

Points of View

Syst. Biol. 51(1):151–155, 2002

Matrix Representation with Parsimony, Taxonomic Congruence, and Total Evidence

DAVIDE PISANI^{1,2} AND MARK WILKINSON²

¹*Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, U.K.; Davide.Pisani@bristol.ac.uk*

²*Department of Zoology, The Natural History Museum, London SW7 5BD, U.K.; marw@nhm.ac.uk*

There is considerable interest, both theoretical and practical, in the construction of phylogenetic supertrees—trees that combine the information of a collection of typically less inclusive source trees. Despite some development of alternative methods (e.g., Gordon, 1986; Lapointe and Cucumel, 1997; Semple and Steel, 2000), thus far the matrix representation with parsimony (MRP) approach to supertree construction, first independently developed by Baum (1992) and by Ragan (1992), is the only approach that has been used in any large-scale practical studies (e.g., Purvis, 1995a; Bininda-Emonds et al., 1999; Lui et al., 2001).

Because supertree methods, including MRP, combine trees rather than data, their use might be seen as analogous to the use of consensus methods in taxonomic congruence studies. However, Bininda-Emonds and colleagues have argued that MRP is not a consensus technique (Bininda-Emonds and Bryant, 1998), that similarities of MRP to taxonomic congruence are superficial (Bininda-Emonds, unpubl.), and that MRP has desirable properties akin to those of analyses of combined data (Bininda-Emonds et al., 1999). Accordingly, MRP has been proposed as a surrogate for total evidence sensu Kluge (1989) (Bininda-Emonds and Sanderson, 2001). If true, that would be regarded as a desirable property by those phylogeneticists who advocate total evidence and would provide a potentially important justification of the method.

Barrett et al. (1991) used a simple hypothetical example to illustrate differences between taxonomic congruence and total ev-

idence. Here we use the same example to demonstrate that MRP is fundamentally similar to taxonomic congruence and that its similarities to total evidence are only superficial. Thus justification of MRP must be sought elsewhere. Furthermore, how MRP resolves conflict among source trees may be problematic. Accordingly, we advocate the development of alternative supertree methods.

MATRIX REPRESENTATION WITH PARSIMONY

MRP makes use of the fact that a tree can be represented by means of a set of binary “characters,” a matrix representation. Following Baum and Ragan (1993), we refer to these “characters” as matrix elements. Matrix representations of different source trees can be readily combined with missing entries in matrix elements corresponding to leaves that are not present in the corresponding source tree. Parsimony analysis of the combined matrix representations of multiple source trees yields one or more most-parsimonious trees that include all the leaves in all the source trees and offers a synthesis of the source trees. We refer to these syntheses as MRP supertrees. Where more than one MRP supertree is found, consensus methods can be used to produce an MRP consensus supertree.

Alternative forms of MRP have been described, varying in the nature of the matrix representation, the relative weighting of source trees or matrix elements (or both), and the type of parsimony analysis used (see Wilkinson et al., 2001, for a review). In the

original approach of Ragan (1992) and Baum (1992) each clade in each of the source trees is represented by a matrix element, with members of the clade scored as 1, nonmembers are scored as 0, and leaves not present in the source tree are scored as ?. Matrix elements are weighted equally, and Fitch or Wagner parsimony (equivalent for this type of data) is used. We refer to this as standard MRP. Variant MRPs have had relatively little exploration, and we focus primarily on standard MRP here.

THE EXAMPLE

Barrett et al. (1991) presented two hypothetical data matrices (Fig. 1) for five taxa (A–D and the root). They analyzed these data separately and simultaneously to contrast the behavior of taxonomic congruence and total evidence. Separate parsimony analyses of these data matrices yielded two MPTs, (((C,D)B)A) and (((B,C)D)A), the strict component consensus of which includes only

the single component shared by these source trees (B,C,D). In contrast, parsimony analysis of the combined data yields a single tree ((A,C)(B,D)) that does not include the component (B,C,D). In this simple example, the total evidence solution contradicts the findings from taxonomic congruence (Fig. 1).

Matrix representations of the most-parsimonious trees from the separate analyses of the two datasets are also shown in Figure 1. MRP is implemented through a combined analysis of these data matrices, resulting in two trees that are identical to the two source trees (Fig. 1). The MRP consensus supertree (obtained by using the strict component consensus method) is therefore identical to the strict component consensus of the source trees. In this simple example, MRP does not emulate total evidence; rather, it produces the same result as taxonomic congruence.

