Syst. Biol. 42(4):576–581, 1993

Missing Data versus Missing Characters in Phylogenetic Analysis

WAYNE P. MADDISON

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

Coding a character for numerical phylogenetic analysis can be problematical when some taxa have various forms of a structure and other taxa lack the structure altogether. A common suggestion (e.g., Sneath and Sokal, 1973:180; Swofford, 1985; Farris, 1988) is to code those taxa with various forms of the structure as having different character states but to code those taxa lacking the structure as having missing data. For example, suppose some taxa have no tails, others have blue tails, and others red tails. The lack of tails in some taxa may be due to loss or to primitive absence. One might expect that the presence or absence of a tail might provide some phylogenetic information and the particular color of tail would provide additional information. These conditions might therefore be coded as two characters, one character of tail presence with states "absent" and "present" and a second character of tail color with states "missing," "blue," and "red." Those taxa that lacked a tail would be assigned "absent" in the first character, but because the tail is absent they could not be coded as having any particular color, and so they would be assigned "missing" in the second character. Thus the three possible tail conditions would be coded as follows:

Observed tail condition	Coding	
	Tail presence	Tail color
absent	absent	missing ("?")
red	present	red
blue	present	blue

As noted by Nixon and Davis (1991) and Platnick et al. (1991), this particular use of the missing data coding is but one of several. A missing data entry in a phylogenetic data matrix might mean that (1) the taxon has some state in the character but it is not known which (the data are missing), (2) the character is simply not applicable to the taxon (the character is "missing"), or (3) the character is polymorphic in the taxon. (Doyle and Donoghue [1986] used the missing data coding in yet another way in their X-coding, which attempted to implement assumptions not then available in common computer programs.) Nixon and Davis (1991) pointed out problems that arise with the third of these usages, i.e., missing data for polymorphisms. My goal here is to describe problems that arise with the second of these usages, i.e., missing data for inapplicable characters. In the example above, those taxa with no tails were coded as having missing data in the tail color character because the character was simply inapplicable to them. Such a coding can lead to unexpected and undesirable results in computer-based phylogenetic analyses using parsimony. The problems arise because the coding can result in ancestors being assigned impossible states, as noted by Platnick et al. (1991).

Suppose that there is a group of 14 species whose phylogenetic relationships are being reconstructed. A data matrix with many characters is obtained and analyzed using a parsimony tree-search program such as PAUP (Swofford, 1990) or Hennig86 (Farris, 1988). All characters other than those concerning tail condition clearly indicate that the most parsimonious tree is consistent with Figure 1, which is fully resolved except for the relationships of the tailed clade on the left side of the tree. Note that there are two major clades, in both of which tailless forms are basal; tailed forms

make up monophyletic groups within both of the major clades. The data concerning tail condition are not able to overturn the relationships shown in Figure 1, but they may help choose among alternative resolutions of the left tailed clade. My concern is how different codings of tail condition will cause us to make different decisions about resolving the left tailed clade.

Figure 2 shows two alternative resolutions of the left tailed clade (2a and 2c, first resolution; 2b and 2d, second resolution) and two alternative codings of tail condition. I will consider only these two resolutions to show how different character codings affect how the resolutions are judged relative to one another. There are other possible resolutions of the left clade, but I will be able to make my points by considering just the two.

If tail condition is split into two characters (Figs. 2a, 2b) as described above, then the first character, presence/absence of tail, agrees equally well with both resolutions (two steps for each of Figs. 2a and 2b). The second character, tail color red or blue, might help us decide relationships within each of the tailed clades.

One might think that Figures 2a and 2b would be equally parsimonious, each with only a single change in tail color in each tailed clade, but available computer programs (e.g., Swofford, 1990) treat taxa with missing data almost as if they were absent from the tree, at least with respect to that particular character. That is, when ancestral states are reconstructed (i.e., HTUs are optimized; Farris, 1970) to count how many steps a tree requires for a given character, the taxa with missing data are transparent to the algorithms (an exception to this can arise when a step matrix is used that violates the triangle inequality, because then shortcuts can be taken when the presence of a taxon breaks a branch in two). This means that the tailed clades will be treated as if they were side by side, with no intervening tailless species, and thus the two tailed clades should agree on their ancestral tail colors to be parsimonious. The number of steps is calculated as two steps for tail color in Figure 2a: one within the left tailed clade and one within the right tailed clade. Because both clades prefer the blue state at their bases, no steps are required between them. In contrast, Figure 2b requires three steps, one in each tailed clade and one between the two major clades. This last step is required between the two major clades because the right prefers blue at its base and the left prefers red at its base.

