Syst. Biol. 51(4):588–598, 2002
DOI: 10.1080/10635150290102339

Increased Taxon Sampling Greatly Reduces Phylogenetic Error

DERRICK J. ZWICKL AND DAVID M. HILLIS

Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, Texas 78712, USA; E-mail: zwickl@mail.utexas.edu and dhillis@mail.utexas.edu

Abstract.—Several authors have argued recently that extensive taxon sampling has a positive and important effect on the accuracy of phylogenetic estimates. However, other authors have argued that there is little benefit of extensive taxon sampling, and so phylogenetic problems can or should be reduced to a few exemplar taxa as a means of reducing the computational complexity of the phylogenetic analysis. In this paper we examined five aspects of study design that may have led to these different perspectives. First, we considered the measurement of phylogenetic error across a wide range of taxon sample sizes, and conclude that the expected error based on randomly selecting trees (which varies by taxon sample size) must be considered in evaluating error in studies of the effects of taxon sampling. Second, we addressed the scope of the phylogenetic problems defined by different samples of taxa, and argue that phylogenetic scope needs to be considered in evaluating the importance of taxon-sampling strategies. Third, we examined the claim that fast and simple tree searches are as effective as more thorough searches at finding near-optimal trees that minimize error. We show that a more complete search of tree space reduces phylogenetic error, especially as the taxon sample size increases. Fourth, we examined the effects of simple versus complex simulation models on taxonomic sampling studies. Although benefits of taxon sampling are apparent for all models, data generated under more complex models of evolution produce higher overall levels of error and show greater positive effects of increased taxon sampling. Fifth, we asked if different phylogenetic optimality criteria show different effects of taxon sampling. Although we found strong differences in effectiveness of different optimality criteria as a function of taxon sample size, increased taxon sampling improved the results from all the common optimality criteria. Nonetheless, the method that showed the lowest overall performance (minimum evolution) also showed the least improvement from increased taxon sampling. Taking each of these results into account re-enforces the conclusion that increased sampling of taxa is one of the most important ways to increase overall phylogenetic accuracy. [Phylogenetic accuracy; phylogenetic error; taxon sampling.]

In recent years, there has been increased interest in estimating large phylogenetic trees of many taxa. Several factors have contributed to this trend. First, it has become computationally feasible to analyze large data sets of many taxa and many characters (e.g., Soltis et al., 1998). Second, there is intrinsic interest in the phylogeny of large groups of organisms, and even interest in eventually producing phylogenetic estimates for the entire Tree of Life (see Hillis and Holder, 2000). Third, many authors have argued that adequate taxon sampling improves phylogenetic estimation, and in some cases may even make otherwise intractable problems tractable (e.g., Wheeler, 1992; Lecointre et al., 1993; Hillis, 1996, 1998; Poe, 1998; Rannala et al., 1998). Most recent authors have argued that both the number of characters as well as the number of taxa sampled are important determinants of phylogenetic accuracy (Swofford et al., 1996). However, since it is computationally much easier to analyze data sets of few taxa than data sets of many taxa, it is tempting to define and investigate a phylogenetic problem with as few taxa as possible.

A recent paper by Rosenberg and Kumar (2001) suggested that sampling few taxa from a large and diverse group carries almost no penalty in terms of the accuracy of the estimated phylogenetic tree, and that reduced taxon sampling thus is not a problem for phylogenetic analysis. A reanalysis (Pollock et al., 2002) of the Rosenberg and Kumar (2001) data supports the opposite conclusion, and suggests that their study design may not have been optimal to investigate the effects of increased taxon sampling. However, Rosenberg and Kumar's (2001) study did raise several interesting hypotheses about the relationships among phylogenetic methodology, taxonomic sampling, and phylogenetic accuracy, and led us to conduct independent simulation studies to test the effects of various aspects of study design on conclusions about the importance of thorough taxon sampling. In this paper, we address five issues related to taxon sampling and its effects on the accuracy of phylogenetic inference. First, we consider the measurement of phylogenetic error across a wide range of taxon sample sizes, and the degree to which error in randomly selected

trees relates to the issue of taxonomic sampling. Second, we address the scope of the phylogenetic problems addressed by different samples of taxa, and methods that can be used to hold the taxonomic scope of a problem relatively constant across different numbers of sampled taxa. Third, we examine the claim of some recent authors that fast and simple tree searches are as effective as more thorough searches at finding near-optimal trees that minimize error. Fourth, we examine the effects of using simple versus complex simulation models on the results of taxon sampling studies. Fifth, we examine whether or not results of taxon sampling studies are dependent on the use of particular phylogenetic optimality criteria.

As we have noted, our current attention to these issues was raised by Rosenberg and Kumar's (2001) study on taxon sampling and its effects on phylogenetic accuracy. As we incorporated many of the aspects of the study design used by Rosenberg and Kumar to address these issues and use their study to discuss several issues of analysis, we begin with a brief description of their study design before discussing our methods.

Rosenberg and Kumar's Study Design

Rosenberg and Kumar (2001) addressed the effects of partial taxon sampling on the error rate of phylogenetic estimation. They used a model tree of 66 taxa that was based on a published study of eutherian mammals (Murphy et al., 2001; Eizirik et al., 2001). Rosenberg and Kumar then conducted simulations under the Jukes-Cantor model of evolution (henceforth JC; Jukes and Cantor, 1969) and varied number of nucleotides and rate of evolution across 50 sets of conditions (which they termed "genes"). Rosenberg and Kumar (2001) stated that their results were similar when a more complex model of evolution (Hasegawa et al., 1985) was used, but they did not show these results. The number of sites per gene was selected from a uniform distribution of 200–3,000 (not 500–3,000 as described in their text; see their Table 1), and the branch lengths of the model tree were scaled by selecting a scaling factor from a gamma distribution with a shape parameter of 1.