Of course, whereas the example concerns trees with identical leaf sets, the most “useful” application of MRP is in combining trees with nonidentical leaf sets. However, this does not detract from the usefulness of the example in demonstrating a fundamental difference between MRP and total evidence. Our use presupposes no view on whether “it is appropriate to abide by a principle of total evidence” (Barrett et al., 1991:491).

SIGNALS

Bininda-Emonds et al. (1999) claimed that MRP emulates total evidence by stripping the data of homoplasy:

Because trees are graphical representation of the signal within a data set, MRP is essentially a parsimony analysis of the different phylogenetic signals within each data set stripped of any confounding noise (i.e., homoplasy). Thus it [MRP] obtains character congruence-like properties such as occasionally indicating clades not found among any of the source trees (analogous to signal enhancement)....

A consideration of the example shows that this result cannot be generally expected. Let us suppose that the total evidence tree is true. In that case, it is the evidence provided by homologies that is stripped from the matrix representations, not the misleading evidence of homoplasies. Furthermore, irrespective of the truth or otherwise of the total evidence tree, the fact that the source trees and their matrix representations conflict means that MRP cannot be stripping all the homoplasy from the source data, because at least one of the source trees must be incorrect.

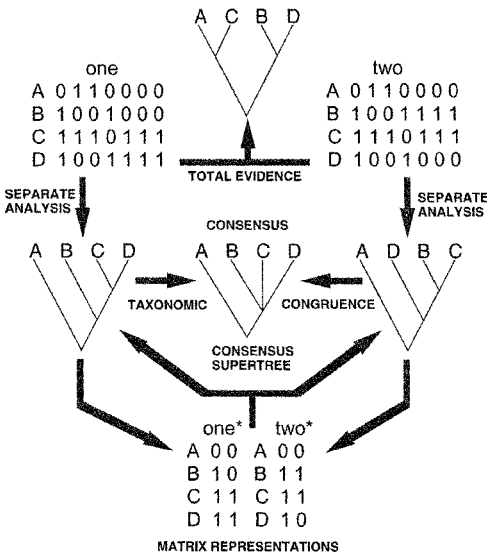


FIGURE 1. Hypothetical example illustrating differences and similarities between total evidence, taxonomic congruence, and MRP (modified from Barrett et al., 1991). Separate analyses of the two data sets (one and two) yield two most-parsimonious trees, from which taxonomic congruence yields a single consensus tree. Total evidence (combined) analysis of the data yields a single most-parsimonious tree that conflicts with the taxonomic congruence consensus tree. Matrix representations of the two most-parsimonious trees from the separate analyses of one and two are shown as one* and two*. Combined analysis of these matrices yields the same two source trees; thus, the MRP consensus supertree is identical to the consensus produced in the taxonomic congruence analysis of the original data.

Phylogenetic signal is a common term that is rarely defined. For the purposes of this paper, we define the principal signal(s) of the data as the tree(s) that the data support (i.e., the optimal trees). Subsignals exist in the data to the extent that subsets of the data support different trees. A dataset contains phylogenetic signal to the degree that the principal signal or any of the subsignals correspond to phylogenetic reality. Clearly, the principal signal may not correspond to phylogenetic signal (the tree could be completely wrong), even when the data include phylogenetic signal (in some subset). We call any phylogenetic signal in the data that is not represented by the optimal tree(s) "weak."

The extent to which optimal tree(s), that is, the principal signal(s), are determined by (and reflect) phylogenetic signal depends on the relative strengths of the multiple signals in the data. The potential advantage of total evidence is that phylogenetic signals from different data partitions, including weak signals, can be combined and thereby enhanced. MRP strips subsignals, leaving only the principal signal of each data set. Combining these principal signals provides no opportunity for the weak signal enhancement that distinguishes total evidence from taxonomic congruence.

MRP can be expected to usefully strip homoplasy to the extent that the source trees are accurate. Equally, however, MRP can be expected to unhelpfully strip homology to the extent that source trees are incorrect. Similar expectations would seem to apply to the behavior of consensus approaches, which are expected to be well-behaved to the extent that the source trees are accurate and to be unreliable to the extent that they are not.

NEW RELATIONSHIPS AND CONFLICT

An important property of MRP is that, in cases where source trees conflict, MRP supertrees may include novel clades that do not occur in any of the source trees (Bininda-Emonds and Bryant, 1998). Bininda-Emonds and colleagues (Bininda-Emonds et al., 1999; Bininda-Emonds and Sanderson, 2001) have likened the appearance of novel clades in MRP to the appearance of such clades in total evidence analyses. However, this similarity is superficial. New relationships that appear in total evidence analyses are caused by the interactions of the subsignals in initially sep-

arate data partitions; as we have seen, however, such subsignals are not included in matrix representations of trees, and their interaction is not a feature of MRP.

