Syst. Biol. 54(5):731–742, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500234583

Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction?

JOHN J. WIENS

Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794-5245, USA; E-mail: wiensj@life.bio.sunysb.edu

Abstract.— Taxon sampling may be critically important for phylogenetic accuracy because adding taxa can help to subdivide misleading long branches. Although the idea that added taxa can break up long branches was exemplified by a study of "incomplete" fossil taxa, the issue of taxon completeness (i.e., proportion of missing data) has been largely ignored in most subsequent discussions of taxon sampling and long-branch attraction. In this article, I use simulations to test the ability of incomplete taxa to subdivide long branches and improve phylogenetic accuracy in situations of potential long-branch attraction. The results show that for most methods and conditions examined, adding taxa that are only 50% complete may provide similar benefits to adding the same number of complete taxa (suggesting that the advantages of increased taxon sampling may be obtained with less data than previously considered). For parsimony, taxa that are less complete (5% to 25% complete) may often have limited ability to rescue analyses from long-branch attraction. In contrast, highly incomplete taxa can be surprisingly beneficial when using model-based methods. The results also suggest the importance of model-based methods in phylogenetic analyses that combine molecular and fossil data. [Combining data; fossils; incomplete taxa; missing data; phylogenetic accuracy; simulations; taxon sampling.]

In recent years, taxon sampling has become a prominent and contentious issue in systematics (e.g., Hillis, 1996, 1998; Kim, 1996, 1998; Graybeal, 1998; Poe, 1998, 2003; Rannala et al., 1998; Soltis et al., 1998; Wiens, 1998; Poe and Swofford, 1999; Rosenberg and Kumar, 2001, 2003; Zwickl and Hillis, 2002; Pollock et al., 2002; Hillis et al., 2003). Many studies have now shown that a major benefit of increased taxon sampling is the potential for added taxa to subdivide long branches (i.e., branches on which many of the included characters have changed), branches which may otherwise "attract" and be erroneously grouped together. The idea that added taxa can potentially subdivide long branches and increase phylogenetic accuracy was exemplified by a study addressing the effects of including relatively "incomplete" fossil taxa in an analysis of extant taxa (Gauthier et al., 1988), where incompleteness refers to the proportion of missing or unknown character states that a taxon bears. However, most subsequent discussions of taxon sampling have focused exclusively on complete taxa (i.e., no missing data), and the question of whether incomplete taxa can improve accuracy by breaking up long branches has been relatively neglected.

Long branches can positively mislead parsimony analyses (Felsenstein, 1978), causing the wrong tree to be estimated with increasing confidence as more characters are added, particularly when there are long terminal branches separated by a short internal branch (a phenomenon called long-branch attraction, or LBA hereafter). Many model-based methods are thought to be much less sensitive to this problem, such as maximum likelihood, neighbor-joining, and Bayesian analysis (e.g., Felsenstein, 1978; Huelsenbeck, 1995; Alfaro et al., 2003). Nevertheless, in some cases, model-based methods may also suffer from the effects of LBA, particularly when the number of characters is limited and/or the model of evolution assumed in the analysis has a poor fit to the processes that generated the data (e.g., Gaut and Lewis, 1995; Huelsenbeck, 1995). Thus, the problem of long-branch attraction is potentially important for all phylogenetic methods.

In their classic study, Gauthier et al. (1988) suggested that a parsimony analysis of morphological data based on living taxa alone gives an unorthodox or even misleading picture of amniote phylogeny, whereas addition of fossil taxa yields a more well-accepted hypothesis of amniote relationships. Gauthier et al. (1988) showed that the change occurs because of the addition of certain key fossil taxa to a critical long branch (the branch leading to mammals). Since then, several simulation studies have shown that adding complete taxa may subdivide long branches and improve phylogenetic accuracy (e.g., Hendy and Penny, 1989; Graybeal, 1998; Rannala et al., 1998), but these studies generally have not considered whether inclusion of incomplete taxa will reap the same benefits.

Taxon completeness is a particularly critical and timely issue. Incompleteness is an obvious problem in analyses of fossil taxa, which may be missing data due to stochastic preservational effects or because whole suites of characters cannot be scored (e.g., molecular data, behavior, soft anatomy). Many researchers are now conducting phylogenetic analyses that combine extensive molecular data sets (i.e., many characters) for living taxa with morphological data sets that include fossil taxa (e.g., Gatesy et al., 2003), raising the question of whether the fossil taxa can have any influence on relationships estimated for the complete, extant taxa.

Incompleteness has become an important issue for strictly molecular analyses as well. As more genes and genomes are sequenced, a striking disparity may be created in the number of characters that are available for different taxa within a given group of organisms. Methods are being developed that can screen databases to find and include complete sets of characters and taxa (e.g., Sanderson et al., 2003), and it seems that many empirical researchers make decisions about including taxa and characters based (at least in part) on the desire to avoid incompleteness in their data matrices.

But does completeness matter? Although sampling may be designed to avoid missing data, Donoghue et al. (1989) suggested that the "incompleteness and

informativeness of taxa are unrelated." When considering the effects of adding incomplete taxa to a phylogenetic analysis, it may be helpful to consider two different aspects of their impact. First, can the incomplete taxa be included and accurately placed in the tree? Second, will they actually improve the accuracy of estimated relationships among the more complete taxa?

Several previous studies have considered whether and how incomplete taxa can be included in an analysis (e.g., Doyle and Donoghue 1987; Wilkinson, 1995; Kearney, 2002; Anderson, 2002) and if they are accurately placed on the resulting tree (e.g., Huelsenbeck, 1991; Wiens and Reeder, 1995; Wiens, 2003a). Despite the observation that adding incomplete taxa may sometimes lead to multiple equally parsimonious trees and poorly resolved consensus trees (e.g., Novacek, 1992), several authors have suggested that the relationship between taxon incompleteness and informativeness is unpredictable (e.g., Donoghue et al., 1989; Wilkinson, 1995; Kearney, 2002). Wiens (2003a) suggested that incompleteness was not important in itself, but rather that the number of complete characters was the critical factor. Using simulations, he found that the negative effects of including incomplete taxa disappeared if the overall number of characters was sufficiently large (at least for relatively short branch lengths).

The second question—whether adding incomplete taxa can improve accuracy for the more complete taxa—has barely been studied. Wiens (2003b) used simulations to show that the level of completeness was potentially important in determining whether or not incomplete taxa improved estimated relationships among the more complete taxa (i.e., highly incomplete taxa often failed to improve accuracy for the more complete taxa). However, that study was relatively limited, in that it examined only parsimony analysis of binary characters.

In the present study, I use simulations to test whether adding incomplete taxa can subdivide long branches and improve phylogenetic accuracy. More specifically, I ask three questions. (1) Can incomplete taxa rescue an analysis from long branch attraction, or is this somehow prevented by their missing data? (2) Do the effects of adding incomplete taxa depend on the overall number of characters in the analysis? (3) Does the impact of incomplete taxa depend on the phylogenetic method used (e.g., parsimony, likelihood, Bayesian, neighbor-joining)?

I also explore the potential tradeoffs between the number of added taxa and their completeness. Given the choice between adding equal numbers of complete and incomplete taxa, it seems obvious that more complete taxa should be preferred (i.e., more data are better), all other things being equal. Less obvious is whether it would be preferable to add data in the form of a few complete taxa or a larger number of less complete taxa. There is no question that a complete taxon potentially can subdivide a long branch, if it is optimally placed along the branch. However, some placements of a complete taxon may not be so effective (e.g., when the added taxon is closer to the tip than the node of the long terminal branch; Poe, 2003). Adding many incomplete taxa may increase the chances that at least one taxon is optimally placed for subdividing the long branch, and may potentially outweigh the disadvantages of their incompleteness.

