Journal of Vertebrate Paleontology 23(2):297–310, June 2003

INCOMPLETE TAXA, INCOMPLETE CHARACTERS, AND PHYLOGENETIC ACCURACY: IS THERE A MISSING DATA PROBLEM?

JOHN J. WIENS
Section of Amphibians and Reptiles, Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213-4080, wiensj@carnegiemuseums.org

ABSTRACT—The problem of missing data is often considered to be the most significant obstacle in reconstructing the phylogeny of fossil taxa and their relationships to extant taxa. In this paper, I review the results of recent simulation studies and present new results that explore how missing data affect phylogenetic accuracy, which is defined here as the success of a method at reconstructing the true phylogeny. Missing data cells are typically added to a phylogenetic analysis in the form of incomplete taxa (e.g., highly fragmentary fossil taxa) or incomplete characters (e.g., a set of DNA sequence or soft anatomical characters in an analysis including living and fossil taxa). These two types of incomplete data affect phylogenetic analyses in two very different ways, suggesting that there is not a single ''missing data problem.'' Recent simulation results show that including incomplete taxa is a problem of including too few characters rather than too many missing data cells—if enough characters are scored in these taxa, even the relationships of highly incomplete taxa (e.g., 95% missing data) can be accurately reconstructed. Including incomplete characters is largely a problem of taxon sampling. Adding incomplete characters can improve accuracy under many conditions, but inadequate taxon sampling in these characters can lead to problems of long branch attraction (which causes methods to reconstruct an incorrect tree). New simulation results show that highly incomplete taxa may have little impact on the relationships estimated for the complete taxa. Thus, adding highly incomplete taxa may not adversely affect relationships among the complete taxa. However, these added taxa may be unable to improve accuracy for the complete taxa if they are too incomplete. These results suggest that analyses which combine data from fossils and molecular data sets can be successful, despite large amounts of missing data. The accuracy of these analyses will depend on adequate sampling of characters for fossil taxa and adequate sampling of taxa for molecular data sets.

INTRODUCTION

Fossil taxa differ from extant taxa in two important ways. First, they are older, and therefore may retain many ancestral states not seen in extant taxa (Gauthier et al., 1988; Donoghue et al., 1989; Huelsenbeck, 1991). Second, they are often incomplete, meaning that states cannot be determined for many characters in these taxa. These taxa must be treated as unknown or ''missing data'' for these characters in phylogenetic data matrices (if these taxa and characters are included). Fossil taxa may be incomplete because there are whole sets of characters that are not preserved in the fossilization process (e.g., DNA sequences, soft anatomy, behavior) or because a certain structure that could potentially be preserved is absent (or so damaged as to be unscorable) in a particular specimen or set of specimens that represents the taxon. The incompleteness of fossil taxa is widely considered to be the most significant obstacle in reconstructing their relationships (Donoghue et al., 1989; Huelsenbeck, 1991) and has led some authors to question their usefulness of including fossil taxa in phylogenetic analyses among major groups of living taxa (Patterson, 1981; Ax, 1987). However, incompleteness is not unique to fossil taxa (Gauthier et al., 1988; Donoghue et al., 1989). Extant taxa may also have many missing data cells, particularly when data sets are combined that do not include identical taxa (e.g., molecules and morphology or different genes; Wiens and Reeder, 1995).

Missing data cells in phylogenetic data matrices are widely considered to be undesirable, and can be eliminated by excluding either the incomplete taxa or the incomplete characters. For example, paleontologists may choose to exclude taxa that have a certain proportion of their characters missing data (e.g., Rowe, 1988; Grande and Bemis, 1998). Similarly, they may exclude characters that cannot be scored in a certain proportion of the taxa (e.g., Livezey, 1989; Smith et al., 1995). Exclusion of incomplete taxa and characters may be considerably more common than is explicitly stated in the literature; in many cases it is obvious that incomplete taxa and characters have been eliminated, but without explanation or justification (e.g., exclusion of highly fragmentary fossil taxa or of soft anatomical characters in analyses that combine fossil and living taxa).

The exclusion of missing data cells seems to be widely practiced, but may come with a cost. Excluding missing data cells requires eliminating incomplete taxa and/or characters from an analysis. There is abundant evidence from simulations that increasing the number of characters increases the probability of reconstructing the true phylogeny under many conditions (e.g., Huelsenbeck and Hillis, 1993; Hillis et al., 1994; Huelsenbeck, 1995; Wiens and Servedio, 1998). Similarly, many simulation studies also show that increasing the number of taxa can improve phylogenetic results (e.g., Graybeal, 1998; Hillis, 1998; Rannala et al., 1998; Wiens, 1998a; but see also Kim, 1996; Poe and Swofford, 1999). The question then becomes: do the advantages of excluding missing data cells outweigh the disadvantages of reducing the number of taxa and/or characters? This question is particularly important because the exact mechanisms (if any) that might cause missing data to be problematic are rarely stated and remain poorly explored.

To address this question, we need a criterion by which to evaluate our results. In other words, how do we know if including or excluding a set of characters or taxa has made the results better or worse? The most important criterion may be how different approaches (e.g., including vs. excluding incomplete characters or taxa) affect phylogenetic accuracy. Phylogenetic accuracy refers to the ability of a method (e.g., parsimony) to correctly reconstruct the true phylogenetic relationships of a group of organisms. Accuracy is a difficult criterion to address, because the phylogeny of most groups of organisms is entirely unknown. However, there are several approaches that can be used to explore accuracy (see review by Hillis, 1995).

These include (1) analysis of data from known, laboratory produced phylogenies of viruses and other organisms (e.g., Hillis et al., 1992, 1994; Wiens and Reeder, 1995), (2) congruence analyses (e.g., Miyamoto and Fitch, 1995; Cunningham, 1997; Wiens, 1998a), comparing the ability of different methods to correctly reconstruct clades that are strongly supported by many different types of evidence (e.g., morphological, molecular, and chromosomal data); and (3) computer simulations, generating character data assuming a given phylogeny and model of evolution and comparing the ability of methods to reconstruct this phylogeny under different simulated conditions (e.g., Huelsenbeck, 1991, 1995; Huelsenbeck and Hillis, 1993; Hillis et al., 1994; Graybeal, 1998; Wiens, 1998a, b, c; Wiens and Servedio, 1998). Each of these approaches has its strengths and weaknesses, but the most widely used approach for evaluating accuracy is simulations (Hillis, 1995).

The advantages and disadvantages of simulations are closely intertwined. Clearly, simulated data sets lack the complexity of character data in the real world, and it is therefore dangerous to extrapolate specific simulation results directly to empirical phylogenetic problems (e.g., using the results of this study to decide that it is always advantageous to include taxa that are missing 50% of their data cells when 100 characters have been sampled). On the other hand, it is the simplicity of simulated data sets that allows one to control, vary, and thereby understand the relevant parameters and mechanisms that affect phylogenetic accuracy. Even if the true phylogeny of some group of naturally occurring organisms were somehow known, a comparison of the phylogenetic accuracy of different approaches using data from this group would be of surprisingly limited value by itself. The results obtained might only be applicable to (for example) that particular tree shape, number of taxa, combination of branch lengths, and type of data.

In the present paper, I review the results of recent simulation studies that address the impact of missing data on phylogenetic accuracy. First, I address the effects of including incomplete taxa. Second, I discuss the effects of adding incomplete characters. I then present new simulation results that compare the accuracy of both approaches simultaneously and address several unresolved questions.

Including Incomplete Taxa

Most of the literature on missing data has focused on the problem of including taxa with many missing data cells. Several authors have noted that including taxa that are highly incomplete (i.e., a high proportion of their characters are missing data) may lead to multiple equally parsimonious trees and poorly resolved consensus trees (e.g., Gauthier, 1986; Nixon and Wheeler, 1992; Novacek, 1992; Wilkinson, 1995; Wilkinson and Benton, 1995; Gao and Norell, 1998). These incomplete taxa may be difficult to place phylogenetically, and their inclusion may obscure otherwise well-resolved relationships among the more complete taxa, at least in some cases. Given these observations, many authors have excluded taxa a priori based on their level of completeness (e.g., Rowe, 1988; Grande and Bemis, 1998; Ebach and Ahyong, 2001). Furthermore, some authors have suggested that the incompleteness of fossil taxa will generally prevent them from changing relationships among the more complete taxa, and that they can be safely excluded when reconstructing relationships among major groups of living taxa (e.g., Patterson, 1981; Ax, 1987). However, other authors have noted that the impact of including an incomplete taxon on a phylogenetic analysis (i.e., leading to multiple trees or changing relationships among the more complete taxa), may be difficult to predict based on its level of completeness alone (e.g., Donoghue et al., 1989; Novacek, 1992; Kearney, 2002) and have developed different exclusion criteria (e.g., Wilkinson, 1995; Anderson, 2001).

Huelsenbeck (1991) used simulations to address the impact of incomplete fossil taxa on phylogenetic accuracy. Specifically, he explored the tradeoffs between advantageous temporal position (i.e., older taxa retaining more ancestral traits) and disadvantageous level of completeness for an eight-taxon tree with 100 characters. He found that including fossil taxa could improve accuracy relative to including complete, extant taxa when the fossil taxa are relatively old, relatively complete, and/or when branches are relatively long (ancient divergences and/or high rates of character change). He also found that including highly incomplete taxa leads to multiple equally parsimonious trees and thereby to decreased phylogenetic accuracy (relative to including complete taxa). He proposed that highly incomplete taxa are problematic because their inclusion increases the proportion of ambiguously resolved ancestral character states for a given node of the tree. However, Huelsenbeck (1991) did not directly address the impact of including versus excluding incomplete taxa, and did not vary the number of characters in his analysis (see below).

