Syst. Biol. 47(4):625–640, 1998

Does Adding Characters with Missing Data Increase or Decrease Phylogenetic Accuracy?

JOHN J. WIENS

Section of Amphibians and Reptiles, Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213-4080, USA; E-mail: wiensj@clpgh.org

Abstract.—Missing data are a widely recognized nuisance factor in phylogenetic analyses, and the fear of missing data may deter systematists from including characters that are highly incomplete. In this paper, I used simulations to explore the consequences of including sets of characters that contain missing data. More specifically, I tested whether the benefits of increasing the number of characters outweigh the costs of adding missing data cells to a matrix. The results show that the addition of a set of characters with missing data is generally more likely to increase phylogenetic accuracy than decrease it, but the potential benefits of adding these characters quickly disappear as the proportion of missing data increases. Furthermore, despite the overall trend, adding characters with missing data does decrease accuracy in some cases. In these situations, the missing data entries are not themselves misleading, but their presence may mimic the effects of limited taxon sampling, which can positively mislead. Criteria are discussed for predicting whether adding characters with missing data may increase or decrease accuracy. The results of this study also suggest that accuracy can be increased to a surprising degree by (1) "filling the holes" in a data matrix as much as possible (even when relatively few taxa are missing data), and (2) adding fewer characters scored for all taxa rather than adding a larger number of characters known for fewer taxa. Missing data can also be eliminated from an analysis through the exclusion of incomplete taxa rather than incomplete characters, but this approach may reduce the usefulness of the analysis and (in some cases) the accuracy of the estimated trees. [Accuracy; missing data; parsimony; simulations.]

Missing data are a common and widely recognized nuisance factor in phylogenetic analyses (e.g., Rowe, 1988; Huelsenbeck, 1991; Platnick et al., 1991; Novacek, 1992; Wiens and Reeder, 1995; Wilkinson, 1995). Given an initial data matrix that is complete (no missing data), missing data are typically added by two (nonindependent) vectors: (1) taxa that are highly incomplete (e.g., a fossil taxon known only from a tooth), or (2) sets of characters that are highly incomplete (e.g., DNA sequences obtained for 4 species in a genus with 12 species). Studies that discuss the effects of missing data on phylogenetic analysis have so far addressed only the impact of including incomplete taxa (e.g., Rowe, 1988; Huelsenbeck, 1991; Novacek, 1992; Wiens and Reeder, 1995; Wilkinson, 1995). These studies have shown that highly incomplete taxa tend to increase the number of shortest trees in a parsimony analysis and to decrease the resolution of consensus trees, and that they may reduce the overall accuracy of estimated trees.

Although incomplete taxa are a common source of missing data in phylogenetic analysis, incomplete sets of characters may be equally common. Many data sets may be scored for only a limited set of taxa in the context of a larger analysis, including molecular data sets, DNA sequences that can be aligned for only a subset of taxa, larval characters, histological characters, and behavioral characters. Unless one reduces the analysis to include only those taxa scored for all or most characters (which may be undesirable because this limits the taxonomic scope of the study and may be impossible if different sets of characters are missing in different sets of taxa), using these characters requires adding missing data entries. To avoid this, some authors have deliberately excluded characters because they contain abundant missing data (e.g., Livezey, 1989; Kress, 1990; Hufford, 1992; Hufford and Dickison, 1992; Smith et al., 1995; Smith, 1996). In contrast, Wiens and Reeder (1995) argued against excluding such characters, on the grounds that incompleteness of characters should not itself be positively misleading. Yet, the justification for either exclusion or inclusion of these characters has not been thoroughly addressed. Previous studies

have suggested that increasing the number of characters should generally increase phylogenetic accuracy (e.g., Hillis et al., 1994) but that increasing the amount of missing data may decrease accuracy (e.g., Huelsenbeck, 1991; Wiens and Reeder, 1995). Do the advantages of increasing the number of characters outweigh the disadvantages of adding missing data entries?

In this study, I used computer simulations to explore the effect of including sets of characters with missing data on the accuracy of parsimony, the most widely used phylogenetic method (Sanderson et al., 1993). For a given set of conditions, a complete data set (no missing data) was simulated and analyzed, after which a second data set, of the same size and type but with a number of data cells rendered missing, was added to it. Accuracy, the similarity between the estimated and true phylogeny, was compared between the first data set alone and the combined first and second data sets, to determine whether accuracy was increased or decreased by adding characters with missing data.

Simulations are an important tool for addressing general questions about phylogenetic accuracy because they provide a context for cases where the true phylogeny and evolutionary parameters are known. Although simulated data sets never capture the complexity of real data sets, the simplicity of simulated data sets allows the parameters that affect the behavior of phylogenetic methods to be controlled and understood. Results that are consistent across a broad range of simulated conditions can then be used to make predictions about how methods may behave in the real world.

Materials and Methods

Unrooted trees of 16 taxa each were simulated. Simulated trees were either fully balanced (symmetric) or unbalanced (asymmetric or pectinate), to span the range of all possible levels of symmetry for unrooted trees of 16 taxa (Fig. 1). Simulated characters were either binary (two states, as are many morphological characters) or had up to four unordered states (simulated DNA sequence data). All changes between states

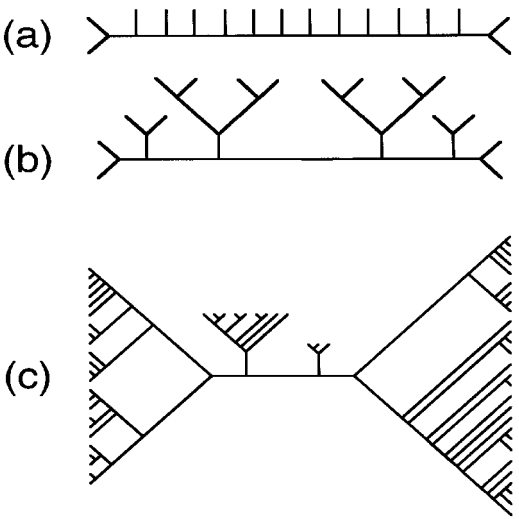

FIGURE 1. Model trees used in the simulations: (a) fully asymmetric, 16 taxon; (b) fully symmetric (balanced), 16 taxon; (c) 64-taxon tree based on hypothesized phylogenies of the lizard family Phrynosomatidae. All trees are unrooted.

were considered to be equally likely (e.g., gains, losses, transitions, transversions). For the purposes of this paper, branch length was considered to be the probability of a change occurring by the end of a branch for a given character. Branch lengths were varied among simulations to assess the impact of different rates of change but were held constant across all characters and all branches of the tree for a given simulation. This assumes an extremely punctuated model of change, as opposed to having divergence increase linearly with time on a rooted tree, but allows the effects of a given branch length to be tested. For binary characters the branch lengths examined were 0.01, 0.05, 0.10, 0.15, and 0.20. Because the presence of four unordered states reduces the probability of homoplastic changes, the DNA sequence data were unsaturated by homoplasy at longer branch lengths than the binary data, and so a broader range of branch lengths was tested (0.01, 0.10, 0.20, 0.30, and 0.40). For both types of characters, the lengths chosen span a broad range of conditions over which trees can be estimated correctly using parsimony (given a finite sample of characters), based on preliminary analyses. With

shorter branch lengths, there are few informative characters, and with longer branch lengths, parsimony has extremely low accuracy (because of homoplasy). One hundred replicated matrices were generated for each set of conditions examined (combination of tree shape, data type, and branch length).

