Cladistics **15**, 363–371 (1999)
Article ID clad.1999.0114, available online at http://www.idealibrary.com on IDEAL®

Character Coding and Inapplicable Data

Ellen E. Strong[1] **and Diana Lipscomb**

Department of Biological Sciences, George Washington University, Washington, DC 20052

Accepted May 15, 1999

Inapplicable character states occur when character complexes are absent or reduced in some of the taxa. Several approaches have been proposed for representing such states in a character matrix so that the inapplicable condition has no effect on the placement of taxa and/or the applicable states are independent and not redundant. Here we examine each of these approaches and demonstrate that all have shortcomings. Coding inapplicables as "?" (reductive coding), although flawed, is currently the best way to analyze data sets that contain inapplicable character states.

INTRODUCTION

To construct trees, systematists translate their observations into characters and states and create a matrix in which each taxon is scored (or coded) for all characters. Not all characters, however, are relevant for all taxa—and it is not always clear how to code these features in a matrix. Inapplicable characters occur when a complex feature is absent or reduced in some taxa. Because complex features are often translated into multiple characters, it is not clear how to code the taxa in which the feature is missing. For example, the salivary glands of mollusks consist of several different characters, yet some mollusks lack salivary glands altogether. Thus, salivary gland characters should be uninformative in determining the placement of these animals, but may provide significant information for classifying those with salivary glands. How can such characters be coded so that they do not affect the placement of taxa that lack them but still provide information about the placement of those which do have the character? Furthermore, will the coding method violate fundamental properties of cladistic characters (such as non-redundancy, independence, etc.)? As we shall show, none of the existing methods is able to satisfy all of these criteria, and even the best method is not implemented satisfactorily in most computer programs.

[1]To whom correspondence should be addressed. E-mail: eestrong@gwu.edu.

ALTERNATIVE CODING METHODS: DESCRIPTION AND DEFICIENCIES

Reductive Coding: Using "?"

In many papers, inapplicable characters are denoted with a "?" (question mark) in a character matrix (reductive coding, Wilkinson, 1995a). Although it seems a reasonable choice, most existing computer programs do not treat "?" codes as neutral placeholders. Instead, "?" is interpreted as being one of the existing states (in other words, they are treated as missing rather than

0748-3007/99 $30.00

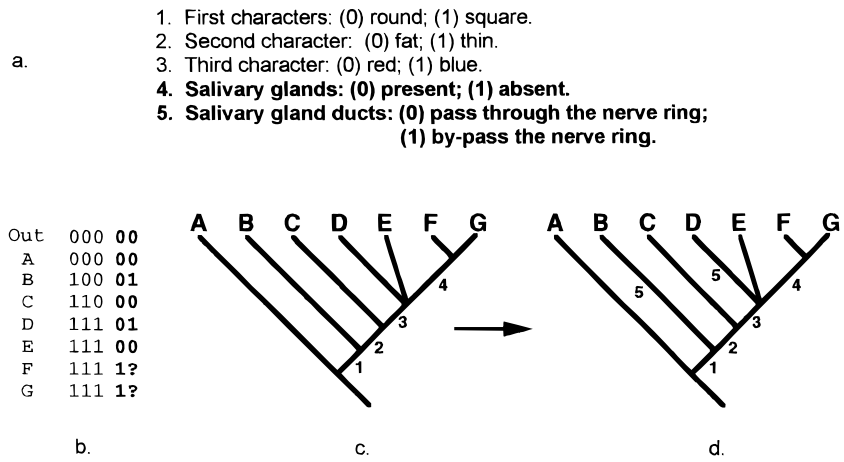

FIG. 1. The characters (a) are reductively coded (b) so that inapplicable characters are denoted with "?" (question mark). Analysis of the characters that are applicable for all of the taxa (characters 1–4) gives the tree in c. If character 5 is added but not allowed to affect the placement of taxa F and G (for which it is inapplicable), the tree in d is obtained.

inapplicable) and this may affect the placement of the taxa for which it is inapplicable. As a matter of bookkeeping, some prefer to use "?" for missing characters and a "-" (dash) for inapplicable characters, but both codes are treated the same by existing computer programs.