For each of the 50 conditions examined (combinations of nucleotide length and rate of evolution), Rosenberg and Kumar (2001) simulated 100 replicates. For each condition they then randomly selected subsamples of 5–50 taxa from the full set of taxa for analysis, and measured error between the true tree and the estimate (see the section below titled Measurement of Error for further description and discussion of error measurement). They evaluated and compared three measures of error: E_G (the proportion of error between the true tree and the estimated tree from the full sample of 66 taxa), E_S (the proportion of error between the true tree and the estimated tree from the subsample of taxa), and E_P (the proportion of error between the true tree and the subsample of taxa, as pruned from the full analysis). For each simulation, they analyzed the data using minimum evolution (ME), uniformly weighted maximum parsimony (MP), and maximum likelihood (ML) criteria, as well as the neighbor-joining (NJ) heuristic. (They used PAUP* [Swofford, 2000] for all analyses; see Swofford et al. [1996] for a description of the methods.) However, they only presented detailed data for the ME criterion, stating that the results for the other criteria were "quite similar." They used a single-tree heuristic search with nearest-neighbor-interchange branch swapping to estimate the optimal trees for each criterion.

Methods

Data Simulation

Our datasets were simulated on the Rosenberg and Kumar model tree (provided by M. S. Rosenberg; see Appendix I) under a number of common evolutionary models using Seq-Gen, version 1.2.5 (Rambaut and Grassly, 1997). Simulation models included JC (Jukes and Cantor, 1969), HKY(Hasegawa et al., 1985), HKY with continuous gamma-distributed rate heterogeneity (Yang, 1993), and the General Time Reversible Model with continuous gamma-distributed rate heterogeneity and a proportion of invariant sites (Lanave et al., 1984; Tavaré, 1986; Yang, 1993; Swofford et al., 1996). A single dataset was simulated under each model. The particular parameter values we used (see Appendix II) are ones estimated by maximum likelihood on a tree obtained by a parsimony search for two of the genes present in the Murphy et al. (2001) dataset (12S rRNA and *cnr 1*, a protein-coding gene). All simulated datasets were 3,000 bases long, as we wished to avoid confounding the issue of sampling additional

taxa by also varying sequence length (see Pollock et al. [2002] for a discussion of the effects of sequence length on taxon sampling). Thus, we selected the upper bound of the sequence lengths examined by Rosenberg and Kumar (2001) for our simulations.

Subsampling

Subsampling of taxa was performed using a C++ program written by one of us (D.J.Z.). Eleven subsample-sizes were selected ranging from 4 to 60 taxa, from the complete model set of 66 taxa (the entire sample of taxa was also evaluated). For each sample-size of taxa, the procedure was as follows:

1. Randomly select a set of taxa using random-number generation.
2. Determine the diameter (the maximum distance between any two taxa in a tree) of this subset based on the branch-lengths of the model tree and place it into the appropriate diameter "bin." (The rationale for this step is discussed in the section titled Taxon Subsamples, in RESULTS AND DISCUSSION. Briefly, our goal was to examine the effects of taxon subsampling across problems that spanned a similar phylogenetic scale.)
3. Repeat 1 and 2, discarding identical subsets, until all bins contain 100 subsamples.

The diameter-bins ranged from 0.10 (the diameter of the smallest quartet) to 0.45 (the diameter of the entire 66-taxon tree), in increments of 0.05 (thus, the smallest bin contained trees of 0.10–0.15 diameter). As the number of taxa in the subsample increased, the number of bins for which possible subsamples exist necessarily decreases. Thus, quartet subsamples (samples of four taxa) covered seven diameter categories (giving 700 subsamples), whereas subsamples of 60 taxa covered only the largest four diameter bins (giving 400 subsamples). The same subsamples were used in all of our analyses.

Analysis

All subsamples were analyzed using PAUP* 4.0b8 (Swofford, 2000). Except for the maximum likelihood analyses and the experiment examining the effect of search thoroughness (see the section titled *Thoroughness of Searches*), all subsets were subjected to a heuristic search with five random stepwise-addition starting trees followed by tree-bisection-reconnection (TBR) branch swapping (see Swofford et al., 1996 for a description of these methods). Due to computational constraints, the likelihood searches were conducted with a single stepwise-addition starting tree, followed by TBR branch swapping using the JC model of evolution. ME searches were conducted using JC and HKY + Γ distance corrections. With both distances measures, we conducted ME searches allowing negative branch-lengths (set to zero for score calculation), and also with branch-lengths constrained to non-negative values. For the ME searches using HKY + Γ distances, the alpha shape parameter was set to its simulation value (0.399). For all searches, all equally optimal trees were retained.

To assess accuracy of reconstruction, the Robinson and Foulds (1981) symmetric distance measure (henceforth RF distance) was calculated between the optimal tree(s) and the model tree pruned to contain the same taxa. When we found multiple equally optimal solutions, we calculated the average RF distance of all solutions to the true tree. The measurement of phylogenetic error is a point of discussion, and is presented below.

RESULTS AND DISCUSSION

Measurement of Error

The measure of error used by Rosenberg and Kumar is fairly standard, and one of us has used the same measure in previous studies (e.g., Hillis, 1996). This measure of error (E) uses the RF distance between the true tree and the estimate, divided by twice the number of internal branches in the comparison (the maximum possible RF distance, or RF_{max}). For any size tree, E ranges from 0 to 1; a value of 1 indicates that no internal branches are shared in common between the true tree and the estimate, and a value of 0 indicates complete agreement between the two trees. As long as the number of taxa in the analysis remains constant (or large), E provides a reasonable measure that combines error from false negatives (branches absent in the estimate but present in the true tree) and false positives (branches present in the estimate but absent in the true tree). However, in comparing the relative error across trees with different numbers of taxa, the measure E has a major drawback. For trees with only four

taxa, for example, there are only three possible solutions. One of the solutions has an RF distance of 0, and the other two have RF distances of 2. Thus, the expected RF distance (RF_{exp}) between a randomly selected topology of four taxa and a true tree is 1.33, not 2 (the value of RF_{max} for trees of four taxa). The expected distance only approaches the maximum distance asymptotically with increasing numbers of taxa (Penny et al., 1982). If one simply examined the error associated with choosing trees at random across a diversity of sample sizes, E would be lowest for trees with few taxa and would gradually increase with taxon sample size. Therefore, this measure does not provide a uniform comparison for evaluating the improvement of phylogenetic methods across varying numbers of taxa. (Note that this is a different problem related to measurement of error in taxon sampling studies than the one discussed by Pollock et al., 2002).

