Cladistics (1991) 7:337–343

ON MISSING ENTRIES IN CLADISTIC ANALYSIS

Norman I. Platnick[1], Charles E. Griswold[2] and Jonathan A. Coddington[2]

[1] *Department of Entomology, American Museum of Natural History,*
Central Park West at 79th Street, New York, NY 10024, U.S.A.
[2] *Department of Entomology, National Museum of Natural History,*
NHB 164, Smithsonian Institution, Washington, D.C. 20560, U.S.A.

Received for publication 15 March 1991; accepted 29 August 1991

Abstract—The exact algorithms of two commonly used parsimony programs, Hennig86 by J. S. Farris and PAUP by D. Swofford, sometimes produce different solutions, and sometimes produce resolutions that are not supported by the data being analysed. The discrepancies apparently involve the treatment of missing entries, which can currently represent unknown data, inapplicable character and/or polymorphic taxa. Each of those potential sources of ambiguity is logically (if not computationally) different; with regard to binary characters, unknown data could be either 0 or 1, inapplicable characters are neither 0 nor 1 and polymorphisms are both 0 and 1. Resolutions that cannot be supported by any possible combination of known state attributions should either be flagged as such or suppressed entirely.

Introduction

In current cladistic analyses, missing entries in data matrices can represent information that is unknown, characters that are inapplicable to the taxon in question and/or polymorphism in terminal taxa. Because these different sources of ambiguity have different implications, it may not be appropriate to treat them as equivalent in all respects.

Take, for example, the case of data that are simply unknown, a situation most commonly encountered when a species is known only from one sex, or only from one life stage, or only from fragmentary fossils. Clearly, the "?" entries for such cases might represent (for a binary character) either a 0 or a 1. In such cases, algorithms should allow for both those possibilities when selecting the most parsimonious cladogram(s). It appears, however, that the two most commonly used parsimony programs may not always allow for these possibilities in the same way. Consider the following data set for five taxa:

Matrix 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	?	?	0	?	0	?	0	?	0	0	0	0	0	0	0
B	?	0	?	0	?	0	?	0	?	1	1	1	?	?	?
C	0	?	?	?	?	1	1	1	1	1	?	?	1	1	?
D	1	1	1	1	1	?	?	1	1	?	1	?	1	?	1
E	1	1	1	1	1	1	1	?	?	?	?	1	?	1	1

Note that for each character, there are two 1s, one 0 and two missing entries. Exact analysis using the Branch and Bound option of Swofford's (1990) PAUP program, version 3.0n, yields 15 cladograms, each 15 steps long. Those 15 cladograms all place

0748-3007/91/040337 + 07$03.00/0

taxon A as the sister group of B–E; together they represent the 15 possible dichotomous resolutions for the other four taxa (B, C, D and E).

When this data set is analysed using the implicit enumeration option of Farris' (1988) Hennig86 program, version 1.5, only a single most parsimonious cladogram is reported. That single solution is the totally unresolved "bush" for five taxa, and it is also reported to require only 15 ateps.

What is the source of this discrepancy in results? In PAUP, "only those characters that have non-missing values will affect the location of any taxon on the tree" (Swofford, 1990). Thus, the cladograms produced by PAUP can be obtained by first noting that taxon A is not united with any other taxon by any "1" entries, which therefore places it at the base of the cladogram. Note next that any of the 15 characters can be placed at a node of any of the 15 possible resolutions for taxa B–E by simply postulating "1" entries as necessary. For example, consider the pectinate resolution A(B(C(DE))). Characters 1, 2, 3, 4, 5 and 15 map immediately as DE synapomorphies. Characters 6, 7, 8, 9, 13 and 14 can be mapped as CDE synapomorphies by presuming that the missing entries for taxa C, D and E are each 1s. Characters 10, 11 and 12 can be mapped as BCDE synapomorphies by presuming that the two missing entries for each of taxa B, C, D and E are all 1s. This can be done for any of the 15 resolutions, without requiring any homoplasy (i.e. in 15 steps). For some of those resolutions, the missing entries of characters 1, 2, 4, 6 or 8 (i.e. those characters in which taxa B or C are known to have 0 entries) must all be assumed to be 1s, which then become the plesiomorphic state.