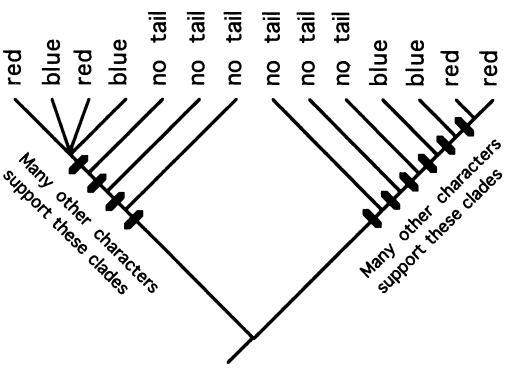

FIGURE 1. Example in which coding an inapplicable character as missing data could lead to an inappropriate conclusion in numerical analysis. Clear support in the data matrix (not shown) for the basic cladistic structure shown is assumed: two major clades with a tailed clade in each, basal to which is a series of groups lacking tails in each major clade. The far right tailed clade is fully resolved; all that remains is to resolve the far left tailed clade. The data on tails are assumed to be the only data that speak to the issue.

In this example, the two tailed clades have influenced each other, even though they are widely separated on the phylogeny, because the intervening taxa with missing data allowed the influence to leak through. This effect is often inappropriate and undesirable. These intervening taxa are tailless, and therefore one would suppose the ancestor of the whole group is tailless and therefore that tail color in each clade should be considered in isolation. It would not make sense to count a change in tail color in Figure 2b at the root node between the two major clades if this ancestor did not even have a tail. The two tailed clades should each be allowed to place either tail color basally without concern for the other tailed clade.

If the analysis were being done by hand, this problem could be noted, the colors in