Wiens and Reeder (1995) followed up the study of Huelsenbeck (1991) by examining the effects of including incomplete taxa when data sets with different numbers of taxa are combined. Using subsampling experiments with molecular data from a known bacteriophage phylogeny (Hillis et al., 1992; Bull et al., 1993), they found that including incomplete taxa tended to decrease the overall phylogenetic accuracy of the estimated trees. However, they also found that this decrease was generally minor (except when taxa were 75% incomplete) and comparable to that found for including complete taxa.

I recently used simulations to address the mechanisms that cause incomplete taxa to be problematic (Wiens, 2002). Given that including highly incomplete taxa can lead to multiple trees, poorly resolved consensus trees, and decreased phylogenetic accuracy, what exactly causes this effect? Two hypotheses are that (1) it is the actual number or proportion of missing data cells (as implied by Huelsenbeck [1991] and other authors) or that (2) it is a problem of subsampling characters, such that too few characters have been scored in the incomplete taxa to accurately place them on the tree. These hypotheses can be distinguished by testing phylogenetic accuracy while simultaneously varying the proportion of missing data in the incomplete taxa and the overall number of characters in the analysis.

The results strongly support the second hypothesis (incomplete taxa are problematic because of the inclusion of too few complete characters rather than too many missing data cells). Given enough characters in the analysis, it is possible to have extremely accurate resolution when including taxa that are only 5% complete and that have nearly 2000 missing data cells each (Fig. 1). Clearly, the amount of missing data itself is not the actual problem. As long as enough characters are sampled in the incomplete taxa to accurately place them on the tree, then the amount of missing data seems to have little impact. This general result appears to be extremely robust to changes in the simulation parameters, including the number of taxa (16 vs. 64), tree shape (fully asymmetric vs. fully symmetric), type of data (binary vs. DNA), different ways of distributing missing data among characters (the same set of characters incomplete in every incomplete taxa vs. incomplete characters selected randomly in each incomplete taxon; Fig. 1), and different ways of distributing incomplete taxa on the tree (randomly selected taxa vs. evenly-spaced on the phylogeny). However, when branch lengths are extremely long it is difficult to increase accuracy when taxa are highly incomplete. The results also show that accurate placement of incomplete taxa is much easier when the missing data cells are confined to the same characters in all

Taxa

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	1	0	1	1	0	1	1	1	1	1
1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0
?	1	1	?	?	1	?	1	1	0	?	?	?	0	?	0
?	0	0	?	?	0	?	0	0	0	?	?	?	1	?	1
?	1	1	?	?	0	?	0	0	0	?	?	?	0	?	0
?	1	1	?	?	0	?	0	0	0	?	?	?	0	?	0
?	0	0	?	?	0	?	0	0	0	?	?	?	1	?	0

Taxa

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
?	1	0	0	0	0	?	0	0	0	0	0	?	0	1	1
0	0	0	?	?	0	?	0	0	1	?	0	?	1	1	1
?	0	0	0	?	0	?	0	0	0	?	0	0	0	?	0
?	0	0	?	0	0	1	0	1	1	0	?	1	1	?	1
1	1	1	?	1	0	0	0	0	0	0	0	?	0	0	0
?	1	1	?	?	1	1	1	1	0	?	?	0	0	0	0
0	0	0	0	0	0	?	0	0	0	0	?	?	1	?	1
1	1	1	1	?	0	0	0	0	0	?	?	?	0	?	0
?	1	1	?	?	0	0	0	0	0	1	1	0	0	0	0
0	0	0	1	1	0	?	0	0	0	?	?	0	1	?	0

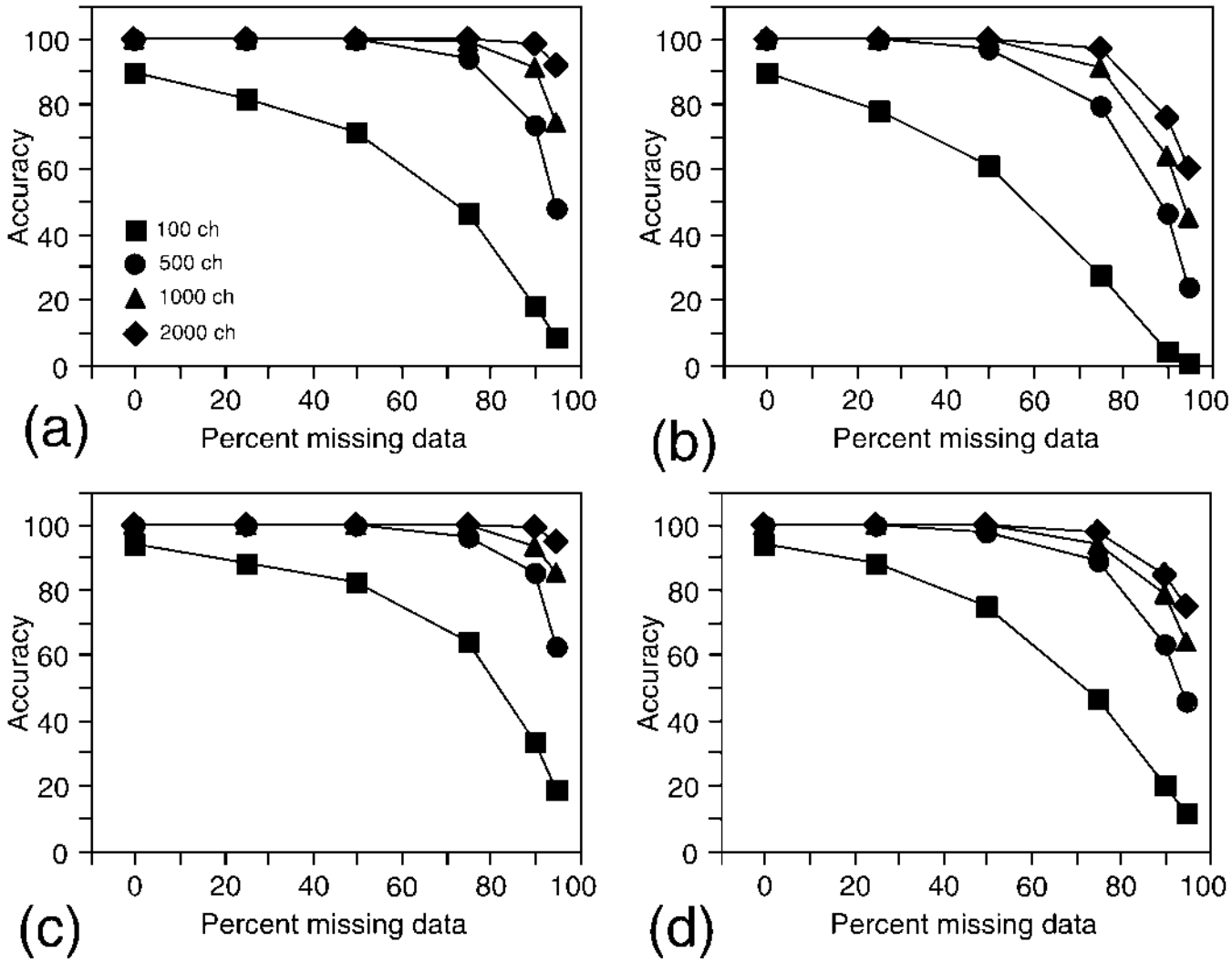

FIGURE 1. Phylogenetic accuracy for highly incomplete taxa depends on the number of characters scored in these taxa, not on the amount or proportion of missing data. Missing data cells are confined to the same set of characters in all incomplete taxa in **a** and **c** (see hypothetical example above graph), and are randomly distributed among characters for each incomplete taxon (**b** and **d**). In a and b, accuracy is measured as the number of correctly resolved nodes, whereas in c and d accuracy is based on the symmetric-difference distance between the true tree and a single randomly chosen tree from among the shortest trees from a given search. Results are based on 16 taxa, binary character data, a fully asymmetric tree, all branches with length 0.05, and with eight incomplete taxa selected randomly. Each point is the average accuracy from 100 replicated matrices. Standard errors for each mean are less than 2.5%, and are not shown. Modified from Wiens (2002).

taxa, rather than being randomly distributed amongst characters in each incomplete taxon.

Including Incomplete Characters

Another way to eliminate missing data cells from a matrix is to exclude those characters that contain any or too much missing data. This exclusion has been advocated by some authors (Livezey, 1989; Smith et al., 1995), but apparently is often used by paleontologists without being stated explicitly. For example, few paleontologists include characters from soft anatomy when analyzing relationships among fossil and recent taxa (but see Gauthier et al., 1988; Trueb and Cloutier, 1991). Yet few authors have explicitly stated why they would expect characters with missing data cells to be problematic, and (in contrast to exclusion of incomplete taxa) there are no empirical observations cited that would support this practice.