Simulations may never capture the diversity and complexity of real character data and taxon sampling scenarios. Furthermore, there may be important tradeoffs between how well simulations match a particular data set and how relevant they may be across diverse data sets. I simulated relatively generalized character data under the classic scenario for long-branch attraction (two long terminal branches separated by a short internal branch; Felsenstein, 1978). The reason for this choice is simple; if the addition of incomplete taxa cannot improve accuracy under this scenario, then it may be very unlikely to do so in more complex cases (e.g., when the effects of adding taxa may be more difficult to predict; Poe and Swofford 1999).

Materials and Methods

Effects of Incompleteness

General simulation strategy.—Simulation methods generally followed previous studies (Wiens, 2003a, 2003b). Programs for simulating data and tallying results from phylogenetic software packages were written in C by the author. A 16 taxon tree was simulated (Fig. 1). Four taxa were chosen to be complete in all analyses (A, H, I, P), and these four taxa were chosen so as to maximize the problem of long-branch attraction when these four taxa are analyzed alone. Thus, two taxa were chosen at the terminal ends of the tree (A, P), and two adjacent taxa were chosen near the "center" of the tree (H, I), to create a combination of a short internal branch and two long terminal branches. The tree shape and number of taxa facilitated creating these optimal conditions for long-branch attraction.

The remaining 12 taxa were designated as incomplete. For each replicate, the data matrix was analyzed with

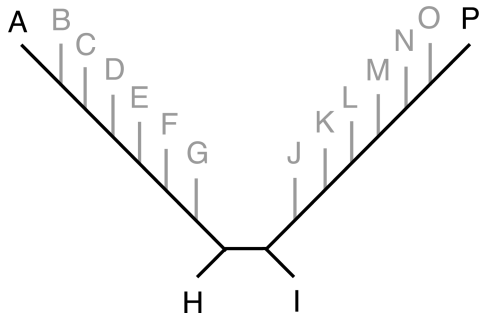

FIGURE 1. Model tree and sampling scheme used in simulations. Taxa A, H, I, and P are complete and included in all simulations. When analyzed alone, the particular combination of long and short branch lengths among these four taxa creates the classic "Felsenstein Zone" problem of long-branch attraction. Taxa B to G and J to O are added in many analyses, but these added taxa have different proportions of their character data replaced with missing data cells ("?"). The goal of this study is to determine if these added taxa can successfully subdivide the long branches (despite their incompleteness) and improve the accuracy of the relationships estimated for taxa A, H, I, and P.

these 12 taxa either excluded or included but only 5% complete (i.e., 95% of characters replaced with missing data), 10% complete, 25% complete, 50% complete, 75% complete, or 100% complete (no missing data). When taxa were made incomplete, it was generally assumed that the same characters were missing in all the incomplete taxa. This is the distribution of missing data one would expect when combining data sets with incomplete overlap (e.g., different genes, molecular and fossil data). However, a set of analyses was also performed in which the set proportion of missing data cells were randomly distributed among characters in each taxon and in each replicate. These two methods of distributing missing data cells represent two extremes in a continuum that might be encountered in real data sets. However, both methods gave similar results, suggesting that "intermediate" distributions might be expected to give similar results as well.

Binary data.—One set of simulations involved parsimony and Bayesian analysis of binary character data (mimicking generalized morphological data). Parsimony analyses were implemented in PAUP* version 4.0b10 (Swofford, 2002), using a heuristic search with 20 random taxon-addition-sequence replicates per search and retaining a maximum of 500 shortest trees.

Accuracy was measured as the number of replicates in which the estimated tree (pruned after each analysis to include only the original four complete taxa) matched the known, simulated true tree for the four complete taxa, divided by the total number of replicates for that set of conditions. When multiple equally parsimonious trees were generated, only a single shortest tree was retained and used to measure accuracy. The use of a single shortest tree should approximate the average accuracy among shortest trees when considered across multiple simulation replicates. Note that because I focused on the accuracy of an unrooted four-taxon tree, an estimated tree was either entirely correct or incorrect, and the number of correct versus incorrect nodes was not an issue. For parsimony analyses, 200 replicates were examined for each set of conditions. The major conditions that were varied were number of characters (100, 500, 1,000, 2,000) and different branch lengths (here defined as the probability of a character state changing from the beginning to the end of a branch). Branch lengths were equal along all branches of the simulated phylogeny, and long-branch attraction was created by excluding taxa (presumably a more common scenario than long branches that are created by dramatic differences in mutation rate or generation time among closely related taxa).

Bayesian analyses of the binary data were conducted using MrBayes, versions 3.0b3 and 3.0b4 (Huelsenbeck and Ronquist, 2001). The model of Lewis (2001) was used, and assuming that all characters were included (coding = "all"). Initial analyses of the simulated data sets suggested that stationarity was achieved consistently before reaching 10,000 generations. Therefore, analyses used a total of 40,000 generations and excluded the first 10,000 as burn-in. Results were then rechecked on the set of conditions with the lowest accuracy (i.e., results potentially affected by an inadequate search) using 400,000 generations and 100,000 generations as burn-in and were found to be extremely similar to those based on 40,000 generations. Trees were sampled every 100 generations. Analyses used four heated chains and default priors. 100 replicates were examined for each set of simulated conditions. The best estimate of phylogeny for each Bayesian analysis was based on the majority-rule consensus of the post burn-in trees, pruned to include only the four complete taxa.

The support (posterior probabilities) for individual branches in the Bayesian trees was not quantified. Although support is obviously an important issue, the more fundamental question is whether the correct tree is consistently estimated. Furthermore, it seems unfair to expect Bayesian analysis to estimate the correct phylogeny with strong support unless the same requirement is made of all the other methods under consideration, and the goal of the study was not to compare branch support values across methods.

The age of the added taxa is thought to be an important factor in determining the costs and benefits of adding incomplete taxa (e.g., Gauthier et al., 1988; Donoghue et al., 1989; Huelsenbeck, 1991). Fossils may be particularly beneficial because they may retain a larger proportion of ancestral states than would living taxa (Gauthier et al., 1988; Huelsenbeck, 1991). In order to span the full range of possible levels of plesiomorphy, simulations were performed in which the added taxa were "living" (i.e., the normal situation in which the terminal taxon is represented by the data at the end of its branch) and in which they were "perfect fossils" (the taxa retain all the character states of their immediate ancestor, such that the states of the terminal taxon are the same as their ancestral node). However, the effect of the age of these taxa may not be particularly important in these simulations, because the added taxa are on relatively short branches to begin with (Fig. 1).

Initially, a broad range of branch lengths were analyzed (0.01, 0.05, 0.10, 0.20). However, for the sake of brevity, only two of these are presented. These represent conditions that were found to be relatively easy for phylogeny reconstruction (0.05, with many informative but slowly evolving characters) and those that are more difficult (0.20, with rapidly evolving characters), based on preliminary results.

DNA sequence data.—Analyses were also performed using simulated DNA sequence data, using parsimony, maximum likelihood, Bayesian analysis, and neighbor-joining. Parsimony, likelihood, and neighbor-joining analyses were implemented in PAUP* (version 4.0b10) and Bayesian analyses were implemented using MrBayes (versions 3.0b3 and 3.0b4).

Simulations initially used a very simple model of sequence evolution, with equal base frequencies and equal rates of change among substitution types and nucleotide positions (JC; Jukes and Cantor, 1969). This allowed all methods (including parsimony) to perfectly match the model assumed by the data and to better isolate the effects of missing data. However, this also made

the Bayesian, neighbor-joining, and likelihood methods potentially insensitive to the problem of long-branch attraction. Therefore, analyses were also performed in which the simulated model of evolution was more complex than the model of evolution assumed by the methods. Presumably, all data sets in the real world will also be more complex than any models used in phylogenetic inference (in terms of the processes of character evolution).

