Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 34–42
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Missing data and the design of phylogenetic analyses

John J. Wiens *

Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA

Received 5 March 2005
Available online 6 May 2005

Abstract

Concerns about the deleterious effects of missing data may often determine which characters and taxa are included in phylogenetic analyses. For example, researchers may exclude taxa lacking data for some genes or exclude a gene lacking data in some taxa. Yet, there may be very little evidence to support these decisions. In this paper, I review the effects of missing data on phylogenetic analyses. Recent simulations suggest that highly incomplete taxa can be accurately placed in phylogenies, as long as many characters have been sampled overall. Furthermore, adding incomplete taxa can dramatically improve results in some cases by subdividing misleading long branches. Adding characters with missing data can also improve accuracy, although there is a risk of long-branch attraction in some cases. Consideration of how missing data does (or does not) affect phylogenetic analyses may allow researchers to design studies that can reconstruct large phylogenies quickly, economically, and accurately.

Keywords: Accuracy; Bayesian analysis; Maximum likelihood; Missing data; Neighbor-joining; Parsimony; Phylogeny; Phylogenetic method; Systematics

1. Introduction

In recent decades, phylogenetic methods have become widely used in medical research. For example, phylogenetic trees are now frequently used to trace the origin, spread, and evolution of viruses and other diseases [1–5]. But just as in any other application of phylogenetics, conclusions that are drawn from phylogenies may hinge on details of data analysis and study design [6,7].

The issue of "missing data" in phylogenetics is an important one, in part because it may (implicitly) determine how phylogenetic studies are designed. In this paper, "missing data" refers to empty cells of a character-by-taxon data matrix that is used in a phylogenetic analysis (Fig. 1). These empty cells are usually indicated by a "?." A common type of situation in which missing data cells would be encountered is in a phylogenetic analysis of (for example) 10 taxa (individuals, species) based on a combined analysis of two gene regions, in which five taxa are lacking data for the second gene. In this case, the taxa that lack data for the second gene would be coded as missing or unknown for those characters. How will these missing data cells affect the phylogenetic analysis? Will their inclusion increase or decrease the chances of reconstructing the correct phylogeny? How will they affect branch support (e.g., bootstrap values, Bayesian posterior probabilities)? In our hypothetical example, should the five taxa that are missing data from the second gene be included at all? Or should all 10 taxa be included, but only for one gene? These are the types of questions that will be addressed in this paper.

It appears that the desire to avoid missing data, whether justified or not, actually determines the design of many empirical phylogenetic studies at a fundamental level (i.e., which taxa and characters are included). Most phylogenetic studies report analyses that are based on relatively complete data sets, with sequences for all genes obtained for all or most taxa. Yet, according to

* Fax: +1 631 632 7626.
E-mail address: wiensj@life.bio.sunysb.edu.

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.04.001

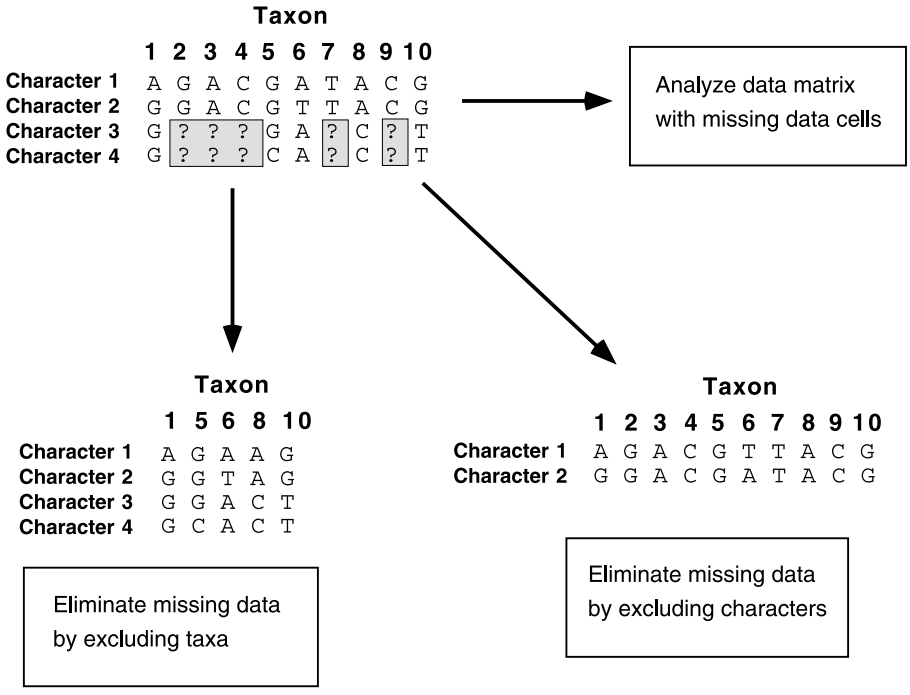

Fig. 1. Hypothetical example illustrating missing data in phylogenetic analysis. Taxa two, three, four, seven, and nine lack data for characters 3 and 4. If the researcher includes all of these data in a single analysis, there will be missing data cells ("?"). A researcher might choose to deal with this situation by deleting these taxa, deleting characters 3 and 4, or by simply including all the characters and taxa. The first two options are based on the implicit assumption that including these five taxa and characters 3 and 4 will somehow be problematic because of the effects of missing data.