I have previously (Wiens, 1998b) used simulations to test whether including characters with abundant missing data increases or decreases phylogenetic accuracy. Two data sets for each 16-taxon tree were simulated, each with 50 characters. The first data set contained no missing data, whereas in the second data set, various taxa (either 4, 8, or 12) were randomly selected to have all 50 characters replaced with missing data. The accuracy of data set 1 alone was then compared to that based on combined analysis of datasets 1 and 2. For these conditions, adding the set of incomplete characters generally increased accuracy (Fig. 2), except when 12 of the 16 of the taxa were incomplete, in which case there was little change in accuracy (on average). The results suggest that increasing the amount of missing data is not harmful, but enough missing data robs the incomplete characters of their ability to improve phylogenetic accuracy. Thus, these results indicate that adding sets of incomplete characters is either beneficial or harmless. These basic results are robust to changes in branch lengths, tree shape, number of taxa and characters, and different ways of distributing missing data among taxa and characters.

Despite the general innocuousness of adding incomplete characters, this study also showed that certain ways of distributing missing data cells among taxa could lead to problems of long-branch attraction (LBA) in the sets of incomplete characters. Long branch attraction (Felsenstein, 1978) typically occurs when there are two or more unrelated, long, terminal branches separated by one or more short internal branches (Fig. 3). In this context, ''long'' means that there is a relatively high probability that each character will change along that branch, given a stochastic model of evolution. By chance, many of the changes that occur on the long branches will be parallel changes shared between the long branches. These parallel changes will be interpreted as synapomorphies by parsimony (Fig. 3). Parsimony (and other methods) will therefore tend to estimate trees in which the long branches are placed together, even though these long branches are not sister clades (Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995; Wiens and Servedio, 1998). The short internal branch separating the long terminal branches makes it easier for this misleading signal to overcome the true signal, because the short branch will have relatively few correct synapomorphies. LBA is a particularly serious problem because adding characters to the analysis only increases the probability that the incorrect tree (placing the long branches together) will be reconstructed, at least when using parsimony (Felsenstein, 1978).

A common way to create long branches is by including distantly related taxa in a phylogenetic analysis. Failing to sample the phylogenetically intermediate taxa along the branches that separate distantly related taxa will typically make these branches ''long'' (e.g., Hendy and Penny, 1989; Graybeal, 1998; Hillis, 1998). Conversely, adding taxa can potentially subdivide

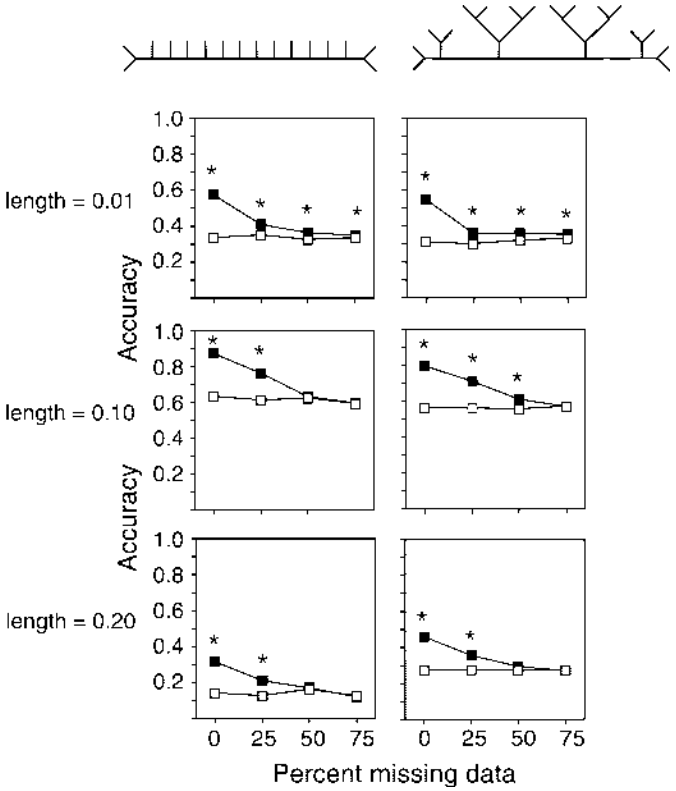

FIGURE 2. Adding sets of incomplete characters can increase phylogenetic accuracy. Each data set contains 50 binary characters and 16 taxa, and each data point is the average accuracy from 100 replicates. □ = data set 1 alone (no missing data); ■ = data sets 1 and 2 combined (with missing data). Accuracy is measured as the proportion of correctly resolved clades (based on a strict consensus tree when multiple equally parsimonious trees are generated from a search). Asterisks denote $P < 0.01$ for paired *t*-tests of accuracy with and without addition of the second data set. Modified from Wiens (1998b).

long branches and ''rescue'' an analysis from the effects of LBA. Long-branch attraction that is caused by limited taxon sampling is a particularly serious concern for molecular studies, because phylogenetically intermediate fossil taxa generally cannot be sampled and because large numbers of characters may not prevent an analysis from being misled (especially when using parsimony). Coding phylogenetically intermediate taxa with missing data cells can also create long branches among the taxa that are complete (Wiens, 1998b), potentially leading to problems of LBA (Fig. 4). This makes the addition of incomplete sets of characters potentially problematic. A limited set of simulations, however, showed that even under conditions where LBA was maximized, the overall accuracy of the trees was not greatly reduced by including incomplete characters (Wiens, 1998b).

Unresolved Questions

Several fundamental questions about the effects of missing data on phylogenetic accuracy have yet to be adequately addressed using simulations. First, given that missing data can be eliminated from a matrix by deleting incomplete taxa or characters, which approach gives the most accurate results, and under what conditions? Or is it better to simply include all the taxa and characters regardless of the amount of missing data? Second, under what conditions does including (or excluding) incomplete taxa increase or decrease phylogenetic accuracy?

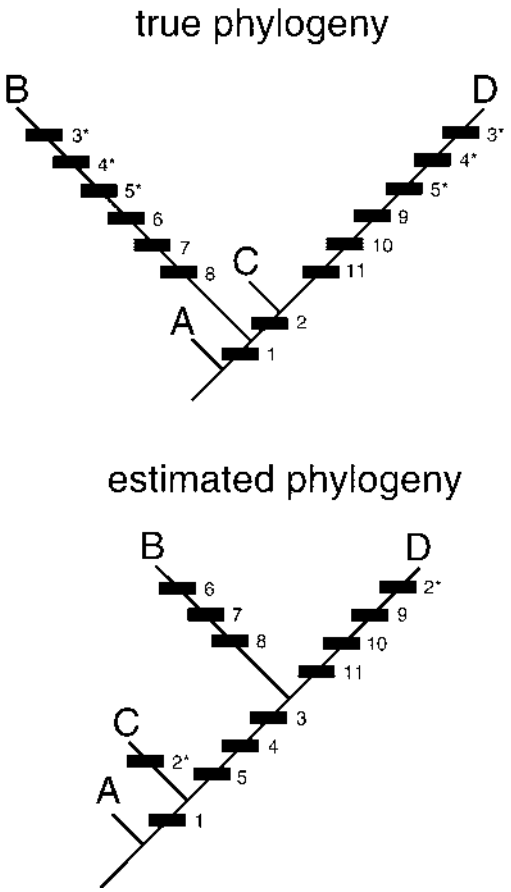

FIGURE 3. Hypothetical example illustrating the problem of long-branch attraction (LBA). Numbered bars indicate character-state changes from 0 to 1 for 11 characters. Taxa B and D are associated with long branches, meaning that there is a high probability of each character changing on these branches. By chance, these long branches accumulate many parallel changes (characters 3–5). A parsimony analysis will treat these parallel changes as synapomorphies and incorrectly group B and D together. Characters that change more than once on a given tree are asterisked. The top tree shows actual character changes, the bottom tree shows reconstructed character changes.

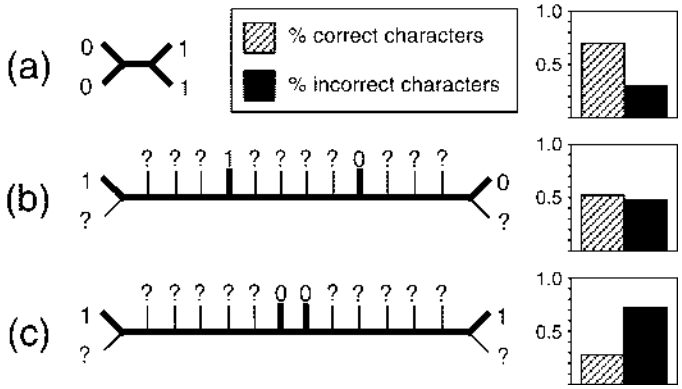

FIGURE 4. Missing data causes long-branch attraction (LBA) by increasing the proportion of parsimony-informative characters that support the incorrect topology versus the correct tree (for the four complete taxa). Each graph is based on a sample of 500 characters. (**a**), four-taxon unrooted tree with all branches of length (0.10). (**b**), 16-taxon unrooted tree with all branches of length 0.10, with 12 taxa incomplete. (**c**), 16-taxon unrooted tree with all branches of length 0.10, with 12 taxa made incomplete so as to maximize LBA. Modified from Wiens (1998b).

Third, how does the accuracy of the complete taxa alone compare to the overall accuracy of the tree? Do the incomplete taxa actually overturn relationships among the complete taxa? Can adding incomplete taxa ''rescue'' an analysis from LBA? Some of these questions were briefly addressed in a limited set of simulations by Wiens (1998b). In this paper, I expand on these simulations.