Previous studies suggest that among-site rate variation is the most important parameter in determining the sensitivity of model-based methods to long-branch attraction; model-based methods may be misled by long-branch attraction if there is among-site rate variation in the data but among-site rate variation is assumed to be absent in the model used in the phylogenetic analysis (e.g., Gaut and Lewis, 1995; Poe, 2003). Therefore, analyses were performed in which among-site rate variation was simulated but not assumed by the phylogenetic methods. The simulated data were modeled to resemble protein-coding sequences. In one set of analyses, the first two characters of every three had branch lengths of 0.02 and 0.02, whereas the third had a branch length of 0.20 (the 10-fold difference in rates was initially chosen based on protein-coding genes in salamanders; Wiens, unpublished data). In another set of analyses, the rate variation was arbitrarily made more extreme, with lengths of 0.01, 0.01, and 0.40. Even more extreme heterogeneity was also explored (0.00, 0.00, and 0.50) but caused all methods to perform very poorly (Wiens, unpublished data). The gamma distribution (Yang, 1993) is often used to simulate among-site rate variation, although it does not mimic any specific process of molecular evolution. The gamma distribution was used to estimate among-site rate variation in the simulated data sets in some of the analyses (see below), following standard practice in empirical studies.

The effects of unequal base frequencies and transition-transversion ratios were also simulated (HKY model; Hasegawa et al., 1985). A set of simulations was conducted assuming a 3:1 transition:transversion ratio and base frequencies of A = 37%, G = 12%, C = 24%, and T = 27% (parameter values based on mammalian sequences and reported by Zwickl and Hillis, 2002). For these analyses, all phylogenetic methods were implemented assuming the Jukes-Cantor model, to address the impact of assuming an oversimplified model of evolution for these parameters. However, for simulations that included among-site rate variation, the likelihood, Bayesian, and neighbor-joining analyses were also modified to incorporate this parameter (using the gamma distribution). Thus, the effects of ignoring different parameters were addressed. For Bayesian analyses, the gamma-shape parameter was estimated individually for each simulation replicate. This could also be done in likelihood analyses (in theory) but would be extremely time intensive and somewhat unnecessary, given that an identical model of evolution is simulated in each replicate. Instead, the gamma shape parameter was estimated for 10 simulated data sets (2,000 characters each) for each of the two patterns of among-site rate variation using likelihood and holding the phylogeny constant. The average estimate for the 10 data sets was then used in the likelihood and neighbor-joining analyses.

Parsimony analyses of the DNA data used a heuristic search with 20 random taxon-addition-sequence replicates per search and retaining a maximum of 500 shortest trees. Two hundred replicates were analyzed for each set of simulated conditions using parsimony and neighbor-joining. Likelihood searches were considerably slower than parsimony analyses. Therefore, only five random addition sequence replicates were used per search, and only 50 replicate matrices were examined for each set of simulated conditions. For Bayesian analyses, 100 replicated matrices were examined. Based on preliminary analyses, each Bayesian analysis used 40,000 generations with the first 10,000 discarded as burn-in. More extensive analyses (increasing the number of generations 10 fold for the set of conditions with the lowest accuracy) gave results that were nearly identical to those based on the smaller number of generations (i.e., all accuracy values within 5%).

Simulations of the DNA sequence data were run using 500, 1000, and 2000 characters. However, the general results were qualitatively similar across different levels of character sampling, and for the sake of brevity only the results for 1000 characters are figured and discussed.

Taxon Number versus Taxon Completeness

Given that adding more data will generally be preferable (e.g., adding eight taxa that are 100% complete is better than adding eight that are 25% complete, all other things being equal), an important question is whether it is better to add a few complete versus a larger number of less complete taxa. To begin to address this question, the simulations of DNA sequence data (JC model) were rerun but this time adding either: (1) eight taxa that are 25% complete, (2) four taxa that are 50% complete, or (3) two taxa that are 100% complete. Thus, the number of complete data cells added was the same in each case, but the three scenarios differed in the number of characters, taxa, and missing data cells. The taxa to be included were randomly selected in each simulation replicate, with the restriction that an equal number of taxa were added to each of the two long branches (Fig. 1). The simulations assessed the effects of these sampling strategies on the accuracy of the estimated trees for the four complete taxa. Again, only results based on 1,000 characters are figured and discussed.

Combining Fossil and Molecular Data

A limited set of simulations was performed to specifically examine the potential consequences of combining morphological (fossil) and molecular data. A set of 100 binary characters for all 16 taxa was simulated to represent the morphological data (branch length = 0.05). DNA sequence data (2,000 or 8,000 characters) were simulated for 16 taxa, but only four taxa were included (A, H, I, P). Thus, combined analyses were performed in which four

taxa were 100% complete (molecular and morphological data) and the other 12 taxa were represented by morphological data only and were either 5% complete (with 2,000 DNA sequence characters) or 1% complete (8,000 DNA characters). DNA sequences were evolved under the HKY model, as described previously, with branch lengths of either (a) 0.02, 0.02, and 0.20 for first, second, and third base positions, respectively; (b) 0.01, 0.01, and 0.40; and (c) 0.20 for all characters. Data were analyzed using equally weighted parsimony (200 replicates) and with Bayesian analysis (100 replicates). Bayesian analysis used the JC model for the molecular data (without among-site rate variation) and the Lewis (2001) model for the morphological data. Bayesian analyses used 40,000 generations with the first 10,000 discarded, and results again were confirmed with analyses using 10 times as many generations. Models and parameters were unlinked in the combined Bayesian analysis. Accuracy was assessed for the four complete taxa when analyzed using the molecular data alone, and with the morphological and molecular data combined. Statistical significance of differences in accuracy were assessed using a paired *t*-test, implemented with the Statview™ software package. Except for the extreme (but realistic) disparity in number of characters between the molecular and morphological data sets, the simulations represent a "worst-case scenario" for the molecular data (i.e., incomplete taxon sampling, faster rates of evolution, imperfect fit between the simulated data and reconstruction model) but a "best-case scenario" for the morphological data (i.e., complete taxon sampling, low rates of change, no missing data within the morphological data set). Thus, these analyses address the question of whether adding morphological or fossil data could potentially improve accuracy in a combined analysis under conditions where this improvement might be expected. Adding fossil taxa obviously could not improve accuracy under conditions where the molecular data set alone always estimates the correct phylogeny (e.g., short branch lengths, perfect fit between simulated model and reconstruction model). It should be noted that these simulations should also be relevant to cases in which morphological and molecular data sets are combined and molecular data are available for only some of the taxa.

RESULTS

Binary Data

For parsimony analysis of binary data, the proportion of missing data (incompleteness) strongly influences whether or not added taxa improved accuracy in cases of LBA (Fig. 2). For both low and high rates of change, the addition of highly incomplete taxa (5% to 10% complete) did not improve phylogenetic accuracy. At low rates of change, adding more complete taxa (50% to 100% complete) did increase accuracy, whereas at high rates of change even taxa that were 50% complete provided little improvement.

Bayesian analysis of the binary data (Fig. 2) showed much less sensitivity to LBA than parsimony at lower rates of change (length = 0.05). However, at high rates of change, Bayesian analyses of the four complete taxa were consistently inaccurate, despite large numbers of characters. To my knowledge, this is the first study to assess the sensitivity of Bayesian analysis to LBA using Lewis' (2001) model. The most fragmentary taxa (5% to 10% complete) had limited ability to rescue Bayesian analyses from LBA, although addition of more complete taxa (25% to 75%) caused dramatic increases in accuracy.

Results were generally similar (for both parsimony and Bayesian analysis) when missing data cells were distributed randomly among characters (Fig. 3a, b). Incomplete taxa that retained their ancestral states (e.g., simulated fossils; Fig. 3c, d) had little impact on the results at low rates of change but did improve results for both methods for relatively complete taxa at high rates of change. The older taxa also increased accuracy for Bayesian analysis when there was high incompleteness and high rates of change. However, the effects of taxon age generally were minor relative to the effects of incompleteness.

DNA Sequence Data

As in the analyses of binary data, highly incomplete taxa (5% to 10% complete) were unable to rescue parsimony analyses of DNA sequence data from LBA (Fig. 4). However, addition of taxa that were 50% complete led to dramatic increases in accuracy in most cases, as did taxa that were 25% complete in some cases.