Sanderson and Driskell [8], the sampling of genes and taxa for a given group of organisms in GenBank (where the DNA sequence data for almost all published molecular phylogenetic studies is archived) is quite sparse. Thus, even though a few taxa are sequenced for most genes (e.g., those with sequenced genomes) most taxa represented in GenBank are sampled for a limited number of genes, and the sets of genes that are sequenced may show only limited overlap across phylogenetic studies. Combining existing gene sequences for any taxonomic group would create a matrix dominated by missing data cells [9]. Thus, empirical studies are designed to avoid missing data, whereas the existing database of available sequences consists of large amounts of missing data. Although it seems unlikely that simply adding missing data cells would have positive benefits on an analysis, missing data cells are added in the form of incomplete taxa or characters. Thus, eliminating missing data cells generally means eliminating non-missing data as well, whether by excluding taxa or excluding characters. The question then becomes: when do the benefits of excluding missing data outweigh the costs of excluding these characters and taxa?

In fact, the "fear" of missing data was an important motivation for the controversial phylogenetic supertree approach [10]. The supertree approach involves combining trees from different phylogenetic studies to extend the taxonomic scope of these individual studies. However, it does not involve the actual analysis of character data, in part to avoid the large amounts of missing data that are associated with combining data sets with different sets of taxa [10]. Similarly, methods are being designed to select sets of genes from sequence databases to avoid or minimize missing data [11].

The desire to avoid missing data extends to more theoretical studies as well. For example, there has been considerable debate in recent years over the merits of sampling more characters versus taxa for a given phylogenetic study [12–23]. However, these papers have consistently assumed that the same characters are scored in all the taxa, and that there are no missing data.

Why are missing data generally avoided? In fact, this is rarely explicitly discussed in molecular phylogenetic studies. Most of the literature on the effects of missing data is associated with the problem of analyzing highly incomplete fossil taxa [24–30], but see [31,32]. These paleontologically oriented studies have suggested that highly incomplete taxa can be problematic in that their inclusion may lead to many equally parsimonious trees with a poorly resolved consensus tree (i.e., the relationships are uncertain). Not only will these incomplete taxa be difficult to place on the phylogeny, but they may also obscure relationships among the more complete taxa (in the consensus tree). Nevertheless, some of these studies have also suggested that there is no obvious relationship between how complete a taxon is and how it will influence an analysis [26,30]. Using simulations, Huelsenbeck [27] concluded that inclusion of highly incomplete taxa

may decrease the probability of reconstructing the true phylogeny, and that the studies which found no obvious relationship between completeness and resolution were not typical.

In this paper, I will review the results of recent studies which have attempted to address the impact of missing data on phylogenetic analysis. Most of these studies are based on computer simulations. Simulations have become the most widely used approach for assessing the accuracy of phylogenetic approaches, that is, the ability of these methods to reconstruct the true phylogeny [33]. The most important advantage of simulations is that they provide a context in which the true phylogeny is known with certainty. Furthermore, simulations allow one to systematically manipulate and thereby understand the parameters that may affect phylogenetic accuracy, including number of taxa, number of characters, proportion of missing data, tree shape, branch lengths, taxon sampling, and rates of evolution. Simulated data sets never match the complexity of data sets in the real world, making it problematic to generalize from a specific simulation result to a specific empirical study (e.g., concluding from simulations that accuracy will always be >90% if the added taxa are >50% complete). However, this unrealistic simplicity can be beneficial in helping to understand the general mechanisms that influence phylogenetic accuracy (i.e., by controlling all potentially confounding variables).

2. Can incomplete taxa be accurately placed on a phylogeny?

Most of the literature on the effects of missing data has focused on the inclusion and exclusion of taxa that are relatively incomplete. There are two fundamental questions regarding the addition of highly incomplete taxa. First, can the incomplete taxa be accurately placed on a phylogeny? In other words, when taxa with abundant missing data are added to a matrix with more complete taxa, is it possible to accurately resolve the true phylogenetic position of these incomplete taxa? Even if the incomplete taxa do not change relationships estimated for the complete taxa, it may still be valuable to resolve the phylogenetic position of all relevant taxa. The second question is whether or not the addition of incomplete taxa will actually improve the estimate of relationships for the more complete taxa alone. For example, there has been considerable debate over whether adding incomplete fossil taxa can change relationships reconstructed for living taxa alone [25–27,34].

I will focus on the first question first. Many authors have noted that including highly incomplete taxa often can lead to poorly resolved phylogenetic relationships [24,28,35,36], but not always [30]. Analyses using simulations [27] and known, laboratory-produced viral phylogenies [31] have also suggested that highly incomplete taxa can lead to reduced phylogenetic accuracy (measured for all of the taxa). What is much less clear is how exactly they produce these negative effects.

There are two basic hypotheses for how the incompleteness of a taxon might negatively impact an analysis. First, the missing data cells themselves may be problematic. For example, Huelsenbeck [27] suggested that the proportion of missing data cells increased the number of ambiguously resolved character states at each node. Similarly, many authors have excluded taxa based on their proportion of missing data cells, given the implicit assumption that the relative or absolute number of missing data cells is somehow critical [37–39].

