Cladistics 26 (2010) 326–339

Cladistics

10.1111/j.1096-0031.2009.00289.x

The impact of missing data on real morphological phylogenies: influence of the number and distribution of missing entries

Francisco J. Prevosti[a,]* and María A. Chemisquy[b]

[a]*División Mastozoología, Museo Argentino de Ciencias Naturales "Bernardino Rivadavia" – CONICET, Av. Angel Gallardo 470, C1405DJR Buenos Aires, Argentina;* [b]*Instituto de Botánica Darwinion (CONICET, ANCEFN), Labardén 200, Casilla de Correo 22, B1642HYD San Isidro, Buenos Aires, Argentina*

Accepted 22 August 2009

Abstract

Here we explore the effect of missing data in phylogenetic analyses using a large number of real morphological matrices. Different percentages and patterns of missing entries were added to each matrix, and their influence was evaluated by comparing the accuracy and error of most parsimonious trees. The relationships between accuracy and error and different parameters (e.g. the number of taxa and characters, homoplasy, support) were also evaluated. Our findings, based on real matrices, agree with the simulation studies, i.e. the negative effect increases with the percentage of missing entries, and decreases with the addition of more characters. This indicates that the main problem is the lack of information, not just the presence of missing data *per se*. Accuracy varies with different distribution patterns of missing entries; the worst case is when missing data are concentrated in a few taxa, while the best is when the missing entries are restricted to just a few characters. The results expand our knowledge of the missing data problem, corroborate many of the findings previously published using simulations, and could be useful for empirical or theoretical studies.

The negative effect of missing data cells (usually represented by a "?") in phylogenetic analyses was first noticed more than 20 years ago (e.g. Gauthier et al., 1988). Since then, several authors have discussed this issue, covering how and why the missing data entries affect the phylogenetic reconstructions under different conditions (Nixon and Davis, 1991; Platnick et al., 1991; Maddison, 1993). Some authors have empirically analysed the effect of the missing data in the context of phylogenetic studies (e.g. Gauthier et al., 1988; Novacek, 1992; Wiens and Reeder, 1995; Wilkinson and Benton, 1996; Flynn et al., 2005; Fulton and Strobeck, 2006), whereas others have proposed methodologies to avoid the problem (Wilkinson, 1995, 2003; Anderson, 2001; Kearney, 2002; Kearney and Clark, 2003), or evaluated whether the inclusion of incomplete taxa could prevent, or at least diminish, the long-branch attraction effect (Wiens, 2005). Cobbett et al. (2007), by contrast, explored whether fossil taxa (usually with many missing cells) were more unstable than recent taxa in phylogenies. Recent studies have evaluated the effect of missing data in phylogenomic or large sequence database projects, and different phylogenetic inference methods (e.g. Driskell et al., 2004; Philippe et al., 2004; Wiens, 2006; Hartmann and Vision, 2008; Wiens and Moen, 2008; Wolsan and Sato, 2009). Other contributions have evaluated the potential change of the results obtained (nodes and branch support) when filling the missing cells (Norell and Wheeler, 2003), studied strategies of taxa and character sampling (Wiens, 2003b; Wiens et al., 2005), or performed simulation studies to analyse the missing data effect under a range of different parameters (e.g. percentage and pattern of distribution of missing data, number of taxa and characters, branch lengths; see Huelsenbeck, 1991; Wheeler, 1992; Wiens, 1998, 2003a,b, 2006; Wiens and Moen, 2008; and references therein cited).

Most of these contributions have shown that a considerable number of missing data in a particular taxon may cause it to change its position (i.e. to "float") without

*Corresponding author:
E-mail address: protocyon@hotmail.com

changes in tree length, thus generating a great number of most parsimonious trees and a poorly resolved strict consensus (the "wildcard" taxa effect; see Nixon and Wheeler, 1992). However, the authors generally agree that the level of incompleteness is not a good predictor of the negative effect of a taxon/character inclusion. In addition, they have shown that the inclusion of taxa or characters with several missing entries usually improves the results (e.g. Novacek, 1992; Wiens, 1998, 2003a,b, 2005, 2006; Wiens and Moen, 2008).

More missing data implies more problems? Simulation-based studies have shown that as the percentage of missing data increases, the accuracy of the phylogenetic methods decreases (Wheeler, 1992; Wiens and Reeder, 1995; Wiens, 2006). These studies have shown results that are invariant to changes in the distribution pattern of missing entries, or the use of binary versus multistate characters.

Does having more taxa or more characters counteract the negative effect of the missing data better? According to previous studies based on simulations, an increase in the number of characters improves accuracy, while long branches decrease the accuracy of phylogenetic methods (Wheeler, 1992; Wiens, 2003a,b, 2006; Wiens and Moen, 2008).

What patterns of distribution of missing entries are expected to be more problematic? Are the most destructive patterns common or rare? A non-random distribution of missing entries was suggested as the most common pattern (Wheeler, 1992; Wiens, 2003a), although these studies explored only a small number of real matrices. Previous studies based on simulations have explored only the effect of two patterns: missing entries distributed in several taxa and in blocks of taxa and characters (e.g. Wiens, 1998, 2003a,b). Consequently, a broader sampling of patterns is needed in order to determine which is the most detrimental.

Do simulations represent real data sets? As stated by several authors (e.g. Wiens, 2005; Wiens and Moen, 2008), the results obtained from simulations come from simple scenarios, with limited kinds of characters and taxon numbers, and only a few patterns of missing data. Thus, extrapolation of these results to empirical and more complex data is not as straightforward, and must be contrasted with more realistic scenarios. The evaluation of published matrices could allow us to explore the effect of missing data in phylogenetic analyses more deeply.

Here we explore the effect of missing data in phylogenetic analyses by using a large number of published morphological matrices, with a wide range of numbers of characters and taxa. We compared the relationship between the effects of the missing data and several parameters: percentage and pattern of distribution of the missing data, type of characters, levels of homoplasy, and phylogenetic signal. Finally, we contrasted the published results, mainly based on simulation analyses, with the exploration of a large sample of real matrices that span a wider range of parameter values.

Our main conclusions are: (i) the missing entries have a negative effect on the cladistic performance of the matrices, and this effect is more pronounced in real matrices, as compared with simulations, due to the higher levels of homoplasy present in the real data sets; (ii) the inclusion of more characters could make the matrices more robust to this problem, meaning that the bias is mainly caused by the lack of information, not just the presence of missing data; and (iii) the most prejudicial distribution of missing data is when missing entries are accumulated in several taxa (i.e. taxon bias), an uncommon pattern in real matrices.

Methods

Here we deal only with morphological matrices and the maximum parsimony method of phylogenetic inference. We restricted our analysis to morphological data for practical reasons, i.e. in order to put a workable limit to the study, and also because we work mainly with morphology and fossil taxa. In addition, it is known that the analysis of real molecular data implies the additional problem of alignment, which can affect the resulting trees (e.g. Martin et al., 2007; Ogden and Rosenberg, 2007a). Moreover, as the treatment of gaps in molecular analyses is not a trivial issue (for further discussion of this issue see Ogden and Rosenberg, 2007b) and the treatment of alignment and gap issues is beyond the scope of this paper, we did not consider molecular data in the analyses.

Data sets

A sample of 354 matrices used in real phylogenetic analyses was compiled. The matrices were obtained from the online databases Cladestore (palaeo.gly.bris.ac.uk/cladestore/default.html) and Treebase (http://www.treebase.org), and both published (e.g. Goloboff et al., 2008a) and unpublished analyses (see supplementary Appendices S1 and S2).

