Cladistics (1996) 12:139–145

FORUM

PROBLEMS WITH "SOFT" POLYTOMIES

Jonathan A. Coddington* and Nikolaj Scharff†

**Department of Entomology, National Museum of Natural History NHB 105, Smithsonian Institution, Washington, DC 20560, USA and †Department of Entomology, Zoological Museum, Universitetsparken 15, 2100 Copenhagen, Denmark*

Received for publication 15 February 1996; accepted 17 May 1996

Abstract — The "soft" assumption attributes polytomies to lack of data, not simultaneous cladogenesis (the "hard" assumption). Most systematists prefer the first interpretation, but most parsimony programs implicitly use the second. Results can thus be inconsistent with initial assumptions. Under certain circumstances that seem especially typical for large data sets treating higher taxa, it may be valid to eliminate both compatible and incompatible polytomous trees from consideration. Consistent treatment of soft polytomies can reduce the ambiguity of cladistic solutions and improve the resolution, and testability, of phylogenetic hypotheses.

Introduction

This note focuses on practical implications of the "soft" interpretation of polytomies in primary, as opposed to consensus, cladograms (Maddison, 1989), and especially the implications of that assumption for tree choice under parsimony. Several other criteria for tree choice within the philosophical framework of parsimony exist (Carpenter, 1988; Coddington and Scharff, 1995; Farris, 1969; Michevich, 1982; Maddison, 1993; Platnick et al. 1991; Wilkinson, 1995a, 1995b). For example, character optimization (Lipscomb, 1992; Swofford and Maddison, 1992), is often a more compelling criterion for choice than topological resolution. A fully resolved tree that makes no sense is still nonsensical. In what follows, however, we set aside issues of optimization to focus on resolution.

We are particularly interested in situations where solution sets contain hundreds or thousands of differentially resolved but equally parsimonious trees. Evaluating such results can be especially difficult. In these cases, the less resolved (more polytomous) trees are either compatible with the more resolved trees in the solution set, or they are not. If compatible, at least one of the more resolved trees is a resolution of a polytomy in an otherwise identical tree. If incompatible, non-polytomous regions of the trees are different.

The following data set is a simple example: nine taxa and seven characters. Six trees result (Fig. 1) of which three are "compatible" polytomies (Trees 1–3), one is "incompatible" (Tree 4), and two are fully resolved (Trees 5–6).

0748-3007/ 96/ 020139+07/ $18.00/ 0

A	0000000
B	1000000
C	1100010
D	1101000
E	1101001
F	1110111
G	1110110
H	1110000
I	1110001

This discussion makes several assumptions. Many workers implicitly or explicitly prefer the soft interpretation; all most parsimonious trees are known (i.e. exhaustive or exact solution sets); and "legitimate" or "valid" cladograms by definition have support at all nodes simultaneously. In other words, topologies that must contain unsupportable (zero-length) branches are discarded because they cannot be completely justified with the available data (Wilkinson, 1995a; Coddington and Scharff, 1995). Tree 6 in Fig. 1 contains a zero-length branch (either HI or HICGF) due to the behaviour of the third character. Tree 5 is the only fully resolved, legitimate topology for the data.

The "hard" polytomy interpretation asserts simultaneous cladogenesis. Hard polytomies are real and ineradicable; they posit speciation patterns supposed to be rare in nature. The interpretation rejects the explanation that internodes have gone undetected due to inadequate sampling of data or taxa. Under the hard polytomy interpretation, all polytomous topologies (Trees 1–4) are legitimate phylogenetic hypotheses for the study taxa, regardless of resolution.

In our experience most systematists do not prefer the hard polytomy interpretation for their work, despite the theoretical possibility of simultaneous cladogenesis. The possibility of more data and more taxa always beckons. Experience shows that today's polytomy is often gone tomorrow. Polytomies in primary cladograms tend to occur where data are scanty or ambiguous, which also favors the soft rather than the hard interpretation. Especially if the phylogenetic research is preliminary or exploratory, if the taxon sample is sparse or based primarily on exemplars, or if higher taxa are terminals, workers are likely to prefer the soft over the hard interpretation of polytomies.

The soft alternative views polytomies as fundamentally temporary and artifactual. Polytomies are both capable of resolution and likely to be resolved or rejected in the future. Significantly, the soft interpretation rejects the polytomy itself as a legitimate phylogenetic hypothesis for the taxa involved. In a mixture of fully resolved (Tree 5) and polytomous most parsimonious trees (Trees 1–4), the soft interpretation says, "none of the polytomous trees are true, but one of the fully resolved topologies could be". If the polytomy itself is allowed as a "resolution," then clearly it is already best supported by data, and the distinction between hard and soft becomes meaningless.