Consider the data set shown in Fig. 1 for a hypothetical group of mollusks. Character 5 is not applicable for the taxa that lack salivary glands, and taxa F and G which lack salivary glands are coded "?" for character 5 in the data matrix (Fig. 1b). Using just those characters that are applicable for all taxa (characters 1–4), a single cladogram is obtained (Fig. 1c). Adding character 5 to the analysis should have no effect of the placement of taxa F and G because it is not applicable to either of these taxa (Fig. 1d).

When this data set is analyzed using the program NONA (Goloboff, 1998), this single expected tree is obtained. Two trees are found when the data set is run through Hennig86 (Farris, 1988) or PAUP* (Swofford, 1998). In these two programs, inapplicable characters are treated as missing characters and "?" is presumed to be either state "0" or state "1" (or sometimes a third unobserved state, but see discussion in Platnick *et al.*, 1991). Therefore, two trees are obtained. One is the expected tree (Fig. 2a); the other (Fig. 2b) results because "?" in F and G is treated as if it is state 1. The second tree (Fig. 2b) is undesirable because the inapplicable character is determining the placement of taxa it should not affect.

The tree in Fig. 2b is a semi-strictly supported tree (Nixon and Carpenter, 1996) because it places taxa on the basis of how "?" or homoplasy (Wilkinson, 1995a) might be resolved, not on the basis of observed character evidence. Because NONA does not allow semi-strict trees, in this case it returns just the expected answer.

The program PHYLIP (Felsenstein, 1995) performed poorly because, in addition to treating all "?" as missing data, polytomies are arbitrarily resolved even when no data support the resolved branches (using the PENNY and MIX routines). For data set 1, therefore, three trees are obtained, none of which is the expected topology (Fig. 3). Given these problems, PHYLIP should not be used when data sets contain inapplicable or missing states.

The analysis of inapplicable characters has a second undesirable consequence: when the resolution of "?"

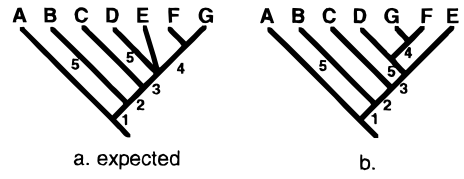

FIG. 2. Analysis of the data in Fig. 1b with Hennig86 and PAUP* gives two trees: (a) the expected tree based on not allowing the inapplicable character to affect the placement of taxa F and G and (b) the tree which results from the programs treating "?" code as a missing state, rather than inapplicable.

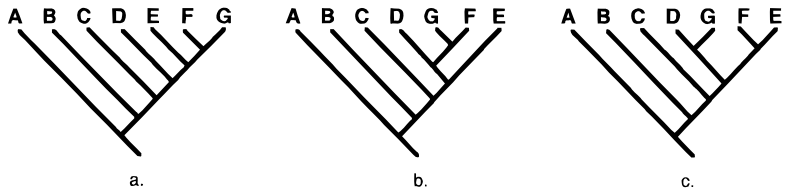

FIG. 3. Analysis of the data in Fig. 1b with PHYLIP gives three trees, none of which is the expected tree.

into a state affects the global optimization of the character at the expense of the local optimization of the clades for which the character is applicable, correct trees may be rejected. This was first discussed by Maddison (1993), but, because not all possible optimizations were discussed in his paper, the issue needs reexamination. The data set in Fig. 4a gives a tree (Fig. 4b) similar to Maddison's Fig. 1 (1993).

An additional character, which is inapplicable for taxa E–J, is added. The states of this new character label the terminal branches in which they occur in Fig. 5. The polytomy uniting taxa A–D can be resolved by this character. Realizing that the character is inapplicable for taxa E–J and, hence, should not be optimized onto the nodes leading to them, the three possible resolutions of the A–D clade are shown in Fig. 6 (contrast to Fig. 2 in Maddison, 1993).

Using NONA, Hennig86, or PAUP*, only the tree in Fig. 6a is obtained when the inapplicable characters are scored as "?." If the character was applicable to all the taxa, this would be correct because this tree reflects the character's global optimization. However, because the character is not globally applicable, the other two optimizations should be considered equally valid.

