*Syst. Biol.* 52(4):528–538, 2003

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150390218330

# Missing Data, Incomplete Taxa, and Phylogenetic Accuracy

JOHN J. WIENS

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

*Abstract*.— The problem of missing data is often considered to be the most important obstacle in reconstructing the phylogeny of fossil taxa and in combining data from diverse characters and taxa for phylogenetic analysis. Empirical and theoretical studies show that including highly incomplete taxa can lead to multiple equally parsimonious trees, poorly resolved consensus trees, and decreased phylogenetic accuracy. However, the mechanisms that cause incomplete taxa to be problematic have remained unclear. It has been widely assumed that incomplete taxa are problematic because of the proportion or amount of missing data that they bear. In this study, I use simulations to show that the reduced accuracy associated with including incomplete taxa is caused by these taxa bearing too few complete characters rather than too many missing data cells. This seemingly subtle distinction has a number of important implications. First, the so-called missing data problem for incomplete taxa is, paradoxically, not directly related to their amount or proportion of missing data. Thus, the level of completeness alone should not guide the exclusion of taxa (contrary to common practice), and these results may explain why empirical studies have sometimes found little relationship between the completeness of a taxon and its impact on an analysis. These results also (1) suggest a more effective strategy for dealing with incomplete taxa, (2) call into question a justification of the controversial phylogenetic supertree approach, and (3) show the potential for the accurate phylogenetic placement of highly incomplete taxa, both when combining diverse data sets and when analyzing relationships of fossil taxa. [Combining data; computer simulations; fossils; incomplete taxa; missing data; phylogenetic accuracy; supertrees.]

The problem of missing data is widely considered to be the most significant obstacle in reconstructing phylogenetic relationships of fossil taxa (e.g., Donoghue et al., 1989; Huelsenbeck, 1991; Anderson, 2001) and in combining data sets (e.g., different genes, morphology) that do not include identical taxa (Wiens and Reeder, 1995; Sanderson et al., 1998). For fossil taxa, only a fraction of the characters that can be evaluated in extant taxa can be assessed (e.g., generally no molecular, behavioral, or soft anatomical characters), and many characters that can be determined in fossil taxa in general may not be scored in individual specimens or taxa because of poor preservation (e.g., many taxa are known only from teeth or vertebrae). A character that cannot be scored for a particular taxon is typically coded as missing or unknown ("?") in a phylogenetic data matrix. Some authors have used the abundance of missing data in fossil taxa to justify exclusion of fossils entirely when reconstructing relationships among groups of living taxa (Patterson, 1981; Ax, 1987). Similarly, the desire to avoid coding taxa with missing data when combining partially overlapping data sets has been used by some authors (e.g., Sanderson et al., 1998) to justify the controversial phylogenetic supertree approach, which involves combining trees from separately analyzed data sets but not combining the data matrices themselves (e.g., Liu et al., 2001). Alternatively, these partially overlapping data matrices may be combined but the incomplete taxa are often excluded (reviewed by Wiens and Reeder, 1995). However, the mechanisms that may cause missing data and incomplete taxa to be problematic have remained unclear (e.g., contrast the different mechanisms proposed to explain the uncertain placement of incomplete taxa by Huelsenbeck, 1991; Wilkinson, 1995; Kearney, 2002). This is particularly true in terms of their effect on phylogenetic accuracy, which I define here as the similarity between the estimated tree and the true phylogeny.

Why are incomplete taxa and their missing data thought to be problematic? Numerous empirical studies, particularly of fossil organisms, have shown that including taxa with many missing data cells can lead to multiple shortest trees and poorly resolved consensus trees (e.g., Gauthier, 1986; Novacek, 1992; Wilkinson and Benton, 1995; Gao and Norell, 1998). These problems seem to be associated with the uncertain placement of the incomplete taxa (e.g., Huelsenbeck, 1991; Nixon and Wheeler, 1992; Wilkinson, 1995). Furthermore, computer simulations (Huelsenbeck, 1991) and analyses of known bacteriophage phylogenies (Wiens and Reeder, 1995) have shown that the increased number of shortest trees and decreased resolution associated with including highly incomplete taxa can lead to decreased phylogenetic accuracy, relative to including the same number of complete taxa. Huelsenbeck (1991:466) proposed that highly incomplete taxa decrease phylogenetic accuracy because their missing data cells increase the percentage of equivocally resolved ancestral characters, which leads to decreased resolution and thus to decreased phylogenetic accuracy.

Given these observations, decisions about whether or not to include taxa are often guided by how much data they are missing (e.g., Rowe, 1988; Grande and Bemis, 1998; Ebach and Ahyong, 2001), for example, excluding taxa that cannot be scored for >50% of the characters. This widely used approach implicitly assumes that incomplete taxa are problematic because of their proportion or absolute number of missing data cells and that the missing data cells are themselves to blame (see also Huelsenbeck, 1991). However, many empirical studies have also found that the impact of incomplete taxa on an analysis—particularly in terms of the number of

shortest trees and resolved clades—may have little to do with their level of completeness (e.g., Gauthier et al., 1988; Donoghue et al., 1989; Novacek, 1992; Wilkinson, 1995; Anderson, 2001; Kearney, 2002). An alternative hypothesis is that the so-called missing data problem for incomplete taxa simply reflects sampling of too few characters in these taxa to accurately place them on the tree. If this were true, the reduced accuracy associated with including highly incomplete taxa would disappear if enough characters were sampled in these taxa, regardless of their number of missing data cells. These hypotheses have never been explicitly tested.

I tested these competing hypotheses (too much missing data versus too few characters) using simulations. Data sets were simulated in which some taxa were complete (no missing data) and others were incomplete. I assessed the ability of parsimony and other methods to reconstruct the true phylogeny for all taxa given different amounts of missing data in the incomplete taxa and different overall numbers of characters. The results demonstrate that the so-called missing data problem for incomplete taxa (in terms of reduced accuracy) is caused primarily by sampling too few characters and not the amount of missing data that they bear. This surprising finding (1) demonstrates why taxa should not be excluded based on their level of completeness alone, (2) suggests a more effective strategy for dealing with incomplete taxa, (3) calls into question a justification for the controversial phylogenetic supertree approach, and (4) shows the potential for accurately resolving the phylogenetic relationships of highly incomplete fossil and living taxa. In this study, I focus on how the incompleteness of taxa may (or may not) affect our ability to accurately reconstruct their relationships rather than on directly comparing the consequences of including versus excluding incomplete taxa (see instead Wiens, 2003).

## Materials and Methods

Simulation methods generally followed those of Wiens (1998), from a study that explored the effects of missing data when adding sets of highly incomplete characters (rather than taxa). For the baseline simulations, unrooted trees of 16 taxa were simulated. Trees were either fully asymmetric (unbalanced; Fig. 1a) or fully symmetric (balanced; Fig. 1b) to span the range of all possible levels of symmetry for 16 taxa. Simulated characters were binary (states 0 and 1). For the purposes of this study, branch length was considered to be the probability of a given character changing state by the end of the branch. Various branch lengths were explored. In a given simulation, all branch lengths were equal to determine how a given length influences the results.