Simulations began with an initial set of 50 characters for 16 taxa, with no missing data. To this data set was added another set of 50 characters (of the same type and evolving at the same rate) with various levels of completeness: (1) 100% complete (no missing data), (2) 75% complete (4 taxa randomly selected to have all missing data entries for all 50 characters in the second data set), (3) 50% complete (8 taxa missing data), and (4) 25% complete (12 taxa missing data). Different ways of adding missing data were also explored, namely, (1) confining the missing data cells to a monophyletic group of taxa (rather than a randomly selected set of taxa), and (2) randomly distributing missing data cells among taxa for each character in the second data set (so that all taxa are equally incomplete on average). "Monophyletic" groups were selected by taking the desired number of taxa, starting at the left side of the trees in Figure 1 (the strict monophyly of these groups depends on how these trees are rooted, but the groups are consistent with monophyly, given an unrooted tree).

The number of taxa (16) was chosen because it is small enough to be computationally tractable but large enough to allow characters to be highly incomplete but still (potentially) parsimony-informative (e.g., 25% complete leaves data in 4 taxa). Fifty characters were used in the initial data set because this number is small enough to allow the addition of characters to potentially improve accuracy (i.e., an accuracy of 100% cannot be improved upon), but large enough that there generally were not huge numbers of equally parsimonious trees to contend with.

A more limited set of analyses tested the effects of including characters with missing data in a 64-taxon case. The model tree used was based on estimated phylogenies for phrynosomatid lizards (Fig. 2). An empirically derived tree was chosen to be more complex and realistic in its shape than the trees used in the 16-taxon case, but the veracity of this tree is not critical (because the true tree is known in the simulations). The higher-level relationships were based on Reeder and Wiens (1996; combined analysis) and species-level relationships were based on the following sources (with some minor modifications): *Phrynosoma* (Montanucci, 1987), sand lizards (*Uma*, *Callisaurus*, *Cophosaurus*, *Holbrookia*; de Queiroz, 1989, combined analysis), *Uta* (Ballinger and Tinkle; 1972, their Fig. 13), *Urosaurus* (Reeder and Wiens, 1996; combined analysis), and *Sceloporus* (Wiens and Reeder, 1997; from combined analysis with all taxa, but only species groups shown). These simulations used an initial set of 100 characters, to which was added another set of 100 characters, which was either (1) 100% complete (no missing data), (2) 75% complete (16 taxa missing data), (3) 50% complete (32 taxa missing data), or (4) 25% complete (48 taxa missing data).

In all the preceding analyses, the number of characters added in the second data set was held constant, but the actual amount of data added was reduced by the missing data entries. Another set of analyses (using 16 taxa) tested the effects of adding different amounts of missing data while keeping the amount of nonmissing data constant. Thus, the second data set contained either (1) 25 characters in all 16 taxa (no missing data), (2) 50 characters for 8 taxa (50% missing data), or (3) 100 characters for 4 taxa (75% missing data). Missing data were distributed to randomly selected taxa for all characters in the second data set. These analyses addressed the question of whether it is better to add fewer characters scored for many taxa in a phylogenetic analysis or to add more characters scored for few taxa.

A limited set of analyses tested an alternative method for excluding missing data entries, i.e., excluding the incomplete taxa. The 16-taxon, asymmetric tree topology was used, with binary characters and intermediate branch lengths (0.05, 0.10, 0.15). The second data set contained either 8 taxa scored for 50 characters, or 4 taxa scored for 100

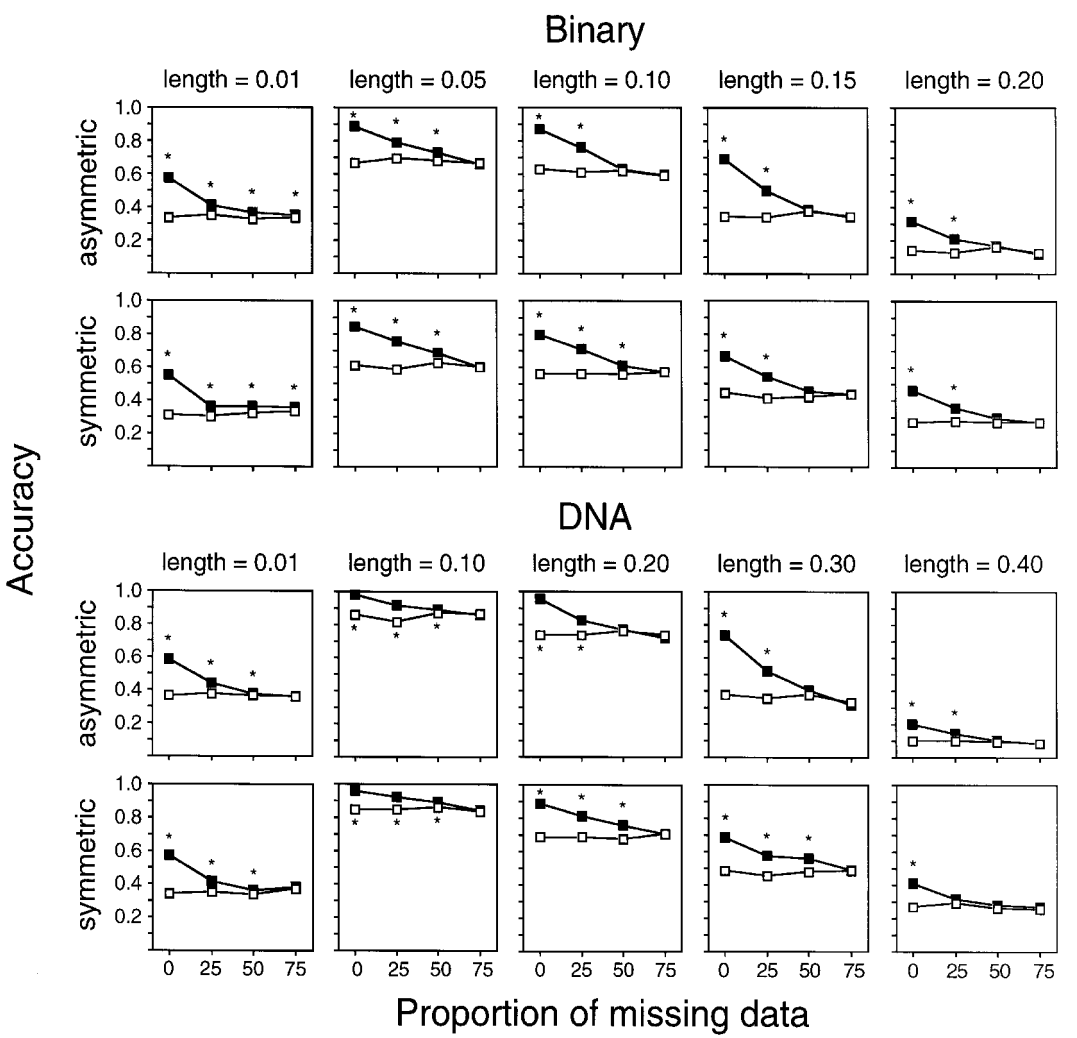

FIGURE 2. Accuracy of parsimony analysis including and excluding characters with missing data in the 16-taxon case. □ = data set 1 alone (no missing data); ■ = data sets 1 and 2 combined (with missing data). Missing data are added by randomly picking taxa to have empty data cells for all characters in data set 2. Each data set contains 50 characters, and each point is the average accuracy from 100 replicates. Asterisks denote $P < 0.01$ for paired *t*-tests of accuracy with and without addition of the second data set.

characters (i.e., keeping the amount of data added constant). Rather than distributing missing data entries randomly on the tree, two general patterns were explored, one in which missing data were distributed evenly on the tree, and another in which the complete taxa were distributed to maximize long-branch effects (Felsenstein, 1978). Although this set of analyses was not exhaustive in terms of the parameter space explored, it did address the accuracy of taxon exclusion relative to character exclusion for favorable and worst-case scenarios (for taxon sampling).