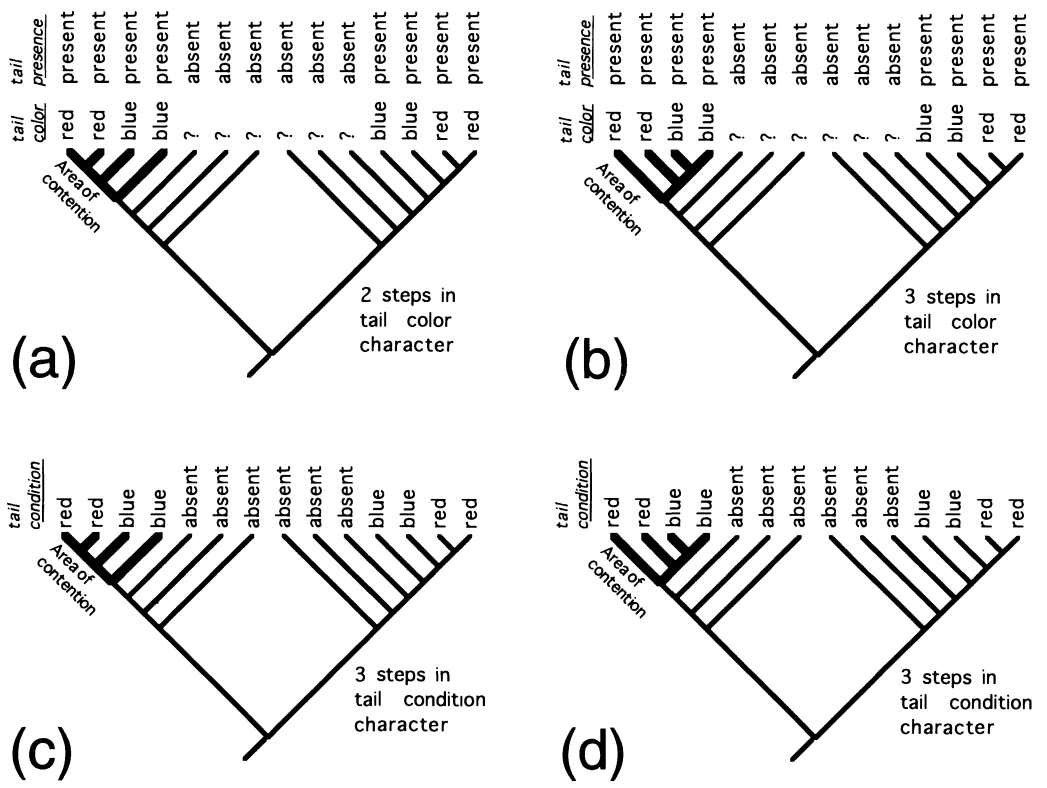

FIGURE 2. Choice among alternative resolutions of the left tailed clade. (a), (b) Tail condition is coded as two characters, presence and color. The presence/absence character costs two steps for each of the two alternatives, but the color character requires an extra step (three versus two) in (b) because of long distance influence from the right tailed clade. (c), (d) Tail condition is coded as a single character, with three states. Both alternative resolutions cost the same number of steps in this character (using Fitch parsimony) because the two tailed clades are insulated by the intervening state "absent."

each tailed clade could be recognized as nonhomologous, and their steps could be counted in isolation. But a computer program, when presented with the missing data coding, will judge the number of steps in the color character as if the ancestors had tails. It will count the resolution in Figure 2a to have fewer steps than that in Figure 2b. As just argued, it may be more appropriate to count the two resolutions as costing the same number of steps (as is done with the coding of Figs. 2c and 2d). The two trees are counted as being unequally parsimonious with the missing data coding, whereas they should have been considered equally parsimonious. The impact of such a mistake may be to cause us to ignore some legitimate alternative cladograms. However, with some data sets such miscounting of steps may lead us to choose a cladogram that should be considered unparsimonious.

The undesirable long-distance influence between the tailed clades is not a fault of the computer programs, because their treatment of missing data is appropriate if the character is applicable to all taxa but its state is unknown in some. Rather, it is the fault of inappropriately coding the tailless species as having missing data in the tail color character.

These problems are not restricted to morphological data. Coding a gap in an aligned protein or nucleotide sequence as being missing data suffers all the same problems.

Not Always a Problem

Coding inapplicable characters as missing data is safe in some circumstances. The problems arise only if there are two or more regions of the cladogram in which the character is applicable separated by regions with taxa for which the character is inapplicable. If there is only a single terminal taxon with the character inapplicable, or if the only taxa with the character inapplicable are all members of a clade, or if the only taxa with the character applicable are all the members of a clade, then the undesirable effects described will not occur. Thus if the character is inapplicable only among the basal outgroups, then the coding of these outgroups as having missing data will not cause problems. However, only analyses that constrain the taxa with the character applicable (or those with the character inapplicable) to remain together will be guaranteed safe. Merely obtaining an output tree that places them together is not enough, because at some point in the computer analysis a tree like the one in Figure 1 may have been examined and rejected inappropriately.

In special circumstances, the long-distance interaction of tailed regions of the cladogram may actually be desirable, as being consistent with certain biological assumptions. Suppose a basal paraphyletic group was tailed, and from it sprang a clade that was basically tailless except for a small group that had reinvented tails. One might want the reconstruction of tail color evolution within the basal tailed group and the secondarily tailed group to influence each other across the intervening tailless region if it is thought that the character was most likely to reappear as it was just before its loss (see Donoghue and Doyle, 1985). However, existing parsimony algorithms could not be used to implement this assumption properly because they would have to adapt if the tree were rerooted within the tailless region as in Figure 1; for such a rooting, no argument of reappearance could be made.

Were the analysis done by hand, problems such as those described in this paper could be recognized and handled. But mental analyses cannot usually consider all alternative trees, and computer analyses are becoming more commonplace. We must therefore consider how to deal with these problems in computer analyses.