The problem discussed above can be easily corrected (as previously noted by Poe, 1998) by standardizing the observed RF distance between the true tree and an estimated tree by the expected RF distance (RF_{exp}) to a randomly selected tree, rather than by RF_{max}. We calculated RF_{exp} exactly for trees up to 10 taxa, and estimated RF_{exp} for larger trees. We used PAUP* 4.0b8 to generate either all possible trees for a particular number of taxa, or a sample of at least 3 million trees for more than 10 taxa. The RF distance was then calculated between each of the random trees in the set to an arbitrary reference tree. The expectation of this distribution was calculated by multiplying the RF distances by the number of trees from the set having that distance to the reference tree, summing these values, and dividing by the total number of trees. We then define adjusted error (E_{adj}) as the RF distance between the true tree and the estimate, divided by RF_{exp}, and absolute error (E) as the RF distance between the true tree and the estimate, divided by RF_{max}. The two measures converge with increasing taxon sample sizes (see Fig. 1). The expected adjusted error for randomly selected trees is 1.0 for all taxon sample sizes; thus, E_{adj} can be used to compare the improvement of a given inference method across a range of taxon sample sizes. Note that E_{adj} can take on values greater than one if the actual distance is greater than would be expected from selecting a random tree of that size, as might occur if a particular method were positively misleading (e.g., Felsenstein, 1978).

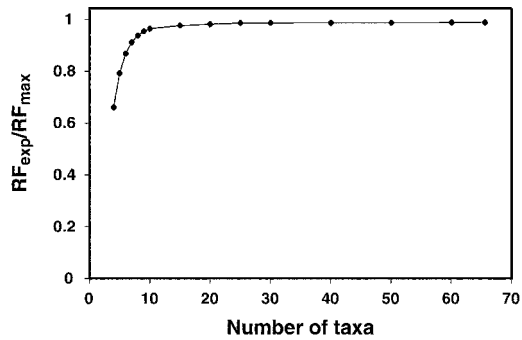

FIGURE 1. Ratio of RF_{exp} to RF_{max} as a function of number of taxa.

Although the effect of evaluating E_{adj} instead of E matters only at small taxon sample sizes, it removes an artifact that otherwise clouds the relationship between phylogenetic error and number of taxa in the analysis. Figure 2 shows a comparison of these two measures (E and E_{adj}) as a function of taxon sample size for the model tree of Rosenberg and Kumar. Although there is a strong decrease in both absolute and adjusted error with increasing sample size, there is an initial increase in absolute error as the samples increase from 4–10 taxa. However, this effect is often eliminated when adjusted error is measured, as in this example. It seems clear that the apparent lower error for the smallest sample sizes is sometimes simply a function of randomly selected trees having a higher

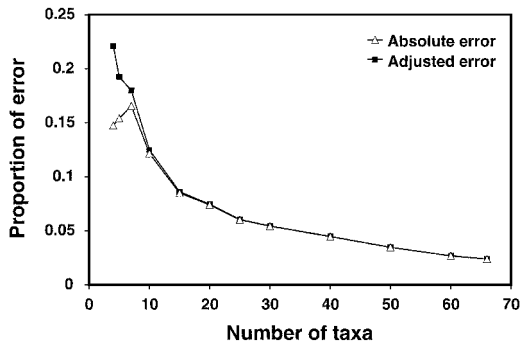

FIGURE 2. Absolute and adjusted error as a function of taxon sample-size for a dataset generated with the HKY model of evolution on the Rosenberg and Kumar (2001) model tree and analyzed under uniformly weighted parsimony. Data points represent the average error over all subsample diameters.