Second, incomplete taxa may be problematic because there are too few characters to accurately place them on the tree [40]. Thus, the placement of these taxa may be either unresolved or incorrect (e.g., if one of the few phylogenetically informative characters that is scored in these taxa happens to be homoplastic).

These hypotheses can easily be tested by simulating phylogenies with highly incomplete taxa and different overall numbers of characters. If the first hypothesis is true, then including highly incomplete taxa will lead to poorly resolved and inaccurate phylogenies regardless of the overall number of characters. If the second hypothesis is true, then phylogenies that include highly incomplete taxa will be accurately reconstructed as long as the overall number of characters is large (because the problem of "too few characters" will disappear).

Simulation results [40] strongly support the second hypothesis (Fig. 2). Phylogenies were simulated with 16 taxa in which eight randomly selected taxa were made incomplete. Taxa were made incomplete by replacing selected sets of characters with missing data cells ("?"), with different proportions of characters made incomplete in each set of replicates. For analyses with a limited number of characters (i.e., 100), the results show the expected pattern from fossil studies and Huelsenbeck's [27] analyses (which also used only 100 characters). Analyses that include highly incomplete taxa produce trees with low accuracy and resolution (Fig. 2). However, when the overall number of characters is high (i.e., 2000, a typical number for a study based on DNA sequence data), the entire tree can be reconstructed correctly even when half of the taxa have 90% of their data cells lacking data. Clearly, the number and proportion of missing data is not important, only the number of complete characters. This general result holds under a variety of circumstances, including different phylogenetic methods (parsimony, likelihood, and neighbor-joining), numbers of taxa (16 and 64), and different ways of distributing missing data among characters (either the same characters missing in all incomplete taxa, or missing data cells randomly distributed among characters in each taxon). One important exception is when branches are extremely