The 354 matrices were classified as palaeontological (having at least one extinct taxon) and non-palaeontological. The percentage of missing data (hereafter referred to as %MISS) of each matrix was calculated. We considered the inapplicable code "–" as missing data because, beyond the theoretical differences, the available algorithms evaluate it in the same way as missing data. The distribution of missing cells in the original matrices was determined by comparing the original distribution with a random distribution in characters and taxa separately using the χ^2 parameter. This was made by contrasting the χ^2 parameter of the distribution of the original missing entries in characters and taxa versus the χ^2 of a random distribution of missing data in the matrices. The χ^2 value of the random distribution was calculated and compared 100 times, by replacing the original missing data with the same number of missing entries randomly distributed. It is important to point out that we use the terms missing data, missing cells or missing entries interchangeably throughout the text.

For the "Missing data analyses" (see below), all those matrices with more than 40 taxa and 100 characters were included, whereas smaller matrices were randomly selected, resulting in 168 matrices for further analysis. We restricted the analysis to this subset of matrices mainly for three reasons: (i) to keep an even sample along the range of characters and taxa number (the complete sample is very skewed to small matrices, see Prevosti and Chemisquy, 2007); (ii) a previous preliminary study of the complete sample gave results similar to those obtained in this paper with a subset of matrices (Prevosti and Chemisquy, 2007); and (iii) to process all this information in reasonable times.

Missing data analyses

For all analyses, the original matrices were modified by replacing the original missing entries with states randomly selected among the possible ones for each character (see Norell and Wheeler, 2003), with an equal probability for each state (i.e. 0.5 for a binary character, 0.33 for a three-state character, and 0.25 for a four-state character), using a specially designed TNT script (Fig. 1b; Appendix S3). As the missing entries present in each original matrix had a particular distribution, it was necessary to replace them to explore different distribution patterns and different percentages of missing cells. Two methods were used to distribute the missing data cells. First, the missing data were introduced randomly among all characters and taxa (this distribution will be referred to as "random"; Fig. 1c). Secondly, the missing data were concentrated in a few characters ("character bias"), taxa ("taxa bias"), or blocks of characters and taxa ("block bias"). These last three analyses are referred to as "bias" analyses. In "character bias", 15 and 50% of the characters were randomly selected to have missing entries (Fig. 1e). The same proportions of taxa were chosen to be incomplete in the "taxa bias" analyses (Fig. 1d). Finally, blocks of cells representing 5 and 25% of the matrix were replaced by missing data ("block bias"; Fig. 1f), corresponding to 25 and 50% of characters and taxa, respectively. In the "block bias" analyses, taxa and characters were previously arranged in a random way to avoid any effect of the original order. In all cases, except for the "block bias" analyses, the probability of each cell being replaced with a missing data entry ("?") varied between 0.15 and 0.9, with intervals of 0.15. Note that when we say that a matrix possesses 50% of missing data, this means that half of its cells were scored with "?" (for example, a matrix of 10 taxa and 10 characters, with a total of 100 entries, has 50 cells with "?"). This mode of assigning missing entries is different from that implemented by other authors (e.g. Wiens, 2006; Wiens and Moen, 2008) who only introduced missing entries in a subset of taxa.

As the replacement of missing cells with random character states can introduce "noise" in the phylogenetic signal (see Wenzel and Siddall, 1999), its effect was tested by comparing the results of the "random" and "block bias" analyses (Fig. 1c and f, respectively) with the results obtained by adding missing data in a "random" and "block bias" distribution, but also retaining the original missing entries (i.e. without replacing them with a random character state; Fig. 1g,h).

All the analyses were performed with TNT (Goloboff et al., 2008b). Searches were carried out with all characters equally weighted, and under implied weighting (Goloboff, 1993) with two different values of the concavity constant *k*: 15 and 100. Heuristic searches were performed using 1000 random addition sequences (RAS) and tree bisection-reconnection (TBR). For matrices with more than 70 taxa, searches were carried out by building 20 RAS trees, swapping each one with TBR, sectorial search (with the default parameters), and 20 iterations of tree-drifting (Goloboff, 1999). The strict consensus was calculated in all the cases.

The strict consensus obtained with each matrix with the missing cells replaced by random character states (Fig. 1b) was considered the "true" phylogeny, and the strict consensus obtained with the matrix with the new missing cells was considered the estimated phylogeny. In analyses where the original missing entries were retained, the strict consensus obtained from the original matrix was considered the "true" phylogeny. In simulation studies (e.g. Wiens, 1998, 2003a,b, 2006; Wiens and Moen, 2008) the effect of the added missing data was measured by comparing the accuracy along different numbers of missing cells and distribution patterns. As the authors of those studies had the true trees (in fact, they built the matrix based on the trees) they consequently could measure how many true nodes were recovered with different amounts of added missing data (and different distributions). This is not possible with real matrices, where the true tree is not known. In this context we can use the trees (the strict consensus in this case) obtained from the matrices without missing cells as the "true" or target tree. This strategy was used in other contexts by using different kinds of jackknife analyses, with the aim to compare search strategies of parameters values (see Goloboff, 1997; Ramírez, 2003; Goloboff et al., 2008a). In an analogous view, we interpret that with the addition of missing cells, the information

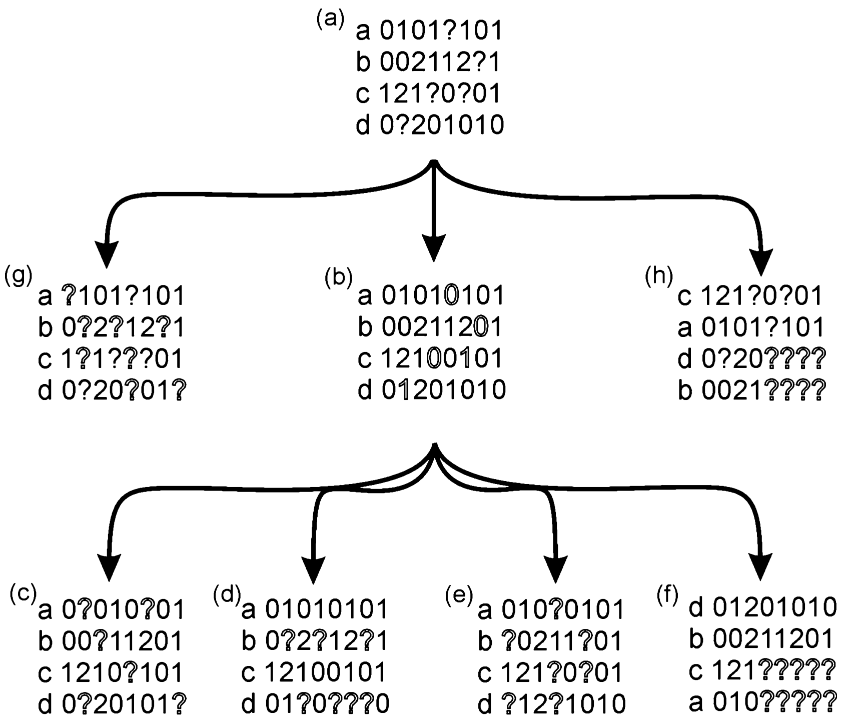

Fig. 1. Procedure used in the missing data analyses. (a) Unmodified matrices with missing entries (?). (b) Replacement of the original missing cells with a random state. (c–f) Introduction of new missing cells in different distribution patterns: (c) "random"; (d) "taxa bias"; (e) "characters bias"; (f) "block bias". (g) "Random" but maintaining the original missing entries. (h) "Block bias" but maintaining the original missing entries.

content present in the matrices is reduced and the unfeasibility of recovering original nodes, or new nodes (i.e. not present in the original tree), is an artefact generated by the missing entries. The use of the reduced strict consensus instead of the strict consensus (see Wilkinson, 2003) could help to separate the effect of the "wild card" taxa. However, we considered these taxa as part of the bias imposed by the missing data, and as our objective was to explore the general effect of the missing entries we preferred to use the strict consensus.