Likelihood, Bayesian analysis, and neighbor-joining generally showed higher accuracy than parsimony under conditions of potential LBA (Fig. 4). However, in cases of high rates of change or mismatch between the simulated and assumed models of evolution (i.e., because of among-site rate heterogeneity), the accuracy of these model-based methods was greatly impaired. In these cases, there typically were dramatic increases in accuracy associated with adding incomplete taxa, even when the added taxa were highly incomplete (10% to 25%). In most cases, taxa that were 50% complete gave results similar to those when the added taxa were 100% complete.

Results were generally similar for DNA sequence data generated under the HKY model (Fig. 5). When among-site rate variation was added to the model assumed by likelihood, Bayesian analysis, and neighbor-joining (for data sets with simulated differences in rates among characters), accuracy was greatly increased, and the potential for added taxa to improve accuracy was greatly diminished (thus, there was seemingly little consequence for ignoring unequal base frequencies and transition:transversion bias in these analyses). Surprisingly, addition of highly incomplete taxa greatly decreased phylogenetic accuracy for the neighbor-joining method under these conditions, even though there was no similar decrease in accuracy under the same conditions when among-site variation was not included in the model.

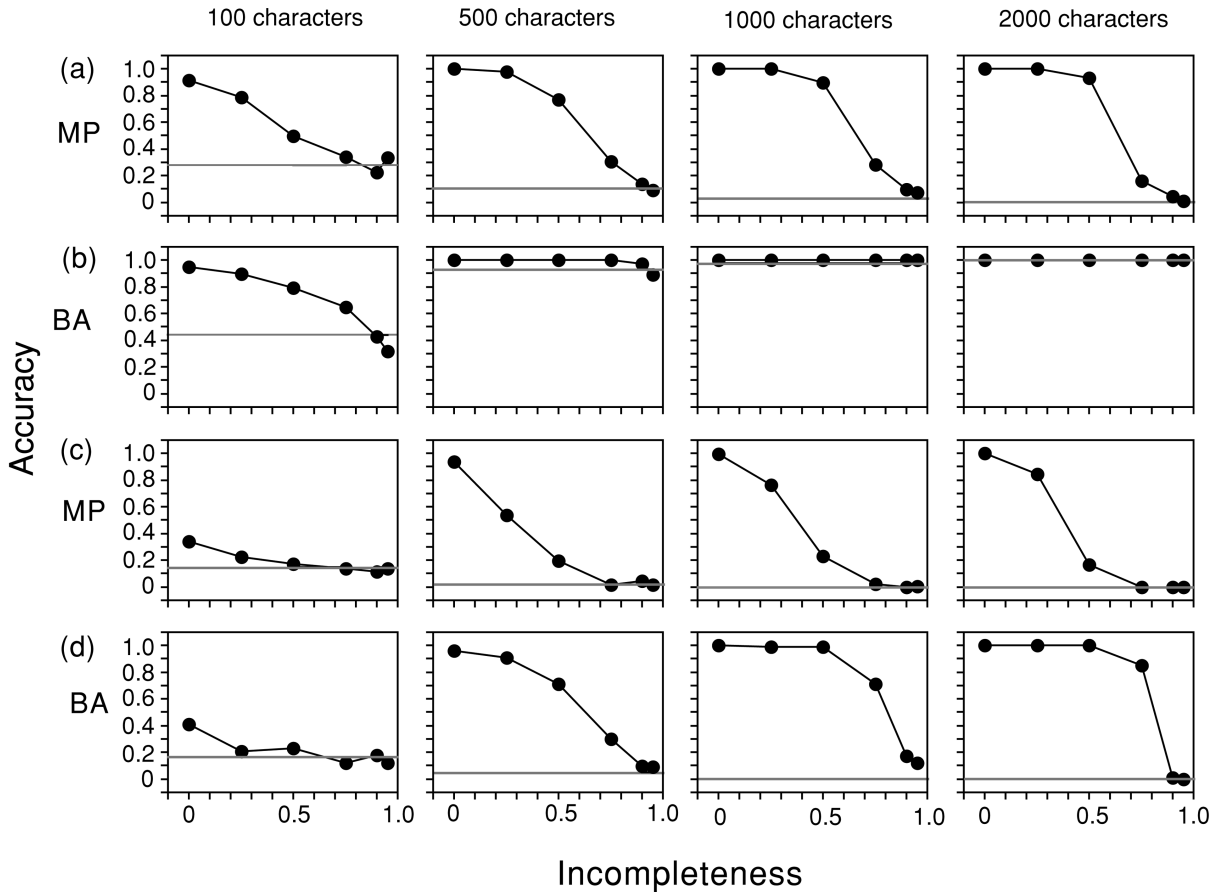

FIGURE 2. The ability of incomplete taxa to subdivide long branches and improve phylogenetic accuracy for parsimony (MP) and Bayesian (BA) analysis of binary character data, for low (branch length = 0.05; a, b) and high (branch length = 0.20; c, d) rates of character change. The gray horizontal line represents the proportion of replicates in which the correct phylogenetic relationships among the four complete taxa ((A, H), (I, P)) are reconstructed for a given set of conditions (accuracy), based on analysis of these four complete taxa alone (see Fig. 1). Filled circles represent accuracy for the four complete taxa after including 12 additional taxa of varying levels of completeness. The same characters are missing in all 12 taxa in a given replicate.

Taxon Completeness versus Number of Taxa

The analyses of taxon completeness versus taxon number showed that both strategies (i.e., few complete taxa versus many incomplete taxa) gave surprisingly similar levels of accuracy for most methods and conditions, and there were few consistent trends favoring one approach over another (Fig. 6).

Combining Morphological (Fossil) and Molecular Data

The analyses of combined data suggest that addition of fossil taxa can potentially improve phylogenetic accuracy despite their relative incompleteness (Table 1). However, this appears to be much more likely for Bayesian analysis than for parsimony. Addition of the morphology-only taxa increased accuracy by 10% to 21% for Bayesian analysis of 2,000 DNA characters and 100 morphological characters. For parsimony analysis, the increase was similar (7%) for conditions of extreme rate heterogeneity, but was negligible for other conditions. When the number of molecular characters was increased to 8000 (perhaps a more realistic number), adding the incomplete (morphology-only) taxa improved accuracy for both parsimony and Bayesian analysis under conditions of extreme rate heterogeneity. Although the increase was not overwhelming (10% to 11%) and was statistically

TABLE 1. Combining molecular (4 taxa) and morphological (fossil) data (16 taxa). Analyses in which the incomplete taxa retained all of their ancestral states gave similar results (Wiens, unpublished data).

Branch lengths (DNA data)	Accuracy of parsimony		Accuracy of Bayesian analysis	
	Molecular	Combined	Molecular	Combined
	Characters = 2,000 molecular, 100 morphological			
0.02/0.02/0.20	0.145	0.190	0.860	0.980**
0.01/0.01/0.40	0.480	0.550	0.570	0.780**
0.20	0	0	0.800	0.900
	Characters = 8,000 molecular, 100 morphological			
0.02/0.02/0.20	0.015	0.030	0.990	1.000
0.01/0.01/0.40	0.275	0.385*	0.690	0.790
0.20	0	0	0.970	1.000

Asterisked values indicate significant increases in accuracy for the combined data (16 taxa) relative to the molecular data alone (4 complete taxa), with one indicating $P \leq 0.05$, and two indicating $P \leq 0.01$.