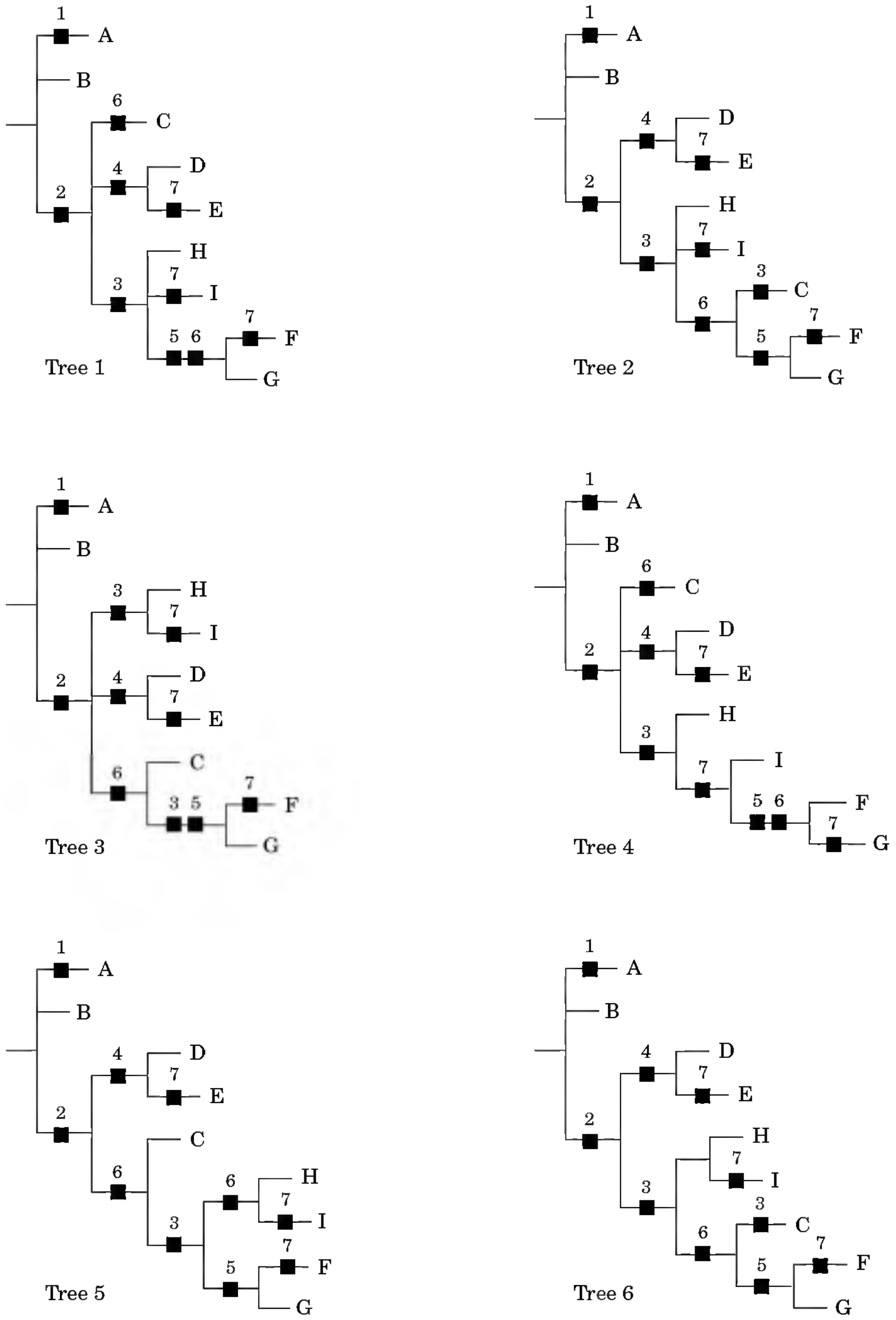

Fig. 1. The six trees implied by the data (see text). Trees 1–3 are "compatible" polytomies, Tree 4 is "incompatible," and Trees 5–6 are fully resolved. Trees 1–5 have character support at all nodes, but Tree 6 must contain a zero-length branch. Under the hard polytomy interpretation (and considering support), Trees 1–5 are legitimate, but under the soft interpretation onlyTree 5 need be considered.