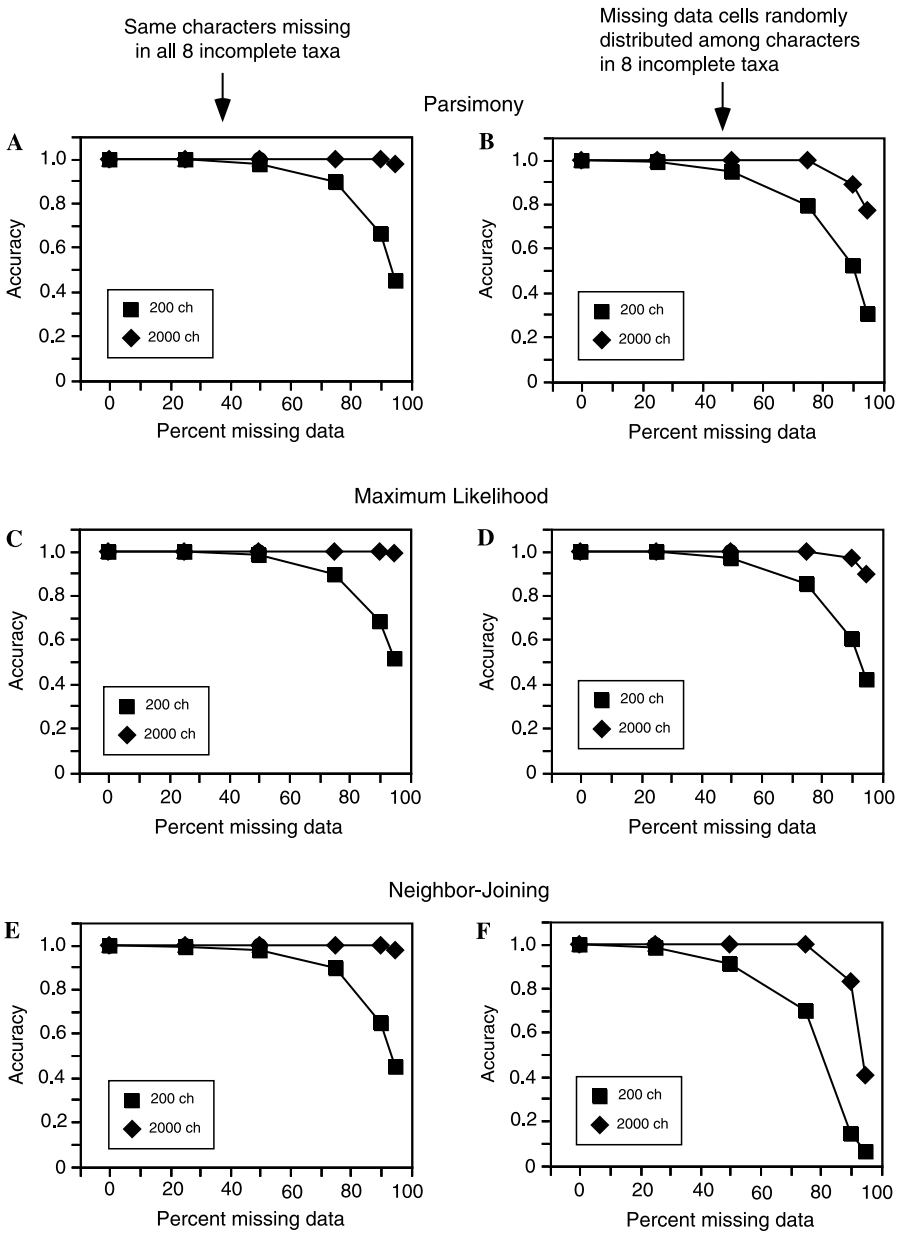

Fig. 2. The phylogenetic relationships of highly incomplete taxa can be accurately resolved if enough characters are included in the analysis. A 16-taxon tree was simulated with characters evolving under the Jukes–Cantor model. In each replicate, eight taxa were randomly chosen to be incomplete. For these taxa, different proportions of their characters were replaced with missing data cells. Missing data cells were either placed into the same set of characters in each incomplete taxon (graphs A, C, and E) or were randomly distributed among characters in the incomplete taxa (B, D, and F). The former treatment (A, C, and E) may be most relevant to empirical studies in which data from different genes are combined. Each data point represents the average of 100 replicates. Figure is modified from [40].

long. Under these conditions, trees with highly incomplete taxa are always difficult to reconstruct, seemingly because the missing data cells exacerbate the problem of long-branch attraction (reduced taxon sampling increases branch lengths among the remaining taxa [17], and if eight taxa are 95% incomplete then the effects may be similar to excluding these taxa entirely).

These results have recently been supported by another set of simulations, this time modeled after a data set of 30,399 characters (129 protein-coding genes) for 36 diverse eukaryote taxa [41]. These authors simulated a combined data set with the same estimated branch lengths, number of characters, and number of taxa as their real data but with some data cells randomly replaced with missing data. When 50% of the data cells were unknown, every branch of the phylogeny was reconstructed correctly in all 100 replicates (using maximum likelihood, the only method they considered). When 90% of the data cells were missing, the correct phylogeny was not always recovered correctly. Nevertheless, the average accuracy

for the 33 internal branches was 89.3%, and nearly half (45%) of the branches were recovered correctly in 95% or more of the replicates.

The general result (that incomplete taxa can be accurately included in phylogenetic analyses) has a number of interesting implications. First, it suggests that the incompleteness of a taxon need not be a constraint on including it in an analysis, as long as the overall number of characters is high. Thus, in an analysis of whole genomes there should be no reason to exclude taxa simply because they are only known from sequences of a single gene. Second, it may explain the conflicting results of empirical studies regarding the effects of adding incomplete taxa. These simulation results suggest that the behavior of incomplete taxa will depend primarily on the amount and quality of evidence that ties them to other taxa on the tree (e.g., the number of complete characters), and not the amount of missing data that they bear. Third, it suggests that there may be little justification for using the controversial supertree approach simply because of the number of missing data cells that are associated with combining data sets that do not overlap fully. Fourth, it suggests that taxon sampling strategies can be devised that do not require sampling every single taxon for every single gene [9]. For example, it may be possible to construct a scaffold of taxa that are scored for many genes and then add many taxa to this scaffold based on taxa scored for one gene [42].

3. Can incomplete taxa be placed with strong statistical confidence?

In some ways, resolving the position of incomplete taxa on a tree is only half the battle. The other half is to have strong statistical support for their placement (e.g., using bootstrapping or Bayesian posterior probabilities). In empirical studies, the true phylogeny is not known, and so other lines of evidence must be used to infer whether or not a taxon has been correctly placed.

My studies of hylid frogs suggest that there is little relationship between the completeness of a taxon and the level of support for its placement on the phylogeny [42]. We reconstructed a "scaffold" of 81 relatively complete taxa based on two mitochondrial genes (12S and ND1), two nuclear genes (POMC and c-myc), and morphology, for a total of 3819 characters. To this scaffold were added 117 additional taxa, most (94 of 117) based on data from the 12S gene alone, but with various taxa having varying levels of completeness (from 6.5 to 100%). We found that each of the incomplete taxa fell into the higher-level clade expected based on previous taxonomy and that the support for the monophyly of these clades remained high (e.g., 100% posterior probabilities). We also quantified the level of support for the placement of each taxon on the tree, based on bootstrap and posterior probabilities. We found no significant relationship between levels of support and levels of completeness (for parsimony, $r^2 = 0.021$, $P = 0.0655$; for Bayesian analysis $r^2 = 0.014$, $P = 0.1385$). However, there was a significant relationship between levels of support for taxa in the combined analysis and their support in the analysis of the 12S data alone (for parsimony, $r^2 = 0.764$, $P < 0.0001$; for Bayesian analysis $r^2 = 0.304$, $P < 0.0001$). This means that the placement of the incomplete taxa was generally determined by the 12S data alone, as well as the level of support for that placement (i.e., when placement of a taxon was well-supported by 12S data alone, it was also well-supported in the combined analysis), regardless of the amount of missing data. This result is fully consistent with those from the simulations [40]. Again, missing data seem to have direct little effect on either the placement of taxa or on levels of support for this placement.

Similarly, the results of Phillippe et al. [41] for a large data set of protein coding genes for eukaryotes also show that incomplete taxa can be placed with high confidence. For example, the four most incomplete taxa in their data set have 56, 60, 61, and 76% missing data, respectively, but the bootstrap values (maximum likelihood) placing them with their respective sister taxa are 100, 92, 98, and 95%. Recent studies based on analyses of large sequence databases for green plants and metazoans have yielded concordant results [9].