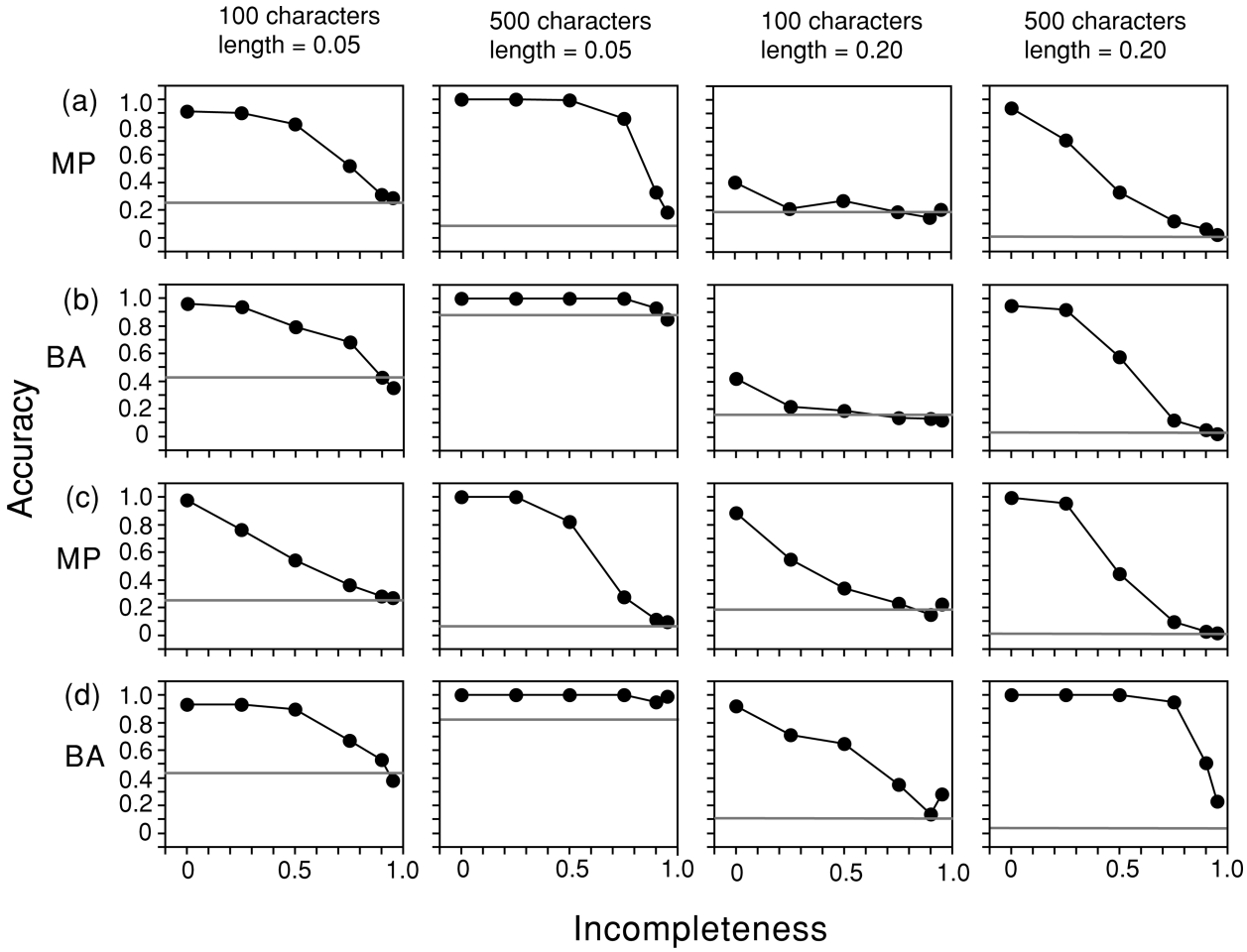

FIGURE 3. The ability of incomplete taxa to subdivide long branches and improve phylogenetic accuracy for parsimony (MP) and Bayesian (BA) analysis of binary character data, when missing data are randomly distributed among characters in incomplete taxa (a, b) and when incomplete taxa retain all of the character states of their immediate ancestor (c, d). The gray horizontal line represents the proportion of replicates in which the correct phylogenetic relationships among the four complete taxa ((A, H), (I, P)) are reconstructed for a given set of conditions (accuracy), based on analysis of these four complete taxa alone (see Fig. 1). Filled circles represent accuracy for the four complete taxa after including 12 additional taxa of varying levels of completeness.

significant only for parsimony, the magnitude is rather surprising because the added taxa were only 1% complete. Results for other conditions show little effect from addition of the incomplete taxa, but highlight the obvious differences in the relative performance of parsimony and Bayesian methods, with Bayesian analysis performing very well (accuracy 97% to 100%) and parsimony very poorly (0% to 3%).

DISCUSSION

Effects of Incompleteness

Can incomplete taxa rescue an analysis from long-branch attraction and thereby improve phylogenetic accuracy? Based on these simulations, the answer clearly is yes. For all methods and most simulated conditions, the benefits of adding taxa that were 50% complete are similar to those for adding taxa that were 100% complete. For analyses using model-based methods, adding taxa that were only 25% complete caused dramatic increases in phylogenetic accuracy in many cases. Under some conditions, similar increases in accuracy were even seen adding taxa that were only 5% to 10% complete (Figs. 4, 5). In analyses of the combined morphological and molecular data, taxa that were only 1% complete caused increases in accuracy in some cases. Thus, the idea that completeness should be the primary determinant of how data are sampled may be worth reconsidering.

Incompleteness is not irrelevant, however, especially when using parsimony. For parsimony analyses that are potentially misled by LBA, highly incomplete taxa (5% to 10%) generally seem unable to effect a rescue. This result makes intuitive sense. In cases of LBA, most of the parsimony-informative characters may support the incorrect tree (Felsenstein, 1978). It seems unlikely that adding taxa which are scored for only a small fraction of these characters could overturn a hypothesis that is strongly supported by a much larger