Given the computational and optimization problems resulting from coding inapplicable characters as "?," it is not surprising that alternative ways of treating inapplicables have been sought. These alternatives are described below.

Composite Coding

In composite coding, a character complex is coded as a single, large multistate character (e.g., Maddison, 1993). In the example in Fig. 1, characters 4 and 5 would be combined into a single character if composite coding is used (Figs. 7a and 7b). The trees obtained when the character is analyzed unordered with NONA, PAUP*, and Hennig86 are shown in Fig. 8.

Even though no inapplicable cells remain in the data matrix, the results suggest the existence of some feature that unites taxa D, F, and G as a clade. This evidence is the alternative orderings of the new multistate character: when "absence" is considered derived from "by-pass the nerve ring," or when "by-pass the nerve ring"

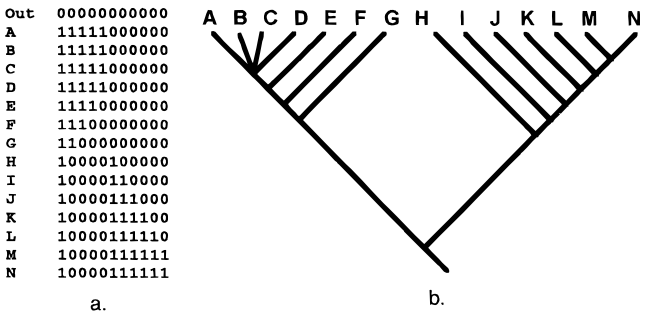

FIG. 4. A data set (a) and the resulting tree (b) designed to give the same result as reported by Maddison (1993).

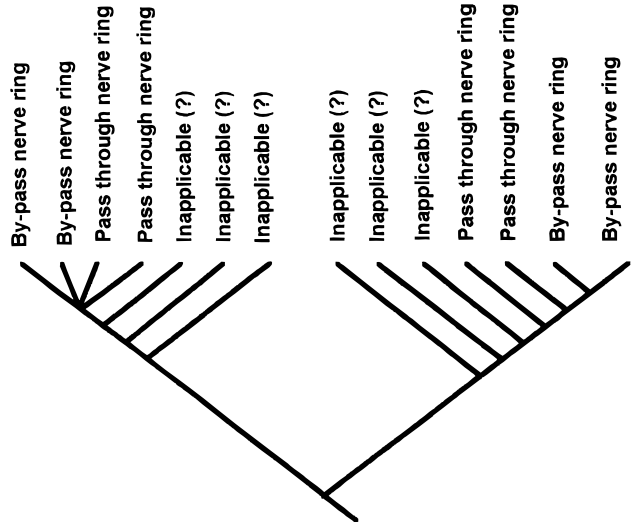

FIG. 5. A new character, which is inapplicable for taxa E–J, is added to the tree shown in Fig. 4.