For a given replicate, a complete matrix was simulated, and then eight taxa were randomly selected to be incomplete. For these taxa, various proportions of their character data were replaced with missing data entries ("?"): 95%, 90%, 75%, 50%, 25%, or 0%. Different overall numbers of characters were also explored, ranging from 100 to 2,000. Simulations showed that this range encompassed conditions under which phylogenetic analyses that included highly incomplete taxa could be very inaccurate or very accurate. The overall number of characters was held constant for each simulation replicate, and so the number of parsimony-informative characters varied depending on the branch length (i.e., the probability of change on each branch, so that the longer the branch length, the greater the number of parsimony-informative characters) and other factors (e.g., number of taxa).

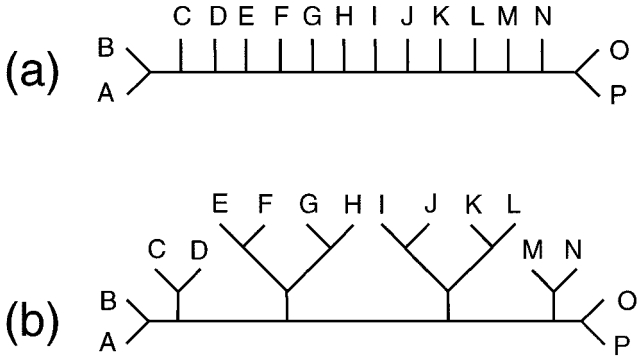


FIGURE 1. Model trees used in the simulations: (a) fully asymmetric; (b) fully symmetric.

Two methods for distributing missing data cells among characters for incomplete taxa were tested. First, the missing data cells were confined to the same set of characters in all incomplete taxa, mimicking the case in which data sets with different numbers of taxa are combined. Second, the prespecified number of missing data cells was distributed randomly among all characters, such that a different set of characters was randomly chosen to be incomplete in each incomplete taxon, mimicking the random preservation of characters in fossil taxa. The first method tends to minimize the number of incomplete characters, whereas the second tends to maximize the number of incomplete characters, even though the number of incomplete taxa and number of missing data cells are the same for each method. Results from the first method may be somewhat more general because even for fossil taxa the preservation of characters is unlikely to be completely random (i.e., hard parts are generally preserved more often than soft parts).

Programs for simulating data and tallying results were written in C by the author. Phylogenetic analyses were performed using Swofford's (2001) PAUP* program, version 4.0b8. Parsimony analyses utilized the tree bisection–reconnection (TBR) option for branch swapping and 20 random-addition sequence replicates per search. The maximum number of shortest trees retained in a search was set to 1,000 to allow searches to be completed in a reasonable amount of time. Restricting the number of shortest trees is a potential source of bias in that the consensus trees may appear to be more resolved than is actually supported by the data. However, this bias seems unlikely to be problematic because use of a single randomly chosen tree from among the shortest trees as a

phylogenetic estimate gives results similar to those from other methods (see below), which suggests that subsampling trees from among the shortest trees does not greatly affect the results.

For a given method and set of conditions, accuracy was measured as the number of clades shared between the true and estimated phylogenies, divided by the total number of clades (number of taxa −2) and averaged across 100 replicated matrices. The proportion of clades shared between trees is equivalent to the symmetric-difference distance between the trees (Penny and Hendy, 1985). The best method for measuring similarity between trees is an unresolved issue, but the metric used here is a standard measure in simulation studies (Hillis, 1995; Rannala et al., 1998), is easy to interpret, and is particularly useful for comparing trees that are relatively similar (Steel and Penny, 1993). The major result of this study is that, under some conditions, phylogenetic analyses that include highly incomplete taxa can consistently recover trees that are identical to the true phylogeny, despite large amounts of missing data. The finding that the true and estimated trees are identical is reflected perfectly by the measure of tree similarity used (Steel and Penny, 1993). However, because the symmetric-difference distance relies on the number of clades shared between the true and estimated trees, it may be sensitive to taxa of highly uncertain placement (e.g., many incomplete taxa). Use of this metric therefore provides a particularly strict test of the hypothesis that accurate phylogenies can be recovered when highly incomplete taxa are included. To directly address the placement of incomplete taxa and their effects on overall accuracy, a set of analyses was also performed in which the accuracy of the estimated trees was evaluated considering the complete and incomplete taxa separately (described below).

Multiple equally parsimonious trees are frequently generated when highly incomplete taxa are included. Accuracy was measured in two different ways when multiple shortest trees were produced (which both gave very similar results). First, accuracy was measured as the proportion of clades correctly resolved in a strict consensus of the shortest trees. Correct resolution of a clade may be the most relevant measure of success for empirical systematists, and it is a conservative measure of accuracy for this study (because it should underestimate accuracy when many taxa are highly incomplete). An alternative measure of accuracy was also explored—using the proportion of clades shared between the true tree and a single tree randomly chosen from among the set of shortest estimated trees. This measure should approximate the average accuracy among shortest trees when averaged across replicates (Rannala et al., 1998), which is recommended when multiple shortest trees are generated in simulation studies (i.e., Hillis, 1995; Rannala et al., 1998). To ensure that the single tree was fully resolved, zero-length branches were not collapsed. For a given method and set of conditions, the standard error of the mean for accuracy was small (consistently <2.5% and typically much smaller), suggesting that 100 replicates are adequate to evaluate accuracy. (These errors were too small to be effectively shown in the figures and are not depicted.)

The robustness of the basic results that were obtained using these simulation methods were explored by varying several parameters. In addition to randomly selecting taxa to be incomplete, a set of analyses was performed in which the incomplete taxa were evenly distributed on the tree (Fig. 1, taxa A, C, E, G, I, K, M, and O), and in another set of analyses the incomplete taxa were confined to a single lineage (Fig. 1, taxa A–H). Analyses were also performed in which the number of incomplete taxa was increased to 12 and decreased to 4. To evaluate the effects of increasing the overall number of taxa, the 64-taxon case was examined to compare to matched conditions in the 16-taxon case. Analyses in the 64-taxon case were extremely time intensive, and only 50 replicates were analyzed for each set of conditions.

A set of analyses was also performed in which DNA sequence data (i.e., up to four unordered states per character) were simulated. These analyses were also used to determine whether results based on parsimony could be generalized to other phylogenetic methods (neighbor joining, maximum likelihood). An extremely simple model of DNA sequence evolution was assumed (Jukes and Cantor, 1969), in which all types of substitutions were equally likely, rates of change were equal across sites, and all bases were at equal frequencies. These assumptions are typically violated in analyses using real DNA sequence data. However, this simple model allowed the assumptions of all three methods (parsimony, distance, and likelihood) to be met and thus allowed a more direct comparison of the effects of incomplete taxa on each method rather than confounding the missing data issue with that of how these different methods deal with more complex data. Furthermore, these complexities (i.e., unequal base frequencies and differences in rates of change among sites or substitution types) should have little direct bearing on the issue of missing data and incomplete taxa. Maximum-likelihood and neighbor-joining analyses assumed a Jukes–Cantor model with no invariant sites or among-site rate variation, and all substitutions and sites were equally weighted in the parsimony analyses. Optimal likelihood trees were sought using TBR branch swapping with five random-addition sequence replicates per search. Because likelihood searches can be extremely slow, only 50 replicated data matrices were examined for each set of conditions using likelihood. This study did not address how missing data may affect calculations of branch lengths or model parameters for likelihood or distance-based methods—this could be an interesting subject for future research.