In the preceding analyses, accuracy was measured as the similarity between the estimated phylogeny (or the strict consensus of multiple equally parsimonious estimates) and the true phylogeny, averaged across 100 replicates. Similarity was measured as the

proportion of nodes in common between the true and estimated trees, by use of the consensus fork index of Colless (1980). This is a commonly used and intuitive measure of accuracy, but it may be sensitive to taxa that are highly unstable in their placement (M. Wilkinson, pers. comm.). Paired *t*-tests were used for assessing the significance of changes in accuracy caused by adding the data set with missing entries (because of the large number of comparisons, only *P* values less than 0.01 are noted).

Alternative measures of accuracy were also explored: (1) using a single, randomly chosen tree (from among the shortest trees) as an estimate, so that all analyses yielded the same level of resolution, and (2) measuring how often the addition of the second data set increased, decreased, or had no affect on accuracy (relative to the accuracy of the first data set alone) out of 100 replicates.

Phylogenetic analyses were performed using Swofford's PAUP*, versions 4.0d55 and 4.0d56, with the heuristic search option, TBR branch swapping, and 20 random-addition sequences per search. Because many thousands of equally parsimonious trees were sometimes generated at the shortest branch length (0.01), the maximum number of trees retained in a search was set to 500, to allow searches to be completed in a reasonable amount of time. This may artificially increase the resolution of the trees (both with and without missing data) but generally only at this branch length. The programs for simulating the data sets and tallying the results were written in C by the author.

In PAUP and PAUP*, missing data entries (usually symbolized by "?") are treated as ambiguous during tree-reconstruction; in other words, they are treated as if any of the observed states could be assigned to the incomplete taxon (Swofford, 1993). Thus, a missing data cell is uninformative in estimating the tree, although other data cells of that character may be informative. The method by which PAUP (and other parsimony programs) treats missing data entries is assumed to be reasonable and noncontroversial (but see Platnick et al., 1991). Missing data cells are assumed to represent an absence of information rather than inapplicability (e.g., the shape of a bone in a taxon where the bone is absent) or polymorphism.

Results

Across all the conditions examined (tree shapes, branch lengths, types of data), adding characters with missing data (in randomly selected incomplete taxa) is generally more likely to increase accuracy than decrease it (Fig. 2). However, the increase in accuracy caused by adding these characters is greatly diminished relative to adding an equal number of complete characters, even when only 25% of the data cells are missing in the second data set. When 50% of the data cells are missing, the increase in accuracy is small, being significantly different from 0 in some cases but not others. With missing data in 75% of the taxa, the accuracy is generally the same as if the incomplete characters were not added at all. This same general pattern is seen (1) in the 64-taxon case (Fig. 3), (2) when missing data entries are confined to a monophyletic group of taxa or randomly distributed among taxa for each character (Fig. 4), and (3) when accuracy is measured by using a single randomly selected, fully resolved tree from each analysis as the estimate (so that all analyses have the same level of resolution), rather than strict consensus trees (Table 1). This pattern is not an artifact of the unstable placement of highly incomplete taxa because the same general results are obtained when all taxa are equally incomplete (Fig. 4).

The same general pattern is also found when the amount of nonmissing data added is standardized in each case (Fig. 5); adding a few characters scored for all taxa, a larger number of characters scored for half the taxa, or a large number of characters scored for only a few taxa. Thus, the results suggest that, given a finite amount of data that can be added to an existing data matrix, it is more beneficial to add fewer characters scored for all the taxa than to add more characters scored for fewer taxa.

The preceding results are all based on comparing the average accuracies of the data sets with and without characters with

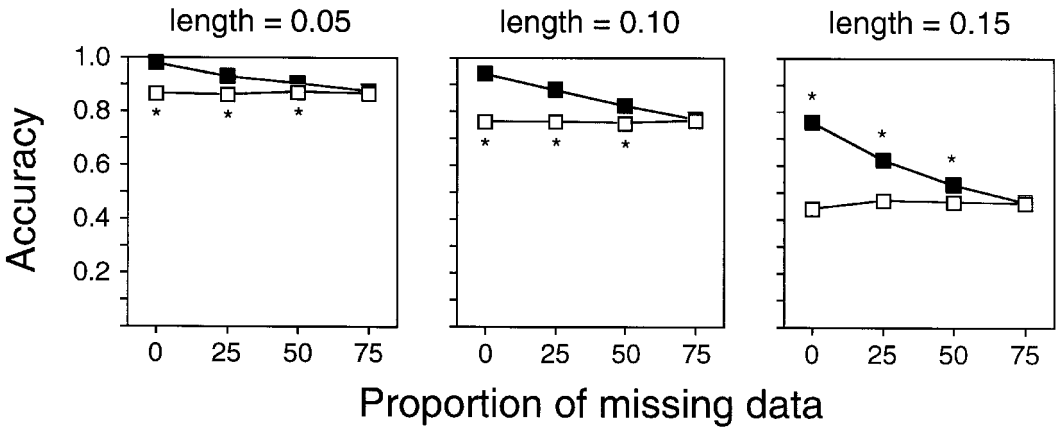

FIGURE 3. Accuracy of parsimony analysis including and excluding characters with missing data in the 64-taxon case. □ = data set 1 alone (no missing data); ■ = data sets 1 and 2 combined (with missing data). Missing data are added by randomly picking taxa to have empty data cells for all characters in data set 2. Each data set contains 100 characters, and each point is the average accuracy from 100 replicates. Asterisks denote $P < 0.01$ for paired *t*-tests of accuracy with and without addition of the second data set.

missing data. Another way of looking at the results is to see how often (what proportion of 100 replicates) adding the second data set increases or decreases accuracy relative to the first data set analyzed alone. Using the latter measure confirms that adding the characters with missing data is generally more likely to increase accuracy than decrease it (Fig. 6). However, in many cases, adding the characters with missing data clearly causes accuracy to decrease; moreover, this seems to occur more often than would be expected for adding a set of complete characters (Fig. 6). This suggests that adding highly incomplete characters can worsen phylogenetic results (contra

TABLE 1. Accuracy of parsimony analysis including and excluding characters with missing data in the 16-taxon case, comparing two ways of measuring accuracy. The first measure uses a strict consensus of the shortest trees as the estimate from a given analysis (as done throughout this paper), and the second uses a single randomly selected, fully resolved tree from among the shortest trees (so that all analyses yield the same level of resolution). The first measure gives lower accuracy overall than does the second, but the general effects of adding characters with missing data are the same. The characters are binary, the tree is asymmetric, there are 50 characters in each data set, and each value is the mean accuracy from 100 replicates.