Solutions

The problems described occur because tail color evolution is considered in isolation from the evolution of tails themselves, and thus a solution should be sought in methods that consider tails and their colors together. One possibility is to code tail condition as a single three-state character that incorporates both tail presence and color. Taillessness would be coded not as missing, but as a third character state (Figs. 2c, 2d). The ancestors intervening between the tailed clades will then be reconstructed as having this third character state, and the tailed clades will be insulated from each other. This coding would be equivalent to the coding of a gap in aligned molecular sequence data as a separate character state (e.g., the gapstate = newstate option in PAUP [Swofford, 1990]).

But how should this three-state character be treated? If it were treated as having unordered character states (Fitch parsimony; Fitch, 1971), then one step would be counted for any change, whether it involved the gain or loss of a tail or a change of tail color. This is how steps are counted in Figures 2c and 2d. This option is reasonable and may often be the best solution.

However, in some cases a systematist may want to assume that gains of a tail are more important and tend to delimit larger monophyletic groups within which a secondary and less important character, tail color, delimits substructure. Gains and losses of tails could be weighted more heavily than changes in color if step matrix algorithms (Sankoff and Rousseau, 1975; Sankoff and Cedergren, 1983; Maddison and Maddison, 1987, 1992; Swofford, 1990) were used. A step matrix indicates the number of steps counted for a transformation between each pair of character states, and the algorithms reconstruct character evolution so as to minimize the sum of steps so counted. Thus, using a step ma-

trix, one might say that a gain or loss of a tail is two steps, whereas a change in tail color is only one step. Lacking developmental or genetic evidence to indicate relative weights, the particular number of steps assigned to each change would be somewhat arbitrary. If an extreme weighting were chosen, e.g., 100 steps for gain/loss and 1 step for change of color, then this would have the effect of choosing a reconstruction that minimizes the number of gains and losses and, secondarily within those areas of the tree with tails, that minimizes changes in tail color. Some systematists might find this effect of giving primacy to gains and losses desirable. However, when achieved with an extreme step matrix, changes in tail color would contribute relatively very little to choice of trees. (Either gain/loss of tails would be given unreasonably high weights or tail color changes would be given unreasonably low weights, depending on the weights given to other characters in the analysis.) One possible adjustment would be to reconstruct ancestral states using such an extreme step matrix but then count the number of changes implied by the reconstruction as the basis for judging the tree (see Maddison and Maddison [1992:112] for a discussion of counting of changes).

Regardless of whether gains and losses are weighted differently from changes in tail color, the approach of coding tail condition as a three-state character has a complication. If another character concerning tails is incorporated into the analysis, for instance tails with scales versus those with hairs, then how is it used? If it is treated as a character with three states, tail absent versus scaled versus hairy, then there will be duplication in the counting of tail gains and losses between this character and the one concerning color, whether they are analyzed using Fitch parsimony, a 2-1 step matrix, or some other step matrix.

In current computer programs, perhaps the best general solution to this problem is to fuse characters that include the same state "tail absent" into a single character. In the example given, the two characters of tail color and tail covering would become a single five-state character with states tail absent, tail blue and scaled, tail blue and hairy, tail red and scaled, and tail red and hairy, and an elaborate step matrix would be used to indicate the cost of various transformations of state. In principle, this practice could be extended to fusing three or more characters. In mental analyses, such a practice is sometimes followed (e.g., Ball and Maddison, 1987:299, character 13). However, if many more distinctions among tails are found (long/short, bumpy/smooth, etc.), it may not be possible to fuse all the characters in existing computer programs because the resulting character could have hundreds of discrete states ("tail red and hairy and short and smooth and twisted" versus "tail red and hairy and short and smooth and coiled" versus . . .). Even if the program could handle a character with so many states, calculations with it using its large step matrix could be unbearably slow.

Perhaps the eventual solution will be to write new algorithms for computer programs that will allow the characters to be coded independently but that will consider interactions between characters and count steps in some characters only on those portions of the tree on which they are applicable. Algorithms that consider interactions between characters have already been devised for certain purposes. For instance, in phylogenetic sequence alignment, insertions and deletions in adjacent nucleotide sites can be counted as being part of the same event (Hein, 1989).

ACKNOWLEDGMENTS

I thank Geeta Bharathan, Alan de Queiroz, Michael Donoghue, Jim Doyle, Junhyong Kim, David Maddison, and David Swofford for stimulating discussion.