The preceding analyses addressed the accuracy of the entire tree, including both complete and incomplete taxa. However, results of previous studies suggest that problems of poor resolution and accuracy that are associated with including incomplete taxa are caused only by the uncertain placement of these taxa and that relationships among the complete taxa may be unaffected (e.g., Huelsenbeck, 1991; Nixon and Wheeler, 1992; Wilkinson, 1995). To address this hypothesis explicitly for the first

time, a set of analyses was performed in which accuracy was compared for (1) all 16 taxa, including the 8 incomplete taxa, (2) the 8 complete taxa alone, by pruning the incomplete taxa from the tree including all 16 taxa, and (3) the 8 incomplete taxa alone, pruning out the complete taxa. For computational simplicity, the incomplete taxa were distributed evenly and consistently on the tree (Fig. 1, taxa A, C, E, G, I, K, M, and O) rather than being distributed randomly. These two ways of distributing incomplete taxa have no appreciable impact on the results (Wiens, unpubl.). Results are reported for the case in which the overall number of characters is relatively low (100), because this is the case in which the accuracies for the complete and incomplete taxa are most likely to differ (i.e., if the entire tree for all taxa is correctly resolved, all taxa will have the same accuracy).

Previous authors have also suggested that highly incomplete taxa are potentially problematic because they can lead to poor phylogenetic resolution (i.e., consensus trees with many polytomies), and Huelsenbeck (1991) suggested that this lack of resolution leads to poor phylogenetic accuracy. However, the relationships among taxon completeness, resolution, and accuracy have not been thoroughly examined. I compared these values for a set of conditions under which incomplete taxa were shown to be problematic (100 characters, binary data, branch length = 0.05, asymmetric model tree, eight incomplete taxa with the same characters incomplete in each one). I measured resolution as the proportion of nodes that are resolved (dichotomous) in a strict consensus of the shortest trees rather than using the number of trees (i.e., Huelsenbeck, 1991), which is a less direct measure of resolution. I then compared resolution to accuracy at different levels of taxon completeness, using the two measures of accuracy described previously (i.e., based on the number of correctly resolved nodes in a strict consensus of the shortest trees, and based on a single tree from among the shortest trees). I also compared resolution and accuracy when analyzing comparable numbers of complete characters alone (e.g., for 5% completeness and 100 characters, analyzing all 16 taxa based on only 5 characters) to further differentiate between the effects of missing data on resolution and limited number of characters.

In this study, I address the effects of incomplete taxa and missing data on phylogenetic accuracy. However, it would be more precise to say that I address how a widely used phylogeny reconstruction package (PAUP*) deals with missing data and how the analytical treatment of missing data by this program may affect phylogenetic accuracy. Other phylogenetic programs may deal with missing data differently, and whether the results of this study can be generalized to other programs will depend on how those programs deal with missing data. Under the parsimony criterion, PAUP* treats a given missing data cell as if it had the most-parsimonious state, given its placement on the tree based on other characters (Swofford, 2003), much like previous versions of PAUP have done. The unknown state in the missing data cell should not affect the placement of that taxon on the tree, supporting the idea that the number of complete cells in an incomplete taxon should determine the impact of that taxon on phylogenetic accuracy more than the amount of missing data it bears. How the sets of complete and incomplete characters and taxa interact to determine phylogenetic accuracy remains uncertain, however. For maximum-likelihood analysis of DNA sequence data in PAUP*, the likelihood for a matrix containing missing data is computed by summing the likelihoods over each possible assignment of A, C, G, or T to each missing data cell (Swofford, 2003). For distance-based methods, there are two options for dealing with missing data in PAUP* (Swofford, 2003). The program can either ignore characters with missing data when calculating distances or can "distribute them proportionately to unambiguous changes" (meaning that distances are calculated as if any of the four bases were an equally likely assignment to the missing data cells; Swofford, 2003). Both options gave similar results for the conditions that I examined (Wiens, unpubl.), and results are presented for only the second (default) option.

## RESULTS

The basic results (Fig. 2) support the hypothesis that the problem of analyzing incomplete taxa stems from including too few characters rather than too many missing data cells. For a limited number of characters (100), accuracy decreases rapidly as the proportion of missing data cells in the incomplete taxa increases, as expected (Huelsenbeck, 1991; Wiens and Reeder, 1995). However, when the overall number of characters is increased, the estimated trees can be perfectly accurate (all clades correctly resolved) even though they include many taxa that are highly incomplete and that have nearly 2,000 missing data cells each. The results also show the importance of the distribution of missing data cells. Characters that are complete (scored across all taxa) increase accuracy much more effectively in highly incomplete taxa than do characters that are scored in only some taxa (i.e., when the missing data cells are randomly distributed among characters in the incomplete taxa).

Similar results are obtained when accuracy is measured based on a single tree from among the shortest trees (Figs. 2c, 2d) and for different tree shapes and branch lengths (Fig. 3). However, when branches are extremely long, the accuracy of trees with highly incomplete taxa remains low, despite the large number of characters (Fig. 3c). Under these conditions, most of the characters are incomplete, and these incomplete characters seem to be affected by long-branch attraction caused by incomplete taxon sampling (as described by Wiens, 1998). Branches are so long under these conditions that even thousands of complete parsimony-informative characters cannot resolve the phylogeny correctly, and these conditions may be very unusual in empirical data sets.

The same basic results (Fig. 2) are also supported using different distributions and numbers of incomplete taxa and a higher overall number of taxa (not shown). Results using simulated DNA sequence data and using

