Cladistics (1994) 10:415–423

PROBLEMS WITH ZERO-LENGTH BRANCHES

Jonathan Coddington and Nikolaj Scharff

Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA and Department of Entomology, Zoological Museum, Universitetsparken 15, 2100 Copenhagen, Denmark

Received for publication 20 January 1994; accepted 1 May 1995

In this note we call attention to a possibly common situation in which frequently used phylogenetic analysis and diagnosis programs may report cladograms that do not have support at every node from the data matrix analyzed to produce them. The problem involves branches that may or may not be interpreted as having zero length, the ambiguity arising from equally applicable but mutually exclusive alternative optimizations.

Platnick et al. (1991) and Wilkinson (1995) commented on manifestations of the problem that involved missing entries and "arbitrary resolutions", respectively. Wilkinson (1995) emphasized that programs can report topologies including zero-length branches even when no missing entries are present. However, Wilkinson posed the problem abstractly, a single character on one tree, addressed only one horn of the dilemma (branches ambiguously of zero or greater length were always assigned zero length), considered solutions only within the philosophical framework offered by PAUP, and offered no advice regarding what practicing systematists could do to detect the problem. In addition, the treatment of the zero-length branch problem by programmers is already more diverse than reported by Wilkinson (1995), and users of the most common packages should be aware of the different interpretations adopted by programmers.

Wilkinson (1995) offered some algorithmic suggestions for a solution, but his solution, although helpful, strikes us as less than ideal. We also focus here on what practicing systematists can do now to detect and avoid the problem. Finally, the solution is not necessarily as simple as eliminating "arbitrary resolutions". Most importantly, Wilkinson (1995) did not consider what happens if branches of ambiguous length are permitted to be non-zero rather than zero-length. A more fundamental issue is the meaning of "acceptable support" for phylogenetic hypotheses, and therefore which hypotheses are worth considering, as Platnick et al. (1991) discussed. That debate is just beginning, but should involve the user community in addition to programmers and methodologists.

Table 1 conveys the problem of zero-length branches more concretely than the example provided by Wilkinson (1995, Fig. 1). Taxa D and E are identical, as are F and G, but this does not affect the point and keeps the matrix simple. **Five** supports the ingroup BCDEFG, **Four** supports CDEFG, **Two** and **Three** support DEFG. **One**, however, is homoplasious, and permits two parsimonious expla-

0748-3007/94/040415+09/$12.00/0

nations: a gain and a loss or a convergence, both of which are two steps. The former is accelerated transformation or "ACCTRAN", which favors homology (or reversals), and the latter is delayed transformation, or "DELTRAN", which favors convergence.

If the matrix in Table 1 is analyzed with Hennig86 (Farris, 1988), PAUP 3.1.1 (Swofford, 1993), or NONA (Goloboff, 1993a) with the ambiguity option set to "amb=" (ambiguous support is considered), all programs report the same three cladograms of six steps (Trees 1–3, Fig. 1). Trees 1 and 2 contain trichotomies, but Tree 3 is fully resolved, and thus might be preferred on the basis of information content because it reports a monophyletic group absent in either of the other two trees. NONA under "amb-" (ambiguous support ignored) reports Trees 1, 2, and 4.

However, the data of Table 1 cannot possibly support Tree 3. The problem lies with **One**. Under DELTRAN optimization, group FG is supported by a parallel gain of **One** (Fig. 2a). Under ACCTRAN optimization group DE is supported by secondary loss of **One** (Fig. 2b). But **One** requires only two steps. Change at taxon

Table 1
Five characters scored for seven taxa

Taxon	One	Two	Three	Four	Five
A	0	0	0	0	0
B	0	0	0	0	1
C	1	0	0	1	1
D	0	1	1	1	1
E	0	1	1	1	1
F	1	1	1	1	1
G	1	1	1	1	1