SIMULATION METHODS

Simulation methods generally followed Wiens (1998b). A 16-taxon fully asymmetric tree was simulated, with binary character data (the majority of characters in most morphological data sets seem to be binary). Use of a fully asymmetric tree facilitated exploring conditions associated with LBA — it is easier to manipulate the relative lengths of the branches connecting complete taxa when using an asymmetric tree, at least for a limited number of taxa. Previous simulations suggest that number of taxa, tree shape, and type of character data (binary vs. multistate) do not greatly impact the results when including incomplete taxa and characters (Wiens, 1998b, 2002). Characters were simulated in two separate datasets. One data set consisted entirely of complete characters and taxa, the other contained taxa and characters selected to be incomplete.

Different numbers of taxa were chosen to be incomplete (4, 8, and 12), representing the addition of sets of characters that were 25, 50, and 75% incomplete (from Wiens, 1998b). Incomplete taxa were distributed in two general ways on the known, model tree: (1) evenly spaced on the tree, and (2) maximizing LBA. The first case was examined as a baseline for all methods, whereas the second case (LBA) should be a worst-case scenario for inclusion of incomplete characters and a best-case scenario for including incomplete taxa (because taxa that are added should tend to greatly increase accuracy by breaking up these long branches). There are several ways that incomplete taxa could potentially be distributed on the tree to create LBA; the distributions used herein were intended to cause the maximum decrease in the overall accuracy of the tree. Note that the first case is not a best-case scenario for all methods, but merely a reasonable ''neutral'' starting point for comparisons.

Different levels of completeness were explored for the incomplete taxa (95, 90, 75, 50, 25 and 0% missing data). These correspond to different ratios of characters in datasets 1 and 2. For example, 95% missing data corresponds to 5 characters in data set 1 (the complete data set) for every 95 characters in data set 2. Missing data cells were distributed among characters in two different ways (Fig. 1). In the first, the missing data cells were confined to the same set of characters in all incomplete taxa (as in an analysis including both fossil and extant vertebrate taxa where one data set consists of osteological characters and the other of soft anatomical characters). In the second, the prespecified number of missing data cells was distributed randomly among all characters in both data sets and different characters were chosen to be incomplete in each taxon (corresponding to the random preservation of parts in fossil taxa). Presumably, in the real world, each character has a different probability of being preserved in a fossil taxon, and the two cases simulated represent two extremes in a continuum of preservation probabilities ranging from purely stochastic to entirely predetermined.

All branches of the 16-taxon trees were set to equal length, in order to facilitate comparing the effects of different branch lengths on the results. For the purposes of this paper, I define branch length as the probability of a character changing state

(i.e., 0 to 1 or 1 to 0) from the beginning of the branch to the end. All characters were assumed to evolve at the same rate, and changes from 0 to 1 and 1 to 0 were considered equally likely. Two different branch lengths were explored (0.05 and 0.20). Previous work (Wiens, 1998b) suggests that these two lengths represent ''easy'' and ''hard'' conditions for accurate phylogeny estimation using parsimony, given 16 taxa and binary character data. Levels of homoplasy are low at a length of 0.05 and high at a length of 0.20. All characters were included regardless of whether or not they were parsimony informative. Under these conditions, about 52% of the characters are parsimony-informative at a length of 0.05 and about 98% are informative at a length of 0.20.

Several different numbers of characters were explored (100, 500, 1000, 2000). Clearly, most paleontological analyses have far fewer than 1,000 parsimony-informative characters. However, exploring these very large numbers of characters allowed some insight into the consistency of the methods under these conditions. A method is consistent under a given set of conditions if it can recover the correct phylogeny with an infinite number of characters (Felsenstein, 1978). Under some conditions (such as LBA), methods will be inconsistent, and converge on the wrong answer as more and more characters are sampled. The unusually large numbers of characters can help distinguish these scenarios, and differentiate errors caused by inconsistency and those caused merely by undersampling characters.

The effect of the temporal position of the incomplete taxa was also explored, following Huelsenbeck (1991). Two extreme conditions were examined, one in which all taxa were living and the other in which the incomplete (fossil) taxa were treated as retaining all the states of their direct ancestors (i.e., no change was simulated from the beginning to the end of the branch for these taxa). The latter represents the optimal temporal position for fossil taxa (Huelsenbeck, 1991).

Two hundred replicates were generated and analyzed for each set of conditions. Previous analyses (Wiens, 1998b) suggest that standard errors in method performance are extremely low, even with only 100 replicates. Simulated datasets were analyzed using Swofford's (2001) PAUP* program, version 4.0b8. Parsimony analyses utilized heuristic searches with TBR branch swapping, and 20 random addition sequence replicates per search.

Four approaches were compared for each simulation replicate: (1) the complete characters (dataset 1) analyzed alone, excluding all incomplete characters (except when missing data are distributed randomly among all characters); (2) all taxa and characters, both complete and incomplete, analyzed together; (3) all taxa and characters included, but with the incomplete taxa pruned from the tree after it is reconstructed (following Swofford and Olsen, 1990; Wiens and Reeder, 1995); and (4) all characters included, but with the complete taxa analyzed alone.

Accuracy was measured as the proportion of clades that are correctly resolved from a given analysis, averaged across the 200 replicated matrices for a given set of conditions. Measuring accuracy when multiple equally parsimonious trees are generated from a search may be handled in several different ways in simulation studies (Hillis, 1995; Rannala et al., 1998). For this study, the accuracy of a given approach was assessed using a single shortest tree from each search. This method gives similar results to basing accuracy on a strict consensus of the shortest trees from a search (see Fig. 1 of this study and Wiens, 1998b, 2002), but should be less biased by highly incomplete taxa that give poorly resolved trees. This approach should also approximate accuracy based on an average of the shortest trees from a given search (Rannala et al., 1998).

RESULTS

Baseline Simulations

The basic results of the simulations are shown in Figure 5, in which all taxa are of equal age (extant) and the missing data cells are confined to the same set of characters in all incomplete taxa. The overall results are not simple to describe, because the relative accuracy of different approaches (e.g., including or excluding taxa or characters) differs considerably depending on the simulated conditions (branch lengths, distribution of incomplete taxa, etc.). Most importantly, the relative accuracy of the approaches depends on whether there is long-branch attraction (LBA) among the complete taxa for a given set of conditions. The presence of LBA is indicated when a method consistently yields highly inaccurate results, even as more characters are added. For example, analyzing the complete taxa alone (method 4) has an accuracy of less than 25% under certain conditions (Fig. 5b, d, bl = 0.20, and 5f), even though 2,000 characters are included and other methods consistently recover the true phylogeny under the same conditions. The presence of LBA depends on the combination of branch lengths (usually high), number of incomplete taxa (many), and the distribution of incomplete taxa on the true phylogeny (creating a combination of long terminal branches and short internal branches).

What is the best approach for dealing with incomplete data based on these results? Under conditions where there is no LBA (i.e., all methods recover the correct tree given enough characters), the most generally accurate method is the one in which the incomplete taxa and characters are all included, but the incomplete taxa are pruned from the tree after the analysis (method 3 in Figs. 5–8). Similar results are obtained from simply excluding the incomplete taxa entirely (method 4), but the pruning method (method 3) performs better when branches are relatively long. When the overall number of characters in the analysis is low and the incomplete taxa have a high proportion of missing data, accuracy based on the complete characters alone (method 1) or including all of the incomplete taxa and characters (method 2) is relatively low.

In contrast, under conditions where there is LBA (Fig. 5f, d where bl = 0.20), analyzing the set of complete characters alone (method 1) gives the most accurate results and excluding incomplete taxa (method 4) tends to give very inaccurate results. Under these conditions, the accuracy of trees in which incomplete taxa are included (method 3) or included and pruned (method 4) is also relatively low, unless the incomplete taxa are relatively complete.

These results can be explained by comparing accuracy based on the complete taxa alone (method 4) to those obtained when including the incomplete taxa (methods 2 and 3). These comparisons also shed light on how incomplete taxa affect overall phylogenetic accuracy and the estimated relationships among the complete taxa. Under conditions where the incomplete taxa have a high proportion of missing data, the overall number of characters is small, and there is no LBA, the accuracy for the entire tree is low when the incomplete taxa are included. However, accuracy is high for the pruned tree and for the tree based on the complete taxa alone. This comparison shows that the low accuracy associated with including incomplete taxa is caused by incorrect (or ambiguous) placement of the incomplete taxa, and that incorrect placement of the incomplete taxa does not adversely affect estimated relationships among the complete taxa.