Branch length	Accuracy measure	Data set	Proportion of missing data (%)			
			0	25	50	75
0.05	consensus	data set 1	0.695	0.662	0.667	0.666
		combined	0.882	0.775	0.721	0.667
	single tree	data set 1	0.799	0.771	0.759	0.769
		combined	0.926	0.850	0.819	0.770
0.10	consensus	data set 1	0.571	0.592	0.610	0.616
		combined	0.873	0.729	0.627	0.612
	single tree	data set 1	0.679	0.702	0.699	0.700
		combined	0.900	0.793	0.722	0.727
0.15	consensus	data set 1	0.363	0.346	0.380	0.351
		combined	0.700	0.536	0.432	0.342
	single tree	data set 1	0.436	0.463	0.464	0.425
		combined	0.742	0.622	0.509	0.419

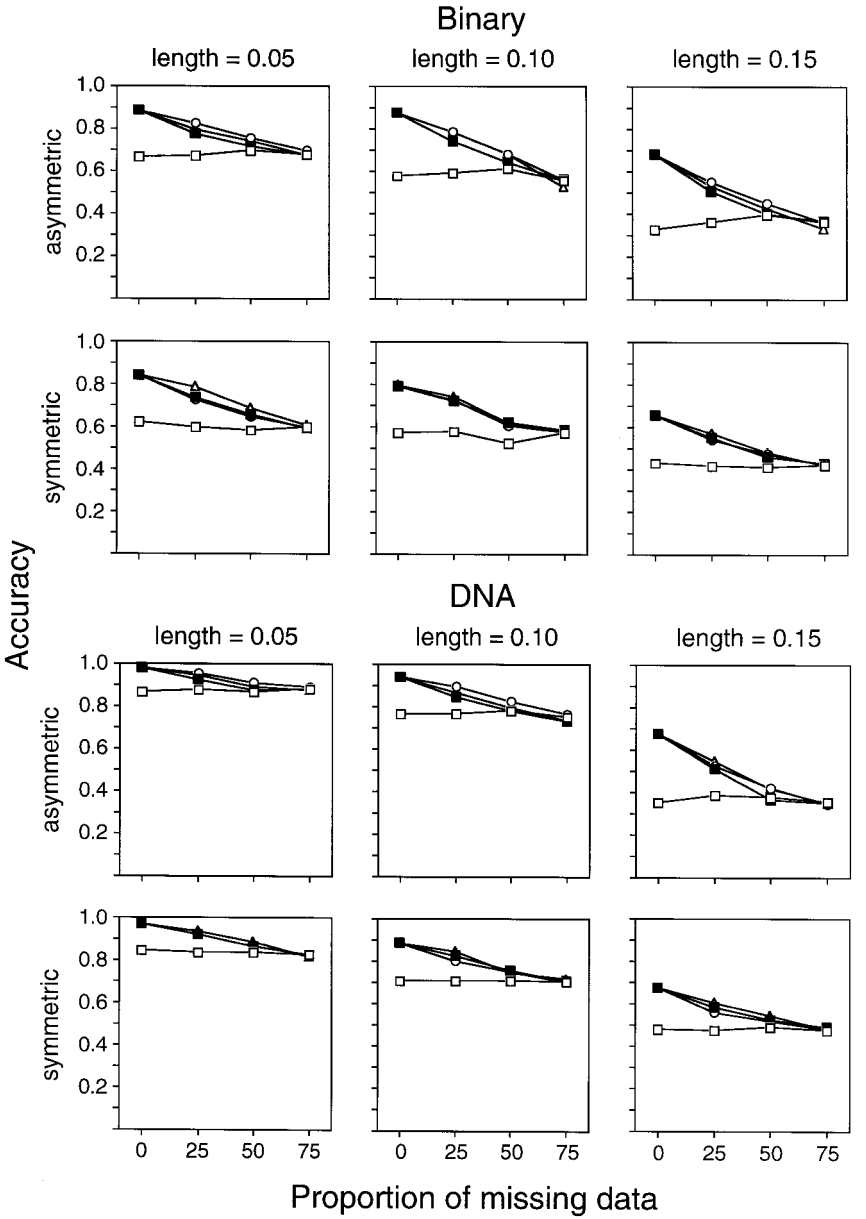

FIGURE 4. Accuracy of parsimony analysis including and excluding characters with missing data in the 16-taxon case, showing that three different ways of distributing missing data cells all give similar results. □ = data set 1 alone (no missing data); ■ = data sets 1 and 2 combined (randomly selected taxa have all characters in data set 2 replaced with missing data); ○ = data sets 1 and 2 combined (missing data confined to monophyletic subset of taxa); △ = data sets 1 and 2 combined (missing data randomly distributed among taxa for each character). Each data set contains 50 characters, and each point is the average accuracy from 100 replicates.

Wiens and Reeder, 1995), at least in some cases. When 75% of the data are missing in the second data set, adding these characters may decrease accuracy as often or more often than it increases accuracy (although in most replicates adding the second data set

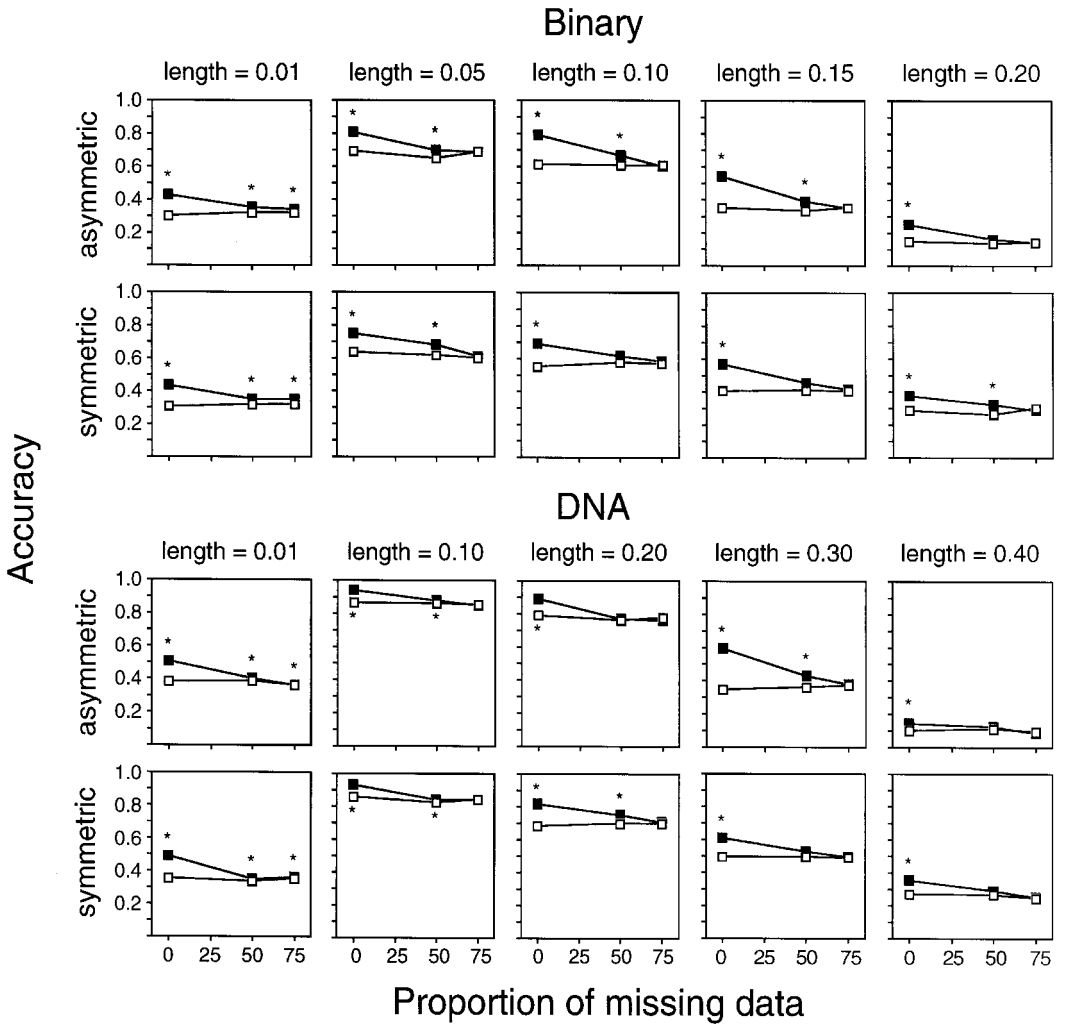

FIGURE 5. Given a finite amount of data that can be added to an analysis, sampling more taxa gives more accurate results than sampling more characters. Four hundred complete data cells are available in data set 2, which are distributed in three ways: (1) 25 characters for all 16 taxa, no missing data; (2) 50 characters for 8 taxa, missing data in 50% of the taxa; (3) 100 characters for four complete taxa, missing data in 75% of the taxa. □ = data set 1 alone (no missing data); ■ = data sets 1 and 2 combined (with missing data). Missing data are added by randomly picking taxa to be missing data for all characters in data set 2. Each point is the average accuracy from 100 replicates. Asterisks denote $P < 0.01$ for paired t-tests of accuracy with and without addition of the second data set.