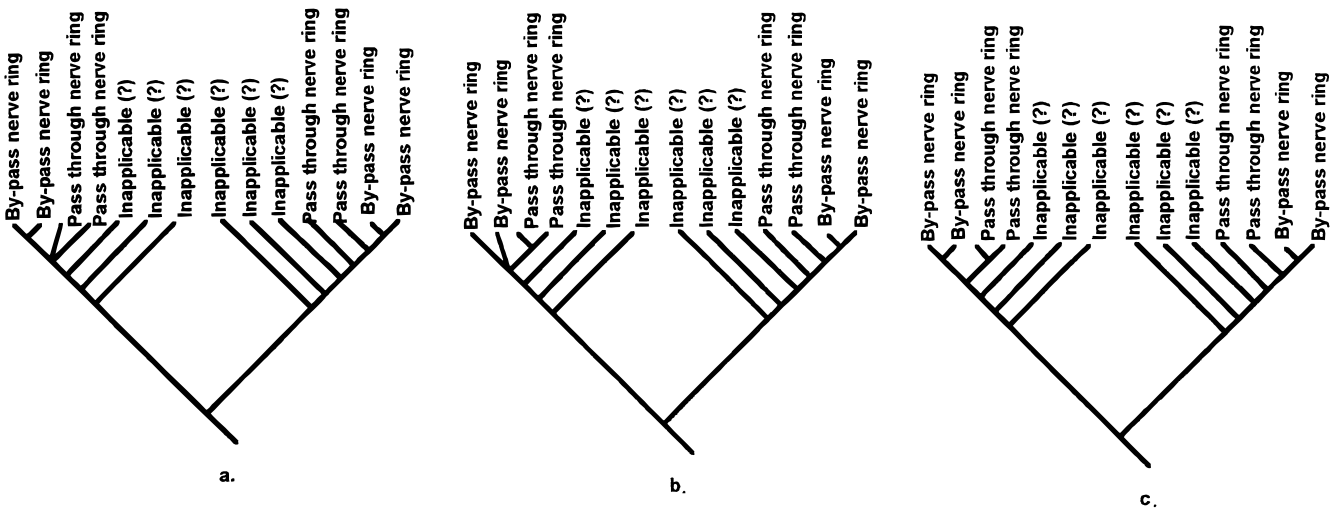

FIG. 6. Choice among alternative resolutions of the polytomy in the trees in Figs. 4 and 5 based on treating the character as inapplicable for taxa E–J. Computer programs such as NONA, Hennig86, and PAUP* treat such characters as missing in E–J, not inapplicable, and so only return the tree in (a). Maddison (1993: p. 578) demonstrated this with a similar example, but incorrectly resolved the remaining polytomy and failed to find the tree equivalent to (c).

is considered derived from "absent," D, F, and G are united. In other words, placement of F and G is determined by the condition of the salivary ducts—a condition which should be inapplicable to them. Thus, coding absence in a multistate character may render inapplicable data informative in determining the phylogenetic relationships when it should not.

Composite coding may succeed in recovering the expected tree, but only when absence is primitive and there are no secondary losses. Under these circumstances, the fact that absence is coded in a multistate character will have no misleading effect on tree construction. However, this clearly requires prior knowledge of the phylogeny. Maddison (1993) and others (e.g., Hawkins *et al.*, 1997) that have relied on this "red tail/blue tail" example have failed to recognize this drawback of composite coding because the example used to demonstrate the utility of the method only considered homoplastic gains and not secondary reversals.

It has also been suggested that partial ordering of Sankoff (composite) characters will cause reductive and composite characters to be analytically equivalent (Wilkinson, 1995b). In this case, partial ordering by the method of intermediates (Mabee and Humphries, 1993) yields a character state tree with each state separated by one step, i.e., unordered. As demonstrated above, the unordered composite does not yield the same result as that obtained from reductive coding. Thus, composite and reductive characters will not be analytically equivalent in all cases.

1. First characters: (0) round; (1) square.
2. Second character: (0) fat; (1) thin.
3. Third character: (0) red; (1) blue.
4. **Salivary glands: (0) present and pass through the nerve ring; (1) present and by-pass the nerve ring; (2) absent.**

a.

Out	000	**0**
A	000	**0**
B	100	**1**
C	110	**0**
D	111	**1**
E	111	**0**
F	111	**2**
G	111	**2**

b.

FIG. 7. The data set shown in Fig. 1 is rewritten as a composite character (a) and the complex character is coded as a single large multistate character (b).

Nonadditive Binary Coding

In non-additive binary coding, every condition (or state) is treated as a separate present/absent character (Pleijel, 1995). When the data set presented in Fig. 1 is converted to this coding, three binary characters replace characters 4 and 5 (Fig. 9). Analysis using NONA, Hennig86, or PAUP* results in two trees, neither of which is the expected topology (Fig. 10a and 10b).