New relationships are expected in supertrees simply because the supertree includes more leaves. In addition, novel clades can arise from the resolution of conflict in the source trees. Conflict within a set of source trees entails that at least one of the source trees is incorrect. We can go further and recognize two logical possibilities, one of which must pertain, in such cases of conflict between source trees. Either (1) the set of source trees includes a subset of accurate trees and a subset of inaccurate trees, or (2) all of the source trees are inaccurate. In either case, if MRP somehow "resolves" the conflict by producing a novel clade, there would seem to be little reason to invest any confidence in the accuracy of the new clade. In the first case, because the novel clade contradicts all of the source trees, including the accurate source trees, this clade will therefore be inaccurate. In the second case, the novel clade could be expected to be accurate only if MRP had the character congruence properties of total evidence, which, as already noted, it lacks. We recommend that if any novel clades *sensu* Bininda-Emonds and Bryant (1998) produced in MRP analyses are included in published MRP supertrees, then they should be identified as such, to highlight their lack of any known justification.

The appearance of novel clades is but one possible outcome of how MRP may "resolve" conflicts among source trees. In addition, MRP may also simply prefer relationships in some source trees over those in others. This is another respect in which MRP might appear more similar to total evidence than to taxonomic congruence. However, the similarity to total evidence is again superficial. Whether taxonomic congruence resolves conflicts or not depends entirely on the type of consensus method used. If strict consensus methods (those that require complete agreement across the source trees for relationships to be included in the consensus tree) are used, then taxonomic congruence will not resolve conflicts among the source trees. In contrast, if, for example, a majority-rule consensus method was used, then taxonomic congruence has the potential to resolve relationships in accordance with the frequency with which the

conflicting relationships occur in the source trees.

Given that MRP provides no opportunity for multiple signals in data partitions to interact, its resolution of conflict (in the case where no novel clades are produced) is more akin to taxonomic congruence with a liberal consensus method than it is to total evidence. Bininda-Emonds and Sanderson (2001) used simulations to show that MRP can approximate total evidence. However, those simulations do not indicate that MRP differs in any way from taxonomic congruence in this respect.

Any method for combining trees that favors relationships that occur most frequently in a set of source trees should tend to produce accurate results, provided the most frequently replicated relationships are accurate. In the absence of similar systematic biases in multiple data partitions, and provided reasonable methods of analysis are used to infer the source trees from the data partitions, then we would expect this condition to be met in the long run. Accordingly, we conjecture that MRP would not differ from taxonomic congruence in its success in emulating the results (but not the method) of a total evidence analysis, given sufficient source trees. What is unclear is whether the power of MRP would differ from that of taxonomic congruence. Both our conjecture and the relative performance of taxonomic congruence should be amenable to further study through simulations.

CONCLUSIONS

The increasing use of MRP to build supertrees is the result, at least in part, of the relative simplicity of the method and the absence of readily implemented alternatives. However, as noted by Wilkinson et al. (2001), applicability in practice should not be confused with acceptability in principle. How MRP "resolves" conflict must be an important consideration in determining the utility of this approach to supertree construction. The fact that MRP lacks the crucial features of total evidence means that the way it "resolves" conflict must be justified in some other way than by the claim that MRP shares these features of total evidence. Some commentators have documented biases in the way MRP resolves conflicts among source trees that relate to the size (Purvis, 1995b; Bininda-Emonds and Bryant, 1998)

and shape (Wilkinson et al., 2001) of the source trees. Although our focus here has been on standard MRP, our conclusions regarding the relationship between standard MRP, taxonomic congruence, and total evidence will necessarily hold for variants of MRP also. Similarly, evidence of biases exists for all variant MRP methods investigated thus far (Wilkinson et al., 2001).

The results of simulation studies suggest a good approximation between MRP and total evidence in some cases, particularly with large numbers of source trees and when appropriate weighting is employed (Bininda-Emonds and Sanderson, 2001). However, we are concerned that MRP may suffer from potentially serious but poorly understood biases and from its potential to produce unjustified new groups. We consider that the properties of MRP (standard and other variants) should be better understood before MRP can be reasonably adopted as a method of choice for supertree construction. It follows that MRP should be used cautiously.