For each analysis, accuracy was measured with the Consensus Fork Index: the number of clades shared by the "true" tree and the consensus tree obtained from the modified data sets (i.e. matrix with added missing cells; "new" tree) divided by the total number of possible clades (CFI; Colless, 1980). The error was measured as the number of nodes of the "new" tree not present in the "true" tree, divided by the total number of nodes of the "new" tree (PE; Ramírez, 2003). Other measurements (e.g. SPR distance, quartet distance; Goloboff, 2007; Hartmann and Vision, 2008) could be used to explore the effect of the missing data (M. Willis, pers. comm.), but we considered that a measure based on recovered "true" or spurious nodes was more explicit. In addition, as these kinds of measures have been used in other studies, they are preferred for comparability.

Comparisons were made contrasting accuracy and error with the number of taxa, the number of informative characters, the percentage of binary characters, the tree imbalance index and the Consistency Index (CI; an estimation of the homoplasy of the matrix). The imbalance index was calculated as the number of not balanced dichotomic nodes (i.e. that have more descents in a daughter branch than in the other) divided by the number of resolved nodes in the tree (excluding polytomies and the root node). In the case of completely dichotomic trees, the number of not balanced nodes is divided by the number of resolved nodes (excluding the root node) minus 1, because in a completely pectinated tree one node (the most terminal) is always balanced; this does not necessarily occur in trees with polytomies. Group support was evaluated by using symmetric resampling with average group frequencies and frequency differences (GC; Goloboff et al., 2003, 2008a), obtained from the original matrices or the modified ones before the introduction of new missing cells.

Correlations between accuracy values and the error rate with the other computed variables were evaluated with Spearman's rank correlation coefficient; this correlation was calculated for each weighting scheme, and each rank of missing cells.

For assessing the effect of the percentage of missing data, the significance of differences between the accuracy values and the error rate of consecutive groups (i.e. successive percentages of missing data with the same distribution) were tested with the bootstrap T test with 5000 repetitions (Manly, 1997), as the data did not fulfil the assumptions of normal distribution and homoscedasticity. The same test was used to compare the effect of different distributions of missing data, and in other comparisons, such as original matrices versus matrices with the missing entries replaced. In the analyses with non-randomly distributed missing entries (i.e. ''bias'' analyses), the percentage of missing entries in the matrix is not necessarily that expected by the probability assigned for the cells. In order to make the comparisons easier, matrices were ordered by the observed percentage of missing data, and divided in ranks comparable with the expected number of missing cells (i.e. those obtained in ''random'' analyses).

Results

Characteristics of the matrices sampled

The complete sample contained 354 matrices, with 5–207 taxa (median: 27) and 4–427 characters (median: 58), but most of the matrices were small, with fewer than 40 taxa or 100 characters.

The proportion of missing data ranged between 0 and 54.31% with most matrices having between 0 and 15% MISS (median: 10.07%). Only 2% of the matrices had a random distribution of missing cells; in 76% of the data sets the missing cells were not randomly distributed, but concentrated on specific characters and/or taxa, whereas in 10% they were randomly distributed only across the characters and the remaining 12% had the missing cells randomly located only in the taxa.

The ''palaeontological'' matrices had a higher percentage of missing cells than ''neontological'' matrices (median: 14.6%, maximum: 54.31%; versus 6.75 and 45.87%, respectively). This difference was statistically significant (*t*: 17.1733 $P < 0.0002$), but the range of percentages of missing entries widely overlapped between these groups of matrices (Appendix S1).

Different proportions of missing data

The general trend observed in all the analyses is that an increase in the percentage of MISS clearly decreases the accuracy and increases the error. At 15% MISS the median accuracy values were between 0.28 and 0.50 and the median error rate was between 0.43 and 0.53; these values reached 0 and 1, respectively, at 90% MISS (Figs 2 and 3a; Table 1). The differences in accuracy and error between successive percentages of MISS were highly significant based on the bootstrap T test (with *P* usually below 0.0002), with the exception of ''character bias 15%'' and ''taxa bias 15%'' (9 versus 15% MISS) and ''character bias 50%'' (30 versus 45% MISS; Appendix S4). The variation in accuracy and error rate was large and ranged almost between 0 and 1, but decreased at high levels of MISS (75–90%; Fig. 2; Table 1). The searches under implied weights showed higher median accuracy values than under equal weights.

Relationship between cladistic performance (accuracy and error rates) and other parameters

In most analyses the number of characters showed a positive moderate to low significant correlation with accuracy values, and a negative low significant correlation with error rate, but some analyses like ''taxa bias 15%'' and ''character bias 15%'' were not significantly correlated with error rate (Fig. 3b; Table 2; Appendix S5).

The number of taxa showed an inverse correlation with the number of characters, with low negative coefficients with accuracy values and low positive coefficients with error rates (Fig. 3c; Table, 2; Appendix S5).

Node support values were positively correlated with accuracy values and negatively correlated with error rates, showing the highest absolute correlation values (Fig. 3d; Table 2; Appendix S5). Homoplasy, expressed as the CI index, presented a pattern of correlation similar but inverse to that observed between support and error rates and accuracy values (Fig. 3e; Table 2; Appendix S5).

The percentage of binary characters in the ''random'' and ''bias'' analyses was sometimes positively correlated with accuracy values ($r \approx 0.18$) and negatively correlated with error rates ($r \approx -0.16$), but in most cases was not significant at $P < 0.01$ (Appendix S5).

The imbalance index showed low or very low correlation coefficients, negative with accuracy values and positive with error rates, and only in a few cases was this significant at $P < 0.05$ (Appendix S5).

In summary, more characters led to higher accuracy values and lower error rates, while the inverse tendency was observed with matrices with more taxa. By contrast, matrices with well-supported nodes showed higher accuracy values, while matrices with elevated levels of homoplasy were correlated with higher error rates. Neither the imbalance index nor the percentage of binary characters showed clear patterns of correlation with accuracy or error rates.

Different distribution patterns of missing data

With 15% MISS, ''random'', ''characters bias 50%'', and ''taxa bias 50%'' showed similar median accuracy