When characters are coded non-additively, spurious clades result because some information has effectively

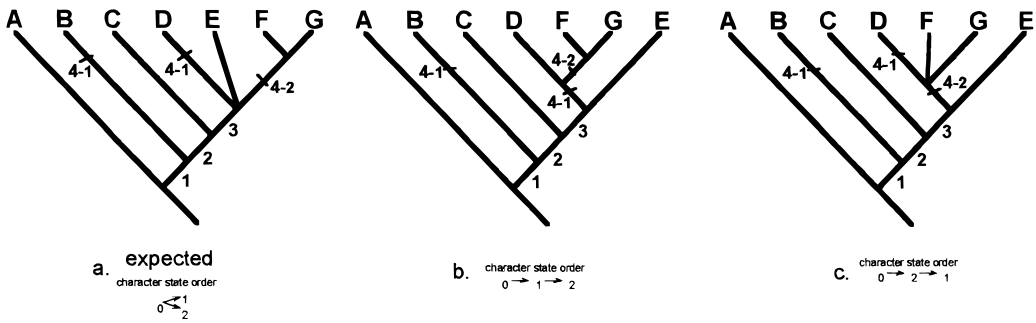

FIG. 8. The data set in Fig. 7 gives three trees when analyzed unordered. Trees b and c result from their state "absent" being alternatively coded as intermediate between "ducts pass through the nerve ring" and "ducts by-pass the nerve ring," and derived from "ducts by-pass the nerve ring," the inapplicable character therefore has an effect on the placement of taxa F and G.

been scored twice. For example, taxa B and D are united (Fig. 10b) because they appear to share two characters: salivary gland ducts that pass through the nerve ring are absent (character 5, state 1) and salivary gland ducts by-pass the nerve ring (character 6, state 1). This is a single observation, not two separate synapomorphies, and non-additive binary coding has made it redundant.

Nonadditive binary coding may also treat non-homologous absence as if it is homologous. Taxa B + D are grouped with F + G in the first tree (Fig. 10a) and D is grouped with F + G in the second tree (Fig. 10b) because they lack salivary gland ducts passing through the nerve ring (character 5, state 1). In this case, state 1 is not the same in B and D as it is in F and G. B and D are coded as state 1 for this feature because their ducts by-pass rather than go through the nerve ring, whereas F and G are coded as state 1 because they lack salivary glands entirely. In short, when nonadditive binary coding is used to eliminate inapplicable codes, the connection between evidence and explanation is lost.

1. First characters: (0) round; (1) square.
2. Second character: (0) fat; (1) thin.
3. Third character: (0) red; (1) blue.
4. **Salivary glands: (0) present; (1) absent.**
5. **Salivary gland ducts pass through the nerve ring: (0) present; (1) absent.**
6. **Salivary gland ducts by- pass the nerve ring: (0) absent; (1) present.**

a.

Out	000	**000**
A	000	**000**
B	100	**011**
C	110	**000**
D	111	**011**
E	111	**000**
F	111	**110**
G	111	**110**

b.

FIG. 9. The data set in Fig. 1 is converted to nonadditive binary coding in which every state is treated as a separate presence/absence character.

Absence Coding

This method divides complex features into multiple characters as in reductive coding, but the inapplicable states are coded as a separate state (e.g., absent) rather than as "?." The absence coding data set is shown in Fig. 11a (for the example presented in Fig. 1). Unordered analysis using NONA, Hennig86, and PAUP* results in two topologies (Figs. 11b and 11c), only one of which is the expected tree.

In the incorrect tree (Fig. 11c), taxa F, G, and D are united, not by a single "absence of salivary glands," but because the non-homologous absence of salivary glands and "absence of ducts passing through the nerve ring" in taxa that have salivary glands have been treated as homologous by this type of coding. Once

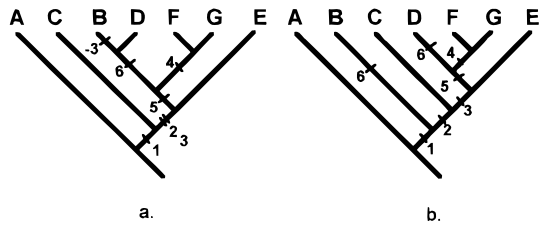

FIG. 10. Analysis of the non-additive binary coded data from Fig. 9 gives two trees. Tree (a) groups F and G with B and D and Tree (b) groups F and G with D because "absence of salivary glands" and "absence of salivary gland ducts by passing through the nerve ring" are incorrectly treated as homologous. Tree (a) also shows that some conditions are coded redundantly: B and D are united because they have "ducts that by-pass the nerve ring" (character 6) and "absence of ducts that pass through the nerve ring." This single observation is a homoplasy, but, because it is counted twice, B with D are grouped together.