Similarly, the single cladogram produced by Hennig86 can only be produced by treating *every* missing entry (including those of taxon A) as a 1. This results in each character having four 1 entries and only a single 0 entry. In the absence of additional outgroup information, Hennig86 then treats 1 as the plesiomorphic state for each character, requiring one autapomorphic change to 0 in each character and consequently providing no resolution in the resulting 15-step cladogram. In fact, Hennig86 includes a preprocessing module that identifies characters, such as the ones in this matrix, that can be interpreted as requiring the same number of steps on all possible cladograms, so that they can be disregarded during the search procedure (Farris, J. S., pers. comm.). Thus, Hennig86 will produce a totally unresolved bush for any data matrix including only character distributions like those above (i.e. with only one 0 and two 1 entries, plus missing entries).

Clearly, in this case neither program is reporting results that are incorrect; each, however, is reporting results that are incomplete. As is well known to its users, Hennig86 declines to resolve the most basal node of a solution on the "strength" of implied polarities only (i.e. on the choice of 0s versus 1s to represent particular character states). However, the difference in results in this example cannot be attributed to that aspect of Hennig86's precision. If one adds a dummy taxon (X), with a 0 entry for each character, and runs the enhanced data, the result is neither the equivalent of the 15 solutions reported by PAUP, nor of the bush reported initially by Hennig86. Rather, a single cladogram results, namely, the pectinate one considered above: A(B(C(DE))). Clearly, the addition of the dummy taxon effectively prevents Hennig86 from treating all the missing entries as plesiomorphic 1s. (For the curious, we can report that adding the dummy taxon has the same effect on PAUP's results, and that no additional change results, for either program, if a second dummy taxon, Y, is added, also with all 0 entries).

It might seem that the problem of incomplete results could be overcome by simply taking the precaution of using both programs. However, there are other resolutions that require only 15 steps for the data of Matrix 1, and which are not reported by either program. For example, the trichotomy A(B(CDE)) is as easily supported by the data as are any of the fully resolved cladograms reported by PAUP. Indeed, the implication of the Hennig86 results is that all partially and fully resolved cladograms for the five taxa can accommodate the data in only 15 steps.

Unknown Data and Resolution

These unexpected differences in results between Hennig86 and PAUP highlight a currently unresolved aspect of cladistic theory. Exactly what is the evidential significance of unknown data, and what role should they play in cladistic analysis? Consider the following data set:

Matrix 2

	1	2	3	4
A	0	0	0	0
B	1	0	0	1
C	1	1	0	0
D	1	1	1	1
E	1	1	1	0

There is, obviously, a single most parsimonious solution, including (once again) the components DE (supported by character 3), CDE (supported by character 2) and BCDE (supported by character 1—only on the assumption, of course, that 0 is plesiomorphic; if 1 is plesiomorphic, character 1 is just an autapomorphy of taxon A, and still requires only one step). Character 4 (state 1) requires parallelism between taxa B and D, resulting in a total length of five steps.

Consider next a modification of this data set, with an additional two taxa (F and G), each of which has one missing entry:

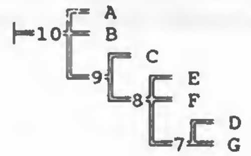
Matrix 3

	1	2	3	4
A	0	0	0	0
B	1	0	0	1
C	1	1	0	0
D	1	1	1	1
E	1	1	1	0
F	1	1	1	?
G	1	1	1	?

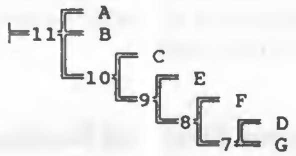
Intuitively, all the available evidence seems to place taxa F and G in a group with taxa D and E, rather than with taxa A, B or C. There seems to be no evidence justifying the placement of taxa F or G as closer to D than to E, or vice versa.