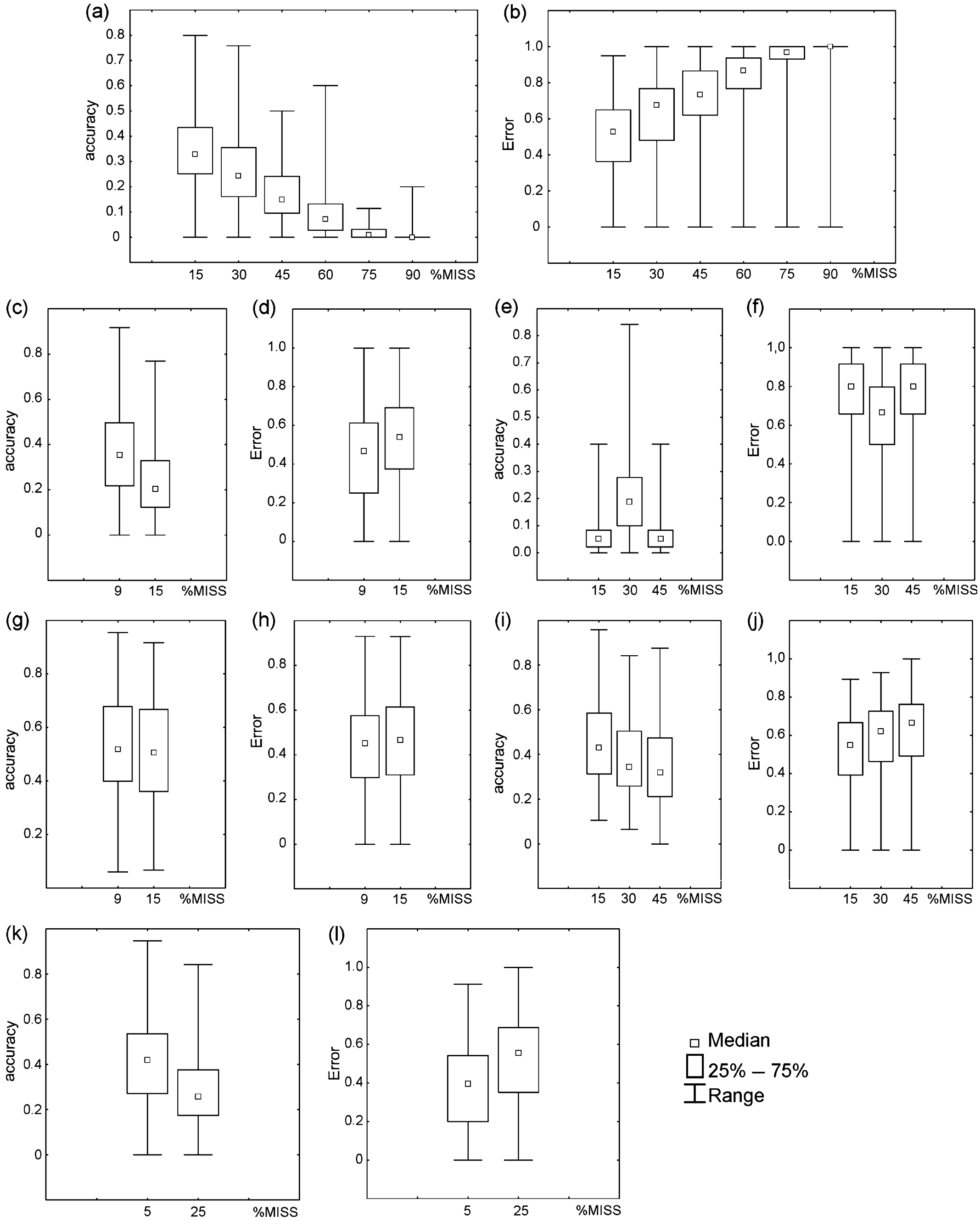

Fig. 2. Accuracy and error with different percentages of missing data (%MISS) and distribution patterns. (a,b) "Random" analysis; (c,d) "taxa bias 15%" analysis; (e,f) "taxa bias 50%" analyses; (g,h) "characters bias 15%" analysis; (i,j) "characters bias 50%" analysis; (k,l) "block bias" analysis. These graphs are based on equal weights searches, but analyses using implied weighting give similar results.

values, higher than for "taxa bias 15%", but lower than for "character bias 15%" (Fig. 4a). These differences were in most cases significant under the bootstrap T test ($P < 0.005$; Appendix S4).

With 30% MISS, "taxa bias 50%" had the lowest median accuracy; "character bias 50%" the highest, and "random" intermediate values, but the only significant differences ($P < 0.0007$) were between "random" and "character bias 50%", and between "taxa bias" and "character bias" analyses (Fig. 4b; Appendix S4). These differences were greater with 45% MISS and highly significant ($P < 0.0002$) between all the pairs (Fig. 4c).

When comparing the "bias" analyses, the general pattern was that "taxa bias 15%" had the lowest

Table 1
Statistics of accuracy (CFI) and error (PE) values. %MISS, percentage of missing cells; K15, K100, implied weighting with K15 and K100, respectively. Originals, analysis keeping the original missing cells

	Equal Weights			K15			K100		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Random									
CFI 15%MISS	0.352	0.327	0.170	0.480	0.475	0.181	0.445	0.435	0.182
PE 15%MISS	0.495	0.529	0.216	0.487	0.509	0.196	0.523	0.553	0.199
CFI 30%MISS	0.266	0.244	0.151	0.357	0.355	0.162	0.332	0.318	0.158
PE 30%MISS	0.609	0.674	0.228	0.604	0.623	0.189	0.631	0.656	0.187
CFI 45%MISS	0.171	0.148	0.109	0.239	0.227	0.133	0.226	0.202	0.137
PE 45%MISS	0.705	0.736	0.218	0.718	0.737	0.167	0.734	0.764	0.164
CFI 60%MISS	0.090	0.072	0.083	0.123	0.105	0.089	0.122	0.109	0.086
PE 60%MISS	0.823	0.870	0.183	0.833	0.862	0.137	0.832	0.855	0.133
CFI 75%MISS	0.020	0.011	0.026	0.026	0.019	0.032	0.027	0.020	0.031
PE 75%MISS	0.911	0.968	0.187	0.945	0.969	0.087	0.941	0.968	0.104
CFI 90%MISS	0.002	0	0.015	0.003	0	0.016	0.003	0	0.016
PE 90%MISS	0.983	1	0.102	0.988	1	0.039	0.982	1	0.064
Character bias 15%									
CFI 9%MISS	0.418	0.405	0.183	0.535	0.526	0.210	0.538	0.517	0.190
PE 9%MISS	0.415	0.444	0.230	0.428	0.430	0.227	0.434	0.453	0.201
CFI 15%MISS	0.395	0.040	0.186	0.491	0.483	0.211	0.508	0.506	0.194
PE 15%MISS	0.454	0	0.222	0.470	0.485	0.232	0.461	0.468	0.206
Character bias 50%									
CFI 15%MISS	0.352	0.333	0.160	0.478	0.463	0.186	0.448	0.432	0.177
PE 15%MISS	0.481	0.521	0.218	0.487	0.500	0.201	0.521	0.550	0.195
CFI 30%MISS	0.297	0.275	0.154	0.392	0.376	0.177	0.379	0.345	0.167
PE 30%MISS	0.562	0.600	0.222	0.569	0.586	0.200	0.583	0.621	0.194
CFI 45%MISS	0.273	0.250	0.159	0.361	0.337	0.182	0.347	0.318	0.180
PE 45%MISS	0.568	0.609	0.223	0.603	0.636	0.201	0.619	0.666	0.201
Taxa Bias 15%									
CFI 9%MISS	0.361	0.346	0.179	0.467	0.461	0.192	0.467	0.461	0.192
PE 9%MISS	0.434	0.470	0.244	0.463	0.479	0.213	0.463	0.479	0.213
CFI 15%MISS	0.234	0.203	0.144	0.333	0.299	0.192	0.312	0.285	0.176
PE 15%MISS	0.514	0.538	0.23	0.496	0.500	0.222	0.532	0.553	0.211
Taxa bias 50%									
CFI 15%MISS	0.352	0.333	0.160	0.478	0.463	0.186	0.448	0.432	0.177
PE 15%MISS	0.481	0.521	0.218	0.487	0.500	0.201	0.521	0.550	0.195
CFI 30%MISS	0.297	0.275	0.154	0.392	0.376	0.177	0.379	0.345	0.167
PE 30%MISS	0.562	0.600	0.222	0.569	0.586	0.200	0.583	0.621	0.194
CFI 45%MISS	0.273	0.250	0.159	0.361	0.337	0.182	0.347	0.318	0.180
PE 45%MISS	0.568	0.609	0.223	0.603	0.636	0.201	0.619	0.666	0.201
Block bias									
CFI 5%MISS	0.419	0.421	0.192	0.561	0.586	0.195	0.550	0.538	0.197
PE 5%MISS	0.379	0.397	0.230	0.389	0.370	0.213	0.405	0.424	0.211
CFI 25%MISS	0.286	0.256	0.160	0.386	0.384	0.178	0.372	0.347	0.181
PE 25%MISS	0.519	0.556	0.233	0.562	0.584	0.203	0.580	0.610	0.208
Originals "Random"									
CFI 15%MISS	0.365	0.367	0.163	0.490	0.503	0.159	0.482	0.475	0.167
PE 15%MISS	0.403	0.446	0.225	0.423	0.433	0.168	0.448	0.443	0.189
CFI 30%MISS	0.348	0.333	0.204	0.439	0.423	0.209	0.430	0.407	0.208
PE 30%MISS	0.456	0.440	0.260	0.470	0.482	0.226	0.486	0.511	0.232
CFI 45%MISS	0.264	0.230	0.189	0.337	0.306	0.200	0.331	0.295	0.203
PE 45%MISS	0.529	0.545	0.273	0.562	0.597	0.240	0.561	0.595	0.249
CFI 60%MISS	0.140	0.105	0.132	0.190	0.144	0.159	0.188	0.144	0.153
PE 60%MISS	0.642	0.741	0.291	0.680	0.722	0.244	0.679	0.761	0.242
CFI 75% MISS	0.035	0.017	0.052	0.057	0.026	0.081	0.052	0.025	0.075
PE 75%MISS	0.861	0.947	0.221	0.853	0.937	0.200	0.854	0.925	0.192
CFI 90%MISS	0.003	0	0.016	0.003	0	0.016	0.003	0	0.016
PE 90%MISS	0.982	1	0.061	0.984	1	0.048	0.970	1	0.109