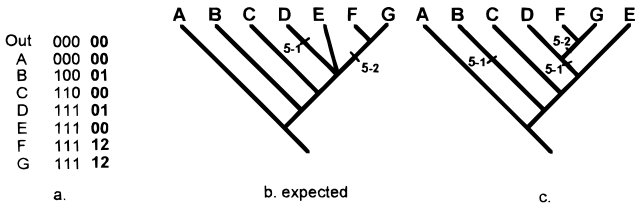

FIG. 11. The data set in Fig. 1 is converted to absence coding (a) in which inapplicable characters are coded as "absent." Analysis of the absence-coded data gives two trees (b, c). Tree (c) groups F and G with B and D because "absence of salivary glands" is incorrectly treated as homologous to and derived from "absence of salivary gland ducts by passing through the nerve ring."

again, by transforming inapplicables into a state of a multistate character, they become informative in determining the relationship of F and G to other taxa.

DISCUSSION

Character coding is the link between observation and explanation. Thus, the ability, or inability, of alternative coding methods to reflect the evidential significance of our observations should be the primary concern in considering alternative methods of coding. Therefore, character coding should result in states that are homologous, independent, and nonredundant (Pimentel and Riggins, 1997).

Wilkinson (1995b) distinguished between biological and logical independence of characters. Biologically dependent characters are those that co-vary because they are caused by the same single biological cause (e.g., features under the control of a single pleiotropic gene). Recognizing biologically dependent characters is of primary importance in cladistic analysis because it relates directly to the concepts of parsimony and corroboration. The requirement that each character, indeed each character state, brings a separate piece of evidence to bear on a hypothesis reflects the fact that the cladogram of choice, the most explanatory hypothesis, is the one that minimizes ad hoc dismissals of evidence. By extension, the severity with which a cladistic hypothesis is tested, and thus the degree to which a hypothesis is corroborated, is directly related to the number of independent falsifiers that are brought to bear on an hypothesis (Farris, 1983).

Logically dependent characters occur when the way in which a taxon is coded for one character partially affects how it will be coded for other characters. This is of particular interest here because the complex of characters that result from inapplicable codes is usually logically dependent. For example, coding the path salivary gland ducts take is dependent, to an extent, on whether or not salivary glands are present. Therefore, "salivary glad ducts: by-pass the nerve ring; pass through the nerve ring" is logically dependent on the character "salivary glands: present; absent." Such logically dependent characters may or may not provide redundant evidence for a clade, depending on the way in which they are coded. Logical dependence of characters does not result in overweighting evidence on a single branch when one character defines a larger clade, while another character logically dependent upon it defines one of its subclades. On the other hand, if states of separate but logically dependent characters are essentially the same observation, these states must not redundantly overweight a single clade of the tree.

When the criteria of homology, biological independence, and non-redundancy are applied to the alternative coding methods for inapplicable characters, some methods are better than others.

Absence coding may fail to meet the criteria of independence and non-redundancy. The absence of the complex character is coded multiple times in the data matrix (see Fig. 11). Because this absence has only one biological cause, this violates the requirement of biological independence of characters. Further, if absence is a synapomorphy, coding it multiple times is redundant and will suggest more evidence for a clade than actually exists.

Non-additive binary coding treats each state as a separate character and gives each of these new characters the states "present" and "absent" (see Fig. 10). Such coding may violate the criterion of homology because absence often applies to more than one condition (absence is not homologous in all of the taxa so coded). If presence is plesiomorphic, taxa scored "absent" may be united by this non-homology (see Fig. 10). Further, absence of the character complex is redundantly coded because taxa that lack the complex are coded "absent" for the character complex, but "absent" again for every character for the different forms of the character complex. Finally, non-additive binary coding

denies homology and the hierarchical relationships between states. The result are cladograms and character interpretations that are absurd and inaccurate representations of our observations.