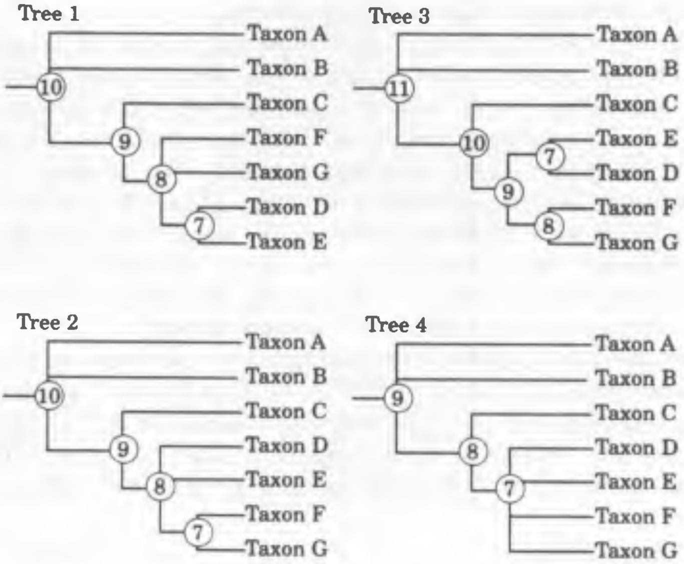

Fig. 1. For the data of Table 1, PAUP, Hennig86, and NONA under "amb=" find trees 1–3. NONA under "amb–" finds trees 1, 2, and 4.

C or node 10 accounts for one step, which leaves only one additional step to support group DE or FG, but not both. Under ACCTRAN the only correct topology is Tree 1, and under DELTRAN, the only correct topology is Tree 2. As working systematists, we only accept cladograms with character change on all branches as phylogenetic hypotheses. All monophyletic groups must have synapomorphies; all branches must have support simultaneously. Tree 3 fails this criterion, even though it is fully resolved and therefore of great interest.

We agree with Wilkinson (1995, and authors cited therein) that reporting Tree 3 provides little or no benefit to users. One could argue that Tree 3 is not rejected by the data, but neither is it supported. While some users may wish to consider trees "not rejected" by data (another interpretation of "acceptable support"), presumably most want to consider only completely supported topologies, or, at least, want the option of filtering output to exclude the former. In our view, Trees 1 and 2 are the only most parsimonious cladograms for these data. Swofford and Begle (1993: 45–47, and figure on p. 48) explained their approach to the problem of zero-length branches, although they did not explicitly mention the case where alternative optimizations of the same character provided the only support for adjacent branches as in Table 1. They outlined three possible rules to deal with zero-length branches (p. 46):

1. Collapse an interior branch if the *minimum* possible length of the branch is zero. That is, if there exists at least one MPR [most parsimonious reconstruction] for every character such that no length is assigned to the branch, the branch is collapsed.

2. Adopt an ancillary criterion for choosing one MPR from the full set of MPRs for each character (or choose one arbitrarily). If no length is assigned to the branch for all characters, then the branch is collapsed.

3. Collapse an interior branch if the *maximum* possible length of the branch is zero. That is, if all MPRs assign zero length to the branch for every character, then the branch is collapsed.

NONA implements Rule 1 to yield Trees 1, 2, and 4 (Fig. 1) under "amb-". The DEFG polytomy of Tree 4 results because both branches could be zero. Reporting

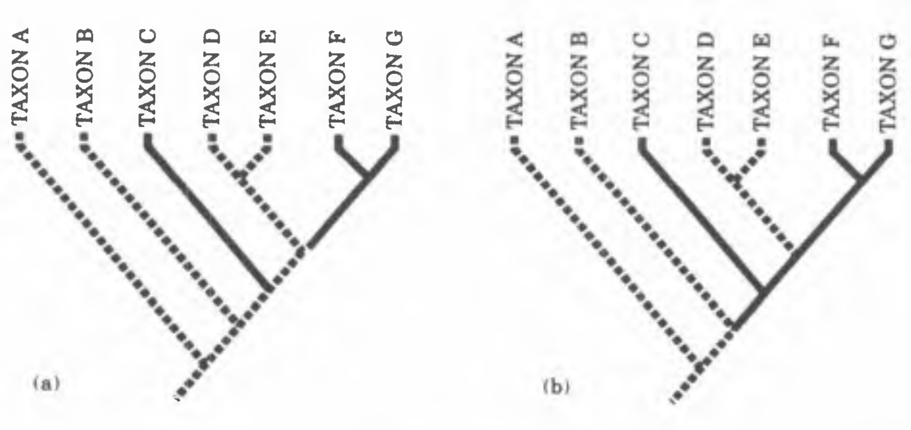

Fig. 2. The two possible ways to optimize character One on Tree 3: (a) delayed transformation, or "DELTRAN"; (b) accelerated transformation, or "ACCTRAN". Both trees are treelength 6. □=0; ■=1.