REFERENCES

BALL, G. E., AND D. R. MADDISON. 1987. Classification and evolutionary aspects of the species of the New World genus *Amblygnathus* Dejean, with description of *Platymetopsis,* new genus, and notes about selected species of *Selenophorus* Dejean (Coleoptera: Carabidae: Harpalini). Trans. Am. Entomol. Soc. 113: 189–307.

DONOGHUE, M. J., AND J. DOYLE. 1985. A method for coding partially ordered multistate characters for

cladistic analysis: Examples from a phylogenetic study of seed plants. Am. J. Bot. 72:950. (Abstr.)
DOYLE, J., AND M. J. DONOGHUE. 1986. Seed plant phylogeny and the origin of the angiosperms: An experimental cladistic approach. Bot. Rev. 52:321–431.
FARRIS, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19:83–92.
FARRIS, J. S. 1988. Hennig86, version 1.5. Published by the author, Port Jefferson Station, New York.
FITCH, W. M. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20:406–416.
HEIN, J. 1989. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol. Biol. Evol. 6:649–668.
MADDISON, W. P., AND D. R. MADDISON. 1987. MacClade, version 2.1. Cambridge, Massachusetts.
MADDISON, W. P., AND D. R. MADDISON. 1992. MacClade, version 3: Analysis of phylogeny and character evolution. Sinauer, Sunderland, Massachusetts.
NIXON, K. C., AND J. I DAVIS. 1991. Polymorphic taxa, missing values and cladistic analysis. Cladistics 7:233–241.
PLATNICK, N. I., C. E. GRISWOLD, AND J. A. CODDINGTON. 1991. On missing entries in cladistic analysis. Cladistics 7:337–343.
SANKOFF, D., AND R. J. CEDERGREN. 1983. Simultaneous comparison of three or more sequences related by a tree. Pages 253–263 *in* Time warps, string edits and macromolecules: The theory and practice of sequence comparison (D. Sankoff and J. B. Kruskal, eds.). Addison-Wesley, Reading, Massachusetts.
SANKOFF, D., AND P. ROUSSEAU. 1975. Locating the vertices of a Steiner tree in an arbitrary metric space. Math. Program. 9:240–246.
SNEATH, P. H. A., AND R. R. SOKAL. 1973. Numerical taxonomy: The principles and practice of numerical classification. W. H. Freeman, San Francisco.
SWOFFORD, D. L. 1985. PAUP: Phylogenetic analysis using parsimony, version 2.4. Illinois Natural History Survey, Champaign.
SWOFFORD, D. L. 1990. PAUP: Phylogenetic analysis using parsimony, version 3. Illinois Natural History Survey, Champaign.

Received 23 January 1992; accepted 11 May 1993

Syst. Biol. 42(4):581–587, 1993

Parsimony Can Be Consistent!

MICHAEL A. STEEL,[1] MICHAEL D. HENDY,[2] AND DAVID PENNY[3]

[1]*Department of Mathematics, University of Canterbury, Christchurch, New Zealand*
[2]*Department of Mathematics, Massey University, Palmerston North, New Zealand*
[3]*School of Biological Sciences, Massey University, Palmerston North, New Zealand*

A desired property of any method for reconstructing evolutionary trees is that it be consistent, i.e., as sequences become longer the method will recover the correct tree with probability tending to 1. In an important development, Felsenstein (1978) showed that the popular parsimony criterion could, with a simple model of evolution of two-state characters, converge to an incorrect tree as the sequences became longer. This problem with parsimony also applies to the compatibility criterion for selecting optimal trees.

The problem with parsimony was originally thought to be limited to cases where lineages had markedly different rates of evolution (Felsenstein, 1978), but the problem was later found under wider sets of conditions. With five or more taxa it can be a problem, even with constant rates of evolution (Hendy and Penny, 1989; Zharkikh and Li, 1993). With six or more taxa with both constant and arbitrarily low rates of evolution (Hendy and Penny, 1989) and with a large but unspecified number of taxa, the standard parsimony criterion may fail to converge to the correct tree even when all edges of the tree have the same expected number of changes (Steel, 1989). Similar examples can be found with four-state character models, such as Kimura's 3ST model (Hendy and Charleston, 1993).

We report here that the original conclusion is too sweeping in that the prob-