Both reductive and composite coding avoid redundancy and character dependence. However, analysis of data coded either way treats the inapplicable state as applicable and homologous to the truly applicable states. This problem can often be circumvented for reductive coding (but not composite coding) by using NONA's option for eliminating semi-strict branches such as those defined by ambiguous resolutions of "?." This does not solve the optimization difficulty described in the examples illustrated in Figs. 4–6, but balanced against the problems imposed by the other methods, reductive coding analyzed using the program NONA appears to be the best choice. Recently, Hawkins *et al.* (1997) made the same conclusion but for different reasons. We believe their reasons were flawed and, because they overlooked some of the problems with reductive coding, following their guidelines could result in the errors described in this paper.

Hawkins *et al.* (1997) concluded that character and character-state definition in general, and coding inapplicables in particular, could be accomplished by adhering to the strict rule that characters are "conditional phrases" and that states are the alternative forms of the condition. They concluded that reductive coding is to be preferred because it is the only method that fits this rule. In Hawkins *et al.'s* example (which is modified from Maddison 1993), animals are observed with red tails and blue tails, or lacking tails. Observing the similarity of the tails induces an hypothesis of homology and this leads to the formation of the conditional phrase: "Red tails and blues tails are homologous as ____." According to Hawkins *et al.*, the only logical phrase to fill in this blank is "as tail color" and to code those animals that lack tails with "?." They also argue (Hawkins *et al.*, 1997: p. 5) that when some taxa have a character and others do not, a presence / absence character is required. Thus, Hawkins *et al.* conclude that reductive coding is the only way to code inapplicable character states because it is the only method that is logical given the theoretical basis of primary homology assessment.

There are three problems with Hawkins *et al.'s* line of reasoning:

First, it is not clear that the conditional phrases can be chosen objectively (i.e., there may be more than one way to "fill in the blank" in the conditional phrase) nor is it clear how it can be applied to true presence / absence characters (e.g., what is the conditional phrase when some taxa do not have the feature at all?).

Second, the reductive method may fail to result in trees that reflect our primary homology assessment, making it difficult to justify the method on those grounds. Consider the example in Fig. 12 (adapted from Hawkins *et al.*, 1997, Table 1). Analysis of these data yields three trees (see also Hawkins *et al.* 1997, Fig. 1). In the tree in Fig. 12d, the node that unites E + F + C + D is supported by the presence of a tail, but no color character state optimizes to that node. As argued above, these optimizations are equally parsimonious and, thus, equally valid. However, if the goal of cladistics is to reflect our observations and the resulting hypotheses of homology, reductive coding may produce topologies that accomplish neither goal. Through reductive coding, logically dependent characters can be decoupled during optimization. The result can be seen in the tree in Fig. 12d: the node uniting E + F + C + D is supported by the observation of unmodified tails. However, we have observed blue tails or red tails. This node suggests the presence of tails with no color or a third unobserved color. This clearly does not conform to our observations nor our hypotheses of homology.

Third, Hawkins *et al.* (1997) failed to convincingly demonstrate that reductive coding is unique in its ability to convey homology information. In their example,

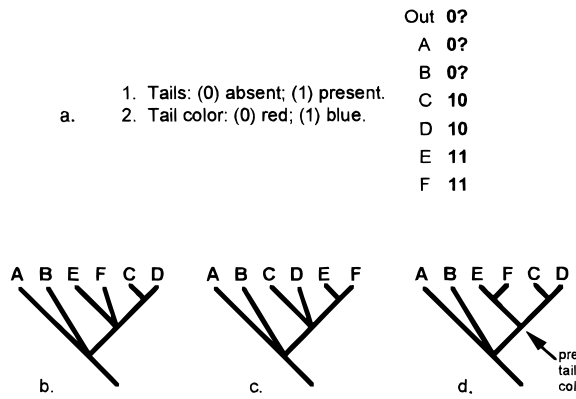

FIG. 12. Reductive coding may result in illogical character optimizations. In this example, presence of a structure optimizes on a tree prior to its color even though it must have some color.

the compositely coded character is analyzed as unordered. The resulting tree topologies and corresponding decrease in resolution lead them to conclude that this character is incapable of conveying information about homology. However, in their example, an ordered multistate character is entirely capable of conveying the information of the original hypothesis of homology. In fact, various authors have stated that all characters should comprise a series of ordered states (Pimentel and Riggins, 1988; Pogue and Mickevich, 1990; Lipscomb, 1992) to convey information about homology. Ordering their composite multistate character would have left them with no criterion for eliminating this method.