has no effect on accuracy, and the average accuracy is not significantly decreased). The same pattern is seen when missing data are distributed randomly among taxa for each character (Fig. 7b). However, when the missing data are restricted to a monophyletic subset of taxa, the tendency for accuracy to decrease is reduced (Fig. 7a).

In a limited set of analyses, I compared the accuracy of excluding incomplete taxa relative to including or excluding incomplete characters. The results (Fig. 8) show that the estimate based only on the complete taxa may be relatively accurate if the complete taxa are distributed evenly on the true phylogeny (Fig. 8a, c), or it may be relatively

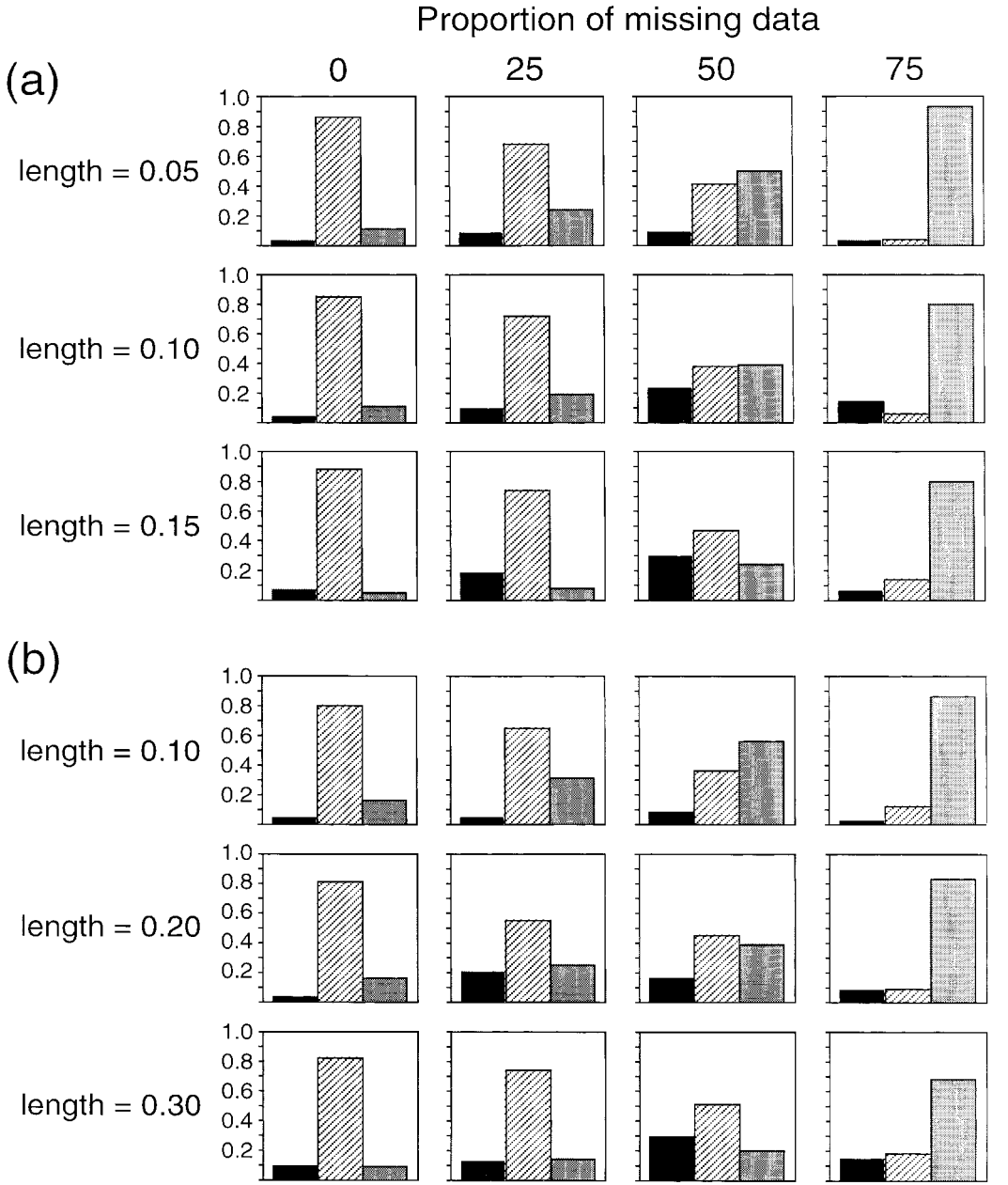

FIGURE 6. Frequency with which adding data set 2 (with missing data) decreases (■), increases (▨), or has no effect on accuracy (▩), relative to data set 1 analyzed alone, based on 100 replicated matrices. Missing data are added by randomly picking taxa to be missing data for all characters in data set 2. Each data set contains 50 characters. (a) Asymmetric model tree, binary characters. (b) Symmetric model tree, DNA sequence characters.

inaccurate if these taxa are distributed so as to maximize long-branch attraction (Fig. 8b, d). However, even when the complete taxa are spaced evenly on the true tree, an analysis restricted to them is consistently less accurate than one that includes all the taxa