Under conditions where there is LBA, accuracy based on the complete taxa alone is consistently low. When incomplete taxa are included but subsequently pruned from the tree (method 3) accuracy is low when the incomplete taxa are highly incomplete (i.e., 75–95% missing data), but can be high when the incomplete taxa are more complete (25–50% missing data). Thus, the

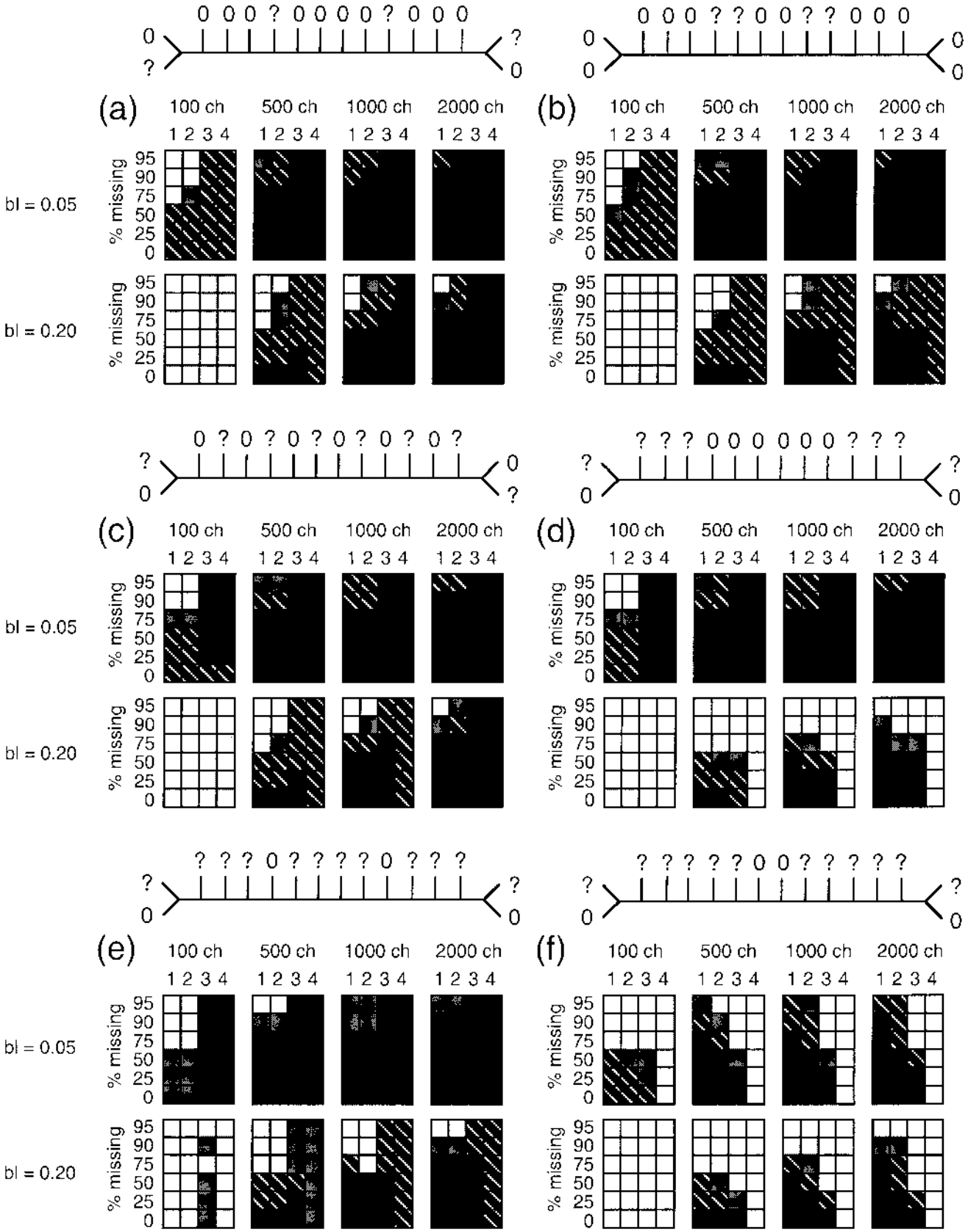

FIGURE 5. Accuracy of different approaches for dealing with missing data, where incomplete taxa are extant and missing data cells are confined to the same characters in all incomplete taxa: 1 = data set 1 alone (no missing data); 2 = data sets 1 and 2 combined (with missing data); 3 = data sets 1 and 2 combined, accuracy based only on complete taxa (incomplete taxa pruned subsequent to the analysis); 4 = data sets 1 and 2 combined, complete taxa only analyzed. Different shadings represent the accuracy of each method (average of 200 replicates): □ = 0–25%; □ = 26–50%; ■ = 51–75%; ■ = 76–94%; ■ = 95–100%. **a** and **b** = four taxa incomplete; **c** and **d** = eight taxa incomplete; **e** and **f** = 12 taxa incomplete. For a, c, and e, incomplete taxa are evenly spaced on the true tree; for b, d, and f, incomplete taxa are distributed on the tree so as to maximize LBA.

addition of incomplete taxa can break up long branches and ''rescue'' an analysis from the effects of LBA. However, doing so requires that these taxa have a certain level of completeness. The exact level of completeness needed to break up these branches varies, depending on the number and distribution of missing data cells and on the overall branch lengths.

In general, adding the set of incomplete characters (method 2) increases accuracy relative to analyzing the set of complete characters alone (method 4). The major exception is when there is LBA in the set of incomplete characters, as indicated by low accuracy in analyses of the complete taxa alone. In these cases, adding the set of incomplete characters may decrease the overall accuracy of the estimated trees. In the scenario where there is LBA and the added characters are highly incomplete (Fig. 3f, bl = 0.20), the negative effects of adding these characters appear to be mostly confined to the complete taxa, because the accuracy for the complete taxa is very low whereas the overall accuracy for the entire tree is surprisingly high.

Effects of Temporal Position and Distribution of Missing Data

In general, the same basic results described above (Fig. 5) are obtained regardless of the temporal position of the incomplete taxa or how the missing data cells are distributed among characters (Figs. 6–8). Nevertheless, some differences are present between the baseline results and those in which these two parameters are varied. First, when the incomplete taxa have the oldest and optimal temporal position (i.e., they retain all the character states of their direct ancestors), methods that include the incomplete taxa have higher accuracy than when all the taxa are extant (Fig. 6). This difference is most obvious at the higher branch length (bl = 0.20).

When the incomplete taxa have their missing data cells randomly distributed among characters (Fig. 7), the results are similar to those in which the missing data are confined to the same characters in all incomplete taxa (Fig. 5). However, the methods that include incomplete taxa tend to be less accurate when the missing data cells are randomly distributed, particularly when the taxa are highly incomplete (see also Fig. 1). This reduction in accuracy is not present in the pruned tree, which suggests that the reduced accuracy is caused by a higher rate of error in the placement of the incomplete taxa. The results in which missing data cells are distributed randomly also differ in that accuracy is somewhat higher for the pruned taxa in some cases where there is LBA (Fig. 7b, f). In these cases, the random distribution of missing data cells may dilute the negative signal caused by LBA (because fewer characters share the same distribution among taxa). Note that under conditions where the missing data cells are randomly distributed among characters, removing missing data cells by analyzing data set 1 alone is not really an option (even though results for data set 1 alone are shown in Figs. 7, 8) because missing data cells are randomly distributed across both data sets.

When the incomplete taxa are direct ancestors and the missing data are distributed randomly among characters (Fig. 8), the results remain similar to those in Figure 5, in which all taxa are extant and missing data cells are confined to the same characters in all taxa. However, accuracy is relatively lower when all taxa and characters are included (presumably caused by the random distribution of missing data among characters) and relatively higher for the pruned approach in cases of LBA, presumably from the combination of beneficial temporal position and the dilution of false signal by the random distribution of missing data cells.

DISCUSSION

What is the Missing Data Problem?

The results of recent simulation studies suggest that there is no single ''missing data problem.'' In fact, there are really two problems that are associated with missing data: limited character sampling (when incomplete taxa are added) and limited taxon sampling (when incomplete characters are added). In both cases, it is not the missing data cells themselves that create these problems. Addition of incomplete taxa and characters can clearly reduce phylogenetic accuracy (relative to excluding them) under some circumstances, just as they can improve accuracy under others. When incomplete characters are added, certain distributions of missing data cells among taxa can reduce accuracy through LBA. When incomplete taxa are added, the overall accuracy of the tree may be decreased by the unresolved or incorrect placement of these taxa, especially when there are few characters in the analysis and the missing data cells are distributed randomly among characters. For both incomplete taxa and characters, the amount of missing data itself is not the critical factor. For example, adding highly incomplete characters may have little negative effect if the missing data cells are evenly distributed among taxa, and highly incomplete taxa can be placed on the tree correctly if many characters are sampled overall.

Implications of New Simulation Results

The new simulation results presented in this study offer insights into how incomplete taxa and characters affect phylogenetic analyses. The results suggest that adding highly incomplete taxa may often have little impact on the accuracy of relationships among the complete taxa. On the positive side, when highly incomplete taxa are added that reduce the overall accuracy of the tree (as in these simulations when only 100 characters are sampled overall), the relationships among the complete taxa may be largely unaffected and may be reconstructed very accurately by excluding or pruning out the incomplete taxa. On the negative side, highly incomplete taxa may have relatively little ability to improve the estimated relationships among the more complete taxa. For example, taxa that are highly incomplete may be unable to break up long branches affected by LBA. The exact level of completeness that determines whether taxa will be effective at subdividing long branches seems to depend on several factors, such as overall branch lengths, temporal position, and the distribution of missing data cells among characters in the incomplete taxa. Thus, the limited completeness of a taxon is not a constraint on whether it can be accurately placed in an analysis, but may be a constraint on whether or not it will improve the estimate of relationships among the complete taxa. Although superficially paradoxical, these findings do make intuitive sense. In theory, a taxon can be correctly placed on the tree by only a single synapomorphy, whereas improving the estimate of phylogeny, especially when ''rescuing'' an analysis from LBA, may require overcoming conflicting signal from the set of complete characters (thus, the relative number of characters is important).