We are optimistic that methods involving more direct combination of source trees can be used to build supertrees. One such method is the MinCut approach of Semple and Steel (2000), which is elegant and efficient but its behavior and utility for phylogeneticists need thorough investigation. Another potential approach could involve extending the puzzling step used in quartet puzzling (Strimmer and von Haeseler, 1996) to the domain of supertree construction. Along similar lines, we envisage fast, heuristic algorithms in which supertrees are constructed by grafting leaves to a selected source tree, the placement of the leaves being determined by their relationships in the other source trees. As with quartet puzzling, repeated applications of such an approach with randomization of the seed tree (and perhaps also the order in which leaves are added or trees are considered) would yield a set of supertrees that could be summarized with a majority-rule consensus. With careful design to avoid biases in the resolution of conflict, the frequencies of relationships in the consensus supertree could be usefully related to the support they receive from the source trees.

ACKNOWLEDGMENTS

We thank Paul Pearson, Joe Thorley, David Gower, Mike Sanderson, and an anonymous reviewer for their

constructive comments and Olaf Bininda-Emonds and Mike Sanderson for providing a preprint of their simulation work.

REFERENCES

- BARRETT, M., M. J. DONOGHUE, AND E. SOBER. 1991. Against consensus. *Syst. Zool.* 40:486–493.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R., AND M. RAGAN. 1993. Reply to A.G. Rodrigo's "A comment on Baum's method for combining phylogenetic trees". *Taxon* 42:637–640.
- BININDA-EMONDS, O. R. P., AND H. N. BRYANT. 1998. Properties of matrix representation with parsimony analysis. *Syst. Biol.* 47:497–508.
- BININDA-EMONDS, O. R. P., J. L. GITTLEMAN, AND A. PURVIS. 1999. Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). *Biol. Rev.* 74:143–175.
- BININDA-EMONDS, O. R. P., AND M. SANDERSON. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Syst. Biol.* 50:565–579.
- GORDON, A. D. 1986. Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labelled leaves. *J. Classif.* 9:335–348.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* 38:7–25.
- LAPOINTE, F. J., AND G. CUCUMEL. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Syst. Biol.* 46:306–312.
- LIU, R. F., M. M. MIYAMOTO, N. P. FREIRE, P. Q. ONG, M. R. TENNANT, T. S. YOUNG, AND K. F. GUGEL. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- PURVIS, A. 1995a. A composite estimate of primate phylogeny. *Philos. Trans. R. Soc. London B* 348:405–421.
- PURVIS, A. 1995b. A modification to Baum and Ragan's method for combining phylogenetic trees. *Syst. Biol.* 44:251–255.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* 1:53–58.
- SEMPLE, C., AND M. STEEL. 2000. A supertree method for rooted trees. *Discrete Appl. Math.* 105:147–158.
- STRIMMER, K., AND A. VON HAESELER. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964–969.
- WILKINSON, M., J. L. THORLEY, D. T. J. LITTLEWOOD, AND R. A. BRAY. 2001. Towards a phylogenetic supertree for the Platyhelminthes? Pages 292–301 in *Interrelationships of the Platyhelminthes* (D. T. J. Littlewood and R. A. Bray, eds.). Chapman-Hall, London.

Received 24 April 2001; accepted 2 September 2001
Associate Editor: M. Sanderson

Syst. Biol. 51(1):155–165, 2002

Geographic Origin of Human Mitochondrial DNA: Accommodating Phylogenetic Uncertainty and Model Comparison

JOHN P. HUELSENBECK AND NIKITA S. IMENNOV

Department of Biology, University of Rochester, Rochester, New York 14627, USA; E-mail: johnh@brahms.biology.rochester.edu

Abstract.—Many biogeographic problems are tested on phylogenetic trees. Typically, the uncertainty in the phylogeny is not accommodated when investigating the biogeography of the organisms. Here we present a method that accommodates uncertainty in the phylogenetic trees. Moreover, we describe a simple method for examining the support for competing biogeographic scenarios. We illustrate the method using mitochondrial DNA sequences sampled from modern humans. The geographic origin of modern human mtDNA is inferred to be in Africa, although support for this hypothesis was ambiguous for data from an early paper.

Debate on the origin of anatomically modern humans has concentrated on two competing hypotheses. The "out-of-Africa hypothesis" argues that modern humans originated in Africa and then migrated to other parts of the world, replacing other species of *Homo* as they spread (Stringer and McKie, 1996). The "regional continuity hypothesis," on the other hand, argues

for a single species of *Homo* widely spread throughout the Old World, with populations connected through gene flow (Wolpoff and Caspari, 1997). Anatomically modern humans are then thought to have originated over a wide geographic area and after any migration event from Africa. The usual approach taken in a molecular test of these competing hypotheses is to collect DNA