4. Can incomplete taxa improve phylogenetic accuracy for complete taxa?

Perhaps the central question in the debate about incomplete taxa is whether or not their inclusion will improve the accuracy for relationships among the more complete taxa. Although there has been considerable debate regarding this issue in the empirical literature (e.g. [25,26,34]) most studies have not directly addressed whether adding incomplete taxa will improve phylogenetic accuracy.

Of course, the issue of adding incomplete taxa brings up a larger question: does adding taxa (whether complete or incomplete) actually improve phylogenetic accuracy? It is generally agreed that there are conditions where adding taxa can improve accuracy. The most obvious case is the classic "Felsenstein Zone" scenario [43,44], in which there are two long terminal branches separated by a short internal branch. In this case, the two long branches accumulate many parallel changes. Parsimony (and other methods) will tend to erroneously group the long branches together based on these parallel changes, a phenomenon called "long-branch attraction." Adding taxa can subdivide these long terminal branches, and greatly increase the chances of estimating the correct tree (e.g. [16,17,45]).

I recently used simulations to explore whether adding incomplete taxa could improve accuracy under these conditions [46]. I simulated a 16-taxon fully asymmetric tree in which 12 of the taxa were removed to create the classic "Felsenstein Zone" scenario among the four remaining taxa. I then added the 12 taxa back into the analysis, but only after replacing parts of their character data with missing data cells. The data were then analyzed using parsimony, neighbor-joining, Bayesian analysis, and maximum likelihood. The resulting trees were then "pruned" to include only the original four taxa, and it was determined whether accuracy was higher or lower with or without the addition of the incomplete taxa. Adding 12 complete taxa can clearly rescue the analysis from long-branch attraction that occurs when the four selected taxa are analyzed alone. The question is whether the 12 incomplete taxa can rescue the analysis as well.

For most methods and conditions, taxa that were only 50% complete were just as beneficial as taxa that were 100% complete (Fig. 3). For parsimony, less complete taxa (5–25% complete) were often unable to rescue an analysis. However, for model-based methods (e.g., Bayesian analysis, likelihood, and neighbor-joining),