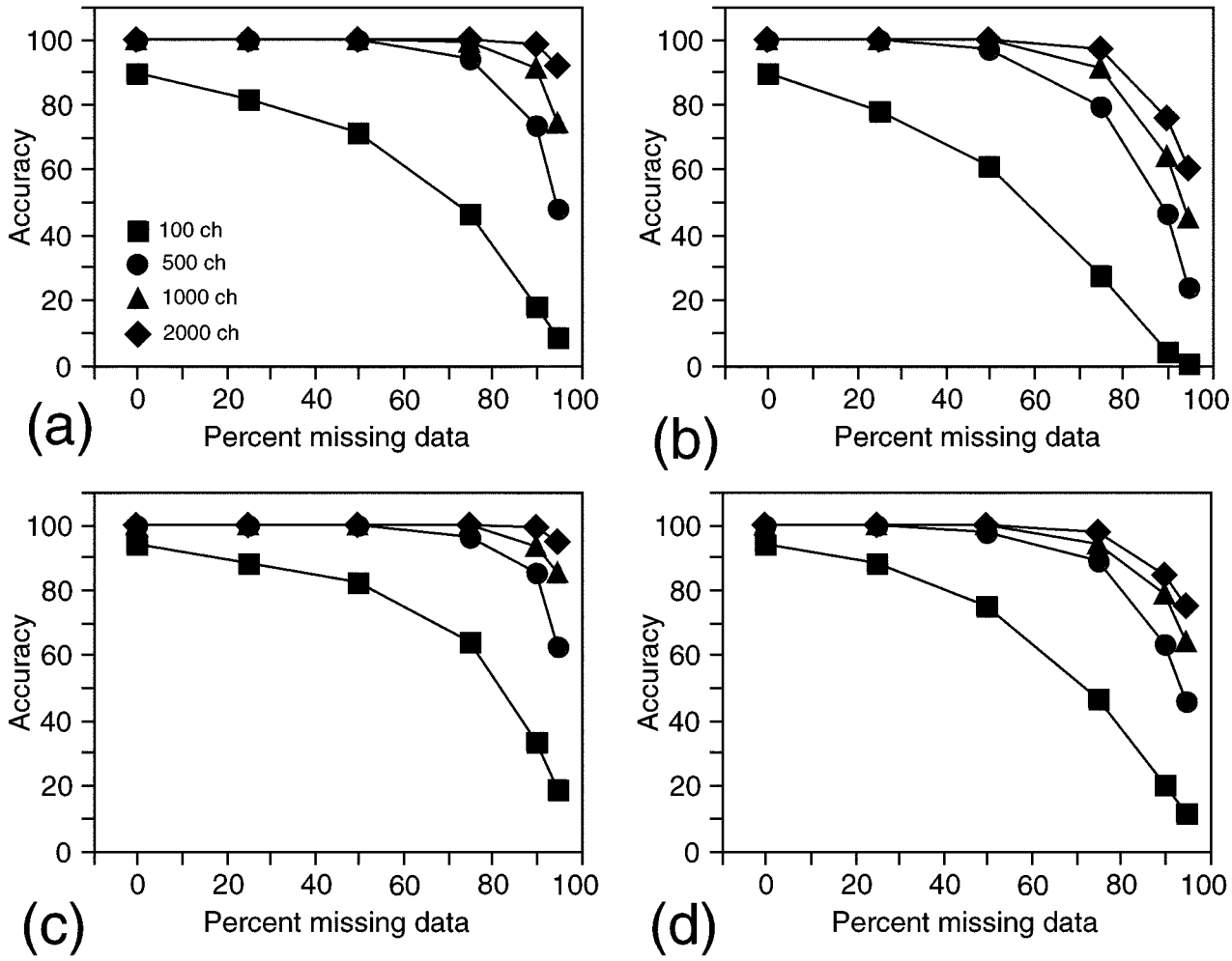


FIGURE 2. Effects of level of completeness and number of characters on phylogenetic accuracy when missing data cells are confined to the same set of characters in all incomplete taxa (a, c) and randomly distributed among characters for each incomplete taxa (b, d). In a and b, accuracy is measured as the proportion of clades that are correctly resolved, whereas in c and d accuracy is based on the proportion of correctly estimated clades in a single randomly chosen tree from among the shortest trees generated by a given search. Results are based on 16 taxa, binary character data, a fully asymmetric tree, and branch length = 0.05 (58% of characters parsimony informative), with 8 incomplete taxa selected randomly.

neighbor-joining and maximum-likelihood analyses are also similar overall (Fig. 4). However, under conditions where 90–95% of the data are missing in the incomplete taxa and the missing data cells are distributed randomly among characters in the incomplete taxa, parsimony tends to outperform neighbor joining and maximum likelihood outperforms parsimony. The speculation by Gatesy et al. (2002) that likelihood and distance methods are generally more sensitive to missing data than parsimony is not supported by these limited results.

The reduced accuracy that results from including incomplete taxa (relative to complete taxa) is associated with incorrect placement of only the incomplete taxa, and not the complete taxa (Fig. 5). Under conditions in which the analysis of all taxa gives relatively inaccurate results (i.e., 95% missing data, 100 characters), the relationships among the complete taxa are estimated almost perfectly. The low accuracy associated with including incomplete taxa results from incorrect or uncertain placement of the incomplete taxa relative to each other and relative to the complete taxa.

Comparisons of resolution and accuracy suggest that decreased resolution is the proximate mechanism by which inclusion of incomplete taxa reduces overall accuracy (Table 1). Under conditions where accuracy is extremely low, very few nodes are resolved. However, among the nodes that are resolved, at least some may be resolved incorrectly (given that the proportion of resolved clades is greater than the proportion of correctly resolved clades). The poor resolution associated with including highly incomplete taxa is seemingly caused by the limited number of complete characters rather than by the missing data per se. The results (Table 1) show that similar levels of resolution and accuracy are obtained when analyzing comparable numbers of complete characters in reduced data sets that lack missing data entries. The low accuracy under these conditions is not merely an artifact caused by poor resolution of strict consensus