Table 1
(*Continued*)

	Equal weights			K15			K100		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Originals "Block Bias"									
CFI 5%MISS	0.458	0.424	0.223	0.596	0.639	0.187	0.582	0.555	0.189
PE 5%MISS	0.272	0.184	0.233	0.296	0.230	0.196	0.308	0.261	0.205
CFI 25%MISS	0.323	0.288	0.176	0.407	0.400	0.178	0.423	0.375	0.181
PE 25%MISS	0.426	0.473	0.219	0.473	0.470	0.205	0.455	0.488	0.208

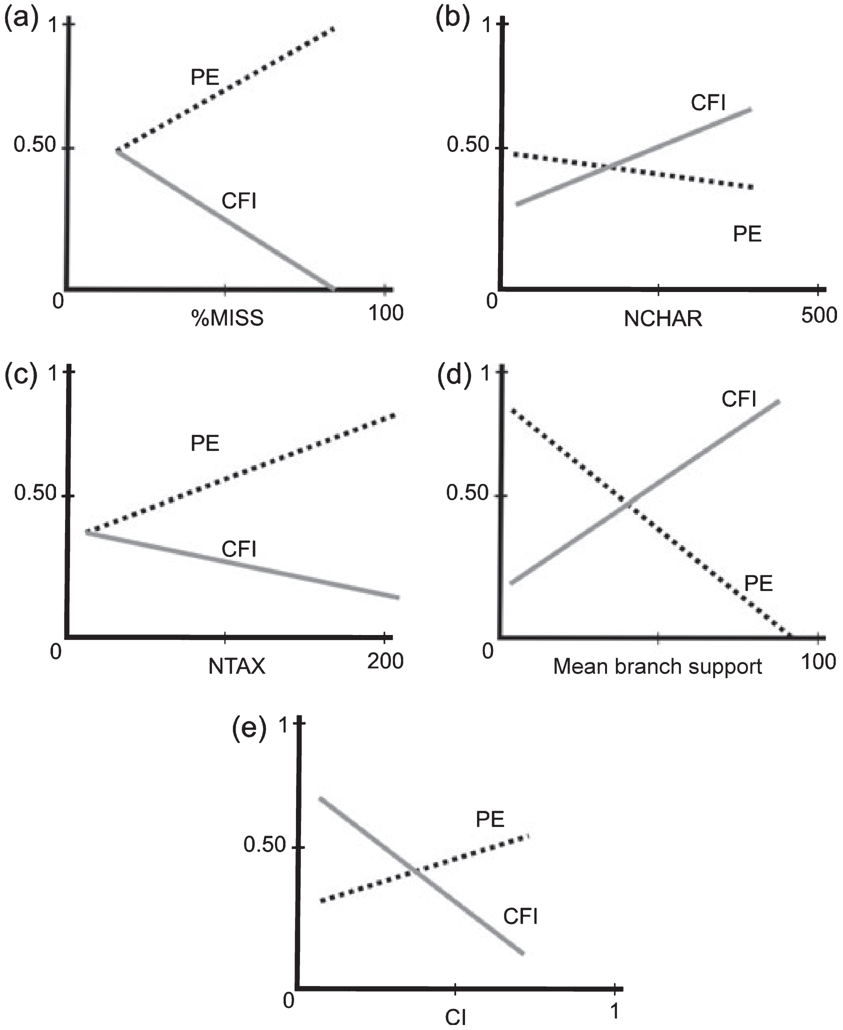

Fig. 3. Graphical representation of the general relationships between accuracy and error with (a) %MISS, (b) number of characters (NCHAR), (c) number of taxa (NTAX), (d) branch support, and (e) homoplasy (CI).

accuracy values, "character bias 15%" the highest, and the other analyses intermediate values (Fig. 4g). This pattern was more evident with 25% MISS than with 5% MISS (Fig. 4h). The only significant differences were between "taxa bias 15%" versus "character bias 15%", "character bias 15%" versus "block bias", and "taxa bias 15%" versus "block bias" (Appendix S4).

The rate of error showed non-significant differences between the different analyses. The only significant differences that remained stable under the different

Table 2
Summary of the pattern of correlations between accuracy (CFI) and error (PE), and number of characters (NCHAR), number of taxa (NTAX), branch support, and consistency index (CI)

	NCHAR		NTAX	
Analysis	CFI	PE	CFI	PE
Random	0.14–0.54***	(30–75%) −0.36/−0.15***	(30% and 60%) −0.16*	(15–60%) 0.17–0.37**
Taxa bias 15%	0.19–0.39***	n.s.	−0.30/−0.17**	(5%) 0.22–0.25***
Taxa bias 50%	0.27–0.46***	(15%) −0.25/−0.16**	(15%) −0.21/−0.12*	0.16–0.37**
Character bias 15%	0.17–0.47**	n.s.	−0.27/−0.17**	0.20–0.29**
Character bias 50%	0.28–0.47***	−0.21/−0.18*	(30%)–0.20**	0.23–0.37***
Block bias	0.20–0.41***	(25%) −0.20/−0.16*	n.s.	0.20–0.27**

	Branch supports		CI	
Analysis	CFI	PE	CFI	PE
Random	(15–60%) 0.55–0.76*** (75%) 0.25–0.32***	(15–60%) −0.78/−0.62*** (75%) −0.42***	(15–60%) 0.11–0.29**	(15–60%) −0.56/−0.34*** (75%) −0.19*
Taxa bias 15%	0.48–0.66***	−0.66/−0.50***	(5%) 0.26–0.36***	−0.39/−0.55***
Taxa bias 50%	0.40–0.73***	−0.76/−0.55***	(15–30%) 0.15–0.37**	−0.52/−0.40***
Character bias 15%	0.61–0.68***	−0.70/−0.48***	0.16–0.46***	−0.57/−0.37***
Character bias 50%	0.62–0.77***	0.77/−0.57***	0.21–0.44***	−0.54/−0.42***
Block bias	(5%) 0.66–0.69***	(5%) −0.71/−0.60***	(5%) 0.20–0.20**	(5%) −0.52/−0.38***

Values in parentheses refer to the percentage of missing data that presented those correlation values. If not specified, the correlation values correspond to the complete range of percentages of missing data explored. Significant at: *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

searches were "random" versus "characters bias 50%" with 45% MISS (Fig. 4d–f,i,j; Appendix S4).

In summary, the analyses with higher accuracy levels were both "character bias", while "taxa bias" analyses showed the poorest cladistic performance, and "random" and "block bias" showed intermediate levels of accuracy. The differences between character bias and block bias became higher at elevated percentages of missing data. The error rate remained constant under the different searches, showing no major differences between the analyses.

Effect of replacement of the original missing entries

The random addition of MISS in the original matrices (i.e. without replacing the original missing entries with random selected states) gave more robust results than the "random" analyses. In most of the analyses, the median accuracy values in the former approach were ≈ 0.09 higher, and the median error rates ≈ 0.10 lower than "random", both of which were significant under the bootstrap T test (P usually < 0.0002; Fig. 5a,b).