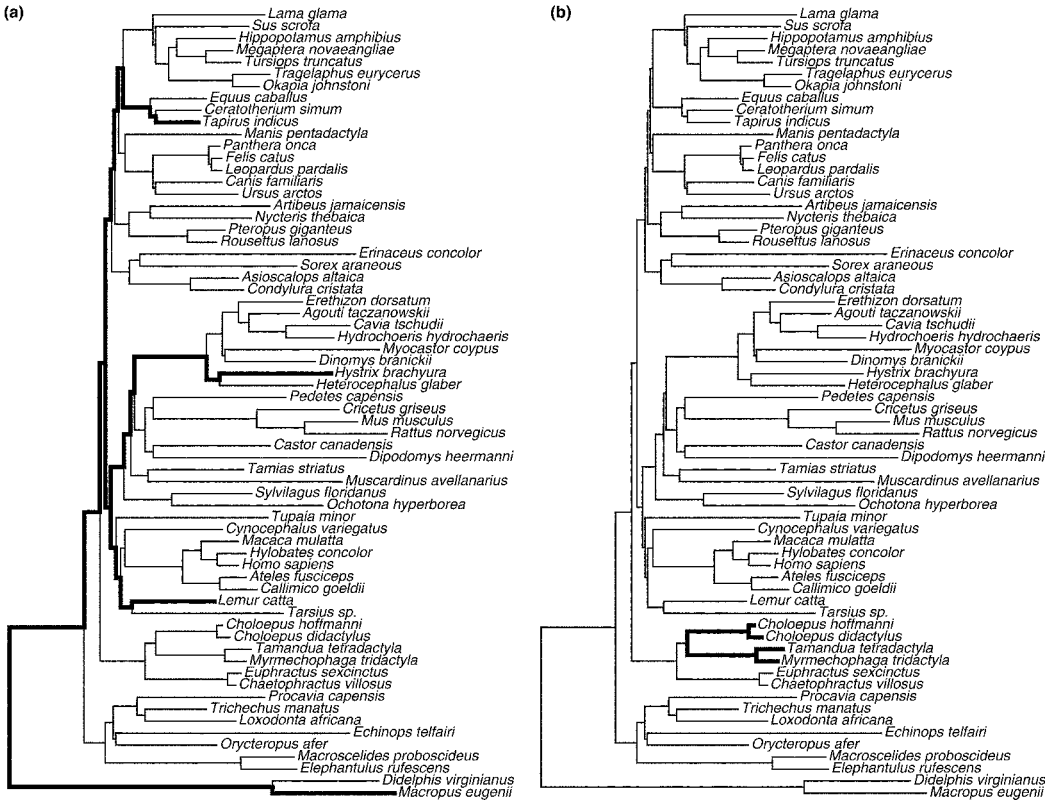

FIGURE 3. Two quartets of taxa sampled from the model tree of Rosenberg and Kumar (2001), shown as heavy lines within the entire model tree. (a) A quartet of large diameter, containing taxa that span the same depth of phylogenetic time as the full data set. (b) A quartet of small diameter, containing taxa that represent complete taxon sampling for a subtree.

probability of matching a limited number of possible internal branches. Thus, for the remainder of our analyses, we present results using E_{adj} rather than E.

Taxon Subsamples

One issue that should be taken into account in studies of taxon sampling relates to the fact that all subsamples from a larger set of taxa do not represent problems of equivalent phylogenetic scale (see Fig. 3). A quartet of taxa may represent widely scattered taxa in a larger tree (Fig. 3a), or it may represent a small, completely sampled subtree from a larger tree (Fig. 3b). One would expect that a quartet of taxa such as that shown in Figure 3b would not present a difficult phylogenetic problem (nor a problem comparable to that posed by all the taxa), whereas the quartet of taxa shown in Figure 3a would present a much harder (but more relevant) problem. Rosenberg and Kumar (2001) selected random subsamples of taxa from the tree shown in Figure 3 with no effort to hold the depth of the phylogeny or the diameter of the sampled tree (the maximum distance between any two taxa) constant. The diameter of the quartet in Figure 3a is approximately equal to that of the full data set, but the diameter is much smaller for the quartet in Figure 3b. Under the Rosenberg and Kumar study design, both sets of taxa would represent equally "incomplete" taxon sampling, and each would be compared against the respective pruned subtrees.

The difficulty of phylogenetic analysis is known to increase with increasing diameter of the underlying tree, especially for trees with small numbers of taxa (e.g., see Huelsenbeck and Hillis, 1993). This point is illustrated for quartets of taxa randomly selected from the model tree of Rosenberg and Kumar in Figure 4. For any given method of analysis, as the diameter of the tree increases, the error also generally increases.