By definition, then, soft polytomies are approximate, inaccurate, mistaken, false, artifactual, ephemeral, etc., especially if otherwise compatible trees in the solution set are more or completely resolved. Soft polytomies imply (only) a set of dichotomous resolutions compatible with the polytomy, any of which would be preferable to the polytomy if data allowed choice between them. Under the soft interpretation trees containing polytomies are hybrids between consensus-like summaries and real cladograms. Synapomorphies are mapped at resolved nodes and branches have length, but it makes no sense to map characters at polytomous nodes under the soft assumption, just as interpreting consensus trees, and the polytomies they contain, as cladograms is a mistake (Miyamoto, 1985).

As noted above, two basic relationships exist between more and less resolved trees when both occur in the same solution set. The polytomies are either "compatible" (Trees 1–3) or "incompatible" (Tree 4) with the more resolved trees (Trees 5–6). A polytomy is compatible if it is more or completely resolved in another, otherwise identical, tree. Thus, Tree 1 is compatible with Tree 5 and Trees 2–3 are compatible with Tree 6. The soft interpretation then excludes a less resolved tree as an hypothesis because the interpretation excludes simultaneously cladogenesis a priori. The interpretation states that any more resolved solution is better, and in this case one or more are at hand. The solution set already contains all those more resolved versions of the polytomy supported by data. The soft interpretation implies that only these hypotheses deserve consideration. The consensus of all the most resolved most parsimonious trees that remain will generally be just as good or better a summary of the solution set than the consensus of all topologies in the solution set.

Some arguments to retain the polytomous, compatible topologies (Trees 1–3) seem initially attractive, but when considered carefully are not compelling. One might argue that retention of the polytomous but compatible tree is "conservative" because the additional dichotomous resolutions absent from the solution set are also most parsimonious for the data. However, the data cannot support any of those trees, and retaining them thus violates the definition of legitimate cladograms presented above. It does not seem conservative to assume that the sole effect of new data or taxa will be to resolve the polytomy in favor of one of these resolutions. Rather, new taxa, new characters or revised codings will probably change the results in other ways, making the previous analysis obsolete. One might also retain the polytomy because the character optimization(s) that permit it seem a better explanation for evolution in those characters than the alternative optimizations that support resolutions of the polytomy. But this argument is just the hard interpretation again. To prefer the character optimizations required by the polytomy implies simultaneous cladogenesis.

The polytomy can also be incompatible, hence embedded in a topology not identical to any of the more resolved trees in the solution set (e.g. Tree 4). Different, equally parsimonious topologies normally demand attention. But in this case the unique topology (the pectinate HIFG component) comes at the price of accepting a false premise. The unique topological moiety of a polytomous cladogram is relevant only if the topology as a whole is legitimate. Just as zero-length branches invalidate trees as legitimate cladograms compared to cladograms that have full simultaneous support, soft polytomies invalidate trees compared to more resolved trees. If the data could support a cladogram containing the unique moiety and a resolution of the polytomy, then the polytomy would be compatible, not incompatible and the above argument applies. Therefore, because the soft interpretation views polytomies as inaccurate, approximate, temporary, false, etc., or a summary of hypotheses not currently supportable by data, topological arrangements that are parsimonious only at the price of including a soft polytomy (Tree 4) are illegitimate. The soft polytomy assumption strongly prefers most resolved most parsimonious trees, and constitutes a basis for tree choice within solution sets containing differentially resolved topologies.

As above, one might argue that the incompatible but polytomous tree implies a set of parsimonious and valid dichotomous trees that new evidence might support. However, such support is absent and speculation again seems futile. Indeed, the only reason to consider polytomies under the soft interpretation seems to be to renege on the soft interpretation itself, i.e. to suggest that the polytomy is hard after all.

Lastly, it may be that all trees in a solution set are polytomous and none are more resolved than any other. Either a polytomy is common to all trees, or all trees are equally but differently unresolved. In these cases the soft interpretation offers no basis for choice, and all trees would seem to be equally bad (or good).

In practice, it is not easy to identify most resolved most parsimonious trees in large solution sets because most tree-finding programs implicitly use the hard interpretation. They calculate length across polytomous nodes and otherwise treat polytomous topologies indistinguishably from fully resolved topologies. In Hennig86 (Farris, 1988) and NONA-PeeWee (Goloboff, 1993) one must examine each topology in the solution set for polytomies, but PAUP (Swofford, 1993) offers a routine to filter a solution set for all polytomous compatible topologies, thus flagging Trees 1–3 (Fig. 1). NONA's default option does not calculate length of polytomies but rather of the underlying "potential" dichotomous trees. No program known to us provides an option to flag or filter polytomous topologies (Trees 1–4), per se.