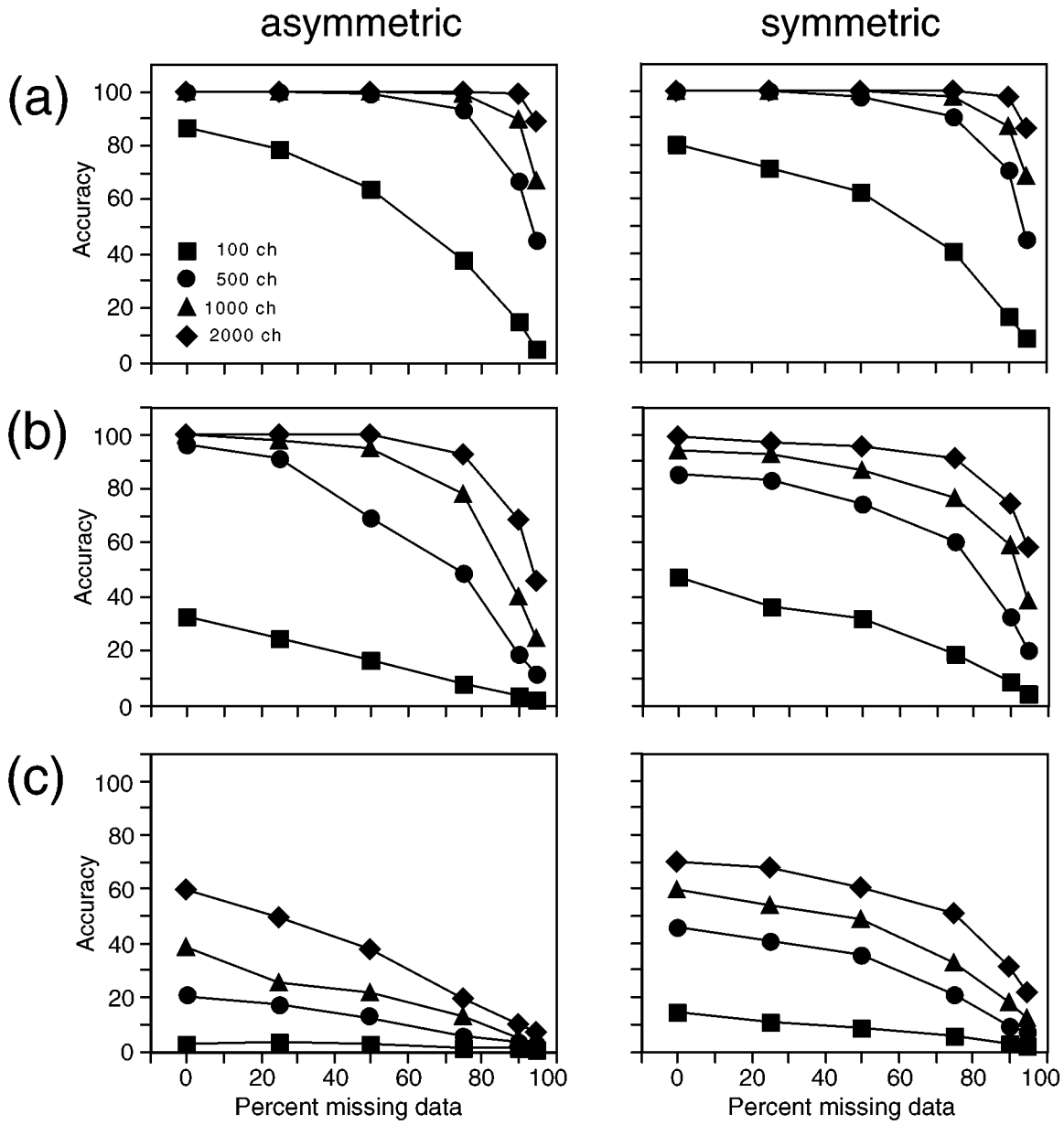


FIGURE 3. Effects of completeness and number of characters on phylogenetic accuracy for different tree shapes and branch lengths: (a) length = 0.10 (approximately 86% of characters parsimony informative); (b) length = 0.20 (98%); (c) length = 0.30 (100%). Results are based on 16 taxa, binary character data, 8 incomplete taxa selected randomly, and missing data cells confined to the same set of characters in all incomplete taxa, with accuracy based on the proportion of clades that are correctly resolved.

trees; accuracy is also low when based on a single shortest tree. If taxa are too incomplete to be localized on the estimated phylogeny, their placement may be largely random in individual trees, leading both to poor resolution and poor accuracy.

## DISCUSSION

The results of this study support the hypothesis that the missing data problem for incomplete taxa is primarily one of including too few characters rather than including too many missing data cells. This result has a simple explanation. The reduced accuracy associated with including incomplete taxa stems largely from their poorly resolved placement (Fig. 4; as hypothesized by Huelsenbeck, 1991), and the number of characters scored in the incomplete taxa is critical for correctly placing these taxa on the tree (Figs. 2, 3). In theory, only a single character may be necessary to correctly resolve the position of an incomplete taxon (i.e., a unique synapomorphy shared with its sister taxon). Increasing the number of characters sampled increases the probability that such

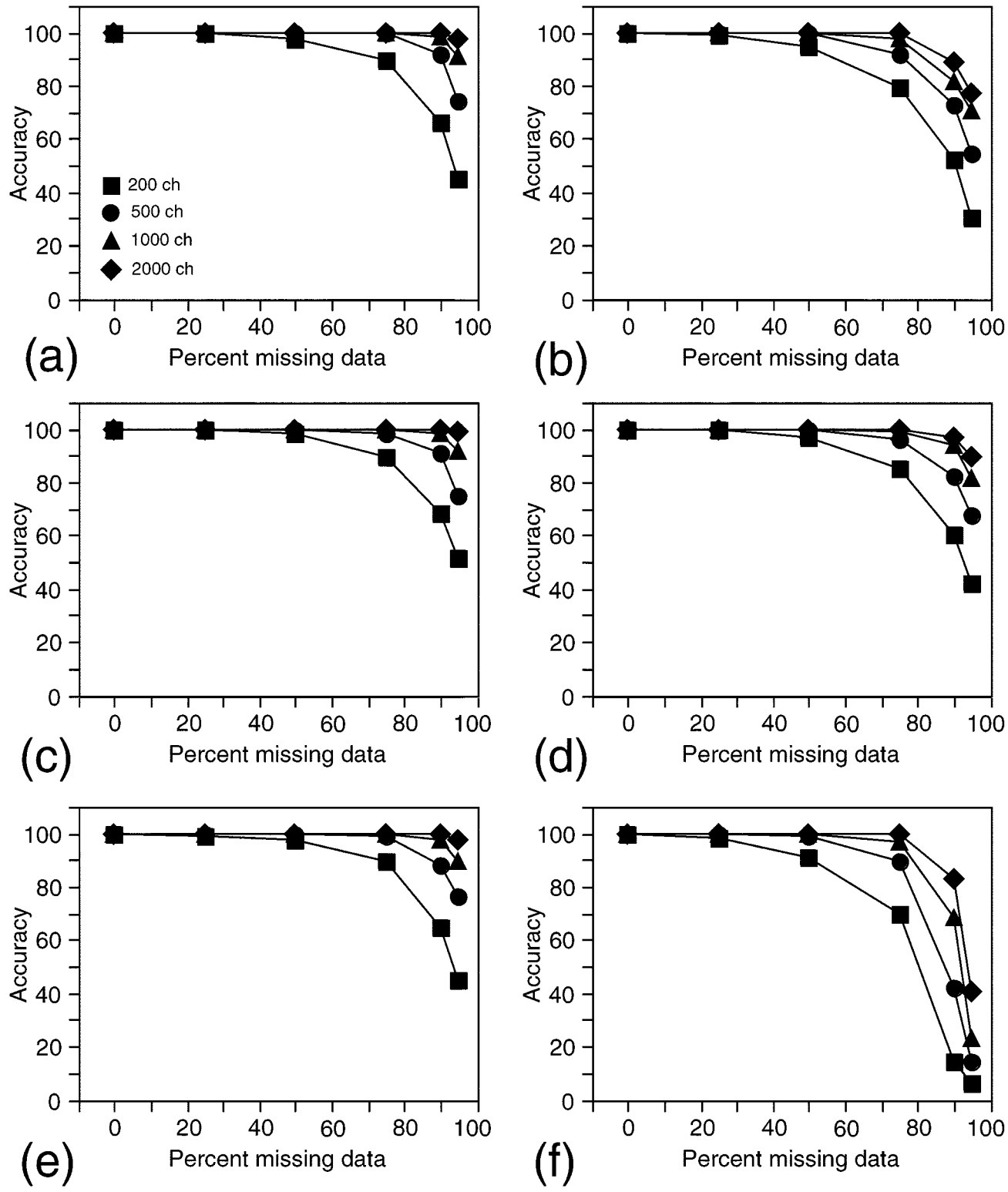


FIGURE 4. Effects of completeness and number of characters on phylogenetic accuracy for DNA sequence data analyzed with parsimony (a, b), likelihood (c, d), and neighbor joining (e, f). Missing data cells are either confined to the same set of characters in all incomplete taxa (a, c, e) or are randomly distributed among characters for each incomplete taxon (b, d, f). For parsimony and likelihood, accuracy is scored as the proportion of clades shared between the true tree and a single tree chosen randomly from among the optimal trees generated by a given search (so that all three methods have the same level of resolution in the estimated trees). The lowest number of characters used in these simulations was 200 because of problems in implementing neighbor joining and maximum likelihood when all four bases are not present. The model tree is fully asymmetric, with branch length = 0.05 (52% of characters parsimony informative).

