCA, as opposed to sampling additional taxa that do not share this MRCA, which is the implicit outcome of Kim's (1996) simulation methods.

The kernel densities from which the speciation times were generated are shown in Figures 2a and 2b for the two sets of conditions of Figures 1a and 1b, respectively. The density is skewed in favor of short terminal branches when the complete set of extant taxa are sampled (Fig. 2a) and becomes more symmetrical, with much longer terminal branches, on average, as the sampling fraction becomes small (Fig. 2b).

The fraction of taxa sampled has an important influence on the accuracy of phylogenetic trees inferred using the MP method (Fig. 3). Figure 3 shows the accuracy of MP implemented with the optimization of Fitch (1971). Accuracy was measured as the average proportion of correctly resolved taxon bipartitions on the (completely bifurcating) estimated tree. Estimated trees were obtained by using the program PAUP* (provided by David L. Swofford), implemented with the stepwise-addition sequence; no branch swapping was done on starting trees. Trees were generated by using the birth–death process with taxon sampling to model cladogenesis with λ and μ chosen to satisfy $\lambda/\mu = 2$ and E[s] equal to either 20 or 200. Each point in the graph is the average for 100 simulated data sets. Because we consider only a small subset of the total number of extant taxa, the terminal branches are extended and the phylogeny becomes more difficult to reconstruct. As is well known, long terminal branches may cause a phylogenetic tree to be estimated incorrectly by

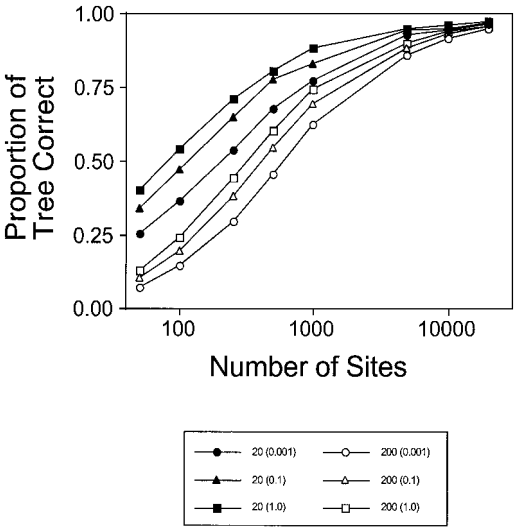

FIGURE 3. Accuracy of maximum parsimony in estimating phylogeny by using of DNA sequences simulated on trees generated under a birth–death process with taxon sampling. Both the proportion of taxa sampled and the absolute number of taxa sampled have strong effects on the accuracy of estimates of phylogeny. Accuracy was measured as the average proportion of correctly resolved taxon bipartitions on the (completely bifurcating) estimated tree. The taxon-sampling fraction (100%, 10%, or 0.1%) was varied as well as the total number of taxa included in the analysis (20 or 200). Speciation and extinction rates were chosen so that the expected number of sampled taxa was either 20 or 200, and $\lambda/\mu = 2$. DNA sequences were simulated under the K80 + Γ model of DNA substitution with $\kappa = 2.0$ and $\alpha = 0.5$ (see text). The tree height (expected number of substitutions from the base of the tree to the tip for a single lineage) was set to be 0.15. Each point on the graph is the average proportion of correct bipartitions for 100 simulated data sets.

the MP method (Felsenstein, 1978), even in the absence of rate variation among lineages (Hendy and Penny, 1989).

The results of our simulations suggest that conditions exist under which phylogeny will be difficult to infer accurately by using the MP method, whether the total number of taxa is large or not. For example, our simulations suggest that a phylogeny for 20 taxa that represent 10% of those descended from an MRCA will be more accurately estimated (on average) than a phylogeny for 200 taxa that also represent 10% of those descended from an MRCA (Fig. 3). Given no changes in the fraction of taxa sampled, the age of the MRCA, and the expected number of sampled sequences, a phylogeny that contains fewer taxa can be more accurately estimated (on average) than one with more taxa. For a given number of taxa, the accuracy of the inferred phylogeny is increased (on average) if the taxa represent a more complete sample of the extant taxa (i.e., a larger fraction of the species descended from a MRCA). Simply including more taxa will not increase the accuracy of the inferred phylogeny (on average) if these additional taxa share a more distant ancestor.

BIAS IN SIMULATION STUDIES OF PHYLOGENETIC ACCURACY

In the course of performing the simulations described in this study, several potential sources of bias became apparent. Although these biases were present in simulation studies that used small numbers of taxa, they appeared to be more severe when large numbers of taxa were considered. The two most important sources of bias were the choice of the model tree on which to simulate data and the choice of a method for measuring the accuracy of an estimated phylogeny. We discuss the methods we chose to apply in our study here and contrast them with the methods used by others in previous studies, in hopes that future studies of phylogenetic accuracy (especially those involving large numbers of taxa) can minimize the biases introduced by the simulation methodology. Of particular concern are biases that favor one method of phylogenetic inference over another. However, most of these can be readily avoided by following our recommendations. We also encourage others to explore these questions in hopes that better solutions might exist.