Tree 4 requires an additional assumption about tree length (see below). Tree 4 omits the potential support of either DE or FG, but correctly conveys the impossibility of support for both. Rule 2 could result in either Tree 1 or 2, for example by using only accelerated or only delayed transformation when collapsing branches, but presumably not both and not Tree 3. Rule 3 is used by PAUP and Hennig86, which only collapse a branch if it is zero length under all reconstructions. As Farris (1988: section 13) put it, "Calculated trees are compressed to assure that they show no arbitrary resolution. On a compressed tree a branch is resolved only if some active character of non-zero weight may show a change in that branch in some parsimonious synapomorphy scheme". Goloboff (1993a and pers. comm.) made Rule 1 the NONA default: "branches that are not supported under any possible optimization are shown as collapsed (this is a stricter interpretation than Hennig86's or PAUP's, which keep a branch if it is supported under some optimizations) . . ." and ". . . if any possible states are shared between ancestor and descendant node, the branch is considered unsupported". The default "amb-" option in NONA collapses a branch if there exists any optimization at all under which it has zero length. "Amb=" collapses more branches than Rule 3. "Amb=" retains a branch if ". . . some states occurying [sic] in the set of possible states for the descendant node are absent in its ancestor (or vice versa) . . ." In other words, "amb=" looks only at whether the ancestral and descendant state sets are identical, collapsing branches if they are. Given identical state sets, it ignores those cases in which different states could be assigned. In contrast, Rule 3 retains a branch if different states could possibly be assigned to the ancestral and descendant nodes (Goloboff, pers. comm.). Rule 3 and "amb=" clearly admit Trees 1 and 2, but why Tree 3?

Tree 3 results from focussing on branches rather than whole trees or cladograms. Nodes 7 and 8 are supported by the data under mutually exclusive optimizations; they cannot be supported simultaneously. Wilkinson (1995) pointed out the same problem. Tree 3 (and Tree 4, see below) are different things from most parsimonious cladograms, perhaps useful if identified as such, but misleading if included indistinguishably among topologies of minimum length and supportable by data.

Like Tree 3, Tree 4 is also not a most parsimonious cladogram for these data; it is more like a strict consensus of possible optimizations. PAUP and Hennig86 evaluate Tree 4 as 7 steps, one step longer than minimal. As a cladogram, Tree 4 implies that either F and G gained **One** independently or that D and E lost it. Goloboff (1993a) emphasized that collapsed topologies like Tree 4 may be longer than the dichotomous trees underlying them, and warned that, in NONA, reported lengths refer to the underlying dichotomous trees *in memory*, not the displayed (collapsed) topology. NONA thus follows a different convention from PAUP or Hennig86 on the relation between reported lengths and displayed topologies. Wilkinson (1995; other authors cited therein) also noted that wholesale collapsing of potentially zero-length branches could result in globally less parsimonious topologies like Tree 4. Wilkinson's algorithm emends Rule 1 to report Trees 1 and 2 but not 4, and for programs that implement Rule 1 like NONA, Wilkinson's suggestions should be considered (unless someone thinks Tree 4 has "acceptable support").

In our view, if Tree 3 supposes too much from the data, Tree 4 supposes too little. It misses the opportunity to explain either the shared 0s or the shared 1s as synapomorphy, both legitimate interpretations of **One**. In more complex cases, Rule 1 (even as amended by Wilkinson) will discard monophyletic groups supported by less than all alternative optimizations. Branches are suppressed even if support is available. Imagine a sixth character in Table 1 with two 1s for F and G and 0s elsewhere. Now **One** is available to support DE and Tree 3 is legitimate, but Rule 1 reports only Tree 2 (as implemented by NONA under "amb-"). Hence, users are presented with only some of the most parsimonious cladograms for their data as well as a number of diagrams that may be longer if interpreted as cladograms. The latter are easily detected in NONA by explicitly saving the collapsed, displayed topologies to disk (the "ksave" option), reading them back into the tree buffer, and filtering them for minimum length (the "best" option). Goloboff (1993a) and Carpenter (in press) prefer Rule 1 because it only yields trees with unequivocal, unambiguous support at all nodes (yet another point of view on "acceptable support"). Issues of length aside, Rule 1 as implemented in NONA does produce legitimate, conservative, phylogenetic hypotheses. For those interested in most parsimonious topologies with unequivocal, unambiguous support at all nodes, even if less resolved, NONA is currently the only place to find them.