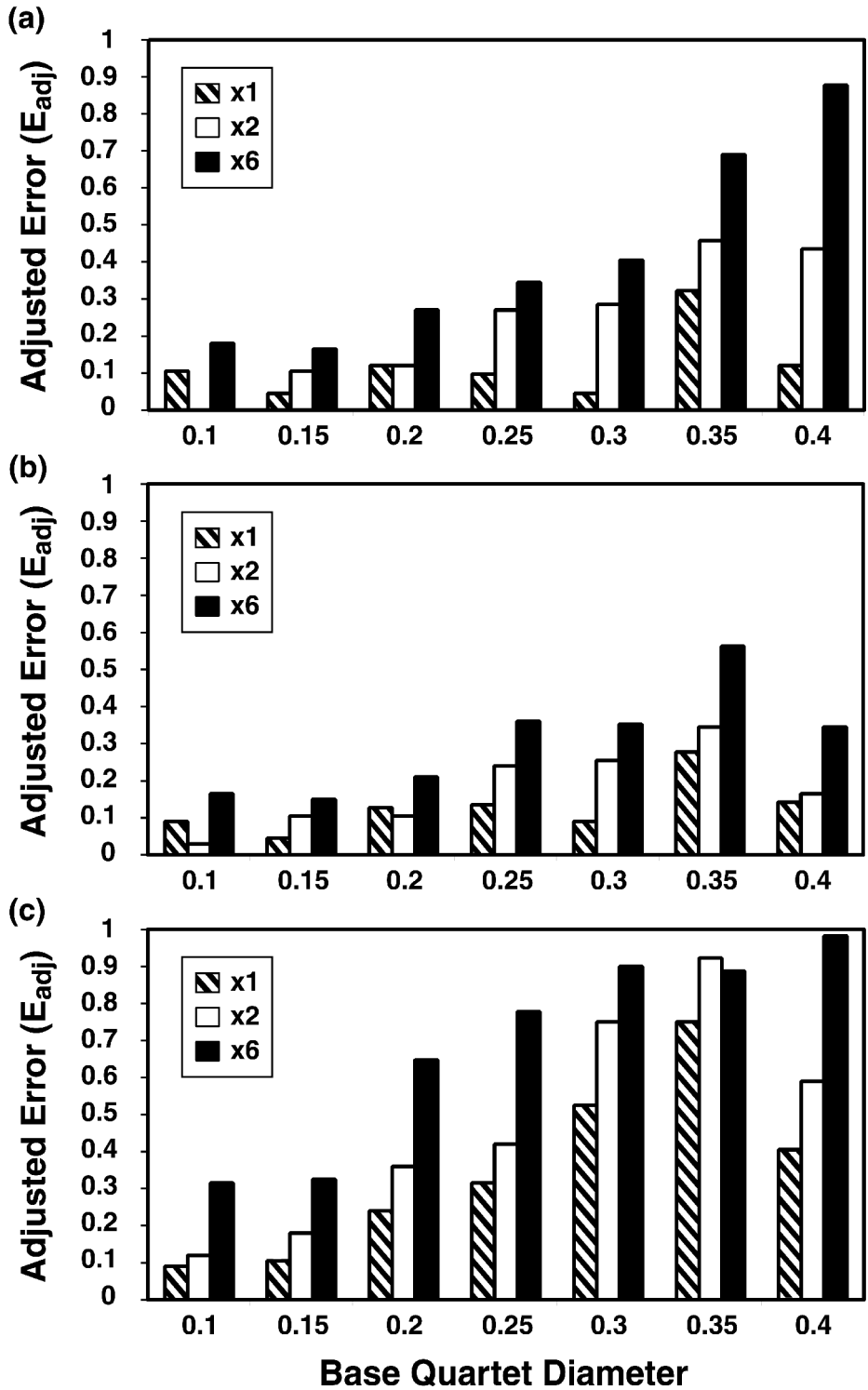

FIGURE 4. Error as a function of tree diameter for quartets of taxa under (a) uniformly weighted parsimony, (b) maximum likelihood, and (c) minimum evolution. The same quartets and data (simulated under the JC model) were used for the analyses shown in (a), (b), and (c). Quartets were sampled from the model tree in the diameter categories listed, and then analyzed for data simulated under the JC model (cross-hatched bars), JC with branch-lengths multiplied by two (white bars), and JC with branch-lengths multiplied by six (black bars).

Thus, any comparison of the effects of subsampling taxa from a larger tree should hold tree diameter relatively constant. Otherwise, one is comparing apples to oranges by comparing simple tree problems (small diameter trees) to hard tree problems (large diameter trees). The usual description of "increased taxon sampling" generally involves adding taxa but keeping the diameter of the sampled tree relatively constant. That is, added taxa lead to a more densely sampled group of interest, not additions of distantly related taxa. If one were interested in studying the phylogeny of mammals, for example, a taxon sample consisting of four primates would not address the relevant problem. If "increased taxon sampling" were used to address this limitation, we would add more mammals rather than birds or beetles. Likewise, comparing error in a tree of 66 mammals (the full tree shown in Fig. 3) to error in a tree of four relatively closely related species (as in Fig. 3b) says little about the importance of taxon sampling on reducing error for the problem of interest (mammalian phylogeny).

We have addressed this issue by consciously selecting subsamples from the full tree with respect to their diameters. Thus, we explicitly either present the average error for subsamples of various sizes over the entire range of possible diameters (e.g., Fig. 2, part b of Figs. 5–8), or compare trees of similar diameters (e.g., part a of Figs. 5–8). For the simpler simulation models, the differences in error between problems of different diameters quickly disappear with larger taxon subsamples, and problems of all diameters show similar error (Fig. 5a). For more complicated models, however, there is a greater difference in difficulty between problems of different diameters for the entire range of taxon numbers (Fig. 5b). However, in both cases, the greatest differences occur with the smallest taxon samples: very small diameter trees of few taxa are clearly easier to estimate than are trees of few taxa that span the full range of the model tree. The net effect of sampling across all possible diameters of trees, then, is to inflate the apparent phylogenetic accuracy of small samples. This higher accuracy is real, but it simply results from examining a smaller portion of the overall phylogenetic tree (e.g., as in Fig. 3b versus 3a).

Tree diameter is not the only consideration that is likely to affect taxon subsampling and its relationship to the accuracy of phylogenetic inference. Difficulty of phylogenetic problems is also related to the size of internal branches in the underlying trees. For this reason, other metrics (such as average distance among taxa, rather than maximum diameter of the tree) might also be considered in analyzing the effect of taxon sampling strategies on phylogenetic accuracy. However, we consider tree diameter (a measure of the depth or scope of the phylogenetic problem) to be the most relevant aspect that should be controlled in such studies.

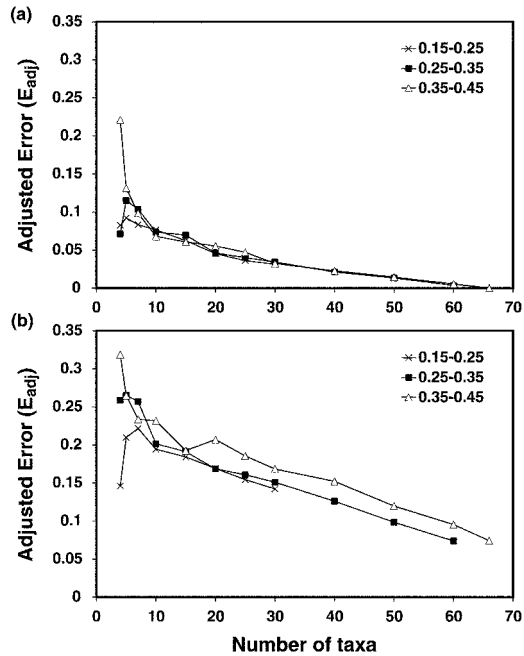

FIGURE 5. The effect of tree diameter on error rate as a function of taxon sample size, under the MP criterion. Data simulated under (a) a simple model (JC), and (b) a more complex model (HKY + Γ).