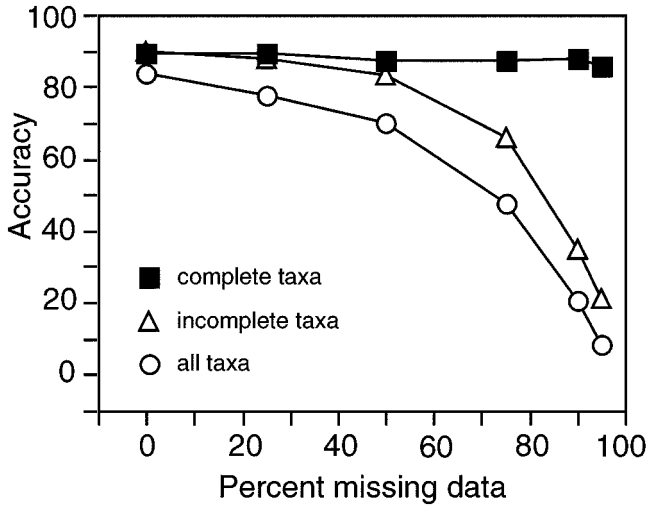


FIGURE 5. Reduced accuracy associated with including highly incomplete taxa (relative to including complete taxa) is caused by incorrect placement of the incomplete taxa. Accuracy is based: ○ = all 16 taxa, including the 8 incomplete taxa; △ = the 8 incomplete taxa alone, pruning the complete taxa from the tree after the analysis; ■ = the 8 complete taxa alone, pruning the incomplete taxa from the tree after the analysis. Results are based on 100 binary characters, a fully asymmetric tree, missing data cells confined to the same set of characters in all incomplete taxa, and branch length = 0.05 (58% of characters parsimony informative), with accuracy based on the proportion of correctly estimated clades in a single randomly chosen tree from among the shortest trees generated by given search. Note that accuracy is lower for trees including all 16 taxa than for those with incomplete taxa alone. Accuracy often decreases with increasing numbers of taxa, even when taxon addition increases accuracy for a fixed set of taxa (Swofford and Olsen, 1990; Wiens and Reeder, 1995). When assessed based on eight taxa (four complete, four incomplete), accuracy is generally intermediate between values for the eight complete taxa alone and values for the eight incomplete taxa alone (results not shown).

TABLE 1. Resolution and accuracy are strongly associated across different levels of completeness (above) and different numbers of characters (below). Above, results are provided for different levels of completeness in analyses that include 8 randomly selected incomplete taxa in an asymmetric 16-taxon tree with 100 binary characters and a branch length of 0.05. The small number of characters and short branch lengths were chosen to increase the impact of incomplete taxa on phylogenetic accuracy, and different tree shapes and types of characters gave similar results. Below, results are provided for the same matrices with the sets of different numbers of characters that are complete for all taxa. Each result is the average of 100 replicates (±1 SE). Resolution is the proportion of clades that are dichotomous in a strict consensus of the shortest trees from a given search. Accuracy (correct resolution) is the number of clades shared between the true tree and a strict consensus of the shortest trees or a single tree randomly selected from among the shortest trees (for a given search, averaged across 100 replicates). Resolution and accuracy are higher for the analyses including all characters (above) because adding characters should generally increase phylogenetic accuracy despite their incompleteness (Wiens, 1998). Standard linear regression of resolution against accuracy for these average values ($n = 6$) gives $R^2 > 0.990$ and $P < 0.0001$ for both measures of accuracy. Regressing accuracy and resolution using individual simulation replicates as data points ($n = 600$) gives similarly strong relationships, with $P < 0.0001$ for all comparisons ($R^2 > 0.920$ for accuracy based on consensus trees, and $R^2 > 0.790$ for accuracy based on a single tree).

| | | Accuracy | |
|---|---|---|---|
| | Resolution | Strict consensus tree | Single tree |
| Completeness (% complete data cells) | | | |
| 5 | 0.122 ± 0.011 | 0.072 ± 0.009 | 0.175 ± 0.013 |
| 10 | 0.263 ± 0.017 | 0.160 ± 0.013 | 0.316 ± 0.018 |
| 25 | 0.554 ± 0.021 | 0.463 ± 0.020 | 0.629 ± 0.016 |
| 50 | 0.788 ± 0.015 | 0.702 ± 0.015 | 0.784 ± 0.013 |
| 75 | 0.886 ± 0.010 | 0.835 ± 0.012 | 0.885 ± 0.010 |
| 100 | 0.915 ± 0.010 | 0.891 ± 0.009 | 0.931 ± 0.007 |
| Number of characters scored | | | |
| 5 | 0.098 ± 0.008 | 0.044 ± 0.007 | 0.108 ± 0.011 |
| 10 | 0.203 ± 0.014 | 0.130 ± 0.012 | 0.256 ± 0.017 |
| 25 | 0.474 ± 0.021 | 0.363 ± 0.018 | 0.507 ± 0.018 |
| 50 | 0.742 ± 0.018 | 0.662 ± 0.018 | 0.759 ± 0.015 |
| 75 | 0.879 ± 0.010 | 0.831 ± 0.012 | 0.891 ± 0.009 |
| 100 | 0.915 ± 0.010 | 0.891 ± 0.009 | 0.931 ± 0.007 |

key characters will be found. If enough characters have been sampled to accurately place all of the incomplete taxa on the tree, the amount of missing data in these taxa seems to have little effect (except when branches are very long). These simulations show that it is possible to estimate trees that are fully resolved and fully correct even when half of the taxa have 90% of their data coded as missing and have nearly 2,000 missing data cells each. Clearly, the missing data cells are not by themselves misleading. Thus, paradoxically, the so-called missing data problem for incomplete taxa can be unrelated to the number or proportion of missing data cells that these taxa bear.