However, Rule 1 does not implement the solution we prefer to the zero-length branch problem because it rejects all branches conceivably of zero-length, not just the topologies that *must* contain such branches. Note that Pee-Wee, NONA's mother program, differs by calculating support of trees as implied weights or fits (Goloboff, 1993b, 1993c), but it adopts the same philosophy of character support and conventions as NONA.

Programmers are obviously aware of the zero-length branch problem in general (e.g. Farris, 1982; Maddison and Maddison, 1992; Swofford and Begle, 1993; Swofford and Maddison, 1987, 1992; Goloboff, 1993b [Pee-Wee documentation serves for NONA as well], Goloboff, 1993d). Farris (1982: 425), for example, pointed out "On occasion . . . this reconstruction method may assign apparent synapomorphies to groups that are in fact unsupported by data . . ." Their algorithms perform as stated.

Having considered the potential benefits to the user of considering Trees 3 and 4, it is fair to consider the costs. For small, clean data sets, the costs are not great if one assumes that character support is verified for every node of all most parsimonious trees and that lengths are checked. The inevitability of a spurious node in Tree 3 will be detected and the tree rejected on grounds of impossibility or parsimony. However, such trees might escape detection if one assumed that the most parsimonious trees reported by software are fully supported by data. Length checks will identify summaries like Tree 4, but the most parsimonious most resolved cladograms discarded by Rule 1 are not recovered. In either case, one can be misled into considering solution sets with cladograms that, in our view, are either unacceptable (contain zero-length branches) or that do not contain all most resolved, most parsimonious trees.

For larger data sets, however, Rules 1 and 3 cause much mischief. Large data sets during initial stages of analysis often result in hundreds or even thousands of trees. Reporting the likes of Tree 3 increases the number of trees reported by as much as one third if there is only one such zero-length branch problem in the data. Rule 1 often results in fewer trees, but omits some most parsimonious cladograms, and it still includes spurious topologies. If so large a fraction of reported trees are spurious in either of the above senses, the goal of identifying acceptable phylogenetic hypotheses to explain the data is impeded. As noted by Wilkinson (1995), consensus procedures based on them will also be affected, as will any other procedure that operates on ensembles of trees.

Large data sets are of special concern because the probability of logically and topologically independent region of cladistic instability increases with size. Regions with multiple solutions that interact in all possible combinations are frequent in large problems. To appreciate the effect, consider a data set we recently analyzed under Rule 3 that had two unstable regions, one with 9 solutions and the other with 43. For the latter, 37 of the possible resolutions had to contain at least one zero-length branch, so that only 54 (=6×9) most parsimonious trees resulted, not 387 (=43×9). The former means less output, less ambiguity, an easier result to comprehend, and a more satisfactory analysis. Rule 1 yielded 21 trees, 19 of which were less resolved and different from the 43 found under Rule 3. While none of these contained zero-length branches, all of them omitted monophyletic groups supportable by data. Rules 1 and 3 will have increasingly pernicious effects as advances in software, hardware, and data collection make large phylogenetic analyses tractable.

We are practicing systematists, not programmers, and our priorities for output may differ. To repeat, "acceptable support" means that cladograms have length and character change is present at all nodes, jointly and simultaneously. At 6 steps, Tree 3 may be parsimonious, but not all nodes are capable of support simultaneously. Tree 4 reports all branches with unambiguous support, but it is not parsimonious. We see great benefit in being able to eliminate both sorts of trees from output, perhaps as a filter applied to trees found under Rule 3. We suggest:

4. Discard all trees that must contain a zero-length branch.

Rule 3 finds all legitimate and some illegitimate most resolved trees. Applied to the results of Rule 3, Rule 4 should filter out illegitimate topologies to give what we think most systematists want: all most parsimonious cladograms supported by data. This rule evaluates the whole tree, not only the branch. For Table 1, it should result in reporting only Trees 1 and 2, not Trees 3 or 4. Rule 4 is not an algorithm (implementation will be complex), but it states clearly what we think a majority of users want.

Pleas from the user community are all very well, but difficulties may arise in implementation even if no controversy attends the suggestion. One difficulty is that an ambiguity may permit many possible optimizations (not just the extremes of accelerated versus delayed transformation). If several such ambiguities exist, the number of combinations to be checked is large. However, the mainframe phylogenetic package PHYSYS, as well as a microcomputer program SHEN (never widely distributed) followed a rule that would have resulted in Trees 1 and 2 but not 3 or 4 (Farris, pers. comm.).