Thoroughness of Searches

Some recent authors have suggested that thorough searches of tree space are not important for improving the accuracy of phylogenetic inference. For instance, Rosenberg and Kumar (2001) stated that the relatively simple tree searches that they performed (single-tree heuristic searches combined with nearest-neighbor-interchange branch swapping) were adequate to find near-optimal trees, and that more thorough searches would not be expected to decrease phylogenetic error. We tested this hypothesis under the MP criterion, and show the results in

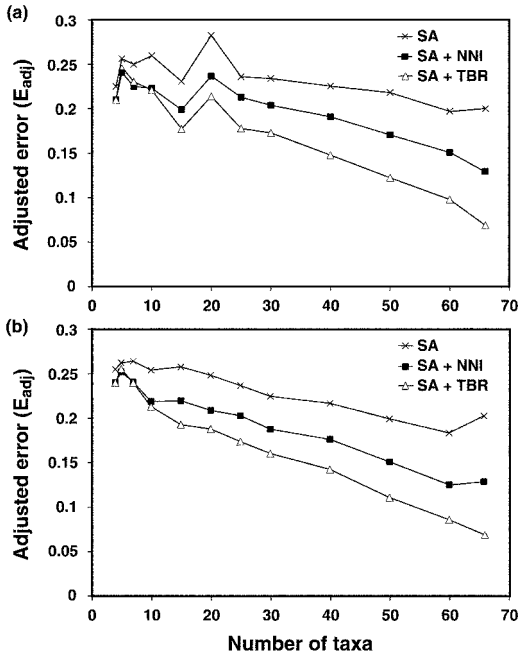

FIGURE 6. The effect of search thoroughness on error rate, as a function of taxon sample size, under the MP criterion. SA: stepwise addition; SA + NNI: stepwise addition plus nearest-neighbor-interchange branch swapping; SA + TBR: stepwise addition plus tree-bisection-reconnection branch swapping. (a) Results for subsamples from the largest diameter bin; (b) results averaged over all diameters. Simple searches result in estimates of greater error than do thorough searches. The importance of thorough searches increases with increasing numbers of taxa. These simulations were conducted under the HKY + Γ model.

Figure 6. Although increased taxon sampling reduced error for all searching methods, the effects were greater for thorough searches than for simple searches, and the importance of thorough searching increased with increasing sample size. Rosenberg and Kumar's (2001:10753) assertion that "[a] more exhaustive, time consuming search is not necessary because it is clear that it does not improve phylogenetic accuracy" is not supported by our analyses. This disagreement may in part be due to the more complex simulation model used in our study (see *Complexity of Simulations*, below).

Figure 6a also illustrates that, under some limited conditions, phylogenetic error can increase through the addition of a small number of taxa. This phenomenon is apparent for trees of large diameter but few taxa (e.g., the left side of Fig. 6a). This appears to correspond to the similar conditions examined by Poe and Swofford (1999), who also noted that phylogenetic error can increase by the addition of one or a few taxa to large diameter trees that contain only a few taxa. However, additional taxa added to the analysis eventually reduced error in these cases in our study (e.g., the right side of Fig. 6a). Therefore, although not all taxon additions result in reduced error, the overwhelming trend appears to be increased phylogenetic accuracy with addition of taxa.

Complexity of Simulations

Rosenberg and Kumar (2001) noted that simulation studies should have a certain advantage in studying the properties of taxon sampling because the true tree is known. Although we agree with this point, the benefit gained by being able to compare inferred trees to a true tree is minimized if the evolutionary model used in simulations is overly simplistic. For instance, the JC model used by Rosenberg and Kumar (2001) incorporates little of the complexity of real sequence evolution. (Rosenberg and Kumar noted that they repeated the study using the HKY model of evolution, but did not show those results and stated that they made no difference in their conclusions.) Simulated sequence data, especially data simulated under a very simple model of evolution, are known to be "easier" to analyze phylogenetically than are data from nature, and the difficulty of the estimation problem increases with increasing model complexity (e.g., Yang, 1996; Rannala et al., 1998; Pollock and Bruno, 2000). The easier the task of inference, the less adding more data (either taxa or characters) should be expected to help. Thus, the effects of taxon sampling would be expected to be least noticeable when analyzing data simulated under a simple model of evolution.

The idea of simulating a number of "realistic" genes and testing the effects of taxon sampling over a variety of parameter values is an appealing one. The fact that the only parameters that were varied in Rosenberg and Kumar's (2001) simulations were the length of the genes and the rate of evolution, however, leaves out much of what is known to vary among genes. No model devised to date fully captures all factors involved in real sequence evolution. Nonetheless, incorporation of some well studied and

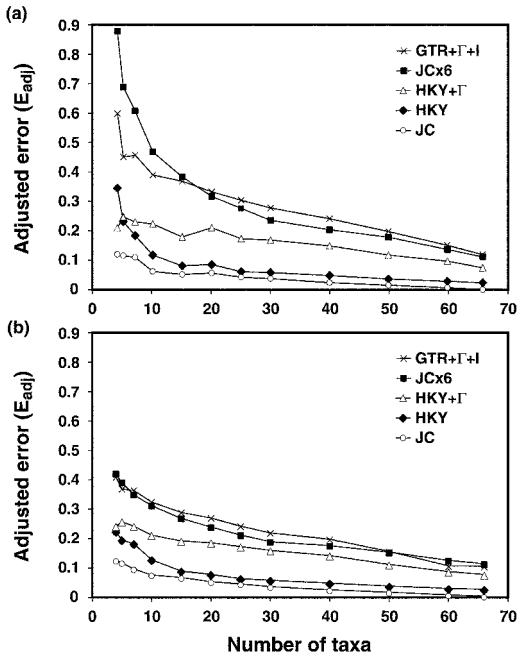

FIGURE 7. The effects of simulation models of varying complexity on error rate, as a function of taxon sample size, under the MP criterion. (a) Results for the largest diameter bin; (b) results averaged over all diameters. "JC × 6" indicates that the Jukes-Cantor model was used, but that rates of evolution were increased sixfold across the tree compared to the model branch-lengths used by Rosenberg and Kumar (2001).

adequately modeled factors should provide a more reasonable assessment of the effects of increased taxon sampling on phylogenetic accuracy. Examples of such factors include variation in substitution rates among sites, differential equilibrium base frequencies, and differential probabilities of substitution. Therefore, we examined the effects of taxon sampling on data simulated under several models of evolution (see Fig. 7). In every case, error was greatly reduced by including increased numbers of taxa in the analyses. However, the overall error increased with increasing complexity of the underlying model of evolution, and taxon sampling provided a greater reduction in the total amount of error for the more complex models. If error reduction is measured as a proportion (rather than as an absolute difference), then this conclusion does not necessarily hold (e.g., error was completely eliminated in the largest taxon samples for the simplest model of evolution; Fig. 7). Note that an increase in error can also be generated by increasing the rate of evolution (and thus overall tree diameter). This is illustrated for the JC model with rates of evolution increased sixfold (Fig. 7).