These results also demonstrate why highly incomplete sets of characters may have little impact on the overall accuracy of estimated trees (Wiens, 1998b), even though highly incomplete characters may be those most likely to show the effects of LBA. If only a few taxa are complete, it appears that relationships among these few complete taxa can be entirely wrong, but that the overall accuracy of the tree can still be relatively high because the majority of taxa are unaffected by LBA (Fig. 3f, bl = 0.20). These results suggest that adding sets of highly incomplete characters may adversely affect the estimated relationships among only a limited number of clades, leaving many clades

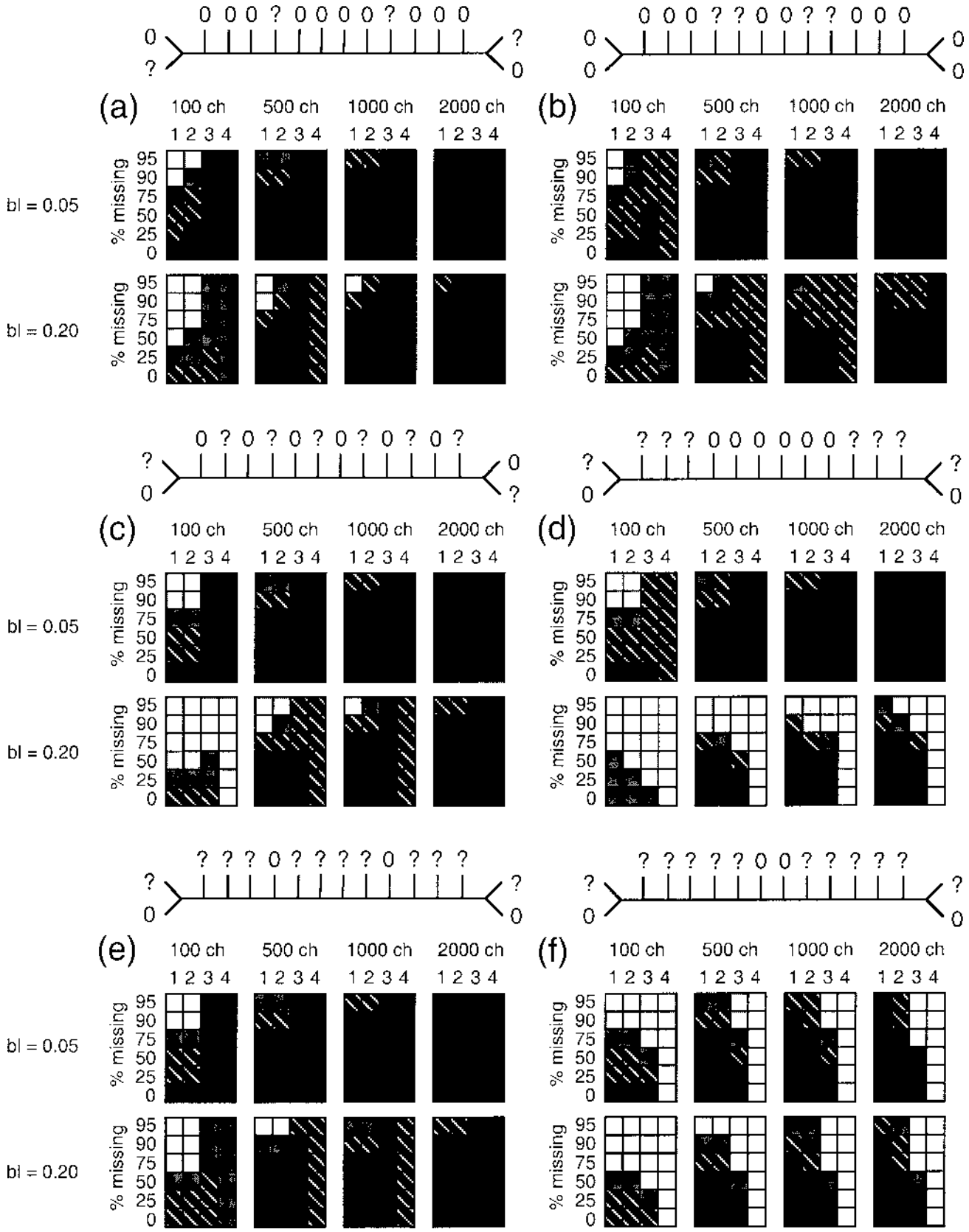

FIGURE 6. Accuracy of different approaches for dealing with missing data, where incomplete taxa have optimal temporal position (retaining all the states of their direct ancestors) and missing data cells are confined to the same characters in all incomplete taxa: 1 = data set 1 alone (no missing data); 2 = data sets 1 and 2 combined (with missing data); 3 = data sets 1 and 2 combined, accuracy based only on complete taxa (incomplete taxa pruned subsequent to the analysis); 4 = data sets 1 and 2 combined, complete taxa only analyzed. Different shadings represent the accuracy of each method (average of 200 replicates): □ = 0–25%; □ = 26–50%; ■ = 51–75%; ■ = 76–94%; ■ = 95–100%. **a** and **b** = four taxa incomplete; **c** and **d** = eight taxa incomplete; **e** and **f** = 12 taxa incomplete. For a, c, and e, incomplete taxa are evenly spaced on the true tree; for b, d, and f, incomplete taxa are distributed on the tree so as to maximize LBA.

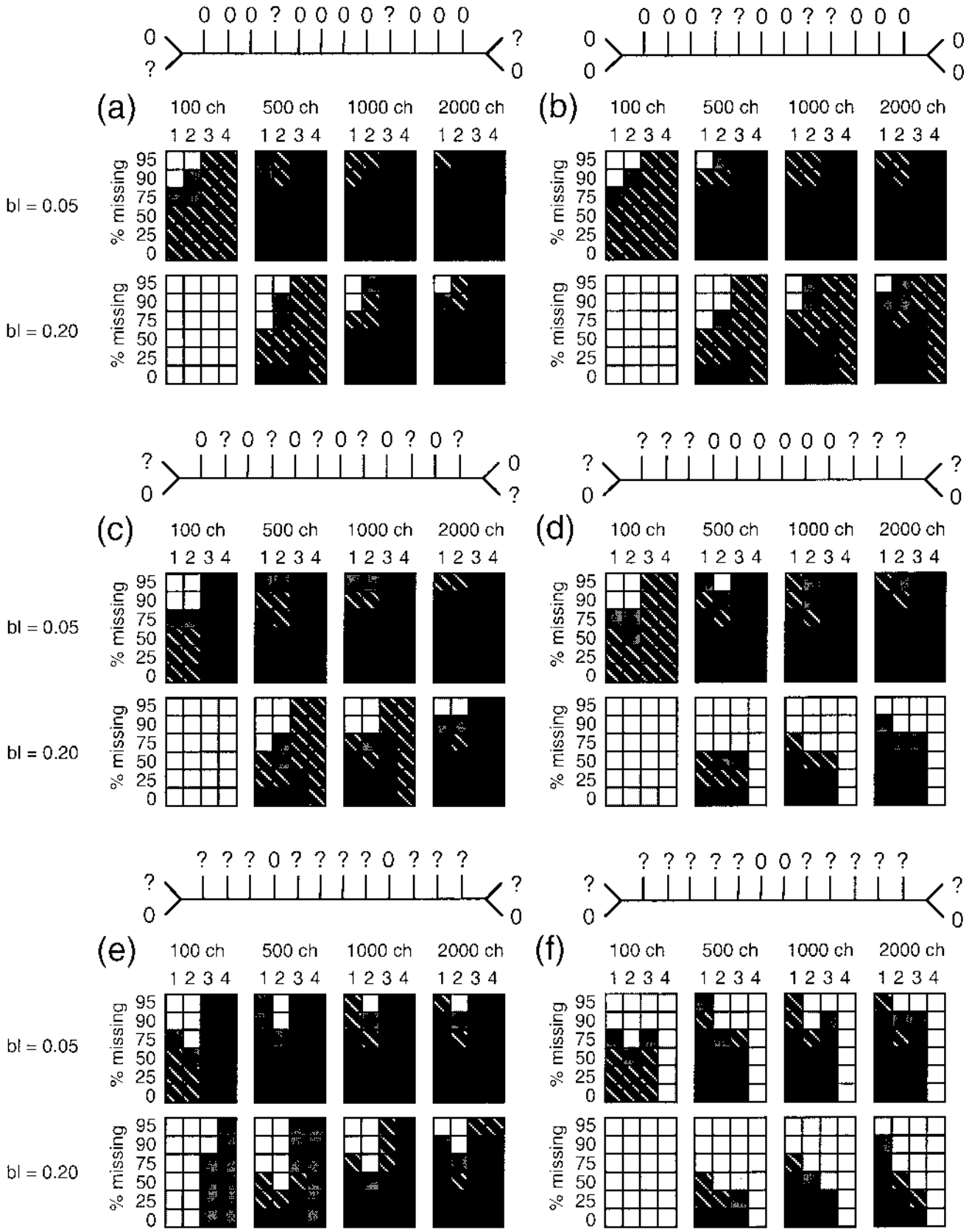

FIGURE 7. Accuracy of different approaches for dealing with missing data, where incomplete taxa are extant and missing data cells are randomly distributed among characters for each incomplete taxon: 1 = data set 1 alone (no missing data); 2 = data sets 1 and 2 combined (with missing data); 3 = data sets 1 and 2 combined, accuracy based only on complete taxa (incomplete taxa pruned subsequent to the analysis); 4 = data sets 1 and 2 combined, complete taxa only analyzed. Different shadings represent the accuracy of each method (average of 200 replicates): □ = 0–25%; □ = 26–50%; ■ = 51–75%; ■ = 76–94%; ■ = 95–100%. **a** and **b** = four taxa incomplete; **c** and **d** = eight taxa incomplete; **e** and **f** = 12 taxa incomplete. For a, c, and e, incomplete taxa are evenly spaced on the true tree; for b, d, and f, incomplete taxa are distributed on the tree so as to maximize LBA.