The first source of potential bias that arises in studies of phylogenetic accuracy based on computer simulations involves the choice of a model tree on which to simulate data. Many phylogenetic methods are biased in favor of certain tree shapes (Huelsenbeck and Kirkpatrick, 1996). If the phylogeny is initially estimated by use of a particular method, and simulations are then performed with this phylogeny, the method that

was used to infer the phylogeny in the first place might be expected to perform better under simulation, since the reconstructed tree is of a type that is easily recovered by that method. This possibility was examined by reconstructing the initial (model) tree for the 228 angiosperm species analyzed by Hillis (1996) from the 18S rRNA genes by another method, the unweighted pair–group method of arithmetic averages (UPGMA; Sokal and Michener, 1958), and then simulating data on that tree and evaluating the performance of MP. The accuracy of the MP method decreased when UPGMA was used to estimate the initial phylogeny (Fig. 4). Although the difference between the two curves showing the MP performance appears slight, we note that roughly twice the amount of sequence data is needed to achieve an accuracy of 95% (i.e., 95% of taxon bipartitions are correct in the inferred tree, on average) when the model tree is determined by UPGMA rather than by MP. In this study, attempting to avoid the bias that may be introduced by choosing a specific tree shape, we generated phylogenies by using a stochastic model of the process of cladogenesis. Phylogenies of many different shapes are generated under the model of cladogenesis and the results are therefore averaged over a sample of the many possible realizations of the process. Ideally, one would also compare the performance of the different phylogenetic methods when trees are generated under several different models of cladogenesis, although we do not pursue that approach here.

The choice of a measure of the accuracy of inferred phylogenies can also influence perceptions of the performance of a phylogenetic method (Hillis, 1995). The general problem is how best to quantify the similarity that an estimated tree bears to the true tree (i.e., the tree on which sequences were simulated). In most simulation studies, the level of similarity between the true tree and an estimated tree is measured by counting the number of taxon bipartitions in common between them. A taxon bipartition is obtained by removing one branch of a phylogenetic tree, thereby dividing the species into those contained within the groups on

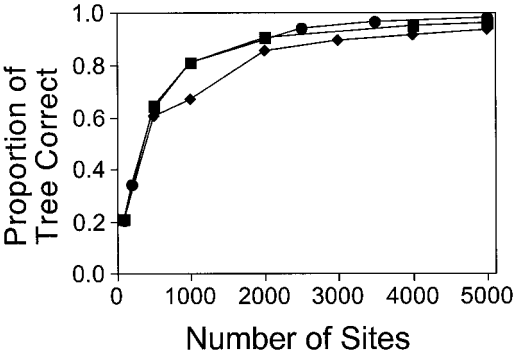

FIGURE 4. The accuracy of maximum parsimony (circles) and neighbor joining determined by using p distances (squares) with the same model tree (of 228 species of angiosperms inferred from rRNA genes by MP) as used by Hillis (1996). Accuracy was measured as the average proportion of correctly resolved taxon bipartitions on the (completely bifurcating) estimated tree. Note that the accuracy of the two methods is almost identical. The accuracy of maximum parsimony when UPGMA was used to estimate the topology and branch lengths for the model tree of the angiosperms is also plotted (diamonds). The accuracy of parsimony is lower in this case, suggesting that a bias may be introduced by the phylogenetic method used to estimate the initial phylogeny to be used as a model tree for the simulations. Each point on the graph is the average proportion of correct bipartitions for 100 simulated data sets.

each side of the deleted branch. For example, for four species (A, B, C, and D) there are three possible taxon bipartitions that may be obtained by removing an internal branch ({AB},{CD}; {AC},{BD}; {AD},{BC}) and four taxon bipartitions that may be obtained by removing an external (or tip) branch ({A},{BCD}; {B},{ACD}; {C},{ABD}; {D},{ABC}). The overall accuracy of a phylogenetic method is often measured as the number of correct taxon bipartitions found on the estimated tree divided by the total number of taxon bipartitions possible for s taxa (Hillis, 1995). We will refer to this measure as the B distance (d_B) between trees. Other metrics of tree similarity are also useful (see Steel and Penny, 1993), especially when the trees compared are not strictly bifurcating, but will not be considered in this paper.