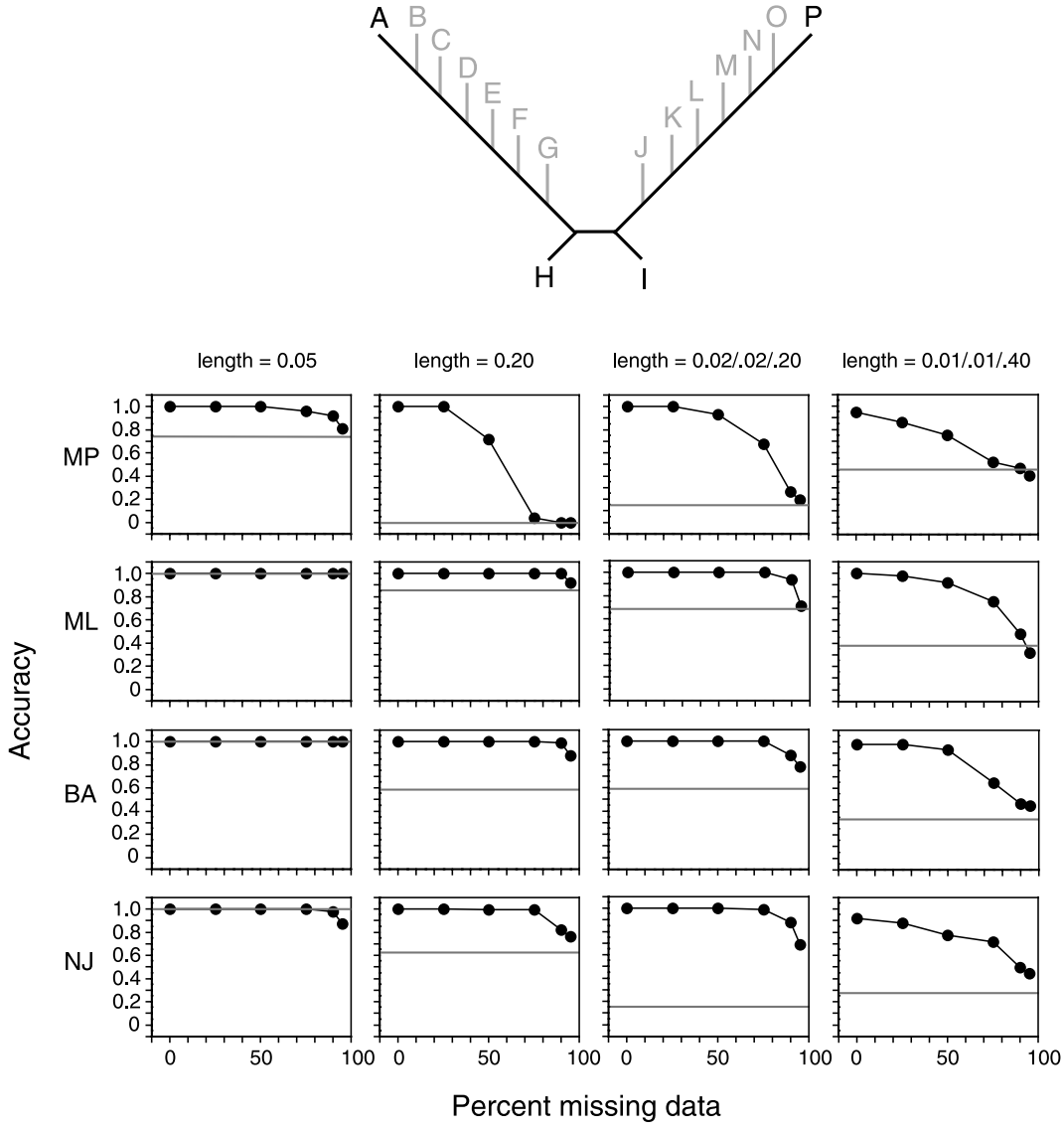

Fig. 3. Incomplete taxa can dramatically improve phylogenetic accuracy by "rescuing" an analysis from long-branch attraction, for parsimony (MP), maximum likelihood (ML), Bayesian analysis (BA), and neighbor-joining (NJ). A 16-taxon tree was simulated with 1000 characters evolving under the Jukes–Cantor model. For the third and fourth columns, the slash after the branch length indicates the rate of change at each simulated codon position. The gray horizontal line represents the proportion of replicates in which the correct phylogenetic relationships among the four complete taxa ((A, H), (I, P)) are reconstructed for a given set of conditions (accuracy), based on analysis of the complete taxa alone. Filled circles represent accuracy for the four complete taxa after including 12 additional taxa of varying levels of completeness. The same characters are missing in all 12 taxa in a given replicate. Each data point represents the average of 200 (MP, NJ), 100 (BA), or 50 (ML) replicated matrices. Model-based methods assumed the Jukes–Cantor model but did not take into account variation in rates of change among sites. Figure is modified from [46].

addition of highly incomplete taxa (10–25% complete) could cause dramatic increases in accuracy. One potential explanation for this pattern is that model-based methods tend to be inherently more robust to the problem of long-branch attraction (e.g. [47,48]). Thus, even the addition of a small amount of data may be enough to "tip the balance" in favor of the correct hypothesis. In contrast, for parsimony, most characters may tend to favor the incorrect hypothesis, making it difficult for a set of characters that support the right tree to overturn the results.

These analyses specifically addressed the "best-case scenario" under which adding taxa could potentially improve accuracy by subdividing long branches. There is considerable debate over how often these conditions might occur, and whether adding characters might be more beneficial instead (e.g. [13–15,19,20,22,23]). Furthermore, there are conditions under which adding taxa and subdividing long branches may not be beneficial and may actually create or exacerbate problems of long-branch attraction (e.g. [18,49]). Clearly, much work remains to be done on the relationships between taxon sampling, missing data, and accuracy. Nevertheless, simulation results thus far suggest that incomplete taxa can potentially improve accuracy for relationships among the complete taxa, at least for those conditions where adding complete taxa improves accuracy. Furthermore, there was little evidence to suggest that adding incomplete taxa would worsen accuracy for the complete taxa, at least for those conditions where adding complete taxa should improve accuracy.

5. Adding characters with missing data

Adding incomplete taxa is not the only way to add missing data cells to a matrix. Given a set of taxa and characters, one could also add a set of characters that are known for only some taxa. For example, given a data set consisting of two genes sequenced for 10 taxa, one could add a third gene known from five of the taxa. Rather than reducing the analysis to five taxa with three genes each (or 10 with two genes each), one could simply retain all 10 taxa and all three genes, but code 5 of the taxa as missing data for the third gene.

There are pros and cons associated with adding incomplete characters. On the positive side, there is abundant evidence that increasing the number of character generally increases phylogenetic accuracy [1,12,44,47]. The major exception is in cases of long-branch attraction [16,43,44,47]. On the negative side, missing data cells may actually create long-branch attraction in sets of incomplete characters [50]. This creates a potential situation in which adding a large number of highly incomplete characters may reduce phylogenetic accuracy.

To try and address this question, I [50] tested the effects of adding 50 incomplete characters to a set of 50 characters in the 16-taxon case using parsimony. I found that under these circumstances, adding the set of incomplete characters generally increased phylogenetic accuracy relative to excluding them (Fig. 4). This increase was sometimes statistically significant. However, the results also showed that as the proportion of missing data in the added characters increased, the ability of the added characters to increase accuracy quickly decreased. When adding a set of characters that consisted of 75% missing data, the change in accuracy was typically slight. Interestingly, the addition of these highly incomplete sets of characters never had a significantly negative impact on the results, but sometimes had a significantly positive impact. Overall, the results suggested that adding characters with missing data was (on average) either neutral or good. The results also suggested that it is more beneficial to add a smaller number of characters with data for more taxa (and less missing data) than a