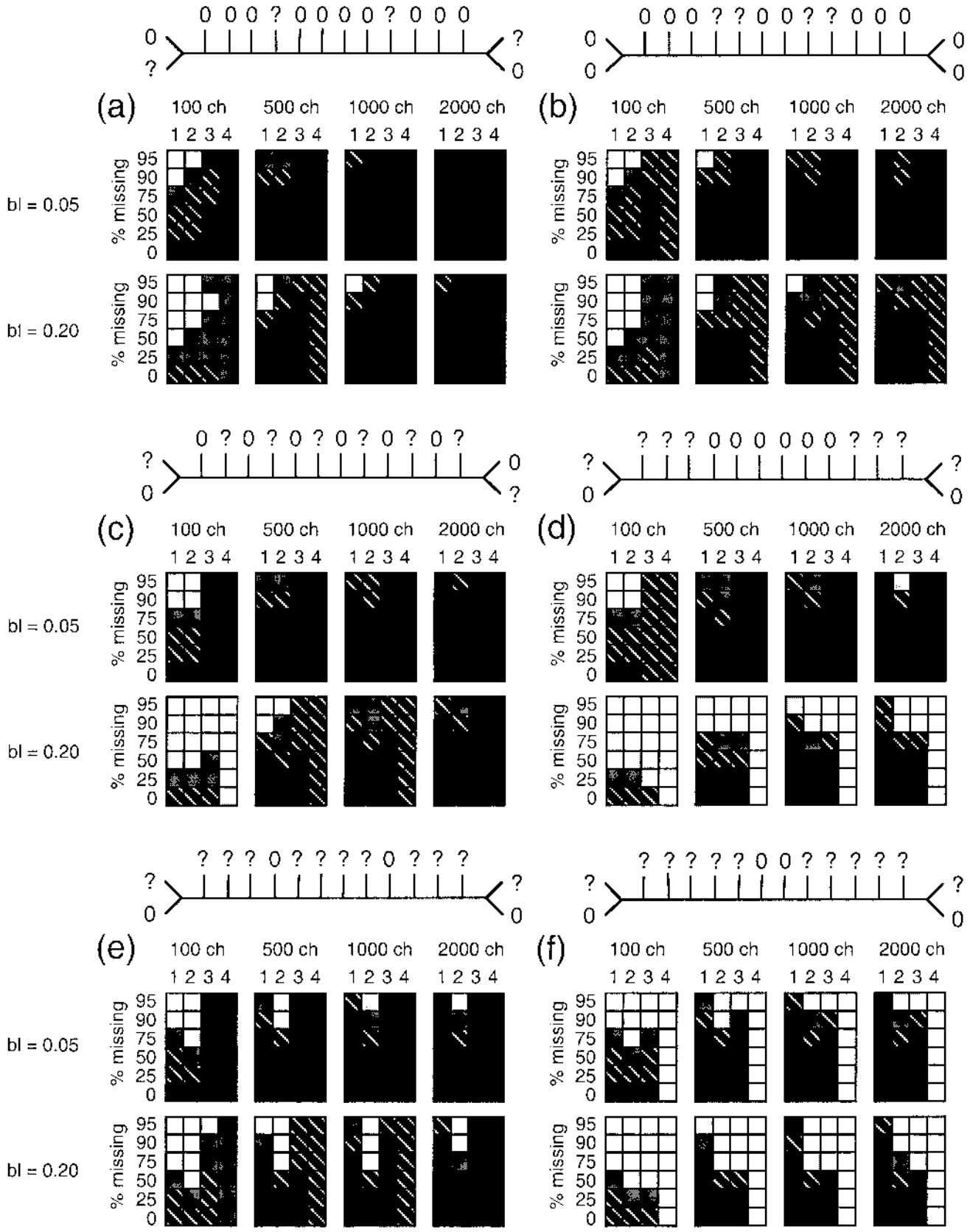

FIGURE 8. Accuracy of different approaches for dealing with missing data, where incomplete taxa have optimal temporal position (retaining all the states of their direct ancestors) and missing data cells are randomly distributed among characters for each incomplete taxon: 1 = data set 1 alone (no missing data); 2 = data sets 1 and 2 combined (with missing data); 3 = data sets 1 and 2 combined, accuracy based only on complete taxa (incomplete taxa pruned subsequent to the analysis); 4 = data sets 1 and 2 combined, complete taxa only analyzed. Different shadings represent the accuracy of each method (average of 200 replicates): □ = 0–25%; □ = 26–50%; ■ = 51–75%; ■ = 76–94%; ■ = 95–100%. **a** and **b** = four taxa incomplete; **c** and **d** = eight taxa incomplete; **e** and **f** = 12 taxa incomplete. For a, c, and e, incomplete taxa are evenly spaced on the true tree; for b, d, and f, incomplete taxa are distributed on the tree so as to maximize LBA.

unaffected. Conversely, when the highly incomplete characters are not affected by LBA, adding these data seems to improve accuracy only for the few complete taxa, and increases the overall accuracy of the tree only slightly (Wiens, 1998b; Fig. 8).

The results presented in this study are based on a relatively limited set of conditions. Based on previous simulation studies, most of the parameters that were not varied in the present study seem unlikely to significantly change the results or conclusions, such as number of taxa, type of character data, other branch lengths, and tree shape (Wiens, 1998b, 2002). On the other hand, taxon sampling regimes are particularly important, and in this study they were designed to contrast scenarios with and without LBA. Scenarios with LBA were intended to highlight cases where adding taxa may have the most beneficial effects. However, some studies have shown that including additional taxa can actually exacerbate or create problems of LBA (e.g., Kim, 1996; Poe and Swofford, 1999). Problems of LBA are typically caused by certain combinations of short and long branches, and adding taxa may subdivide and shorten some long branches in such a way as to make other long branches attract. Thus, in empirical studies, the observation that an incomplete taxon is added and relationships among the complete taxa are changed may not necessarily indicate that accuracy has been improved. Although simulation studies have so far yielded conflicting results concerning the potential benefits of adding taxa, most studies have added taxa in a fairly limited set of branch length scenarios (i.e., Graybeal, 1998; Hillis, 1998; Poe and Swofford, 1999). This is an area in need of further study.

Recommendations for Empirical Studies

Finally, given these results, how should empirical workers deal with missing data in phylogenetic analyses? The results suggest that choosing the best approach (e.g., including or excluding taxa or characters) is not simple, because the relative accuracies of different approaches can vary dramatically, depending on parameters that may be difficult to estimate with empirical data sets (e.g., the distribution of incomplete taxa on the true phylogeny). However, these results do provide some basis for devising strategies for empirical studies.

In this study, the most generally accurate method for dealing with missing data was the one in which all characters and taxa are included, but the incomplete taxa are pruned from the tree after the analysis. This pruning approach can be accomplished easily (using PAUP*) by generating trees based on all the data and then deleting taxa with the ''prune deleted taxa from trees'' option. The pruning approach has the advantage of including all taxa to help subdivide long branches (unlike approaches which exclude taxa a priori), without allowing the ambiguous or incorrect placement of the highly incomplete taxa to obscure relationships among the more complete taxa. Surprisingly, this approach is rarely (if ever) used in empirical phylogenetic analyses. A disadvantage of this approach is that it does not attempt to address the relationships of the incomplete taxa (Wiens and Reeder, 1995). Thus, the pruning approach may be most useful when attempting to generate an accurate phylogeny for a select set of taxa, rather than for every species in the clade of interest.

Recent simulation results (Wiens, 2002) suggest that highly incomplete taxa can be included and accurately placed in phylogenetic analyses, given enough overall characters in the analysis. In fact, the level of completeness seems to be a poor criterion for deciding whether or not to include a taxon (Donoghue et al., 1989; Novacek, 1992; Kearney, 2002), and alternate approaches to taxon deletion have been developed that do not depend on the amount of missing data alone (Wilkinson, 1995; Anderson, 2001). A much better criterion may be the number of characters that can be scored in the incomplete taxa, particularly those characters that can be scored consistently for all taxa in the analysis. But even though highly incomplete taxa can be included and accurately placed on reconstructed phylogenies, they may not be able to improve the estimated relationships among the more complete taxa. Nevertheless, including and resolving the relationships of these incomplete taxa may be an important goal in itself, given that we would someday like a complete and accurate picture of the entire Tree of Life for both living and fossil taxa. The best way to include these taxa and resolve their relationships may be to increase the number of characters for which they are scored (Wiens, 2002). Although adding characters may not be possible in some cases, extracting more information from the characters that are already available may also improve the chances of accurately reconstructing the relationships of these highly incomplete taxa. For example, more information (and phylogenetic resolution) may be extracted from many morphological characters by coding and analyzing them directly as continuous variables rather than qualitative characters (Wiens, 2001).