In our analysis, we found no significant difference between the performance of the

MP and NJ methods in estimating phylogeny on the model tree of Hillis (1996) when we used the average proportion of identical bipartitions between the true tree and the inferred tree (the *B* distance) as our accuracy criterion for both methods. Figure 4 compares the performance of MP and NJ in terms of *B* distances. In contrast, Hillis's (1996) results appear to suggest that MP outperforms NJ for short sequences (see his Fig. 2). The explanation for this discrepancy lies in the treatment of polytomies on the estimated tree. If one possible resolution of a polytomy on the estimated tree is consistent with a taxon bipartition on the true tree, then we would consider the accuracy to be $1/b$, where b is the number of possible resolutions of the polytomy, whereas Hillis (1996) considered the accuracy to be 0.5 (another possible solution is to treat the accuracy as zero for that resolution because the method failed to correctly identify the partition). Hence, for sequences of length zero, MP (or any other method that allows polytomies) would be either $1/b$ or 0.5 correct, depending on which measure of accuracy is used. Because the number of unresolved polytomies typically increases with an increase in s, the discrepancy between the two measures of tree similarity matters most when large numbers of taxa are studied. These different ways of resolving polytomies normally would not be a source of bias except that some methods, such as NJ, will either (1) arbitrarily resolve some branches to have a small, but nonzero, length because of stochastic error in the process of substitution or rounding errors in computer memory, or (2) be represented in computer memory as a strictly bifurcating tree even if some branches are effectively zero in length (the program PAUP* stores MP trees with zero length branches as polytomies but stores NJ trees as binary trees even when branch lengths do not differ significantly from zero). By considering polytomies as half correct, one will tend to favor methods that present unresolved nodes as polytomies rather than arbitrary resolutions, although neither approach contains more information about phylogeny. Our method of averaging over all possible resolutions of a polytomy renders the results from MP and NJ analyses directly comparable, because averaging over many replicate simulations is equivalent to generating an arbitrary resolution for an unresolved node. Figure 5 illustrates the relationships among the various measures of tree similarity, showing how the measure chosen can affect the perceived accuracy of a phylogenetic method.

Our recommendation is that phylogenetic accuracy be measured as the average proportion of correct taxon bipartitions (*B* distances) over all possible resolutions of a polytomy. An alternative approach would be to choose one of the possible resolutions by assigning an equal probability to each; this would again allow a direct comparison

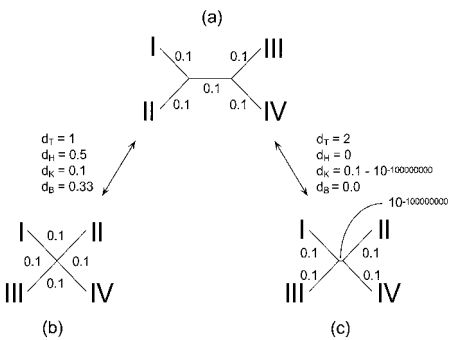

FIGURE 5. Different measures of tree similarity can give different impressions of phylogenetic accuracy. In this case, the "true" tree, shown in (a), is strictly bifurcating, with all branches being 0.1 times the expected substitutions per site in length. The two different estimated trees shown in (b) and (c) are very similar. The estimated tree of (b) is a star tree, whereas the estimated tree of (c) is bifurcating, but with an internal branch that is very short ($10^{-100000000}$). A useful tree similarity metric should assume nearly identical values when comparing each of the trees in (b) and (c) with that in (a) because the "true" length of the internal branch in tree (c) is negligible. The different measures of accuracy are (1) d_T (Robinson and Foulds, 1979), the number of branch contractions and expansions necessary to transform one tree into another; (2) d_H (Hillis, 1996), the number of taxon bipartitions in common between the true and estimated trees (with polytomies treated as half correct); (3) d_B, the number of taxon bipartitions in common between the true and estimated trees (with the average accuracy calculated for all possible resolutions of polytomies); and (4) d_K (a modification of the Kuhner and Felsenstein, 1994), a measure proportional to the amount of branch length that must be contracted and expanded to transform one tree into another.

of the accuracy obtained for MP and NJ because, in the case of nodes with no statistical support, the dichotomy presented by NJ is essentially a random resolution of the node. Another, possibly better, solution would be to consider a metric of tree similarity (such as that used by Kuhner and Felsenstein, 1994) that takes into account similarities in branch lengths between trees as well as similarities in topology. The Kuhner and Felsenstein metric sums the total length of branch that must be expanded or contracted to transform one tree into another. This metric has the advantage that it does not appear to favor any particular phylogenetic method in the way that the polytomies are resolved. We note, however, that if accuracy is measured with the Kuhner and Felsenstein metric, MP is strictly inconsistent (i.e., the metric does not converge to the value expected for complete agreement between trees) since the MP method will tend to underestimate the total expected number of substitutions along any branch.