When comparing the effect of the addition of MISS with a "block bias" distribution, the only effect observed was that maintaining the original MISS produced significantly lower error rates ($P < 0.02$; Fig. 5c,d). However, the original matrices approach had more missing entries than the "block bias" approach, because the new "?" were placed in sites without the original "?". Therefore, the approach in which the original missing entries are retained possessed higher accuracy values relative to the percentage of MISS present in the matrices, which agrees with the comparison of the "random" analyses.

The correlations between accuracy values and error rates with other parameters showed patterns similar to those observed in the analyses in which the MISS were replaced, with the exception that in "random" the percentage of binary characters was never correlated with accuracy values or error rates (Appendix S5).

Discussion

Number and distribution of missing data in the real matrices

Our general results, obtained with 354 matrices from real phylogenetic analyses, agree with general knowledge regarding missing data. It is clear that the missing data distribution in most matrices is not random, but concentrated on some characters and/or taxa; although our test for the distribution of missing data does not evaluate the randomness of the distribution in characters and taxa at the same time (i.e. if MISS are in a block composed of taxa lacking information for the same characters), it is very likely that this happens in most of the matrices that have the original missing entries not randomly distributed in characters and taxa. In fact, a non-random distribution of the missing entries is expected due to the non-random preservation of anatomical structures or tissues in fossils, or the combination of different sources of phylogenetic characters in different taxon sampling (e.g. osteology and soft anatomy, or morphology and DNA; Wheeler, 1992;

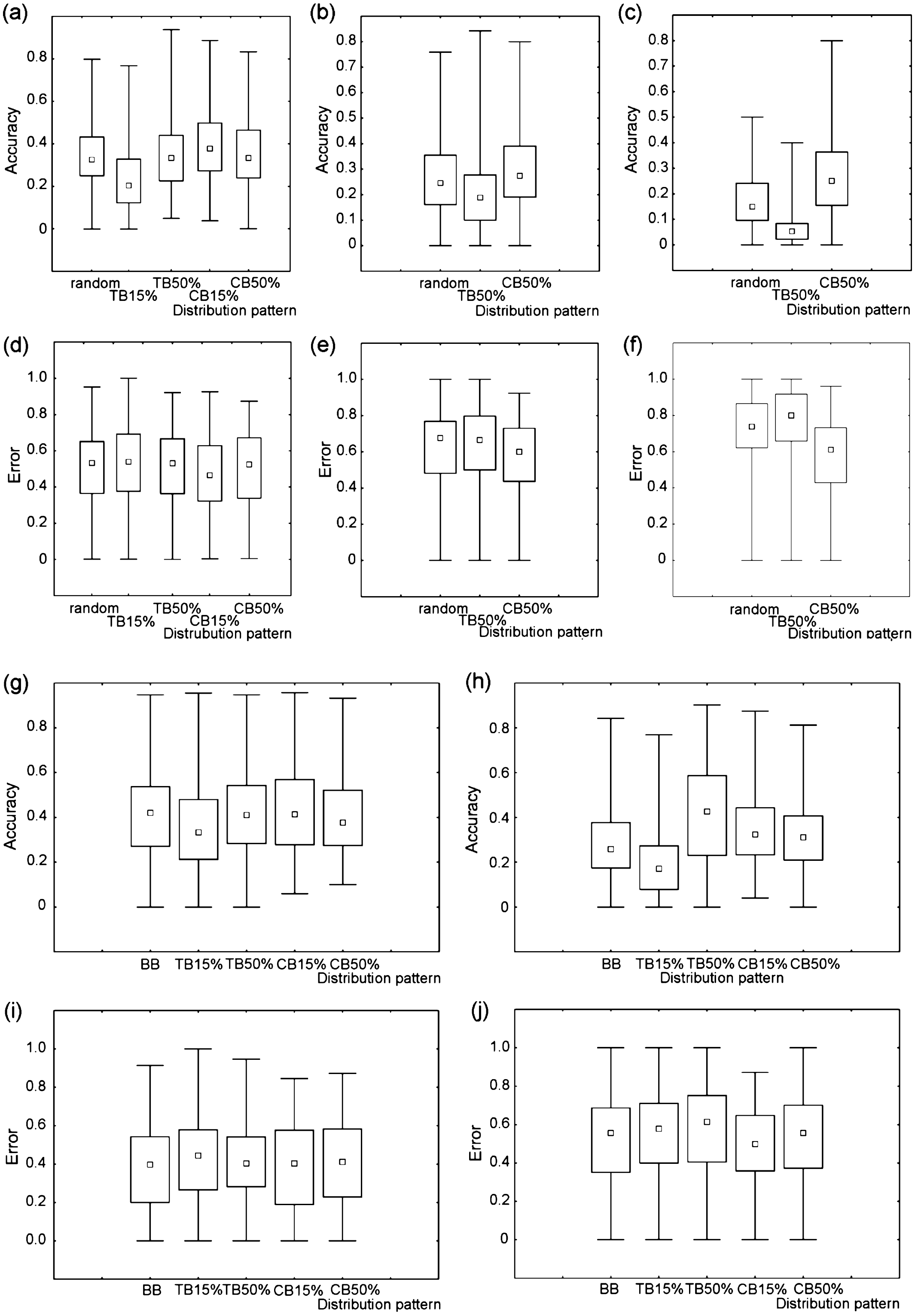

Fig. 4. Effect of patterns of distribution of missing cells on accuracy and error rate. (a) CFI with 15% missing data; (b) CFI with 30% missing data; (c) CFI with 45% missing data; (d) PE with 15% missing data; (e) PE with 30% missing data; (f) PE with 45% missing data; (g) CFI with 5% missing data; (h) CFI with 25% missing data; (i) PE with 5% missing data; (j) PE with 25% missing data; TB15%, "taxa bias 15%"; TB50%, "taxa bias 50%"; CB15%, "characters bias 15%"; CB50%, "characters bias 50%"; BB, "block bias". Graphs only show results under equal weights.

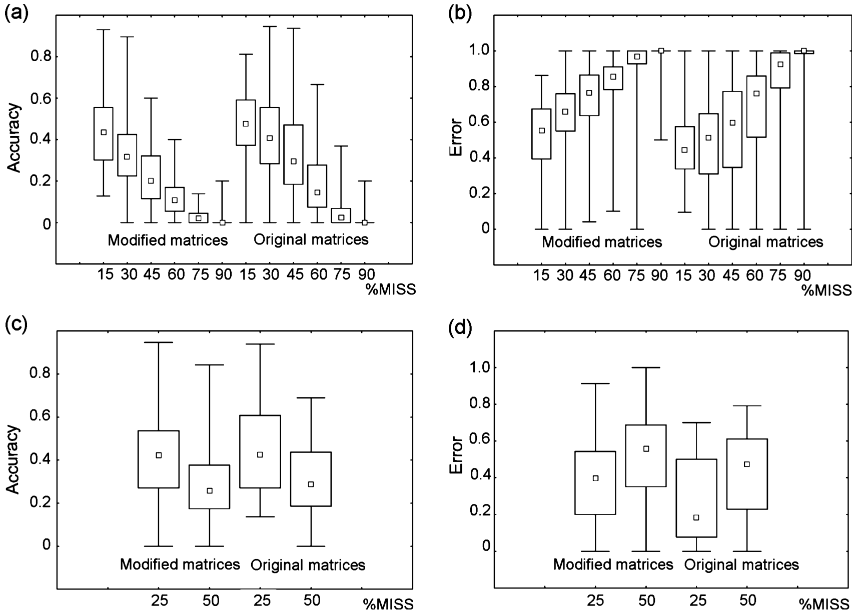

Fig. 5. Comparisons among original and modified matrices in the missing data analysis. (a,b) "Random" analyses; (c,d) "block bias" analyses. Original matrices: analysis that retains the original missing cells; modified matrices: analyses where the missing cells present in the matrices were replaced by random states. For other abbreviations see legend to Fig. 2. Graphs only show results under equal weights.