Consistent treatment of polytomous trees in differentially resolved solution sets obtained under the soft interpretation has important practical effects. First, polytomous trees (certainly compatible polytomies) should be excluded from successive weighting procedures. Successive weighting computes fits of characters to polytomies, which is illogical under the soft interpretation. Second, polytomous trees should be excluded from consensus procedures, particularly if the polytomies are incompatible. The strict consensus procedure treats polytomies as hard. The consensus of a polytomy and even one of its implied resolutions is the polytomy itself. Here available resolution is actually discarded because illegitimate topologies remain in the solution set as the consensus is computed. Under the soft interpretation polytomous trees should be excluded from any procedure that operates on trees or especially on ensembles of trees, such as successive weighting, data and/or taxon permutation techniques, diagnosis, tree-tree metrics, tree-length histograms, tree shape analyses, statistics calculated from populations of trees, etc.

Excluding polytomous trees has the beneficial effect of decreasing the number of trees to consider and making interpretation and presentation of results easier. Excluding polytomous trees also makes sense if one wishes to maximize the number of clades detected, information content of the cladogram or the falsifiability of the phylogenetic hypothesis. Under any of these criteria, most resolved most parsimonious trees are preferable to less resolved trees. It seems that the soft interpretation of polytomies supports this preference as well.

To summarize, we assume that many workers a priori prefer the soft interpretation of polytomies in primary cladograms. However, most cladistic procedures that operate on ensembles of trees effectively assume the hard interpretation. This makes for a generally unappreciated inconsistency between a priori assumptions and those implicit in common analytical procedures. To be consistent one should either adopt the hard polytomy interpretation at the outset and consider all legitimate trees regardless of resolution (Trees 1–5, Fig. 1), or exclude all compatible (Trees 1–3), and even incompatible (Tree 4), polytomous trees from a solution set containing more and less resolved trees. If all trees in the solution set are polytomous, the most resolved most parsimonious trees should be preferred. As noted above, other criteria may well be judged more important than resolution when considering multiple equally parsimonious trees, but consistent treatment of soft polytomies can reduce the ambiguity of cladistic solutions and improve the resolution, and testability, of phylogenetic hypotheses.

Acknowledgments

We thank Jim Carpenter, Mike Donoghue, Gustavo Hormiga, Darlene Judd, Diana Lipscomb, Wayne Maddison, Kevin Nixon, Dawn Southard and Mark Wilkinson for helpful conversations and/or comments on previous drafts. This work was supported by the Scholarly Studies, Neotropical Lowlands, and Biodiversity Programs of the Smithsonian Institution (to JAC) and Danish National Research Council Grants 9300106 and 9502155 (to NS).

REFERENCES

Carpenter, J. M. 1988. Choosing among multiple parsimonious cladograms. Cladistics 4: 291–296.

Coddington, J. A. and N. Scharff. 1995. Problems with zero-length branches. Cladistics 10: 415–423.

Farris, J. S. 1969. A successive approximations approach to character weighting. Systematic Zoology 18: 374–385.

Farris, J. S. 1988. Hennig86, version 1.5 manual/ software and MSDOS program. Distributed by Dr A. G. Kluge, University of Michigan, Ann Arbor, Michigan.

Goloboff, P. A. 1993. NONA, ver. 1.16 (a bastard son of Pee-Wee) 32-bit version. Program and documentation. Distributed by Dr J. S. Carpenter, American Museum of Natural History, New York, New York.

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

Maddison, W. P. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5: 365–377.

Maddison, W. P. 1993. Missing data versus missing characters in phylogenetic analysis. Systematic Biology 42: 576–581.

Mickevich, M. F. 1982. Transformation series analysis. Systematic Zoology 31: 461–478.

Miyamoto, M. M. 1985. Consensus cladograms and general classifications. Cladistics 1: 186–189.

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. Smithsonian Institution, Washington, DC.

Swofford, D. L. and W. P. Maddison. 1992. Parsimony, character-state reconstructions, and evolutionary inferences. In: R. L. Mayden (ed.), Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford University Press, Palo Alto, California, pp. 186–283.

Wilkinson, M. 1995a. Arbitrary resolutions, missing entries and the problem of zero-length branches in parsimony analysis. Systematic Biology 44: 108–111.

Wilkinson, M. 1995b. Coping with abundant missing entries in phylogenetic inference using parsimony. Systematic Biology 44: 501–514.