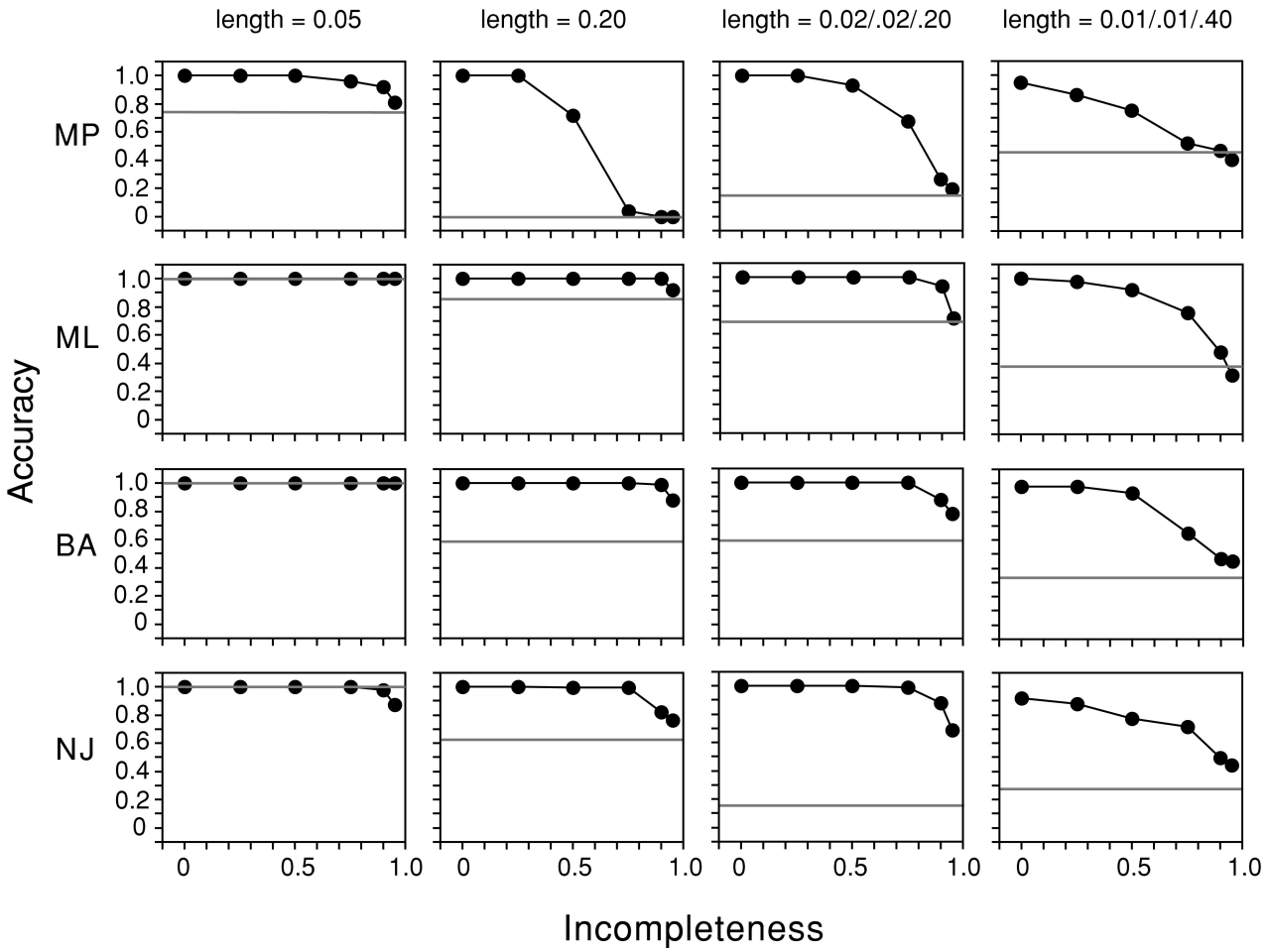

FIGURE 4. The ability of incomplete taxa to subdivide long branches and improve phylogenetic accuracy for simulated DNA sequence data (1,000 characters, Jukes-Cantor model), using parsimony (MP), likelihood (ML), Bayesian analysis (BA), and neighbor-joining (NJ) methods. The gray horizontal line represents the proportion of replicates in which the correct phylogenetic relationships among the four complete taxa ((A, H), (I, P)) are reconstructed for a given set of conditions (accuracy), based on analysis of these four complete taxa alone (see Fig. 1). Filled circles represent accuracy for the four complete taxa after including 12 additional taxa of varying levels of completeness. The same characters are missing in all 12 taxa in a given replicate.

number of characters. This hypothesis also explains why this pattern remains similar across different numbers of characters. In these cases, it is the relative number of characters that is critical, not the absolute number. In contrast, the accurate placement of incomplete taxa seems to depend critically on the absolute number of characters scored in these taxa, rather than their relative completeness (Wiens, 2003a).

Bayesian analysis, likelihood, and neighbor-joining appear to be less sensitive to LBA than is parsimony analysis, even when the models assumed by these methods are incorrect (e.g., Huelsenbeck, 1995; Alfaro et al., 2003; this study). They may be only marginally impacted by the effects of long branches, and therefore may be more easily influenced by the addition of small amounts of data. This may explain why even highly incomplete taxa are able to dramatically increase accuracy (in some cases) using these methods. Nevertheless, there were also many cases in which taxa that were highly incomplete (5% to 10%) had little effect on accuracy.

The study by Gauthier et al. (1988) offered the classic case of the importance of incomplete fossil taxa in phylogenetic analysis, and might seem to be in conflict with the results of this study (i.e., in showing that adding highly incomplete taxa may be of limited benefit in parsimony analyses). In their study, addition of certain critical (though incomplete) fossil taxa overturned relationships based on more complete taxa, and did so in a way that suggests the tree that includes the fossil taxa may be the more accurate estimate of phylogeny. However, it is important to note that the critical incomplete taxa in the Gauthier et al. (1988) study had at least 50% of their character data present. Taxa that are 50% complete were also highly beneficial in parsimony analyses in many of the simulation results of this study. Thus, there is no apparent conflict between the results of this study and those of Gauthier et al. (1988).

Interestingly, there were almost no conditions encountered in which the addition of incomplete taxa had a substantially negative impact on phylogenetic

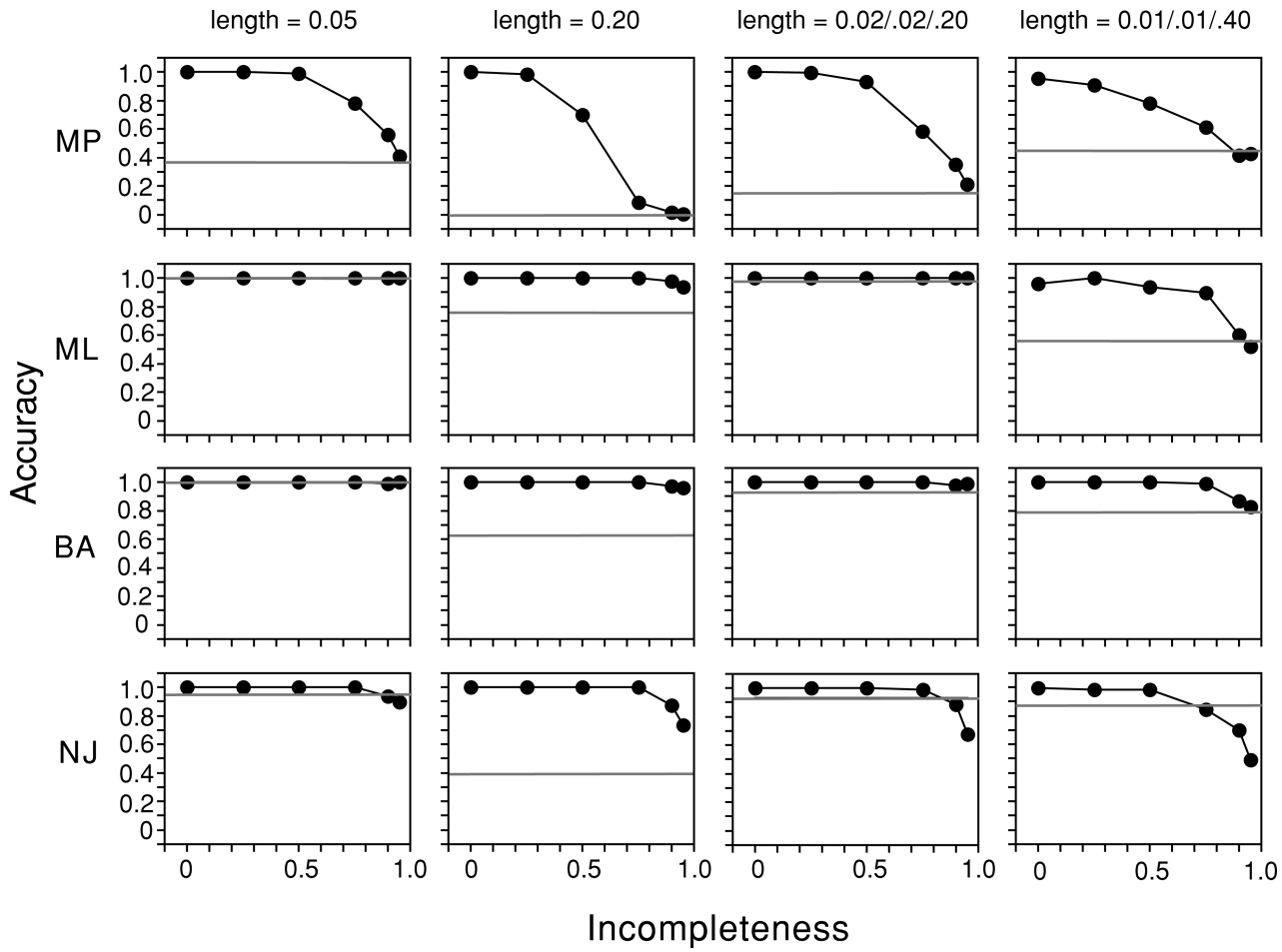

FIGURE 5. The ability of incomplete taxa to subdivide long branches and improve phylogenetic accuracy for simulated DNA sequence data (1,000 characters, HKY model), using parsimony (MP), likelihood (ML), Bayesian analysis (BA), and neighbor-joining (NJ) methods. Data are simulated under the HKY model but analyzed assuming the JC model. Among-site rate variation was incorporated into the ML, BA, and NJ analyses using the gamma-shape parameter (for the two sets of simulations incorporating rate heterogeneity). The gray horizontal line represents the proportion of replicates in which the correct phylogenetic relationships among the four complete taxa ((A, H), (I, P)) are reconstructed for a given set of conditions (accuracy), based on analysis of these four complete taxa alone (see Fig. 1). Filled circles represent accuracy for the four complete taxa after including 12 additional taxa of varying levels of completeness. The same characters are missing in all 12 taxa in a given replicate.

accuracy for the four complete taxa (with the exception of neighbor-joining, and only under a limited set of conditions, Fig. 5). This result suggests that, under conditions in which adding taxa will be helpful, the addition of incomplete taxa is likely to be either helpful or harmless. Of course, the results of this study do not guarantee that the effects of adding incomplete taxa will always be positive or neutral, because there are conditions under which adding taxa may be detrimental to phylogenetic accuracy (e.g., Poe and Swofford, 1999; Poe, 2003).