Wiens, 2003a). Thus, one would expect the most common pattern of distribution of missing entries to be the one in which some taxa lack data for the same characters (i.e. our "block bias" analysis). In spite of this predominance, the other patterns evaluated here are also present in several of the matrices included in this study.

A second area of agreement is that the palaeontological matrices possess more missing data than the matrices comprising only recent taxa (e.g. Wilkinson, 1995). In our sample the first group (almost half of the matrices analysed) had significantly higher median and mean %MISS. However, the range of values observed in both groups overlapped widely, and some neontological matrices showed a considerable level of MISS that could reach 45% that of the matrix. Thus, as noted by other authors (e.g. Gauthier et al., 1988; Flynn et al., 2005; Fulton and Strobeck, 2006) the missing data problem is not exclusively related to the inclusion of fossils in a phylogenetic analysis (cf. Cobbett et al., 2007).

Impact of the number of characters, branch support and homoplasy

The main pattern observed when analysing the effect of missing data in real matrices is that increasing the number of missing data decreases the performance of the phylogenetic analyses (i.e. less accuracy and more error). As observed by other authors in empirical or simulated analyses (e.g. Huelsenbeck, 1991; Novacek, 1992; Wiens, 1998, 2006), such a pattern is independent of the distribution of the missing data and the kind of weighting scheme used (i.e. equal or implied weighting). This pattern is not explained by the number of missing cells *per se*, but rather by the number of scored characters per taxon (Wiens, 2003a,b). Our results support this claim, as we found a positive correlation between the number of characters and accuracy and a negative correlation with error rate in most of the analyses, both using four different distribution patterns and varying the weighting method.

It is interesting to note that branch supports showed the same pattern as that of the number of characters, although with stronger correlation values, thus implying that matrices with high mean branch supports will have higher accuracy values and lower error rates. The similar behaviour shown by the number of characters and branch support reflects the positive relationship between them, as previously noted by several authors (e.g. Sanderson and Donoghue, 1996; Bremer et al., 1999; Siddall, 2002). This means that the inclusion of more characters may produce a more robust phylogenetic signal, providing additional "resistance" to the missing data bias. A similar pattern has been observed between support values and "phylogenetic noise" (i.e. more "noise" is needed to lose nodes with higher support values; Wenzel and Siddall, 1999), and in the positive correlation of jackknife support values with the missing entry replacement data analysis (MERDA) index (r: 0.56–0.89 $P < 0.03$; calculated here using the tables from Norell and Wheeler, 2003). Consequently, more characters may lead to more synapomorphies that could help to rescue the clades from the negative impact of the missing entries, or from other kinds of matrix "perturbations".

The relationship between the number of characters and cladistic performance (expressed as the consensus fork index and the error rate) along with the neutral or negative correlation between the number of taxa and accuracy (or positive with the error rate) suggests that, in this context, sampling more characters instead of sampling more taxa will be a better way to temper the effect of missing data. Several authors are in agreement that adding more characters is the best strategy for obtaining more robust phylogenies in general (e.g. Wheeler, 1992; Bremer et al., 1999; Poe and Swofford, 1999; Mitchell et al., 2000; but see Graybeal, 1998; Pollock et al., 2002). The presence of missing entries on the characters appears to have no influence on the benefit obtained from adding more characters (cf. Wiens, 1998, 2003a,b; Kearney, 2002; Kearney and Clark, 2003), and thus our results do not support the idea of excluding characters because they have missing cells.

Homoplasy shows an inverse pattern, in which less homoplasy (i.e. higher CI) is related to higher accuracy values and lower error rates, and vice versa. This is expected, because more homoplasy may reflect more incongruence in a matrix, thus implying a less robust phylogenetic signal; in fact, CI is inversely correlated with branch support values. One interesting issue is that homoplasy increases with the number of characters (Archie, 1989, 1996; Sanderson and Donoghue, 1996; this contribution), but this increase is not strong enough to interfere with the positive correlation between the number of characters, branch support, and cladistic performance.

Although the replacement of the original missing entries present in the matrices by randomly chosen character states adds a random "noise", the only consequence we observed was a lower accuracy and higher error rate compared with the analyses performed without replacing the original missing entries. Despite the fact that the phylogenetic analyses performed with the modified matrices (i.e. with the missing entries replaced) underestimated the phylogenetic performance of the original morphological matrices, they still could be considered as conservative indicators of phylogenetic performance. Other patterns related to the effect of different percentages and distributions of missing data and the relationship between accuracy and error rates with other parameters were similar between analyses with the original matrices and analyses with the matrices with missing entries replaced.

Effect of the pattern of distribution

Our analyses suggest that some patterns of distribution of missing data are more detrimental than others. Concentration of missing entries in a few taxa ("taxa bias 15%") showed the worst performance, while concentration in a few characters ("characters bias 15%") showed the least detrimental effect. The other patterns of distribution (i.e. "character bias 50%", "random", "block bias", and "taxa bias 50%") were placed between these extremes, although a wide overlap in accuracy and error rate values was observed between the different distributions. An explanation for this is that concentrating most of the missing cells in a few taxa could deplete the information necessary to place them correctly in a tree, or even that some of those taxa can become redundant (Wilkinson, 2003). If this is the case, this taxon could float in the tree and collapse several nodes in the consensus tree, behaving as wildcard taxa (Nixon and Wheeler, 1992; Wilkinson, 1995, 2003; Wiens, 2003a). In contrast, a character with 90% missing entries only loses information, and this could lead to node loss or a slight increase in error rate, although the expected impact is much less extreme (cf. Wiens, 1998, 2003a). Fortunately, the "worst" pattern ("taxa bias") is not the predominant one in the real matrices surveyed here. However, if one has to deal with redundant or wildcard taxa that may be generated by this missing data distribution pattern, there are several tools available, such as the "safe taxonomic reduction" approach or the reduced consensus (see Wilkinson, 1995, 2003).

Comparison with simulations

Our results are similar to those obtained using simulated matrices (e.g. Huelsenbeck, 1991; Wheeler, 1992; Wiens, 1998, 2003a,b, 2006). However, it is noteworthy that here we explored a more heterogeneous data set (as it came from real analyses) and a wider range of parameter values (e.g. homoplasy, branch support). One apparent discrepancy is that the absolute values of accuracy are much lower here than those obtained in other analyses, but this could be explained because previous authors have not used very long branches to simulate the data, an thus their matrices have less homoplasy than the real matrices studied here. In fact, studies that have explored a wider range of branch lengths have obtained results similar to ours, when using higher branch length values (see figs 2–5 in Wiens, 1998; fig. 3 in Wiens, 2003b). This is related to the negative correlation observed here between accuracy and homoplasy (CI), and the positive correlation between CI and the error rate (see above). In the analyses where the original missing entries were preserved we generally obtained better results than in the analyses where the missing data were replaced by a random character state. This better performance may be another side of the negative relationship between homoplasy and performance under the presence of missing data. The approach in which the original missing entries were maintained showed lower homoplasy, higher branch support, and higher accuracy than with the alternative approach.

Our results are also in accordance with recent simulations that have shown that the "taxa bias" approach has clearly lower accuracy than the "block bias" approach, especially with few characters (500 or fewer) and more than 70% missing cells (see Wiens, 2003a,b figs 2 and 4). It is noteworthy that the simulations mentioned above explored limited patterns of missing data distribution, namely the "taxa bias" and "block bias" of the present study.