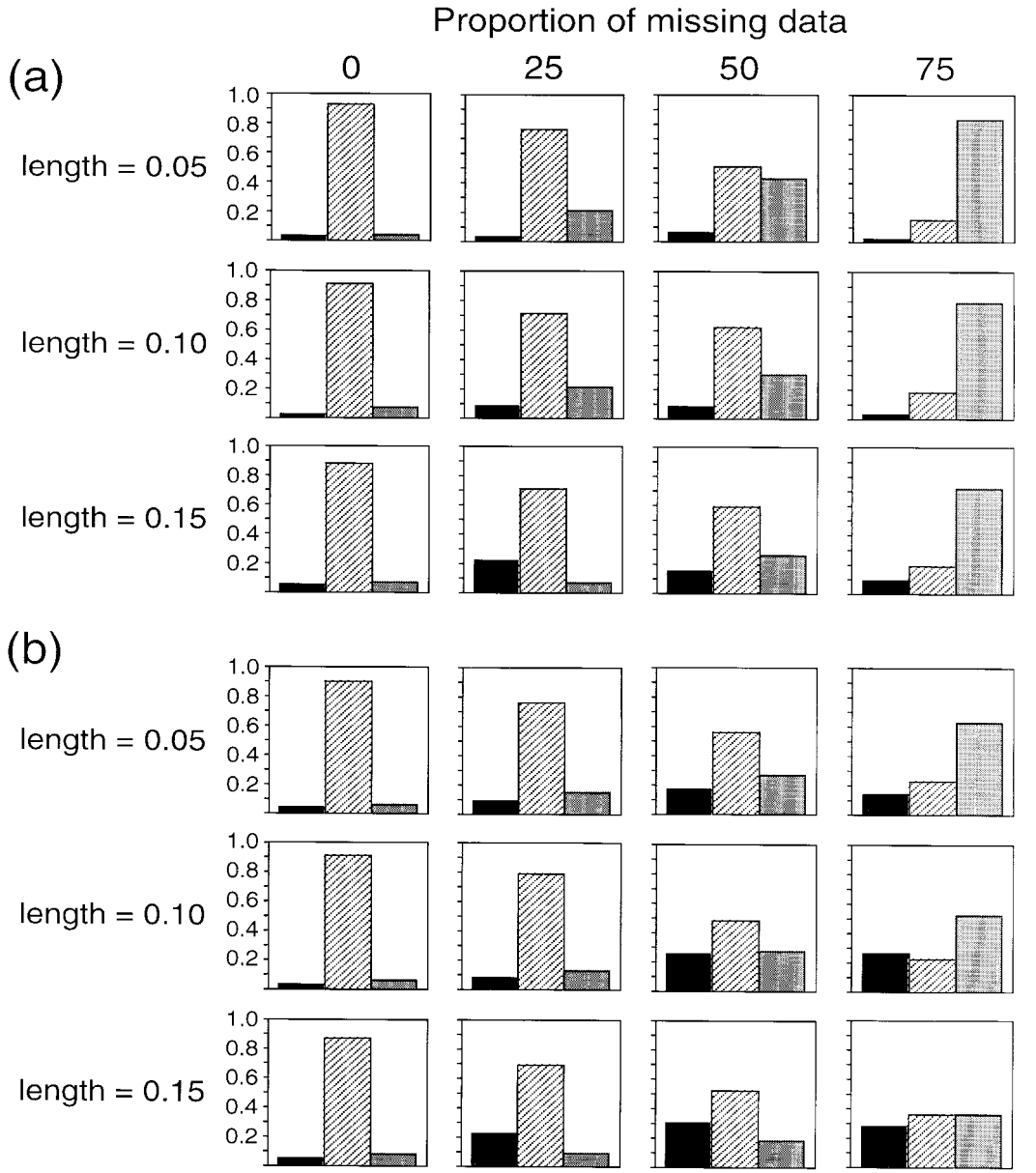

FIGURE 7. Frequency with which adding data set 2 (with missing data) decreases (■), increases (▨); or has no effect on accuracy (▩), relative to data set 1 analyzed alone, based on 100 replicated matrices. Each data set contains 50 characters. (a) Asymmetric model tree, binary characters, missing data added to a monophyletic subset of taxa. (b) Asymmetric model tree, binary characters, missing data added to a randomly selected set of taxa for each character.

but prunes incomplete taxa from the shortest trees. In other words, when the measurement of accuracy is standardized so that the same number of taxa are compared in each case, the analysis including more taxa always gives the better estimate. This analysis also demonstrates that missing data cells can be distributed in such a way that including

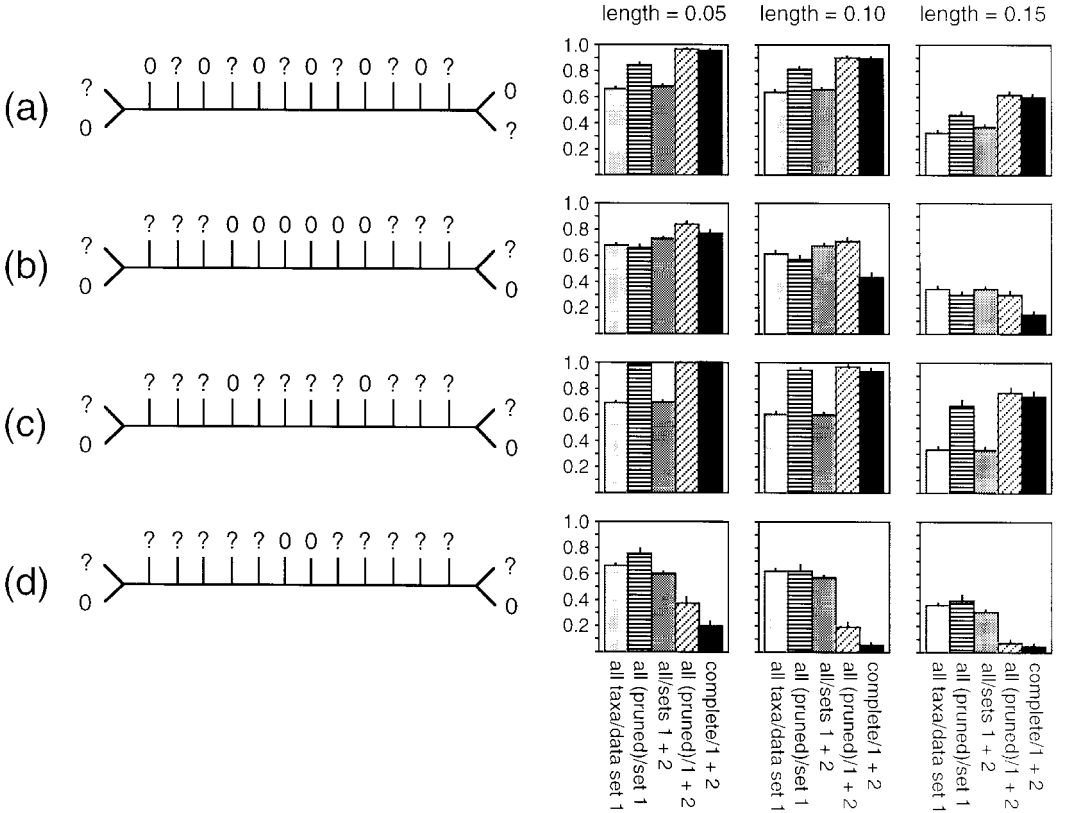

FIGURE 8. The effects of including and excluding incomplete taxa on phylogenetic accuracy. ■ = 16 taxa, data set 1 alone; ▤ = 16 taxa, data set 1 alone, accuracy based only on those taxa that are complete in data set 2; □ = 16 taxa, combined data sets; ▨ ≒ 16 taxa, combined data sets, accuracy scored for complete taxa only; ■ = combined data sets, incomplete taxa excluded. All characters are binary. Each bar is the average accuracy from 100 replicates, and the line above each bar is the standard error. (a) Eight taxa complete in data set 2, incomplete taxa distributed evenly on tree, 50 characters in each data set. (b) Eight taxa complete in data set 2, incomplete taxa distributed to maximize long-branch effects, 50 characters in each data set. (c) Four taxa complete in data set 2, incomplete taxa distributed evenly on tree, 50 characters in data set 1, 100 characters in data set 2. (d) Four taxa complete in data set 2, incomplete taxa distributed on tree to maximize long-branch effects, 50 characters in data set 1, 100 characters in data set 2.

incomplete characters can consistently decrease accuracy (Fig. 8d).

DISCUSSION

How Can Missing Data Decrease Phylogenetic Accuracy?