Discussion

Increased sampling of taxa from within a monophyletic group increases the average accuracy of reconstructed phylogenies because it decreases the average lengths of external (tip) branches. In general, our results here support the finding of Hillis (1996) that large phylogenies may be accurately reconstructed. Hillis's results are important because they strongly contradict the common view that phylogenies for large numbers of species are difficult, or impossible, to resolve accurately. However, although our analysis enabled us to separate the effect of the taxon sampling fraction from that of the total number of taxa sampled, we did not explore several other potential determinants of the accuracy of phylogenetic methods that might be equally important. Additional factors that could affect phylogenetic accuracy include at least the overall rate of nucleotide substitution, differences in substitution rates among lineages, more complex models of DNA substitution, nonindependence of the substitution process among lineages and among sites, and errors in sequence alignments. These factors, and others, can all potentially decrease the accuracy of a phylogenetic method.

By considering models of cladogenesis that incorporate taxon sampling and other processes known to affect the distribution of ancestral speciation times on a phylogenetic tree, we can begin to identify those cases in which phylogenies for large numbers of species may be accurately reconstructed and also those cases in which they may not. The results of the limited simulations done so far, which have been aimed at quantifying the accuracy of phylogenetic trees reconstructed for large numbers of taxa, suggest we may be optimistic that large phylogenies are not as difficult to accurately reconstruct as was once thought.

Acknowledgments

This paper benefited from discussions with M. Slatkin and D. Swofford. This research was supported by a Natural Sciences and Engineering Research Council (NSERC) of Canada postdoctoral fellowship awarded to B.R., a Miller Postdoctoral Fellowship awarded to J.P.H., a grant to R.N. from the Danish Research Council, and a NIH grant GM40282 awarded to M. Slatkin. We thank D. Swofford for providing the program used for many of the simulations. We thank the NOW project at the University of California at Berkeley, and especially E. Brewer, for making available the many SUN Ultrasparc workstations used to perform the simulation analyses.

References

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qui, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandian, and V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. 80:528–580.

Cummings, M. P., S. P. Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12:814–822.

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

Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20:406–416.

GRAYBEAL, A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst. Biol. 43:174–193.

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. 1996. Inferring complex phylogenies. Nature 383:130–131.

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

HUELSENBECK, J. P., AND M. KIRKPATRICK. 1996. Do phylogenetic methods produce trees with biased shapes? Evolution 50:1418–1424.

KENDALL, D. G. 1948. On the generalized birth-and-death process. Ann. Math. Stat. 19:1–15.

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

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

KRINGS, M., A. STONE, R. W. SCHMITZ, H. KRAINITZKI, M. STONEKING, AND S. PAABO. 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30.

KUHNER, M. K., AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

NEE, S., R. M. MAY, AND P. H. HARVEY. 1994. The reconstructed evolutionary process. Phil. Trans. R. Soc. London B 344:305–311.

RANNALA, B. 1997. Gene genealogy in a population of variable size. Heredity 78:417–423.

ROBINSON, D. F., AND L. R. FOULDS. 1979. Comparison of weighted labelled trees. Pages 119–126 *in* Lecture notes in mathematics, (A. Dold and B. Eckmann, Eds.). Volume 748. Springer-Verlag, Berlin.

SOKAL, R. R., AND C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 28:1409–1438.

STEEL, M. A., AND D. PENNY. 1993. Distributions of tree comparison metrics—some new results. Syst. Biol. 42:126–141.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, AND A. C. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

YANG, Z, AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

Received 6 November 1997; accepted 22 February 1998
Associate Editor: D. Cannatella

Syst. Biol. 47(4):710–718, 1998

Interpreting Sister-Group Tests of Key Innovation Hypotheses

ALAN DE QUEIROZ

Department of Environmental, Population and Organismic Biology and University Museum, University of Colorado, Boulder, Colorado 80309-0334, USA; E-mail: dequeiro@stripe.colorado.edu

The idea that a particular trait can increase the diversification rate of a group has a long tradition in evolutionary biology (e.g., Simpson, 1953; Hecht, 1963; Mayr, 1969). Vrba (1980) made an important contribution to testing such "key innovation" hypotheses by noting that one could compare the diversities (numbers of species) of a clade that possesses the trait of interest and of its sister group that lacks the trait. Such a comparison controls for clade age because sister groups are the same age by definition (Mitter et al., 1988), and also reduces the number of possible confounding factors because the two groups share the same evolutionary history up to the point at which they diverged.

Mitter et al. (1988) introduced an approach for statistically testing key innovation hypotheses. Their protocol consists of defining a key trait (or adaptive zone, which could include entrance into a new environment) independently of the recognition of specific groups that possess the trait, and subsequently making multiple diver-