The results of this study clarify two related issues surrounding the inclusion of incomplete taxa. First, under the expected conditions where including incomplete taxa leads to inaccurate results (for shorter branch lengths), the reduced accuracy is associated with incorrect placement of the incomplete taxa alone (Fig. 5). In fact the estimated relationships among the complete taxa are seemingly unaffected by inclusion of the incomplete taxa. Second, this reduced accuracy is associated with poor resolution (i.e., trees with many polytomies; Table 1). However, contrary to previous hypotheses (e.g., Huelsenbeck, 1991), poor resolution is not caused by the missing data cells themselves but rather by the limited number of characters scored in these taxa. This hypothesis is supported by the similar levels of resolution in trees with incomplete taxa and those with comparable numbers of complete characters analyzed alone (Table 1). An insufficient sampling of characters in an incomplete taxon may lead to poor accuracy both (1) through incomplete resolution (i.e., if no characters can place the taxon in a specific clade, then the true tree cannot be correctly resolved and the average accuracy among the multitude of possible trees will be low) and (2) by increasing the chances that the taxon is spuriously placed on the tree by one or more homoplastic characters.

The distinction between the hypotheses of "too few characters" and "too much missing data" may be subtle, but identifying the correct mechanism that causes incomplete taxa to be problematic may have important implications for empirical studies. A common practice in many studies is to exclude taxa based on their proportion of missing data cells (e.g., Rowe, 1988; Grande and Bemis, 1998; Ebach and Ahyong, 2001), even if this is not always

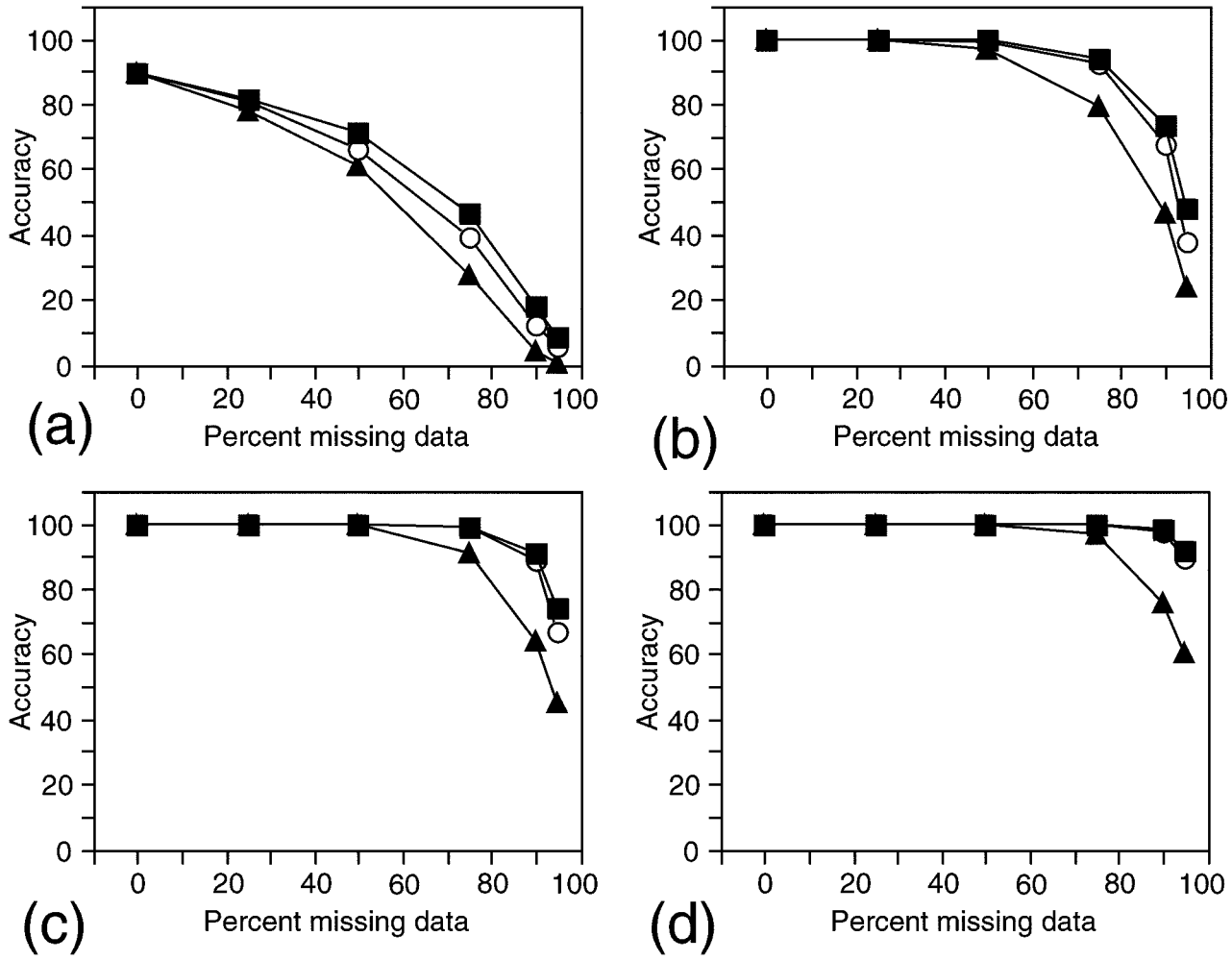


FIGURE 6. Phylogenetic accuracy in analyses that include incomplete taxa closely matches accuracy based on analysis of the set of complete characters alone: ○ = complete characters only (data set 1 alone), no missing data; ■ = data sets 1 and 2, missing data cells confined to same set of characters in all incomplete taxa (Fig. 2a); ▲ = data sets 1 and 2 combined, missing data cells randomly distributed across all characters in each incomplete taxon. (a) 100 characters; (b) 500 characters; (c) 1,000 characters; (d) 2,000 characters. Results are based on 16 taxa, binary character data, a fully asymmetric tree, and branch length = 0.05, with 8 incomplete taxa selected randomly.

explicitly stated. The results of the present study show that the proportion of missing data cells in the incomplete taxa is a poor predictor of their impact on phylogenetic accuracy. A much better predictor is the number of characters that can be scored in the incomplete taxa. The overall accuracy for trees that include incomplete taxa seems to be closely related to accuracy based only on the characters that can be scored in all taxa (Table 1; Fig. 6). Results of previous empirical studies have also suggested that the incompleteness of a taxon may be unrelated to its impact on a phylogenetic analysis in terms of resolution (e.g., Gauthier et al., 1988; Donoghue et al., 1989; Novacek, 1992; Wilkinson, 1995; Anderson, 2001; Kearney, 2002). However, these studies did not address phylogenetic accuracy or the effects of different numbers of characters on resolution. The simulation results of the present study, which show that accurately resolving the placement of incomplete taxa depends on sampling characters and not on the amount of missing data, suggest an underlying explanation that may reconcile the idiosyncratic findings of these empirical studies.

The phylogenetic supertree method combines trees derived from separate analyses of different data sets rather than combining actual data matrices. Sanderson et al. (1998) suggested that an important justification for the supertree method is that it avoids the necessity of coding taxa with large numbers of missing data cells when data matrices with incompletely overlapping taxa are combined into a single matrix. If these data matrices are combined, then the incomplete taxa are often excluded from the combined analyses. The results of the present study suggest that including taxa that are incomplete in a combined analysis is unlikely to be problematic as long as there are sufficient characters in one broadly sampled data set to allow the position of these taxa to be resolved (see also Bininda-Emonds and Sanderson, 2001). Thus, combining data sets with different numbers of taxa into a single "supermatrix" may be possible (and effective)

under a greater variety of circumstances than previously thought.