The major conclusion of this study is that adding characters with missing data generally increases phylogenetic accuracy. However, adding these characters can decrease accuracy in some cases. This result is somewhat counterintuitive, because missing data cells cannot introduce new homoplasy to an analysis. Given this, how is it possible that adding sets of characters with missing data can produce estimates less accurate than if these characters were complete or were excluded? The answer may be due largely to the fact that taxa with missing data are effectively unsampled for those characters. Limited taxon sampling can increase the lengths of the branches connecting the sampled taxa, and long branches can cause phylogenetic methods to be positively misled (e.g., Swofford et al., 1996). By effectively increasing the branch lengths among the "complete" taxa, an abundance of missing data can increase

the proportion of characters that support incorrect topologies (Fig. 9), even when branch lengths are equal among lineages (Fig. 9b). Furthermore, if the missing data are distributed in such a way that there are effectively two long, unrelated terminal branches separated by a short internal branch (the situation described by Felsenstein, 1978), then the number of characters supporting incorrect trees will greatly outnumber the characters supporting the correct trees (Fig. 9c). Thus, not only the amount of missing data is important, but also how these data are distributed among taxa; this is why missing data are less likely to be misleading when they are confined to a monophyletic subset of taxa (Fig. 7). When missing data are distributed randomly among taxa (as in many of the simulations of this study), accuracy may be increased or decreased by adding these incomplete characters, depending upon the individual replicate.

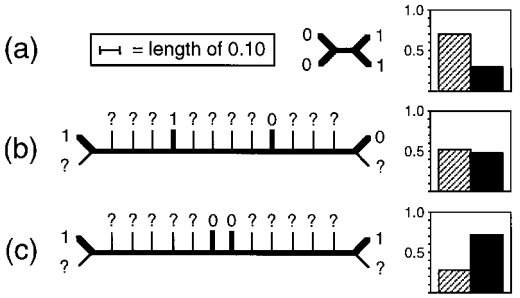

FIGURE 9. Missing data can increase the proportion of parsimony-informative characters that support incorrect topologies (■) versus the correct tree (▨). Each graph is based on a sample of 500 characters. (a) Four-taxon unrooted tree with all branches of length 0.10; (b) 16-taxon unrooted tree with all branches of length 0.10, with missing data added to 12 taxa (4 taxa complete); (c) 16-taxon unrooted tree with all branches of length 0.10, with missing data added to 12 taxa (4 taxa complete), with incomplete taxa chosen to maximize potential long-branch attraction.

Long-branch attraction through incomplete taxon sampling is a possible mechanism by which adding characters with missing data can give misleading results. However, the taxon sampling/long branch problem is probably not the only cause for the decreases in accuracy sometimes caused by adding characters with missing data, because these decreases still may occur (albeit less frequently) when the missing data cells are confined to a monophyletic subset of taxa, and these decreases happen more often than would be expected if complete characters were added (Fig. 7a).

Another effect of abundant missing data entries is to decrease the number of parsimony-informative characters (relative to a complete data set of the same size), especially when data are missing in 75% of the taxa (Table 2). The decrease in informative characters may explain why the addition of sets of characters with very large amounts of missing data tends to have little overall effect on phylogenetic accuracy. However, the relatively small decreases in the number of informative characters with 25–50% missing data suggest that a loss of informative characters is insufficient to explain why characters with missing data increase accuracy so little relative to complete data. Instead, it appears that many characters with abundant missing data are parsimony-informative but highly ambiguous (i.e., consistent with many trees but not with all trees), relative to characters that are complete.

Implications for Empirical Studies

Given the results of this study, should an empirical investigator include or exclude a set of characters with abundant missing data? In general, the results of this study suggest that it should be more beneficial to include these characters (and more costly to exclude them). Nevertheless, in some cases accuracy is decreased by adding characters with missing data, and there are conditions where this can be expected to happen frequently (Figs. 8d, 9). Is there any way to predict whether adding a set of characters with missing data is more likely to increase or decrease phylogenetic accuracy?

In general, when the missing data cells are confined to a monophyletic subset of taxa, the addition of the incomplete data set is less likely to cause a decrease in accuracy (Fig. 7a) and may be relatively safe. Conversely, when missing data cells are distributed so as to create long-branch attraction among the sampled taxa (i.e., distantly related species

TABLE 2. Replacing data cells with missing data entries reduces the number of parsimony-informative characters. Values are the numbers (range) of characters that are parsimony-informative for a given matrix (out of 50 total) averaged across 10 replicated matrices. Missing data are distributed randomly among taxa.

			Proportion of missing data (%)			
Tree shape	Character	Branch length	0	25	50	75
Asymmetric	binary	0.05	28.2 (18–34)	26.1 (20–30)	20.4 (18-26)	6.6 (3–11)
		0.10	43.2 (40–45)	40.2 (33–44)	32.5 (27–39)	11.0 (5–18)
		0.15	47.6 (45–50)	45.1 (39–50)	39.7 (38–42)	14.4 (9–18)
Symmetric	DNA	0.10	38.1 (28–44)	32.8 (28-37)	24.7 (21–32)	7.1 (3–14)
		0.20	48.7 (47–50)	47.3 (46–49)	40.6 (37–46)	9.9 (6–18)
		0.30	49.8 (49–50)	49.4 (48–50)	45.8 (42–48)	7.1 (6–10)

or certain combinations of closely and distantly related species), then their addition is more likely to cause a decrease in accuracy (Fig. 8d). In many cases, however, it may not be known how the taxa with missing data are related to each other.

If there is a long-branch problem caused by incomplete taxon sampling, the complete and incomplete sets of characters may yield strongly conflicting estimates of phylogeny when analyzed separately (for those taxa present in both data sets), even if these data sets share identical underlying phylogenetic histories, branch lengths, and models/processes of character evolution. Given this, the absence of conflict between the trees may indicate that inclusion of these incomplete characters will be relatively safe.

In some cases, changes in the number of equally parsimonious trees caused by adding incomplete characters might be useful for predicting their impact on accuracy. Examining some of the previous results (Fig. 2) more closely suggests that when the addition of characters with missing data increases the number of shortest trees (relative to the complete data alone), accuracy generally either decreases or remains the same, an indication that it may be safe to exclude these characters (Table 3). Conversely, when the number of shortest trees decreases or remains the same, accuracy generally either increases or remains the same, suggesting that it is safe to include these characters. Although these results are based on a limited set of analyses under very simplified conditions, they make intuitive sense for at least two reasons: (1) Adding characters should generally increase resolution, and the added characters must conflict substantially with the original set of characters to decrease resolution; and (2) accuracy and resolution are not independent (i.e., a data set that yields a large number of equally short trees must be

TABLE 3. Changes in the number of shortest trees caused by adding sets of characters with missing data may predict changes in accuracy. The data presented are the number of replicates (out of 100) in which there is an increase, a decrease, or no change in the accuracy and number of trees caused by adding the second data set, relative to the accuracy and number of trees from analyzing the first data set alone. There are 50 characters in each data set, and missing data are randomly distributed among 8 of the 16 taxa in data set 2. Model I = binary characters, asymmetric tree, length = 0.10. Model II = DNA characters, symmetric tree, length = 0.20.

		Change in accuracy (no. replicates)		
Model	No. of trees	Increases	Decrease	No change
I	increases	1	18	12
	decreases	32	4	7
	no change	5	1	20
II	increases	5	18	13
	decreases	25	2	9
	no change	10	1	17

generating a large number of incorrect trees, because only one tree can be correct).

In summary, these criteria may be useful in deciding whether or not to include a set of characters with abundant missing data cells in empirical studies. One should remember, however, that, according to these simulations, including these characters (on average) either increases or has little effect on accuracy under a wide variety of conditions—which suggests that incomplete characters should generally be included.