Limitations of Simulations

Some important limitations of these simulations should be reiterated. In this study, I focused on the classic worst-case scenario for long-branch attraction, in which there are two long terminal branches separated by a short internal branch (Felsenstein, 1978; Huelsenbeck and Hillis, 1993). Conversely, this may also be the best-case scenario for added taxa to subdivide long branches and improve accuracy (see Fig. 3 of Poe, 2003). To my knowledge, incomplete taxon sampling is problematic primarily because of the misleading effects of long branches (see also Rannala et al., 1998).

In these simulations, all of the incomplete taxa that are added help to subdivide a single pair of long branches. In the real world, a more likely scenario is that a large number of incomplete taxa would be added to a data set including a large number of complete taxa, which might include several misleading long branches (or none). It is also possible that none of the added taxa would actually fall along one of the long branches. An important problem in simulating the more realistic scenario is that the results may be far more difficult to interpret.

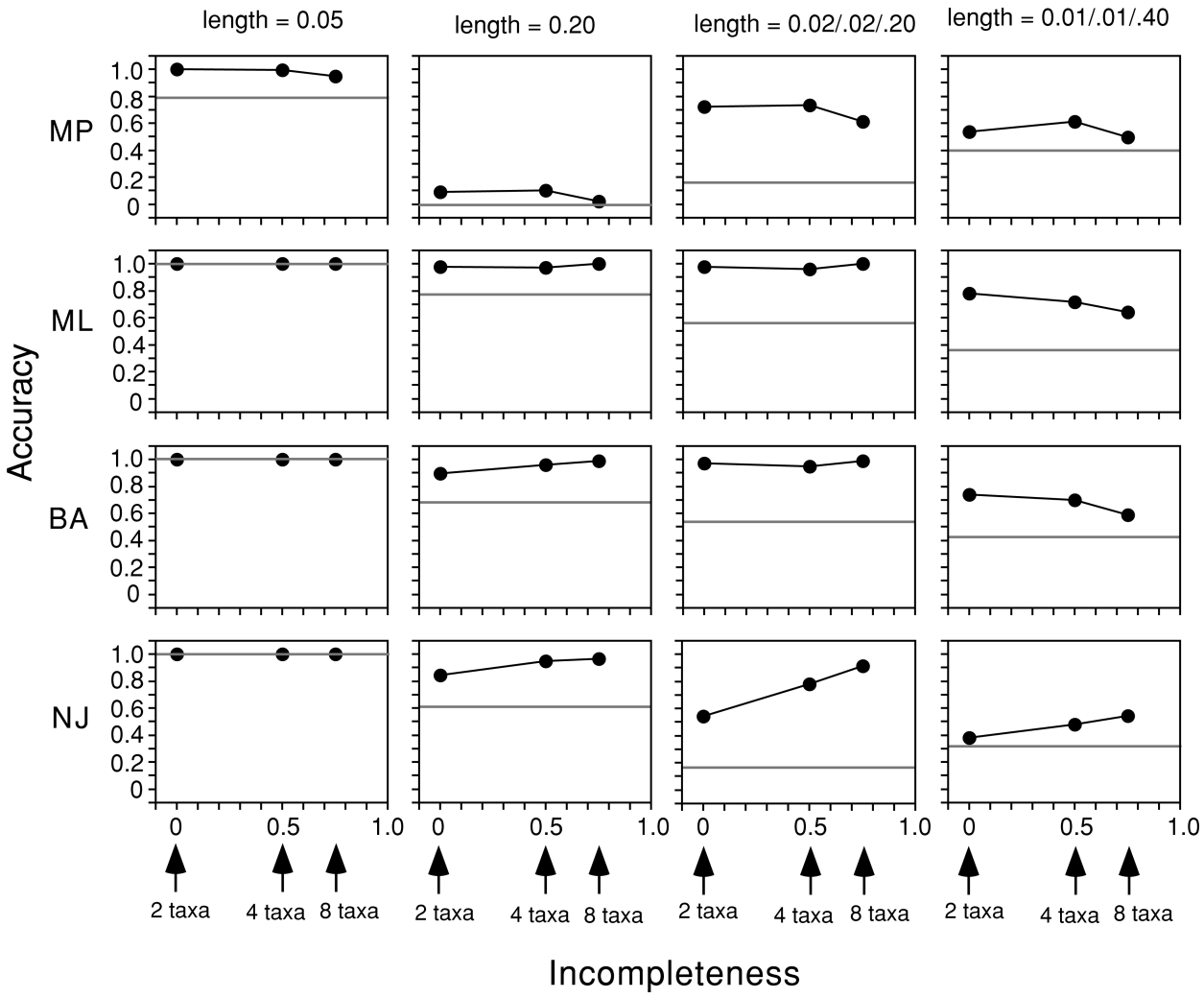

FIGURE 6. Relative benefits of adding many incomplete taxa versus few complete taxa for simulated DNA sequence data (1,000 characters, Jukes-Cantor model), using parsimony (MP), likelihood (ML), Bayesian analysis (BA), and neighbor-joining (NJ) methods. The gray horizontal line represent the proportion of replicates in which the correct phylogenetic relationships among the four complete taxa ((A, H), (I, P)) are reconstructed for a given set of conditions (accuracy), based on analysis of these four complete taxa alone (see Fig. 1). Filled circles represent accuracy for the four complete taxa after including and then pruning out additional taxa. Three strategies for adding taxa are analyzed, eight taxa that are 25% complete (75% incomplete), four taxa that are 50% complete, and two taxa that are 100% complete.

For example, if adding incomplete taxa failed to improve phylogenetic accuracy, it might be because (1) the added taxa were too incomplete; (2) the added taxa did not fall along any long branches; (3) there were no problematic long branches; or (4) there were long branches and the added taxa did fall along them, but added taxa do not improve accuracy under that particular combination of branch lengths (e.g., Poe and Swofford, 1999). Furthermore, several scenarios might apply to different parts of the tree. The simpler simulations employed in this study show that, under conditions where adding taxa should improve accuracy, incomplete taxa can also be helpful despite their missing data.

The character data simulated in this study are also far simpler than those encountered in empirical data sets. Rather than trying to incorporate all the potential complexities of real data (which is not practical), I instead simulated the general scenario in which the data analyzed are more complex than the model assumed in the phylogenetic analysis, focusing on among-site rate variation, unequal base frequencies, and biased transition:transversion ratios. Admittedly, model mismatch might involve some other factor or set of factors in real data sets, given that these parameters can be included using model-based methods. It is also worth noting that long branches can be problematic for model-based methods (and incomplete taxa can be helpful) even when there is a perfect fit between the model generating the data and the model assumed by the method (i.e., when branches are very long). Conversely, these model-based methods can also be extremely accurate in the Felsenstein Zone even when they assume a model that ignores some of the complexities of the simulated data (Fig. 5; lengths of 0.02, 0.02, and 0.20)

Sampling Taxa versus Characters

The results of this study have implications for one of the more critical (and pragmatic) questions in the recent phylogenetic literature: Is it better to add more characters or more taxa to improve phylogenetic accuracy? Among prior studies the results have been mixed, with some studies supporting more characters (e.g., Poe and Swofford, 1999; Rosenberg and Kumar, 2001) and others supporting more taxa (e.g., Graybeal, 1998; Zwickl and Hillis, 2002).