The results of this study also suggest a general solution to the problem of including highly incomplete taxa—increasing the number of characters scored in these taxa. Unlike approaches in which the researcher strives to identify and eliminate incomplete taxa that are potentially problematic or uninformative (i.e., Wilkinson, 1995; Anderson, 2001), increasing the number of characters offers the potential to address and resolve the relationships of all of the taxa of interest. Of course, adding characters may be difficult in some cases, particularly for highly fragmentary fossil taxa. In these cases, a related strategy may be to extract as much information as possible from the characters that can be scored. One way to accomplish this goal is by treating morphological characters as continuous rather than qualitative (Wiens, 2001). Many morphological characters are intrinsically continuous and quantitative and are merely made discrete through the language of description (e.g., describing the length of a structure as "long" vs. "short"), which entails considerable loss of information (Wiens, 2001). Analyzing these characters directly as continuous utilizes the maximum information possible (Wiens, 2001).

Finally, and most importantly, the present study shows that the amount or proportion of missing data present in a taxon need not be a limitation on its accurate phylogenetic placement. Therefore, these results suggest that it may be possible to reconstruct accurate phylogenies for many more living and fossil taxa than previously supposed, regardless of their level of completeness.

The results of this study suggest that limited number of complete characters may be the most important factor limiting the accurate placement of incomplete taxa. Nevertheless, when branches are long and the rate of character-state change is very high, analyses with highly incomplete taxa may yield inaccurate results despite large numbers of characters (e.g., Fig. 3c), suggesting a somewhat different mechanism in these cases. Results of previous simulations (Wiens, 1998) suggest that replacing known data cells with missing data cells may exacerbate the effects of long-branch attraction among the complete taxa, mimicking the effects of limited taxon sampling. Even in this case, the missing data cells are not themselves misleading; analysis of the complete taxa alone may give similarly inaccurate results under these conditions (Wiens, 1998, 2003). Instead, the limited number of known characters in the incomplete taxa seems to limit their ability to "rescue" the analysis from long-branch attraction among the complete taxa (i.e., a small number of complete characters is unable to overturn a hypothesis supported by many incomplete characters). In these cases, the effects of the limited number of complete characters in the incomplete taxa may interact with long-branch attraction to reduce phylogenetic accuracy.

The goal of this study has been identify the general mechanisms by which missing data may affect phylogenetic accuracy, and not to generate simulated data that match the complexity of empirical data. Therefore, as always, readers should be appropriately cautious about directly extrapolating specific simulation results to specific results in the real world (e.g., assuming that a taxon with >100 characters and 50% missing data will have no impact on phylogenetic accuracy based on these simulations). A safer approach may be to use parametric simulations to test the effects of incompleteness for a specific set of conditions encountered in the real world.

## ACKNOWLEDGMENTS

I thank O. Bininda-Emonds, P. Fortey, M. Kearney, B. Livezey, A. Purvis, J. Rawlins, M. Servedio, and M. Wilkinson for comments on drafts of the manuscript.

## REFERENCES

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

AX, P. 1987. The phylogenetic system: The systematization of organisms on the basis of their phylogenesis. Wiley, New York.

BININDA-EMONDS, O. R. P., AND M. J. SANDERSON. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree reconstruction. Syst. Biol. 50:565–579.

DONOGHUE, M. J., J. A. DOYLE, J. GAUTHIER, A. G. KLUGE, AND T. ROWE. 1989. The importance of fossils in phylogeny reconstruction. Annu. Rev. Ecol. Syst. 20:431–460.

EBACH, M. C., AND S. T. AHYONG. 2001. Phylogeny of the trilobite subgenus *Acanthopyge* (*Lobopyge*). Cladistics 17:1–10.

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. Am. Mus. Novit. 3230:1–51.

GATESY, J., C. MATTHEE, R. DESALLE, AND C. HAYASHI. 2002. Resolution of a supertrees/supermatrix paradox. Syst. Biol. 51:652–664.

GAUTHIER, J. 1986. Saurischian monophyly and the origin of birds. Mem. Calif. Acad. Sci. 8:1–47.

GAUTHIER, J., 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. Soc. Vertebr. Paleontol. Mem. 4:1–690.

HILLIS, D. M. 1995. Approaches to assessing phylogenetic accuracy. Syst. Biol. 44:3–16.

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

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

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

LIU, F.-G., M. M. MIYAMOTO, N. P. FREIRE, P. ONG, M. R. TENNANT, T. S. YOUNG, AND K. F. GUGEL. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291:1786–1789.

NIXON, K. C., AND Q. D. WHEELER. 1992. Extinction and the origin of species. Pages 119–143 *in* Extinction and phylogeny (M. J. Novacek and Q. D. Wheeler, eds.). Columbia Univ. Press, New York.

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

PATTERSON, C. 1981. Significance of fossils in determining evolutionary relationships. Annu. Rev. Ecol. Syst. 12:195–223.

PENNY, D., AND M. D. HENDY. 1985. The use of tree comparison metrics. Syst. Zool. 34:75–82.

RANNALA, B., J. P. HUELSENBECK, Z. YANG, AND R. NIELSEN. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

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

SANDERSON, M. J., A. PURVIS, AND C. HENZE. 1998. Phylogenetic supertrees: Assembling the tree of life. Trends Ecol. Evol. 13:105–109.
STEEL, M. A., AND D. PENNY. 1993. Distributions of tree comparison metrics—Some new results. Syst. Biol. 42:126–141.
SWOFFORD, D. L. 2001. PAUP*: Phylogenetic analysis using parsimony (* and other methods), version 4.0b8. Sinauer, Sunderland, Massachusetts.
SWOFFORD, D. L. 2003. http://paup.csit.fsu.edu/paupfaq/faq.html.
SWOFFORD, D. L., AND G. J. OLSEN. 1990. Phylogeny reconstruction. Pages 411–501 *in* Molecular systematics, 1st edition (D. M. Hillis and C. Moritz, eds.). Sinauer, Sunderland, Massachusetts.
WIENS, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.
WIENS, J. J. 2001. Character analysis in morphological phylogenetics: Problems and solutions. Syst. Biol. 50:689–699.
WIENS, J. J. 2003. Incomplete taxa, incomplete characters, and phylogenetic accuracy: Is there a missing data problem? J. Vertebr. Paleontol. 23:297–310.
WIENS, J. J., AND T. W. REEDER. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. 44:548–558.
WILKINSON, M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44:501–514.
WILKINSON, M., AND M. J. BENTON. 1995. Missing data and rhynchosaur phylogeny. Hist. Biol. 10:137–150.

*First submitted 22 September 2002; reviews returned 23 November 2002; final acceptance 24 March 2003*
*Associate Editor: Mike Steel*