Thus, no coding method bears a majority of advantages. In reductive coding, redundant absence states are replaced with question marks, but the expressions relative to presence are only partially independent: presence in one character requires presence in the other but cannot predict which state will appear in the other. In addition, some topologies may not conform to our observations nor contain the information of our putative homology statements.

CONCLUSIONS

In light of the preceding examples, it is clear that reductive coding analyzed with NONA results in tree topologies that best reflect the information content of our observations. However, the fact remains that reductive coding is currently susceptible to errors of optimization that may lead to inappropriate homology statements. Because inapplicables are treated as missing, a globally parsimonious tree may contain local solutions (regions of applicables) that are suboptimal. This will lead to results that are incomplete. It is important to realize that this will occur only under restrictive conditions: homoplastic gains must be separated by intervening primitive absence, and one or both regions of applicables must be supported by only ambiguous change. In addition, this is not only an optimization problem, but also an homology problem. The fact that there are homoplastic gains separated by regions of inapplicables should raise a red flag from the point of view of homology. These clades and the characters supporting them, indeed any clade supported exclusively by homoplasies, should be reexamined and considered carefully; the characters supporting these nodes may require character redefinition, thereby eliminating the problem. If no criterion can be found to redefine these characters, these clades will require optimization by hand to ensure that all local optima are discovered.

Computational drawbacks aside, it is clear that inapplicables, rather than reflecting a character coding problem, in fact are an attempt to represent the hierarchical structure of our data: they are simply placeholders required by the linear restrictions imposed by a square data matrix. Misguided transformations of our data succeed only in obscuring, rather than revealing, hierarchical relationships between characters and consequently between taxa.

ACKNOWLEDGMENT

Support from NSF Grant DEB-9712463 to Diana Lipscomb is gratefully acknowledged.

REFERENCES

Farris, J. S. (1983). The logical basis of phylogenetic analysis. *In* "Advances in Cladistics, Vol. 2" (N. Platnick and V. Funk, Eds.), pp. 7–36. Columbia Univ. Press, New York.

Farris J. S. (1988). Hennig86. Port Jefferson, New York.

Felsenstein, J. (1995). PHYLIP (Phylogeny Inference Package) Version 3.57c. Univ. Washington, Seattle, WA.

Goloboff, P. (1998). Nona. Fundacion e Instituto Miguel Lillo, Miguel Lillo 205, 4000 S. M. de Tucumon, Argentina.

Hawkins, J. A., Hughes, C. E., and Scotland, R. W. (1997). Primary homology assessment, characters and character states. *Cladistics* **13,** 275–283.

Lipscomb, D. L. (1992). Parsimony, homology and the analysis of multistate characters. *Cladistics* **8,** 45–65.

Mabee and Humphries (1993). Coding polymorphic data: Examples from allozymes and ontogeny. *Syst. Biol.* **42,** 166–181.

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

Nixon, K. C., and Carpenter, J. M. (1996). On consensus, collapsibility, and clade concordance. *Cladistics* **12,** 305–321.

Pimentel, R. A., and Riggins, R. (1987). The nature of cladistic data. *Cladistics* **3,** 201–209.

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

Pogue, M. G., and Mickevich, M. F. (1990). Character definitions and character state delineation: The bete noire of phylogenetic inference. *Cladistics* **6,** 319–361.

Pleijel, P. (1995). On character coding for phylogeny reconstruction. *Cladistics* **11,** 309–315.

Simmons, N. B. (1993). The importance of methods: Archontan phylogeny and cladistic analysis of morphological data. *In* Primates and Their Relatives in Phylogenetic Perspective (R. D. F. MacPhee, Ed.), pp. 1–61. Plenum Press, New York.

Swofford, D. (1998). PAUP*. Sinauer Associates, Sunderland, MA.

Wilkinson, M. (1995a). A comparison of two methods of character construction. *Cladistics* **11,** 297–308.

Wilkinson, M. (1995b). Arbitrary resolutions, missing entries, and the problem of zero-length branches in parsimony analysis. *Syst. Biol.* **44,** 108–111.