The relationship between accuracy and other parameters is similar to that obtained in the simulations; Wiens (1998, 2003a,b) used two model trees, one fully asymmetric and the other fully symmetric, and obtained similar results in both kinds of trees. This is in agreement with our results, as we found no significant correlations between the imbalance index and accuracy. Regarding the number of states of the characters, we also obtained results similar to those obtained with the simulation analyses, indicating a lack of influence on accuracy.

Conclusion

Our analyses are relevant to current knowledge of the "missing data problem" in morphological phylogenetic analyses as we have explored a very large sample of real matrices, and a wider and more heterogeneous range of variables (e.g. size, homoplasy, phylogenetic signal) than that used in simulations and previous studies. As a result, we found the same negative effect of missing data entries as that previously reported. However, our results showed lower average values than the simulations due to higher levels of homoplasy present in the real data sets. We also noted that the inclusion of more characters could make the matrices more robust to this bias, indicating that the problem is mainly a lack of information, not just the presence of missing data *per se*. Regarding the distribution of missing data, the most prejudicial case is when missing entries are accumulated in several taxa (i.e. taxon bias); fortunately, this is one of the less common patterns in real matrices.

Acknowledgements

We thank Lone Aagesen, Silvana Sede, Matthew Willis, and Juan C. Fernicola for comments on drafts of the manuscript; Marcos Mirande for helping with the final version of the manuscript and three anonymous reviewers for their helpful comments; to Victoria Gonzalez Eusevi for checking the English, Pablo Goloboff and Diego Pol for their help with scripts and commentaries; to colleagues and friends who sent us matrices; to the Willi Hennig Society for the free version of TNT; and to CONICET for financial support.

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.

Archie, J.W., 1989. Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. Syst. Zool. 38, 253–269.

Archie, J.W., 1996. Measures of homoplasy. In: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: The Recurrence of Similarity in Evolution. Academic Press, New York, pp. 153–206.

Bremer, B., Jansen, R.J., Oxelman, B., Backlund, M., Lantz, H., Kim, K.-J., 1999. More characters or more taxa for a robust phylogeny-case study from coffee family (Rubiaceae). Syst. Biol. 48, 413–435.

Cobbett, A., Wilkinson, M., Wills, M., 2007. Fossils impact as hard as living taxa in parsimony analyses of morphology. Syst. Biol. 56, 753–766.

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

Driskell, A.C., Ané, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172–1174.

Flynn, J.J., Finarelli, J.A., Zehr, S., Hsu, J., Nedbal, M.A., 2005. Molecular phylogeny of the Carnivora (Mammalia): assessing the impact of increased sampling on resolving enigmatic relationships. Syst. Biol. 54, 317–337.

Fulton, T.L., Strobeck, C., 2006. Molecular phylogeny of the Arctoidea (Carnivora): effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Mol. Phylogenet. Evol. 41, 165–181.

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

Goloboff, P., 1993. Estimating character weights during tree search. Cladistics 9, 83–91.

Goloboff, P., 1997. Self-weighted optimization: tree searches and character state reconstructions under implied transformation costs. Cladistics 13, 225–245.

Goloboff, P., 1999. Analyzing large datasets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Goloboff, P., 2007. Calculating SPR distances between trees. Cladistics 23, 1–7.

Goloboff, P., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M., Szumik, C., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goloboff, P., Carpenter, J., Arias, J.S., Miranda Esquivel, D., 2008a. Weighting against homoplasy improves phylogenetic analysis of morphological data sets. Cladistics 24, 1–16.

Goloboff, P., Farris, J., Nixon, K., 2008b. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786. Available at http://www.zmuc.dk/public/phylogeny.

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

Hartmann, S., Vision, T.J., 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8, 95.

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

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

Kearney, M., Clark, J., 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. J. Vert. Paleont. 23, 263–274.

Maddison, W.P., 1993. Missing data versus missing characters in phylogenetic analysis. Syst. Biol. 42, 576–581.

Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall, London.

Martin, W., Roettger, M., Lockhart, P.J., 2007. A reality check for alignments and trees. Trends Genet. 23, 478–480.

Mitchell, A., Mitter, C., Regier, J.C., 2000. More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). Syst. Biol. 49, 202–224.

Nixon, K.C., Davis, J.I., 1991. Polymorphic taxa, missing values and cladistic analysis. Cladistics 7, 233–241.

Nixon, K.C., Wheeler, Q.D., 1992. Extinction and the origin of species. in: Novacek, M.J., Wheeler, Q.D., (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 119–143.

Norell, M.A., Wheeler, W.C., 2003. Missing entry replacement data analysis: a replacement approach to dealing with missing data in paleontological and total evidence data sets. J. Vert. Paleont. 23, 275–283.

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

Ogden, H.T., Rosenberg, M.S., 2007a. Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. Syst. Biol. 56, 182–193.

Ogden, H.T., Rosenberg, M.S., 2007b. How should gaps be treated in parsimony? A comparison of approaches using simulation Mol. Phylogenet. Evol. 42, 817–826.

Philippe, H.E., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

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

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

Pollock, D.D., Zwickl, D.J., McGuire, J.A., Hillis, D.M., 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51, 664–671.

Prevosti, F.J., Chemisquy, M.A., 2007. Entradas faltantes, pesos implicados y reconstrucciones filogenéticas. VII Reunión Argentina de Cladística y Biogeografía, San Isidro (Buenos Aires, Argentina). Darwiniana 45, 31S–33S.

Ramírez, M.J., 2003. The spider subfamily Amaurobioidinae (Araneae, Anyphaenidae): a phylogenetic revision at generic level. Bull. Am. Mus. Nat. Hist. 277, 1–262.

Sanderson, M.J., Donoghue, M.J., 1996. The relationship between homoplasy and confidence in phylogenetic trees. in: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: The Recurrence of Similarity in Evolution. Academic Press, New York, pp. 67–89.

Siddall, M.E., 2002. Measures of support. In: DeSalle, R., Giribet, G., Wheeler, W.C., (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser Verlag, Basel, pp. 80–101.

Wenzel, J.J., Siddall, M.E., 1999. Noise. Cladistics 15, 51–64.

Wheeler, W.C., 1992. Extinction, sampling and molecular phylogenetics. in: Novacek, M., Wheeler, Q. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 205–215.

Wiens, J.J., 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47, 625–640.

Wiens, J.J., 2003a. Incomplete taxa, incomplete characters and phylogenetic accuracy: is there a missing data problem? J. Vert. Paleont. 23, 297–310.

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

Wiens, J.J., 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54, 731–742.

Wiens, J.J., 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39, 34–42.

Wiens, J.J., Moen, D.S., 2008. Missing data and the accuracy of Bayesian phylogenetics. J. Syst. Evol. 46, 307–314.

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

Wiens, J.J., Fetzner, J.W.. Jr, Parkinson, C.L., Reeder, T.D., 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54, 719–747.

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

Wilkinson, M., 2003. Missing entries and multiple trees: instability, relationships, and support in parsimony analysis. J. Vert. Paleont. 23, 311–323.

Wilkinson, M., Benton, M.J., 1996. Sphenodontid phylogeny and the problems of multiple trees. Phil. Trans. R. Soc. Lond. B Biol. Sci. 351, 1–16.

Wolsan, M., Sato, J.J., 2009. Effects of data incompleteness on the relative performance of parsimony and Bayesian approaches in a supermatrix phylogenetic reconstruction of Mustelidae and Procyonidae (Carnivora). Cladistics. doi: 10.1111/j.1096-0031.2009.00281.x.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. List of matrices used in the missing entries analyses, with their percentage and pattern of distribution of missing cells, and references.

Appendix S2. Morphological matrices used in this contribution.

Appendix S3. TNT scripts used for the analyses.

Appendix S4. Statistical comparison of the introduction of different proportions of missing data (%MISS) and between different patterns of distribution of missing cells.

Appendix S5. Correlations between accuracy (CFI) and error (PE) with other parameters.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.