Hennig86, however, obtains six equally parsimonious solutions of length 5 for the enlarged data. As expected, it does not resolve the basal component including all taxa other than A. The CDEFG component occurs in all six cladograms, but only one of the six includes an unresolved DEFG component. The other five equally parsimonious solutions include components such as DF, DG and DFG:

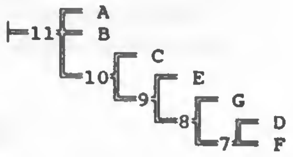
tree 0



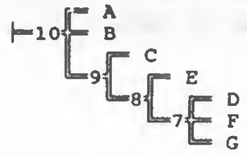
tree 1



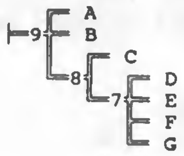
tree 2



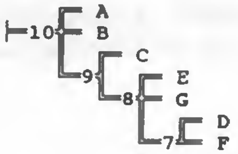
tree 3



tree 4



tree 5



There are only two missing entries in Matrix 3. If one assumes that they may eventually be represented by states 0 or 1 (rather than by some state not yet known, or by a polymorphism, or that the character will prove to be inapplicable to taxa F and G), there are only four possible assignments of states to the missing entries. If character 4 is represented in both taxa F and G by state 0, there is a single most parsimonious cladogram (tree 4). If character 4 is represented in both taxa F and G by state 1, there is a single most parsimonious cladogram (tree 3). If character 4 is represented in taxon F by state 1 and in taxon G by state 0, there is a single most parsimonious cladogram (tree 5). And if character 4 is represented in taxon F by state 0 and in taxon G by state 1, there is a single most parsimonious cladogram (tree 0).

In other words, two of the six equally parsimonious cladograms produced for Matrix 3 (trees 1 and 2) cannot in fact be supported by any conceivable assignment of 0 and 1 states to the missing entries. Ironically, they are the two most fully resolved solutions, and may therefore be the ones preferred by investigators! Indeed, the problem becomes worse as the proportion of missing entries in the matrix increases. Consider, for example, an eighth taxon, with entries identical to those of taxa F and G; intuitively, adding that taxon should only increase the size of the unresolved component already including taxa D–G. If one adds that eighth taxon to Matrix 3, however, not six but 26 equally parsimonious cladograms are reported; adding a ninth taxon with identical entries increases the number of cladograms reported to 150, and adding a tenth taxon with identical entries increases the number of cladograms reported to 1082! This problem is not confined to Hennig86; PAUP reports the same results for each of these analyses (with the additional resolution of a subbasal BCDEFG component in all cladograms).

It is possible, of course, that in this example taxa F and G might turn out to have character states that do not occur in taxa A–E. Consider the possibility that character 4 might turn to have three states: 0, 1 and 2. We then have five additional cases to examine: where F and G have 0 and 2, or 2 and 0, respectively; where F and G have 1 and 2, or 2 and 1, respectively; and where F and G both have state 2.

However, this new multistate character might be considered either as additive or as non-additive. We doubt that any systematists would feel particularly comfortable deciding, in advance, how to treat a multistate character when only two states have actually been observed. It seems difficult to test homology hypotheses about unknown states. But let us consider both possibilities.

For the case where F has state 0 and G has state 2, running the character as additive results in a single cladogram, tree 0. Running the character as non-additive results in two equally parsimonious cladograms, trees 0 and 4. Similarly, if F has state 2 and G has state 0, tree 5 is produced if the character is additive, and trees 4 and 5 are produced if the character is non-additive. Neither case supports the more highly resolved trees 1 or 2.

Consider, however, that F has state 1 and G has state 2. Running the character as additive yields tree 3, but running it as non-additive produces three equally parsimonious solutions: trees 3, 5 and the fully resolved tree 2 (of length 6). Flipping those states, if F has state 2 and G has state 1, tree 3 is again produced if the character is additive. In this case, however, treating the character as non-additive yields three equally parsimonious solutions: trees 3, 0 and the fully resolved tree 1 (also of length 6).

So it might seem as if the programs are producing reasonable results, if one allows that missing entries might turn binary characters into multistate ones, and that the programs can themselves choose to treat those multistate characters as non-additive. However, even if one makes those allowances, the results are not correct. There is a fifth possible case, where taxa F and G each have state 2 for their missing entry. In that case, treating the character as additive yields a unique solution, but it is not among the six originally produced for the data. It is the completely resolved, pectinate cladogram where taxa F and G, which share state 2, are sister groups at the tip of the cladogram. Treating the character as non-additive again produces three equally parsimonious solutions. These are tree 3, the new pectinate resolution, and another new resolution which clusters F and G as sister taxa but leaves them in a trichotomy with D and E.

Moreover, of course, any of these cases involving an unobserved state 2 must necessarily add at least one step to the length of every solution. We conclude, therefore, that the possibility of unknown states does not justify the way these programs are handling missing entries.