Effects of Optimality Criteria

We also tested whether or not the effects of taxon sampling on phylogenetic accuracy are dependent on the optimality criterion examined. We present the results for analyses conducted with uniformly weighted parsimony, minimum evolution (with both JC and HKY + Γ distances), and maximum likelihood in Figure 8. Increased sampling of taxa reduces error for all of the methods, so this basic result does not appear to depend on the optimality criterion selected. However, increased taxon sampling appears to be least important for the ME criterion. Both MP and ML show more rapid improvement with

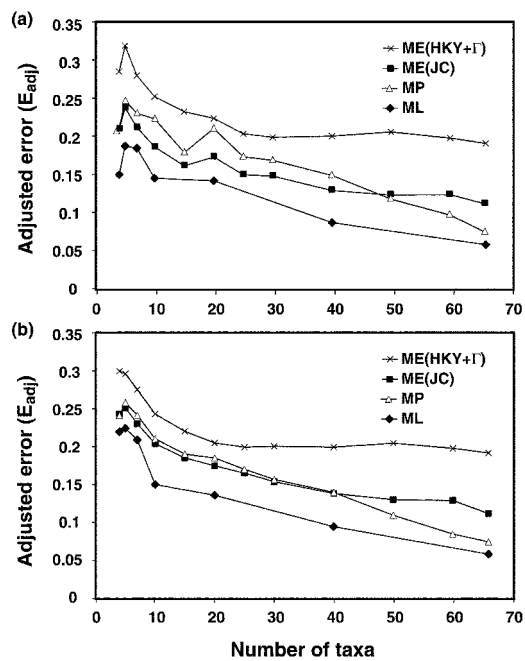

FIGURE 8. The relationship between error and taxon sample size for four optimality criteria. ME(HKY + Γ): minimum evolution with HKY + Γ distances and branch-lengths constrained to non-negative values; ME(JC): minimum evolution with JC distances and negative branch-lengths allowed; MP: uniformly weighted parsimony; ML: maximum likelihood, under the JC model. The results for the two versions of ME searches shown represent the best and worst combinations of distances measured and branch-length constraints (i.e., highest and lowest error) that we examined; the intermediate combinations are not shown for simplicity. (a) Results for the largest diameter bin; (b) results averaged over all diameters. These simulations were conducted under the HKY + Γ model.

increasing taxon sample sizes compared to ME, and as a result both MP and ML show lower error than does ME for the larger taxon samples. Given this behavior, it does not appear advisable to use the ME criterion for trees of many taxa, or to use the ME criterion exclusively in studies of the effects of taxon sampling on phylogenetic accuracy.

Conclusions

All of the simulation analyses that we conducted agree on one point: increased taxon sampling has a clear and strongly positive effect on the accuracy of phylogenetic analyses. This conclusion supports the finding of most other previous studies on the importance of thorough taxon sampling in phylogenetic analysis, but it is in stark contrast to the recent paper by Rosenberg and Kumar (2001) on this topic. Although reanalysis of the Rosenberg and Kumar (2001) data also demonstrates that increased taxon sampling results in increased accuracy of the inferred trees (Pollock et al., 2002), there are a number of reasons why the study design of Rosenberg and Kumar clouded this overwhelming pattern. Our results suggest that studies of the effects of taxon sampling on phylogenetic accuracy should closely consider several aspects of study design. For instance, investigators should consider (1) how error is measured (so as not to bias conclusions as a result of randomly selecting correct trees for small samples), (2) appropriate strategies for sampling taxa (to keep the scope of the phylogenetic problem reasonably constant), (3) the complexity of tree searches (to ensure that near-optimal trees are found for all samples of taxa), (4) the complexity of evolutionary models used in simulations (to ensure that the problems are reasonably realistic), and (5) the different effects of increased taxon sampling on different optimality criteria. When these aspects of study design are incorporated (either separately or together) into an analysis of the effects of taxon sampling on phylogenetic accuracy, the importance of maximizing number of taxa examined becomes overwhelmingly clear.

Although thorough taxon sampling appears to be highly advantageous for phylogenetic analysis, it is not a panacea. Obviously, systematists must also be concerned with collecting enough data, and with collecting data that express appropriate levels of variation for the problem at hand. In some cases, increased taxon sampling will simply not be possible (because of lack of extant taxa to sample, for instance). In many cases, however, the best option available for increasing the accuracy of a phylogenetic analysis will be increased taxon sampling. Biologists should avoid the temptation to define a phylogenetic problem with as few taxa as possible; the additional effort to sample taxa more broadly will almost always result in more accurate (as well as useful) estimates of phylogeny.

Acknowledgments

We thank the members of the Phylogenetics Discussion Group at the University of Texas for initial discussions about this paper. The paper also benefited from discussions with David Cannatella, Mark Holder, Jim McGuire, Steve Poe, and David Pollock. D.J.Z. was supported by an NSF IGERT fellowship in Computational Phylogenetics and Applications to Biology (DGE-0114387), and this research was supported by a National Science Foundation ITR grant (EIA-0121680).