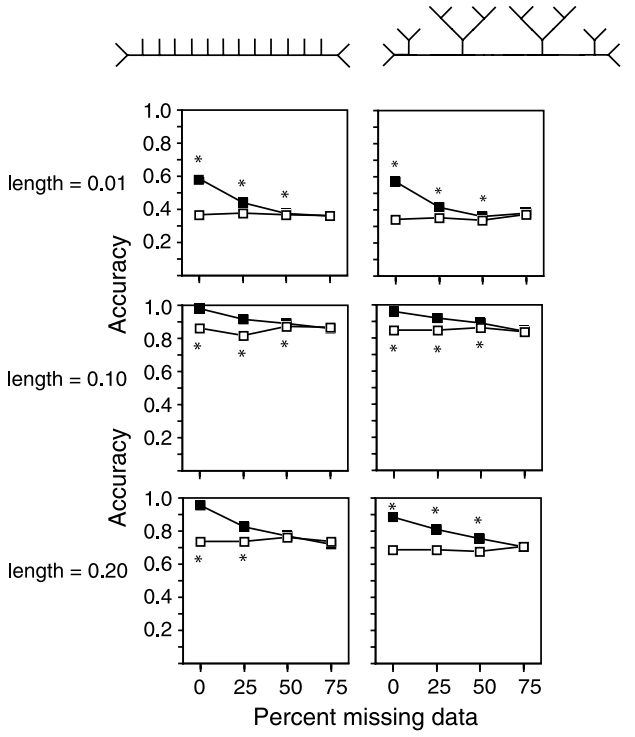

Fig. 4. Adding characters with missing data can increase phylogenetic accuracy, but increasing incompleteness robs characters of their benefits. A 16-taxon tree was simulated with two data sets (50 characters each) of DNA sequence data evolving under the Jukes–Cantor model. The open squares represent accuracy based on parsimony analysis of the first data set alone (50 complete characters). The closed square represents accuracy based on the combination of the first data set and a second data set, in which various taxa are randomly selected to have all 50 of their characters replaced with missing data cells (12 taxa for 75% missing data, eight taxa for 50%, four taxa for 25%, and zero taxa for 0%). Each square represents the average accuracy from 100 replicated data matrices. The figure is modified from Fig. 2 of [50].

larger number of characters with for fewer taxa (and more missing data).

This analysis was certainly not exhaustive. For example, only parsimony was tested and other phylogenetic methods were not. The number of characters in each data set was generally equal and relatively small (50 per data set). An important question is what happens when the number of incomplete characters is much larger than the number of complete characters (in fact, the scenario may be similar to that of adding highly incomplete taxa to break up long branches). Furthermore, the rate of evolution of each set of characters was equal in these simulations, and differences in rates may greatly change the costs and benefits of adding incomplete characters. High rates of change increase the potential for long branch attraction, so adding a set of incomplete characters that are evolving quickly may be quite dangerous, whereas adding a set of incomplete but slowly evolving characters should be beneficial or harmless. These and other questions are clearly in need of further study.

6. Conclusions, implications, and future research

Perceptions about how missing data cells will affect phylogenetic analyses may strongly influence the design of empirical phylogenetic studies (i.e., which taxa and characters are included versus excluded). In this paper, I have tried to review current knowledge of the possible effects of missing data. Recent simulations show that there is little evidence to support excluding taxa based simply on the amount or proportion of missing data that they bear. The placement of highly incomplete taxa in a phylogeny can be resolved with perfect accuracy (based on simulations) and with strong support statistical support (based on empirical analyses). The critical factor determining their placement is seemingly the characters which are present in these taxa (i.e., their number and quality) not the ones that are absent.

Considering these results, it should be possible to design phylogenetic analyses which will resolve higher-level phylogeny with large numbers of slow-evolving characters and then place numerous taxa on this "scaffold" using a smaller number of more rapidly evolving characters [42]. This design should allow optimal resolution of both higher and lower level phylogenies, without the time and expense required to obtain data for every character in every taxon (although this would obviously be preferable given adequate resources).

Despite extensive discussion about the pros and cons of sampling taxa versus characters, recent authors have not considered whether adding incomplete taxa offer the same benefits as complete taxa. In many simulations [46], taxa that are only 50% complete may be able to subdivide long branches and improve accuracy as well as those that are 100% complete. This may have important implications for the "economics" of sampling design (increasing taxon sampling may be much "cheaper" if only half the data are necessary).

Simulations also suggest that characters that contain missing data may still be able to improve phylogenetic accuracy. Nevertheless, abundant missing data may rob these characters of their potential benefits, and the combination of high rates of change and incompleteness may lead to long-branch attraction in some cases.

The effects of missing data on phylogenetic analysis are clearly in need of additional study. For example, no simulation studies have yet addressed how widely used Bayesian methods perform with abundant missing data. Although preliminary simulation analyses (Wiens, unpubl.) suggest that incomplete taxa in Bayesian analyses perform much as they do in parsimony and maximum likelihood, this needs to be studied more extensively. It is hoped that phylogenetics will consider how missing data does or does not affect phylogenetic analyses, rather than letting unstated assumptions determine the exclusion of data or the design of phylogenetic studies.

Acknowledgments

I thank Rob DeSalle and Neil Sarkar for kindly inviting me to contribute this review. Two anonymous reviewers provided helpful comments that improved the paper. My recent work on the effects of missing data has been supported by US National Science Foundation Grant DEB 0334923 to J.J.W.