Adding sets of incomplete characters may increase or decrease accuracy, depending on whether or not there is LBA, but it is hard to tell if there is LBA without knowing aspects of the true tree in advance. In the simulations presented here, LBA is most likely when the added characters are highly incomplete. Fortunately, these are also conditions where adding the incomplete characters will have the least effect on the overall accuracy of the tree. Adding highly incomplete sets of characters seems to affect mostly those taxa that are complete, for better or for worse. Thus, the greatest change in accuracy will come from adding sets of characters that apply to more taxa, and increased taxon sampling should generally decrease chances of LBA by decreasing the average lengths of branches connecting the included taxa (but see Poe and Swofford, 1999). It should also be noted that the simulations in this study were designed to include a worst-case scenario for including incomplete characters, and that this scenario may be relatively unusual. For example, simulations based on randomly distributing incomplete taxa on 16 and 64-taxon phylogenies (Wiens, 1998b) suggest that adding incomplete characters either improves or has little effect on accuracy, and should be either beneficial or mostly harmless. In no case did adding incomplete characters significantly decrease accuracy, but under many conditions adding incomplete characters significantly increased accuracy.

Conclusions and Prospectus

The long term goal of phylogenetics, both neontological and paleontological, is to reconstruct an accurate phylogeny for all species of living and fossil organisms. The problem of missing data has been considered to be the major obstacle to accurately reconstructing the phylogeny of fossil taxa and their relationships to living taxa. Recent simulation studies show that there is not a single missing data problem. Instead there are potentially two problems, depending on how missing data cells are added to a phylogenetic data matrix. Adding incomplete taxa can be problematic because of sampling too few characters in these taxa to accurately place them on the tree, whereas adding incomplete characters can be problematic because limited taxon sampling for these characters may cause long-branch attraction. In neither case are the missing data cells by themselves misleading, and the number and proportion of missing data cells may be a poor indicator of cases where adding incomplete taxa or characters will be problematic. Identifying the mechanisms that may cause incomplete taxa and characters to be problematic is an important step in devising effective solutions.

The major conclusions from the new simulation results of this study are as follows. First, including highly incomplete taxa (such as fragmentary fossil taxa) may decrease the overall accuracy of estimated trees, but generally does not adversely af-

fect relationships among the complete taxa. Second, when analyses of complete taxa alone are misled by inadequate taxon sampling and LBA, adding incomplete taxa may subdivide long branches and ''rescue'' the analysis, but the ability of incomplete taxa to do so depends on their level of completeness (even though the accurate placement of these incomplete taxa does not). Third, adding sets of incomplete characters to a set of complete characters generally increases or has little effect on accuracy, unless the incomplete characters are affected by LBA. When the added characters are highly incomplete their effects (both negative and positive) may be confined largely to the complete taxa.

Given these results, what are the prospects for reconstructing an accurate phylogeny of all fossil and living taxa, particularly one based on both molecular and morphological data? Several studies have combined data from molecular and morphological characters from fossil and living taxa to estimate higher-level relationships of some groups (e.g., Eernisse and Kluge, 1993; Wheeler et al., 1993; O'Leary, 1999; Gao and Shubin, 2001; Sun et al., 2002), which requires coding fossil taxa as missing for the many characters in the molecular data sets. Recent simulation results suggest that the relative incompleteness of fossil taxa should not limit their accurate phylogenetic placement in such combined analyses. Instead, accuracy may be mostly limited by the number (and informativeness) of the characters that can be scored in these taxa. If the fossil taxa can be accurately placed in an analysis of the morphological data alone, they should be accurately placed in the combined analyses as well, regardless of their relative level of incompleteness when the molecular data are added. However, the simulation results also suggest that highly incomplete fossil taxa may have little ability to influence relationships among the more complete taxa. Thus, even though it may be possible to accurately place fossil taxa in a combined-data analysis, the fossil taxa may be ineffective at ''rescuing'' an analysis that has been misled by limited taxon sampling and associated LBA in the molecular data. In summary, the accuracy of phylogenetic analyses that combine fossil taxa and molecular characters should not be limited by the missing data cells alone. Instead, the success of these combined-data analyses may hinge on how accurate the molecular and morphological data sets are when analyzed separately, with the sampling of characters in the fossil taxa and the sampling of taxa for the molecular data sets being especially important.

ACKNOWLEDGMENTS

I thank Jim Clark and Peter Makovicky for inviting me to contribute participate in the symposium on missing data at the 2000 meetings of the Society for Vertebrate Paleontology in Mexico City, and thanks to Maureen Kearney for delivering my oral presentation there. I am grateful to Olaf R. Bininda-Emonds and Maureen Kearney for helpful reviews of the manuscript.

LITERATURE CITED

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

Ax, P. 1987. The Phylogenetic System: The Systematization of Organisms on the Basis of Their Phylogenesis. Wiley, New York, pp.

Bull, J. J., C. W. Cunningham, I. J. Molineux, M. R. Badgett, and D. M. Hillis. 1993. Experimental molecular evolution of bacteriophage T7. Evolution 47:993–1,007.

Cunningham, C. W. 1997. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. Systematic Biology 46:464–478.

Donoghue, M. J., J. A. Doyle, J. Gauthier, A. G. Kluge, and T. Rowe. 1989. The importance of fossils in phylogeny reconstruction. Annual Review of Ecology and Systematics 20:431–460.

Ebach, M. C., and S. T. Ahyong. 2001. Phylogeny of the trilobite subgenus *Acanthopyge* (*Lobopyge*). Cladistics 17:1–10.

Eernisse, D. J., and A. G. Kluge. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution 10: 1,170–1,195.

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

Gao, K., and M. A. Norell. 1998. Taxonomic revision of *Carusia* (Reptilia: Squamata) from the Late Cretaceous of the Gobi Desert and phylogenetic relationships of anguimorphan lizards. American Museum Novitates 3230:1–51.

———, and N. H. Shubin. 2001. Late Jurassic salamanders from northern China. Nature 410:574–577.

Gauthier, J. 1986. Saurischian monophyly and the origin of birds. Memoirs of the California Academy of Sciences 8:1–47.

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

Grande, L., and W. E. Bemis. 1998. A comprehensive phylogenetic study of amiid fishes (Amiidae) based on comparative skeletal anatomy, an empirical search for interconnected patterns of natural history. Society of Vertebrate Paleontology Memoirs 4:1–690.

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

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Systematic Zoology 38:297–309.

Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. Systematic Biology 44:3–16.

———. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Systematic Biology 47:3–8.

———, J. J. Bull, M. E. White, M. R. Badgett, and I. J. Molineux. 1992. Experimental phylogenetics: generation of a known phylogeny. Science 255:589–592.

———, J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science 264:671–677.

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

———. 1995. The performance of phylogenetic methods in simulation. Systematic Biology 44:17–48.

———, and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Systematic Biology 42:247–264.

Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Systematic Biology 51:369–381.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Systematic Biology 45:363–374.

Livezey, B. C. 1989. Phylogenetic relationships and incipient flightlessness of the extinct Auckland Islands Merganser. Wilson Bulletin 101:410–435.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. Systematic Biology 44:64–76.

Nixon, K. C., and Q. D. Wheeler. 1992. Extinction and the origin of species; pp. 119–142 *in* M. J. Novacek and Q. D. Wheeler (eds.), Extinction and Phylogeny. Columbia University Press, New York.

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

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

Patterson, C. 1981. Significance of fossils in determining evolutionary relationships. Annual Review of Ecology and Systematics 12:195–223.

Penny, D., and M. D. Hendy. 1985. The use of tree comparison metrics. Systematic Zoology 34:75–82.

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

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

Rowe, T. 1988. Definition, diagnosis, and origin of Mammalia. Journal of Vertebrate Paleontology 8:241–264.

Smith, A. B., G. L. J. Patterson, and B. Lafay. 1995. Ophiuroid phylogeny and higher taxonomy: morphological, molecular, and paleontological perspectives. Zoological Journal of the Linnean Society 114:213–243.

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. 2001. PAUP*: Phylogenetic Analysis Using Parsimony* (* and Other Methods), Version 4.0b8. Sinauer, Sunderland, Massachusetts.

———, and G. J. Olsen. 1990. Phylogeny reconstruction; pp. 411–501 *in* D. M. Hillis and C. Moritz (eds.), Molecular Systematics. Sinauer, Sunderland, Massachusetts.

Trueb, L., and R. Cloutier. 1991. A phylogenetic investigation of the inter- and intrarelationships of the Lissamphibia (Amphibia: Temnospondyli); pp. 223–313 *in* H. P. Schultze and L. Trueb, (eds.), Origins of the Higher Groups of Tetrapods: Controversy and Consensus. Cornell University Press, Ithaca, New York.

Wheeler, W. C., P. Cartwright, and C. Y. Hayashi. 1993. Arthropod phylogeny: a combined approach. Cladistics 9:1–39.

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

———. 1998b. Does adding characters with missing data increase or decrease phylogenetic accuracy. Systematic Biology 47:625–640.

———. 1998c. Testing phylogenetic methods with tree-congruence: phylogenetic analysis of polymorphic morphological characters in phrynosomatid lizards. Systematic Biology 47:411–428.

———. 2001. Character analysis in morphological phylogenetics: problems and solutions. Systematic Biology 50:689–699.

———. 2002. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology. ?1

———, and T. W. Reeder. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Systematic Biology 44: 548–558.

———, and M. R. Servedio. 1998. Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. Systematic Biology 47:228–253.

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

———, and M. J. Benton. 1995. Missing data and rhynchosaur phylogeny. Historical Biology 10:137–150.

Received 6 June 2001; accepted 6 July 2002.