References

Eizirik, 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.

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

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160–174.

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., and M. T. Holder. 2000. Reconstructing the Tree of Life. Pages 47–50 *in* New technologies for the life sciences: A trends guide. Supplement to Trends Journals, Dec. 2000.

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

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.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

Lecointre, G., H. Philippe, H. L. Van Le, and H. Le Guyader. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. 2:205–224.

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.

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

POE, S. 1998. The effect of taxonomic sampling on accuracy of phylogenetic estimation: A test case of a known phylogeny. Mol. Biol. Evol. 15:1086–1090.

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

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.

RAMBAUT, A., AND N. C. GRASSLY. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phyogenetic trees. Comput. Appl. Biosci. 13:235–238.

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.

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, 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.

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

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

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lec. Math. Life Sci. 17:57–86.

WHEELER, W. 1992. Extinction, sampling, and molecular phylogenetics. Pages 205–215 *in* Extinction and phylogeny (M. J. Novacek and Q. D. Wheeler, eds.). Columbia University Press, New York.

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. 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42:294–307.

First submitted 27 December 2001; reviews returned 28 April 2002; final acceptance 21 May 2002
Associate Editor: Keith Crandall

APPENDIX I. MODEL TREE OF ROSENBERG AND KUMAR (2001)

As provided by M. S. Rosenberg. The tree is in NEXUS format, and can be viewed in PAUP* (Swofford, 2000).

((((((((((((*Megaptera novaeangliae*:0.0143, *Tursiops truncatus*:0.0186):0.0157, *Hippopotamus amphibius*:0.0314):0.0043, (*Tragelaphus eurycerus*:0.02, *Okapia johnstoni*:0.0143):0.0343):0.0071, *Sus scrofa*:0.05):0.0014, *Lama glama*:0.06):0.0171, ((*Ceratotherium simum*:0.0243, *Tapirus indicus*:0.0229):0.0029, *Equus caballus*:0.03):0.0157):0.0014, ((((*Felis catus*:0.0057, *Leopardus pardalis*:0.0057):0.0014, *Panthera onca*:0.0057):0.03, (*Canis familiaris*:0.0357, *Ursus arctos*:0.0443):0.0014):0.0171, *Manis pentadactyla*:0.0643):0.0014):0.0014, ((*Artibeus jamaicensis*:0.0643, *Nycteris thebaica*:0.0543):0.0114, (*Pteropus giganteus*:0.02, *Rousettus lanosus*:0.0157):0.0314):0.0071):0.0014, ((*Erinaceus concolor*:0.1157, *Sorex araneous*:0.0843):0.0057, (*Asioscalops altaica*:0.0257, *Condylura cristata*:0.0286):0.0329):0.0086):0.0043, (((((((((*Cavia tschudii*:0.0343, *Hydrochoeris hydrochaeris*:0.0257):0.02, *Agouti taczanowskii*:0.0271):0.0057, *Erethizon dorsatum*:0.0343):0.0086, (*Myocastor coypus*:0.0829, *Dinomys branickii*:0.0486):0.0014):0.0086, (*Hystrix brachyura*:0.06, *Heterocephalus glaber*:0.05):0.0071):0.0386, ((((*Mus musculus*:0.0286, *Rattus norvegicus*:0.0443):0.0257, *Cricetus griseus*:0.0443):0.0557, *Pedetes capensis*:0.0714):0.0043, (*Castor canadensis*:0.0629, *Dipodomys heermanni*:0.1143):0.0043):0.0057):0.0014, (*Tamias striatus*:0.0514, *Muscardinus avellanarius*:0.1043):0.0086):0.0043, (*Sylvilagus floridanus*:0.0429, *Ochotona hyperborea*:0.0814):0.0257):0.0071, ((((((*Hylobates concolor*:0.0157, *Homo sapiens*:0.0114):0.0086, *Macaca mulatta*:0.02):0.01, (*Ateles fusciceps*:0.0143, *Callimico goeldii*:0.02):0.02):0.0329, *Cynocephalus variegatus*:0.0543):0.0029, (*Lemur catta*:0.0443, *Tarsius sp.*:0.0814):0.0086):0.0014, *Tupaia minor*:0.0829):0.0014):0.0029):0.0029, (((*Choloepus hoffmanni*:0.0029, *Choloepus didactylus*:0.0071):0.0329, (*Tamandua tetradactyla*:0.0143, *Myrmechophaga tridactyla*:0.0114):0.0371):0.0057, (*Euphractus sexcinctus*:0.0071, *Chaetophractus villosus*:0.0043):0.0443):0.0243):0.0086, ((((*Trichechus manatus*:0.0329, *Loxodonta africana*:0.0486):0.0043, *Procavia capensis*:0.0686):0.0129, (*Echinops telfairi*:0.1257, *Orycteropus afer*:0.0543):0.0014):0.0043, (*Macroscelides proboscideus*:0.0286, *Elephantulus rufescens*:0.03):0.0729):0.0114):0.04, (*Didelphis virginianus*:0.0571, *Macropus eugenii*:0.0657):0.1414):0.0;

APPENDIX II. PARAMETER VALUES USED IN SIMULATIONS

1. HKY:
   Transition/Transversion ratio: 2.93
   Base frequencies: A:0.37, C:0.24, G:0.12, T:0.27
2. HKY + continuous gamma rate heterogeneity:
   Same as HKY, plus shape parameter for gamma distribution: 0.399
3. GTR + continuous gamma rate heterogeneity + invariant sites:
   Rate matrix:

A → C: 3.297	A → G: 12.55	A → T: 1.167
	C → G: 2.060	C → T: 13.01
		G → T: 1.00

   Base frequencies: A:0.1776, C:0.3336, G:0.2595, T:0.2293
   Shape parameter for gamma distribution: 0.8168
   Proportion of invariant sites: 0.5447