References

[1] Hillis DM, Huelsenbeck JP, Cunningham CW. Application and accuracy of molecular phylogenies. Science 1994;264:671–7.
[2] Bush RM, Bender CA, Subbaro K, Cox NJ, Fitch WM. Predicting the evolution of influenza A. Science 1999;286:1921–5.
[3] Hillis DM. Origins of HIV. Science 2000;288:1757–9.
[4] Achtman M and 16 other authors. Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci USA 2004;101:17837–42.
[5] Rambaut A, Posada D, Crandall KA. The causes and consequences of HIV evolution. Nat Rev Genet 2004;5:52–61.
[6] Posada D, Crandall KA. Simple (wrong) models for complex trees: a case from the Retroviridae. Mol Biol Evol 2001;18:271–5.
[7] Posada D, Crandall KA. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 2001;18:897–906.
[8] Sanderson MJ, Driskell AC. The challenge of constructing large phylogenetic trees. Trends Plant Sci 2003;8:374–9.
[9] Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ. Prospects for building the Tree of Life from large sequence databases. Science 2004;306:1172–4.
[10] Sanderson MJ, Purvis A, Henze C. Phylogenetic supertrees: assembling the tree of life. Trends Ecol Evol 1998;13:105–9.

[11] Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol Biol Evol 2003;20:1036–42.
[12] Hillis DM. Inferring complex phylogenies. Nature 1996;383:130–1.
[13] Hillis DM. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 1998;47:3–8.
[14] Kim J. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst Biol 1996;45:363–74.
[15] Kim J. Large-scale phylogenies and measuring the effects of phylogenetic estimators. Syst Biol 1998;47:43–60.
[16] Graybeal A. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 1998;47:9–17.
[17] Rannala B, Huelsenbeck JP, Yang Z, Nielsen R. Taxon sampling and the accuracy of large phylogenies. Syst Biol 1998;47:702–10.
[18] Poe S, Swofford DL. Taxon sampling revisited. Nature 1999;398:299–300.
[19] Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci USA 2001;98:10751–6.
[20] Rosenberg MS, Kumar S. Taxon sampling, bioinformatics, and phylogenomics. Syst Biol 2003;52:119–24.
[21] Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 2002;51:664–71.
[22] Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 2002;51:588–98.
[23] Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol 2003;52:124–6.
[24] Gauthier J. Saurischian monophyly and the origin of birds. Mem Calif Acad Sci 1986:1–47.
[25] Gauthier J, Kluge AG, Rowe T. Amniote phylogeny and the importance of fossils. Cladistics 1988;4:105–209.
[26] Donoghue MJ, Doyle JA, Gauthier J, Kluge AG, Rowe T. The importance of fossils in phylogeny reconstruction. Annu Rev Ecol Syst 1989;20:431–60.
[27] Huelsenbeck JP. When are fossils better than extant taxa in phylogenetic analysis? Syst Zool 1991;40:458–69.
[28] Novacek MJ. Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. Syst Biol 1992;41:58–73.
[29] Wilkinson M. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst Biol 1995;44:501–14.
[30] Kearney M. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst Biol 2002;51:369–81.
[31] Wiens JJ, Reeder TW. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst Biol 1995;44:548–58.
[32] Bininda-Emonds O, Sanderson MJ. An assessment of the accuracy of MRP supertree construction. Syst Biol 2001;50:565–79.
[33] Hillis DM. Approaches for assessing phylogenetic accuracy. Syst Biol 1995;44:3–16.
[34] Patterson C. Significance of fossils in determining evolutionary relationships. Annu Rev Ecol Syst 1981;12:195–223.
[35] Wilkinson M, Benton MJ. Missing data and rhynchosaur phylogeny. Hist Biol 1995;10:137–50.
[36] Gao K, Norell MA. Taxonomic revision of *Carusia* (Reptilia: Squamata) from the Late Cretaceous of the Gobi Desert and phylogenetic relationships of anguimorphan lizards. Am Mus Nov 1998;3230:1–51.
[37] Rowe T. Definition, diagnosis, and origin of mammalia. J Vertebr Paleontol 1988;8:241–64.
[38] Grande L, Bemis WE. 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 1998;4:1–690.
[39] Ebach MC, Ahyong ST. Phylogeny of the trilobite subgenus *Acanthopyge* (*Lobopyge*). Cladistics 2001;17:1–10.
[40] Wiens JJ. Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 2003;52:528–38.
[41] Phillipe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 2004;21:1740–52.
[42] Wiens JJ, Reeder TW, Fetzner JW, Parkinson CL, Duellman WE. Hylid frog phylogeny and sampling strategies for species clades. Syst Biol (in press).
[43] Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 1978;27:401–10.
[44] Huelsenbeck JP, Hillis DM. Success of phylogenetic methods in the four-taxon case. Syst Biol 1993;42:247–64.
[45] Hendy MD, Penny D. A framework for the quantitative study of evolutionary trees. Syst Zool 1989;38:297–309.
[46] Wiens, JJ. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol (in press).
[47] Huelsenbeck JP. The performance of phylogenetic methods in simulation. Syst Biol 1995;44:17–48.
[48] Alfaro ME, Zoller S, Lutzoni F. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol Biol Evol 2003;20:255–66.
[49] Poe S. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst Biol 2003;52:423–8.
[50] Wiens JJ. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst Biol 1998;47:625–40.