The algorithm proposed by Wilkinson (1995) also ignores this complicating factor. Wilkinson proposed an algorithm based on the minL and maxL routines in PAUP (the "xsteps c" function in Hennig86 and the "minimum" function in NONA are equivalent), which report minimum and maximum lengths of branches. After identifying all branches with minL=0, Wilkinson suggested collapsing them one at a time, while re-evaluating support at all remaining branches at each step to find the set of all trees that lack arbitrary resolutions. This solution improves on Rule 1, already implemented by NONA by eliminating the Tree 4 problem discussed above, although Wilkinson (1995) did not discuss NONA. It finds a set of trees free of zero-length branches, but it will not find the set of maximally resolved trees with support at all nodes. As noted above, if another character unambiguously supports FG (or DE) of Tree 3 a disparity arises. Although minL is still 0, **One** is now available to support DE (or FG) and both groups are simultaneously supportable; Tree 3 is now legitimate. Because Wilkinson's solution picks minL=0 as a criterion while ignoring other optimizations in which minL>0, it would not report Tree 3 in this case. The solution is not as simple as sequentially collapsing branches where minL=0, but rather finding one possible way a potentially zero-length branch may have length while all remaining branches on the tree simultaneously retain length.

At present, therefore, users of the common tree-finding packages cannot exclude trees containing zero-length branches from the results while simultaneously obtaining all of the most parsimonious trees supportable by the data. Most users either rely on tabular diagnostic output to assess their results or import trees into graphical programs like MacClade 3.0 (Maddison and Maddison, 1992) and Clados 1.2 (Nixon, 1992). In PAUP, to get a complete list of potential apomorphies or branch lengths for Tree 3, one must diagnose the tree twice, once under the ACCTRAN assumption and once under DELTRAN. In the former, no mention is made of the branch between nodes 9 and 7 and, in the latter, no mention is made of the branch between nodes 9 and 8 (nodes numbered as in Tree 3, Fig. 1). Omitted nodes have no support. Still, the user must check output to look for *omitted* branches, which can be time-consuming. Hennig86's diagnostic options are more limited and only report ambiguities (listing all possible states at each node). It does ı.ot flag or otherwise make obvious zero-length branches. NONA's output under "amb=" for Tree 3 reports either 0 or 1 changing to 0 for node 7 and either 0 or 1 changing to 1 for node 8. Either style of output flags the potential problem, but does not distinguish it from ambiguities that are not fatal to the tree as a hypothesis. For any of these programs, detecting no support for either DE or FG in any tree, much less dozens or hundreds or trees, would be onerous.

We regard the obligation to inspect visually every tree for zero-length branches as a less desirable alternative than excluding them from the results in the first place. However, MacClade 3.0 does not make the detection of zero-length branches very easy either. Under the "unambiguous" option for "Trace All Changes" MacClade paints the support for either DE or FG as ambiguous, which it certainly is. Under the "almost all possible changes" option, which might be expected to flag instances in which the same change has been mapped twice, the program report change in **One** at both nodes. It is easy to miss that the single step at each node is the same and thus impossible. The "All possible changes" option lists the same change at both nodes, but it does so for all ambiguous changes everywhere on a tree, and clearly can be no reliable guide to zero-length branches. Difficulty of detection is compounded because MacClade declines to optimize characters at a polytomous node. However, zero-length branches can be detected in MacClade by first setting "Resolving Options" to "show all most parsimonious changes . . .", then setting "Trace All Changes Options" to "unambiguous changes" to determine which nodes have ambiguous support. The "All Changes Options" is switched to "Almost all Possible Changes" and all ambiguous branches are noted if the only support derives from the same character, then the "Equivocal Cycling" routine is used with the "Option+R" key combination to cycle through all most parsimonious reconstructions for that character, while the number of steps available to explain changes in that portion of the tree are recorded, thereby making it possible to discern when a change has been counted twice. Nothing changes graphically at a zero-length branch—the display looks the same as at nodes with ample support. The user must ferret out the problem. However, MacClade can be used to locate and verify zero-length branch problems tree by tree for the entire solution set of most parsimonious trees, and its ability to reconstruct all most parsimonious reconstructions is an important check. Maddison and Maddison (1992: 106–110) emphasize that there can be many more optimizations than those found by pure ACCTRAN and DELTRAN procedures. Of course, MacClade never offered in the first place to flag nonsensical trees; it accepts any input tree.