A surprising result of this study is the extent to which even a limited amount of missing data can rob a set of characters of their ability to improve phylogenetic accuracy (Fig. 3), relative to a set of complete characters of the same size. The implication is that there is a clear benefit to "filling the holes" in a data matrix, even when relatively few taxa are missing data. The results also imply that, given a finite amount of data that can be added to an existing matrix, it may be far more beneficial to sample a small set of characters for all the taxa than to seek out a large number of characters for a smaller sample of taxa (Fig. 5). This result underscores the value of morphological data in phylogeny reconstruction, data that generally offer a limited number of characters relative to molecular data sets but that may often be obtained from a more complete sampling of taxa (Hillis, 1987).

An Alternative Method for Excluding Missing Data: Deleting Incomplete Taxa

This study has dealt primarily with the consequences of including versus excluding sets of characters with missing data while holding the number of taxa constant. An alternative way of reducing the amount of missing data may be to exclude those taxa that are scored for only one set of characters. However, this approach has at least two important disadvantages. First, it limits the taxonomic scope of the study. Even if a phylogenetic analysis were guaranteed to yield the correct phylogeny for 4 taxa, this phylogeny may be considerably less useful than having a hypothesis for 16 taxa that may be only partly correct (Wiens and Reeder, 1995). The second disadvantage is that deleting taxa may decrease accuracy, a possibility confirmed by some of the results of this study (Fig. 8). This result may depend on how the taxa are sampled, but even when the complete taxa are spaced evenly on the tree (a favorable scenario for exclusion), an analysis restricted to the complete taxa may be less accurate than one that includes all taxa but prunes incomplete taxa from the shortest trees (i.e., standardizing the number of taxa between analyses). Using subsampling experiments with viruses and lizards, Wiens and Reeder (1995) found that inclusion of incomplete taxa seemed to decrease slightly the overall accuracy of estimated trees. However, they did not standardize their results for the number of taxa, and their two study groups included only closely related taxa, groups where long branches and incomplete taxon sampling are less likely to be misleading. Although it may be useful in some cases (e.g., Wilkinson, 1995), the practice of excluding taxa to minimize missing data is potentially problematic—not only because it restricts the taxonomic scope of the analysis, but also because it invokes the same problem (limited taxon sampling) that makes characters with missing data potentially misleading. These results further support the practice of including more taxa rather than avoiding relatively incomplete taxa, but this choice remains a problem in need of further study.

Conclusions and Caveats

Systematists may use a variety of criteria for including or excluding characters in a phylogenetic analysis. This study tests the consequences of including characters with many missing data cells to provide some basis for deciding whether to include or exclude these characters in empirical studies. The results generally support the inclusion of these characters, but they also suggest that such characters should not be added uncritically, and some criteria are suggested that may help predict whether including these characters will be advantageous or hazardous in a specific case. Of course, from the philosophical perspective of "total evidence," one could argue that all characters

should always be included (Kluge, 1989); in fact, this view is not really contradicted by these particular results (i.e., on average, including does give more accurate estimates). Even if one accepts this view and considers "to include or not to include" a moot question, the results show a surprising benefit for "filling the holes" in a data matrix as much as possible and for designing studies to emphasize complete sampling of taxa versus greater sampling of characters.

Because the results are based on simulations, they should be taken with some caution. Simulated data sets are greatly simplified relative to those in the real world. Major simplifying assumptions of this study include (1) all characters evolve at the same rate within and between data sets, (2) branch lengths are equal among lineages, and (3) the distribution of missing data is independent of changes in the characters. Nevertheless, it is the simplicity of these simulations that makes it possible to isolate the effects of missing data and other relevant parameters on phylogenetic accuracy. Furthermore, the major results of this study are robust to changes in a number of parameters and methods, including tree shape, number of taxa, number of characters, branch lengths, data type (binary vs. four unordered states), different ways of distributing missing data, and different ways of measuring accuracy. Given the robustness of these results to a variety of simulated conditions, it seems likely they will apply to many empirical data sets as well.

Finally, this study has been limited to the effects of missing data on parsimony analysis and, more specifically, to how they affect parsimony analysis when treated as ambiguities. Although the majority of phylogenetic studies use parsimony (e.g., Sanderson et al., 1993) and the treatment of missing data by current parsimony algorithms appears reasonable, the conclusions of this study may not apply to other phylogenetic methods (e.g., distance, likelihood) or other ways of handling missing data with parsimony.

Acknowledgments

I thank Brad Livezey for suggesting that I do a simulation study to test the consequences of adding characters with missing data. David Swofford kindly allowed me to use and publish results with test versions of his PAUP* program. I thank David Cannatella, Phil Chu, Brad Livezey, Steve Poe, Maria Servedio, Mark Wilkinson, and two anonymous reviewers for useful discussion, advice, reviews, and/or comments on the manuscript.

References

Ballinger, R. E., and D. W. Tinkle. 1972. Systematics and evolution of the genus *Uta* (Sauria: Iguanidae). Misc. Publ. Mus. Zool. Univ. Mich. 145:1–183.

Colless, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: A reappraisal. Syst. Zool. 29:288–299.

de Queiroz, K. 1989. Morphological and biochemical evolution in the sand lizards. Ph.D. Dissertation, Univ. of California, Berkeley.

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

Hillis, D. M. 1987. Molecular versus morphological approaches to systematics. Annu. Rev. Ecol. Syst. 18:23–42.

Hillis, D. M., 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? Syst. Zool. 40:458–469.

Hufford, L. 1992. Rosidae and their relationships to other nonmagnoliid dicotyledons: A phylogenetic analysis using morphological and chemical data. Ann. Mo. Bot. Gard. 79:218–248.

Hufford, L., and W. C. Dickison. 1992. A phylogenetic analysis of Cunoniaceae. Syst. Bot. 17:181–200.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38:7–25.

Kress, W. J. 1990. The phylogeny and classification of the Zingiberales. Ann. Mo. Bot. Gard. 77:698–721.

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

Montanucci, R. R. 1987. A phylogenetic study of the horned lizards, genus *Phrynosoma*, based on skeletal and external morphology. Contr. Sci. Nat. Hist. Mus. Los Angel. Cty. 113:1–26.

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

Platnick, N. I., C. E. Griswold, and J. A. Coddington. 1991. On missing entries in cladistic analysis. Cladistics 7:337–343.

Reeder, T. W., and J. J. Wiens. 1996. Evolution of the lizard family Phrynosomatidae as inferred from diverse types of data. Herpetol. Monogr. 10:43–84.

Rowe, T. 1988. Definition, diagnosis, and origin of Mammalia. J. Vertebr. Paleontol. 8:241–264.

Sanderson, M. J., B. G. Baldwin, G. Bharathan, C. S. Campbell, C. Von Dohlen, D. Ferguson, J. M. Porter, M. F. Wojciechowski, and M. J. Donoghue. 1993. The growth of phylogenetic information and

the need for a phylogenetic data base. Syst. Biol. 42:562–568.

SMITH, A. B., G. L. J. PATTERSON, AND B. LAFAY. 1995. Ophiuroid phylogeny and higher taxonomy: Morphological, molecular, and paleontological perspectives. Zool. J. Linn. Soc. 114:213–243.

SMITH, J. F. 1996. Tribal relationships within Gesneriaceae: A cladistic analysis of morphological data. Syst. Bot. 21:497–513.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, version 3.1. Illinois Natural History Survey, Champaign.

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

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

WIENS, J. J., AND T. W. REEDER. 1997. Phylogeny of the spiny lizards (*Sceloporus*) based on molecular and morphological evidence. Herpetol. Monogr. 11:1–101.

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

Received 2 September 1997; accepted 30 November 1997
Associate Editor: D. Cannatella