An important assumption in these debates is that the benefits of increased taxon sampling come only from adding taxa that are 100% complete (no missing data). However, the results of this study show that it may be possible to reap the benefits of increased taxon sampling without having data for all characters for all taxa. In fact, adding taxa that are 50% complete may show similar benefits to adding complete taxa under many conditions. Under many circumstances, taxa that are only 25% complete (or less) were also beneficial, particularly for model-based methods. Thus, the rewards of increased taxon sampling might be obtained far more cheaply (in terms of the actual data required) than has been considered in previous studies.

As an example, a limited sample of results from this study show that accuracy can be increased as much or more by adding taxa that are only 10% complete than by doubling the number of characters for the complete taxa (Table 2). Yet, adding these highly incomplete taxa requires only 30% as much data as doubling the number of characters for the complete taxa. It should be acknowledged, however, that the conditions simulated in this study might be expected to favor sampling of taxa over characters, at least for parsimony.

If the full benefits of increased taxon sampling could be obtained by including taxa that are only 25% or 50% complete, then one might expect that the benefits of adding a larger number of incomplete taxa should be greater than those for adding fewer complete taxa. Yet, in the analyses comparing the effects of taxon completeness versus taxon number, the results are similar for many conditions and methods (Fig. 6). However, when analyzing the data using model-based methods, addition of either complete or incomplete taxa gives almost perfect accuracy for most of the simulated conditions, which tends to greatly equalize the results. This is not true for conditions of extreme rate heterogeneity, conditions where highly incomplete taxa had more limited ability to increase accuracy, even when using model-based methods (Fig. 4).

TABLE 2. The effects of increasing the number of taxa versus characters on phylogenetic accuracy. The results show that adding incomplete taxa can potentially provide greater benefits with less data than increasing the number of characters. In these simulations (described in Fig. 4), adding 12 taxa that are only 10% complete increases accuracy as much or more than doubling the number of characters in the four complete taxa. However, adding the 12 incomplete taxa requires the addition of only 1,200 data points, whereas doubling the number of characters in the complete taxa requires 4,000.

	Accuracy			
Sampling strategy	MP	ML	BA	NJ
Branch length = 0.20				
4 complete taxa, 1,000 characters	0	0.860	0.590	0.630
4 complete taxa + 12 incomplete taxa, 1,000 characters total	0	1.000	1.000	0.820
4 complete taxa, 2,000 characters	0	1.000	0.790	0.760
Branch lengths = 0.02/0.02/0.20				
4 complete taxa, 1,000 characters	0.155	0.690	0.590	0.155
4 complete taxa + 12 incomplete taxa, 1,000 characters total	0.265	0.940	0.880	0.880
4 complete taxa, 2,000 characters	0.045	0.560	0.640	0.060

How Will the Tree of Life Be Reconstructed?

Many researchers presently are undertaking concerted efforts to resolve major branches of the Tree of Life, combining the efforts of both neontologists and paleontologists. However, it remains uncertain whether adding fossil taxa to combined analyses of molecular and morphological data will have any impact on relationships estimated for the living taxa, given the large number of molecular characters available for living taxa and the resulting incompleteness of the fossil taxa in the combined data matrix.

I conducted a limited set of simulations to address the issue of combining molecular and morphological data, and the results show that this combination may be surprisingly useful, despite the dramatic differences in the size of the data sets. The simulations suggest that adding fossil (morphology-only) taxa might significantly improve accuracy even when the fossil taxa are only 1% to 5% complete (assuming conditions in which the molecular results actually need improvement). Of course, fossil taxa may offer many other benefits besides their impact on the relationships estimated among living taxa.

The results also underscore the dramatic differences in the performance of these phylogenetic methods; under the conditions examined, the combined Bayesian analyses were highly accurate, whereas the parsimony analyses were not. In some ways, this is an obvious result, as the sensitivity of parsimony methods to LBA has long been known (e.g., Felsenstein, 1978). Adding fossil taxa to help subdivide long branches is a sensible strategy, but so is application of a method that is less sensitive to LBA. Applying both strategies simultaneously is now possible, given new likelihood models for morphological data (Lewis, 2001) and software packages that allow sophisticated models of evolution to be applied to molecular and morphological data separately in a combined analysis (Huelsenbeck and Ronquist, 2001; Nylander et al., 2004). So far, phylogenetic analyses combining molecular and paleontological data have only had the option of using parsimony (e.g., Eernisse and Kluge, 1993; O'Leary, 1999; Sun et al., 2002; Gatesy et al., 2003). Although much additional work is needed, these simulation results strongly suggest that model-based methods (e.g., Bayesian, likelihood) may be highly advantageous in analyses that combine fossil and molecular data.

ACKNOWLEDGMENTS

I thank A. Driskell, R. Geeta, P. Herendeen, M. Kearney, D. Moen, S. Smith, P. Soltis, and P. Stephens for helpful comments on the manuscript. This research was supported by U.S. National Science Foundation grant DEB 0334923 to J.J.W.

REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Anderson, J. S . 2001. The phylogenetic trunk: Maximal inclusion of taxa with missing data in an analysis of the Lepospondyli (Vertebrata, Tetrapoda). Syst. Biol. 50:170–193.

Donoghue, M. J., J. A. Doyle, J. Gauthier, A. G. Kluge, and T. Rowe. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431–460.

Doyle, J. A., and M. J. Donoghue. 1987. The importance of fossils in elucidating seed plant phylogeny and macroevolution. Rev. Palaeobot. Palynol. 50:63–95.

Eernisse, D. J., and A. G. Kluge. 1993. Taxonomic congruence versus total evidence, and the phylogeny of amniotes inferred from fossils, molecules and morphology. Mol. Biol. Evol. 10:1170–1195.

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

Gatesy, J., G. Amato, M. Norell, R. DeSalle, and C. Hayashi. 2003. Combined support for wholesale taxic atavism in gavialine crocodylians. Syst. Biol. 52:403–422.

Gaut, B. S., and P. O. Lewis. 1995. The success of maximum likelihood in the four-taxon case. Mol. Biol. Evol. 12:152–162

Gauthier, J., A. G. Kluge, and T. Rowe. 1988. Amniote phylogeny and the importance of fossils. Cladistics 4:105–209.

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

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape spit by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

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., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Huelsenbeck, J. P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. 40:458–469.

Huelsenbeck, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

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

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. Munro, ed.). Academic Press, New York.

Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: Mistaken assumptions and conclusions. Syst. Biol. 51:369–381.

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 effects of phylogenetic estimators. Syst. Biol. 47:43–60.

Lewis, P. O. 2001. A likelihood approach to inferring phylogeny from discrete morphological characters. Syst. Biol. 50:913–925.

Novacek, M. J. 1992. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. Syst. Biol. 41:58–73.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

O'Leary, M. A. 1999. Parsimony analysis of total evidence from extinct and extant taxa, and the cetacean-artiodactyl question. Cladistic 15:315–330.

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

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. Nature 398:299–300.

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.

Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

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

Rosenberg, M. S., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

Sanderson, M. J., A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20:1036–1042.

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.

Sun, G., Q. Ji, D. L. Dilcher, S. Zheng, K. C. Nixon, and X. Wang. 2002 Archaefructaceae, a new basal angiosperm family. Science 296:899–904.

Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony*, version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Wiens, J. J. 1998. The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis: A simulation study. Syst. Biol. 47:381–397.

Wiens, J. J. 2003a. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiens, J. J. 2003b. Incomplete taxa, incomplete characters, and phylogenetic accuracy: What is the missing data problem? J. Vert. Paleontol. 23:297–310.

Wiens, J. J., and T. W. Reeder. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. 44:548–558.

Wilkinson, M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501–514.

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

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

First submitted 15 September 2004; reviews returned 8 March 2005; final acceptance 27 April 2005

Associate Editor: Pam Soltis