Some systematists, including the first author, requested that Henning86 output resolutions be supportable by potential combinations of state assignments for missing entries. Earlier parsimony programs, such as PHYSYS, did not attempt to do so. It now appears that trying to squeeze additional resolution from data sets by considering possible state attributions is simply ill-advised. Constraining the possible state attributions to those that add no homoplasy seems unrealistic, at best; if the observed data contain some homoplasy, the missing entries are likely to conceal homoplasy as well. Although there may be cases (biogeographic or coevolutionary studies, for example) where investigators need to know which additional resolutions require no additional steps, we suggest that the default option in future programs should suppress all unsupported components.

Until such suppression is available, we suggest that in any case where there are fully and partially resolved cladograms of the same length, users should check the character optimizations at each node carefully, to ensure that no nodes are supported only by mutually exclusive optimizations of the same character(s).

Inapplicable Characters and Polymorphism

Not all instances of question marks in current data sets reflect unknown character information, however. Many characters are simply inapplicable to particular taxa. In a data set including several spiders as well as some scorpions and mites, for example, characters relating to details of the abdominal spinnerets (a synapomorphy of spiders) are simply inapplicable to scorpions and mites. It is inappropriate, therefore, for parsimony programs to treat those entries in the same manner as unknown data. Because scorpions lack spinnerets entirely, a character such as "number of segments in the anterior lateral spinnerets" is clearly irrelevant to them. Attempting to assign states of this character to scorpions, on the basis of parsimony, is a fool's errand. Because such inapplicable characters are coded, in current parsimony programs, in exactly the same way (i.e. by the same symbol—?) as unknown data, those programs may produce resolutions, analogous to those we have just seen, which are (in this case) not only supported by no existing characters, but are not even potentially supportable by those characters. To put it baldly, those resolutions may depend on assignments of 0s and 1s that are simply impossible.

At first glance, it might seem that terminal taxa that are polymorphic for a given character could be analysed in a manner exactly like unknown data (i.e. coded as "?"). Under such analyses, polymorphic taxa are added to the cladogram in that position most strongly supported by the non-polymorphic characters, and without increasing the length of the solution. In realistic terms, of course, at least one additional step is implied between the polymorphic taxon and its nearest node, allowing for the *de novo* acquisition of (or reversal to) one or more different character states by some (and only some) members of the terminal taxon.

Coding polymorphisms in this manner, however, algorithmically eliminates the possibility that the polymorphism is itself a trait that was inherited from a common ancestor. Imagine, for example, a data set in which only two (out of many) terminal taxa show a particular polymorphism. Imagine also that other, non-polymorphic characters place those two taxa as sister groups. Standard coding of the polymorphisms as missing entries would overlook entirely the possibility that the polymorphism was acquired by the common ancestor of those two taxa (i.e. in one step), and passed on to both its descendants (which would themselves no longer require an "internal" step to account for the existing diversity).

We conclude, therefore, that unknown data, inapplicable characters, and polymorphic terminal taxa are logically (if not necessarily computationally) different. For a binary character, for example, unknown data could be *either* 0 or 1, inapplicable characters are *neither* 0 nor 1, and polymorphisms are *both* 0 and 1. It is possible, of course, that a character whose state is currently unknown in a particular taxon might prove to be inapplicable to that taxon, or polymorphic in it. Fortunately, the first of those possibilities need not affect the choice of most parsimonious topologies (on the assumption that the "inapplicable" state might be reached from any other state in one step), and polymorphism should presumably not be invoked before it is observed to occur (see Nixon and Davis, 1991, for a detailed discussion of the problems posed by coding polymorphisms as missing entries).

Acknowledgments

We thank Jim Carpenter, Joel Cracraft, Jorge Crisci, Steve Farris, Darrel Frost, Pablo Goloboff, Chris Humphries, Diana Lipscomb, Mary Mickevich, Kevin Nixon, Quentin Wheeler and Ward Wheeler for their comments on a draft of the manuscript.

REFERENCES

Farris, J. S. 1988. Hennig86, version 1.5. Computer program and documentation. Port Jefferson Station, New York.

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

Swofford, D. 1990. PAUP version 3.0. Illinois Natural History Survey, Illinois.