Clados 1.2 also detects zero-length branches and, thus far in our experience, it does so in a reasonably efficient and user-friendly fashion although, as in MacClade, one must still examine every ambiguously supported node on every tree. First, it accepts trees of any topology, and does not decline to optimize characters at polytomies. Second, it provides a reasonably ergonomic and efficient toggle between ACCTRAN and DELTRAN optimization, which speeds the work. Third, the display changes in an obvious manner, either because support (the tick-mark) at the zero-length branch disappears, or because the branch itself collapses when the optimization toggle is invoked (if the DICHOT toggle is set to "0"—do not show unsupported branches). While Clados may lack some features of MacClade, it can be used to detect zero-length branches more easily and efficiently. Note that Clados implements only pure accelerated and delayed transformation; it will not report all most parsimonious reconstructions. Under certain conditions, it may both miss and mistakenly identify zero-length branches.

Thus, we know of no foolproof ways too detect automatically zero length branch problems in output. However, we now routinely use Clados to check all most parsimonious trees suggested to us by software as potential hypotheses. We also check trees in MacClade and indeed prefer it for some aspects of character optimization. We use Hennig86, PAUP, and NONA/Pee-Wee to find and check trees, but all of these programs differ slightly in implicit assumptions, definitions of support, and thus what their output means (e.g. Platnick et al., 1991; Nixon and Davis, 1991). Either more explicit explanation of assumptions and their practical implications, or perhaps trick data sets like Table 1, are necessary to help the user interpret the many "extra" trees permitted by currently implemented rules. For example, in one recent experience, only one of 14 trees found survived the check in Clados for zero-length branches. As any cladist can attest, and any evolutionary biologist should appreciate, there's nothing quite like finding only one most parsimonious tree.

Acknowledgements

We are very grateful for the comments of Jim Carpenter, Steve Farris, Gustavo Hormiga, David Maddison, Kevin Nixon, Rod Page, Joe Slowinski, Dave Swofford, and especially Pablo Goloboff on earlier versions of the manuscript.

REFERENCES

CARPENTER, J. M. (In press). Phylogeny and biogeography of *Polistes*. *In*: E. Turillazzi and M. J. West-Eberhard (eds). Natural History and Evolution of Paper Wasps. Oxford University Press.

FARRIS, J. S. 1982. Simplicity and informativeness in systematics and phylogeny. Syst. Zool. 31: 413–433.

FARRIS, J. S. 1988. Hennig86, ver. 1.5. Microcomputer program available from author, 41 Admiral St., Port Jefferson Station, New York 11776.

GOLOBOFF, P. A. 1993a. NONA. Noname ver. 1 (a bastard son of Pee-Wee) 32-bit version. Program and documentation. Computer program distributed by J. M. Carpenter, Dept. Entomology, American Museum Natural History, New York.

GOLOBOFF, P. A. 1993b. Pee-Wee. (P)arsimony and (I)mplied (W)eights ver. 2.0 (32 bit version). Program and documentation. Computer program distributed by J. M. Carpenter, Dept. Entomology, American Museum Natural History, New York.

GOLOBOFF, P. A. 1993c. Estimating character weights during tree search. Cladistics 9: 83–91.

GOLOBOFF, P. A. 1993d. Character optimization and calculation of tree lengths. Cladistics 9: 433–436.

MADDISON, W. P. AND D. R. MADDISON. 1992. MacClade, analysis of phylogeny and character evolution, ver. 3.0. Sinauer Associates, Inc., Sunderland, MA, pp. 1–398.

NIXON, K. C. 1992. Clados ver. 1.2. Program and documentation. Distributed by author, P.O. Box 270, Trumansburg, NY 14886, pp. 1–42.

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.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic Analysis Using Parsimony, Ver. 3.1. Illinois State Natural History Survey, Champaign, IL.

SWOFFORD, D. L. AND D. P. BEGLE. 1993. User's Manual for PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1. Available from authors, Smithsonian Institution, Washington, DC.

SWOFFORD, D. L. AND W. P. MADDISON. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87: 199–299.

SWOFFORD, D. L. AND W. P. MADDISON. 1992. Parsimony, character-state reconstruction and evolutionary inference. *In*: R. L. Mayden (ed.). Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford Univ. Press, Palo Alto, 186–223.

WILKINSON, M. 1995. Arbitrary resolutions, missing entries and the problem of zero-length branches in parsimony analysis. Syst. Biol. 44: 108–111.