Syst. Biol. 53(2):356–359, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490440396

Trees Versus Characters and the Supertree/Supermatrix "Paradox"

OLAF R. P. BININDA-EMONDS

Lehrstuhl für Tierzucht, Technical University of Munich, Alte Akademie 12, 85354 Freising-Weihenstephan, Germany; E-mail: Olaf.Bininda@tierzucht.tum.de

In a pair of recent articles, Gatesy and colleagues (Gatesy et al., 2002, 2004; also Gatesy and Springer, 2004) have strongly criticized several recent supertree studies. In so doing, they have pointed out important, but correctable, shortcomings in how the supertree approach was applied in specific instances, and have helped to fine-tune the methodology of this comparatively young field.

However, their equally strong critiques of the MRP method, if not the supertree approach as a whole, derive from a faulty basis for comparison. Gatesy et al. (2004:347) state (correctly) that primary character data are "the ultimate source data for both supertree and supermatrix analyses" (also p. 347), and use this statement to justify comparing both approaches on this level. However, because any connection between the primary character data and the supertree analysis is highly indirect—a feature of supertree construction that they also criticize—it is invalid to judge supertrees according to criteria designed for character-based phylogenetic reconstruction. Instead, the (MRP) supertree approach should be judged with respect to the data that it uses directly, namely the phylogenetic hypotheses presented in the source trees. As I hope to show, recognizing that supertree and supermatrix analyses operate at different levels blunts most of Gatesy et al.'s criticisms of the supertree approach, thereby resolving the "paradox" they mention in their earlier paper.

SOURCE TREE COLLECTION AND DATA DUPLICATION

As part of our efforts to construct a supertree for all extant species of mammal, we drew up a list of guidelines to help us decide which source trees were suitable for inclusion (summarized in Bininda-Emonds et al., 2003, 2004). These guidelines were based on the same two major issues raised independently by Gatesy and colleagues: data duplication and source tree quality.

As noted by Gatesy et al. (2004), our rules still allow for the duplication of the primary character data among source trees. However, we do not hold this to necessarily be problematic. Duplication can occur at this level and still result in independent phylogenetic hypotheses because a phylogenetic tree is composed of more than the data going into it (Bininda-Emonds et al., 2003, 2004). All assumptions made in the analysis (e.g., the alignment, any weighting schemes, the model of evolution used) as well as the form of the analysis itself (i.e., the optimization criterion used) can impact on the resultant phylogeny. We raised the example previously where different assumptions of rooting for virtually the same data set gave very different hypotheses about the phylogenetic relationships among cetaceans (see Bininda-Emonds et al., 2003). Another cogent example of the effect any auxiliary assumptions can have on our phylogenetic hypotheses is the detailed study of Maddison et al. (1999) on the phylogeny of carabid beetles, where different manipulations of the same base data set produced very different trees. Even the large molecular supermatrix of Madsen et al. (2001) yielded a different set of relationships when reanalyzed under a different set of assumptions by Malia et al. (2003).

In short, our guidelines specified a level of primary data duplication that we held still resulted in reasonably independent phylogenetic hypotheses. Others will undoubtedly disagree, including Gatesy et al., for whom all primary data duplication is problematic. In the end, what is important is for the researcher to assess data independence in the supertree analysis at the appropriate level, and this is at the level of the source tree and not the primary character data. Moreover, as we stressed, the rules were not designed to be applied literally and inflexibly, but to be interpreted according to the data at hand and the specific question being asked (Bininda-Emonds et al., 2004:277). This is in line with conventional phylogenetic analyses, where hard-and-fast rules with respect to which data to include, how to process them (e.g., aligning molecular data or scoring morphological data), and how to weight or analyze them are extremely rare.

THE THEORETICAL BASIS OF MRP SUPERTREE CONSTRUCTION

Gatesy et al. (2004; also Gatesy and Springer, 2004) argued that MRP lacks a logical basis and, as such, constitutes a systematic "black box" that is inappropriate for phylogeny reconstruction. In part, their perception of the lack of a logical basis to MRP derives from their attempts to judge it according to inappropriate criteria. However, they also reiterate previous criticisms (e.g., Rodrigo, 1993, 1996; Slowinski and Page, 1999) that the use of parsimony as an optimization criterion in MRP is unfounded because any "homoplasy" on a supertree cannot be interpreted in a biologically meaningful way (i.e., as instances of convergence, parallelism, or reversal).

However, incongruence in a supertree analysis is simply that, and there is no reason to equate it with homoplasy. In its purest form, the principle of parsimony makes no statements regarding either homoplasy or incongruence having to be biologically interpretable. It merely asserts that the preferred hypothesis is the one that minimizes the number of *ad hoc* assumptions (i.e., the simplest possible solution, loosely speaking). As such, the use of parsimony in MRP has the same logical basis as that for analyzing character data, namely to find the solution with the minimum amount of incongruence (as measured by the objective function of a parsimony analysis) to the data being analyzed. Homoplasy is instead a *post hoc* explanation that biologists use to explain incongruence in character data, the same as when specific instances of incongruence are held to represent faulty hypotheses of homology on the part of the investigator. Because (MRP) supertree analysis does not analyze character data, there is no need to invoke the idea of homoplasy, nor require incongruence to have a biological meaning (although it can in supertree methods such as gene-tree parsimony; Slowinski and Page, 1999).

Gatesy et al. (2004) noted that MRP supertrees at times variously resemble or contradict the results of either supermatrix or taxonomic congruence analyses, and use this "inconsistent" behavior as evidence for the black-box nature of MRP. The flaw in the argument is seen easily: one could use it to show that parsimony is also a black box because it produces results that are sometimes closer to phenetic methods like NJ and sometimes to probabilistic methods like ML or Bayesian analysis. The reality is that different methods will converge on the same answer at different times because of the nature of the data being analyzed and not because of any black-box qualities to the method.

Nor does the fact that most conventional character-based support measures (e.g., bootstrap frequencies or Bremer support) are invalid when applied to MRP supertrees invalidate the entire approach or cast its logical basis into doubt (as implied by Gatesy et al., 2004). Instead, it merely argues that appropriate supertree-specific support measures be developed that operate at the level of trees and not characters. Several such measures already exist: triplet- and quartet-fit similarity measures (Page, 2002; Piaggio-Talice et al., 2004), or the QS index (Bininda-Emonds, 2003).

HIDDEN SUPPORT

The inability of all supertree methods to account fully for hidden support in the character data is an accepted limitation, but a necessary tradeoff, of the combining of tree topologies in a supertree approach. As such, the validity of any novel clades in a supertree analysis is open to question (Pisani and Wilkinson, 2002; Gatesy et al., 2004). Fortunately, however, such clades appear to be exceptionally rare, at least for MRP supertrees. Simulation results indicate that novel clades occurred predominantly, but still at a frequency of $<0.2\%$, when just two source trees were combined (Bininda-Emonds, 2003). This merely indicates that phylogenies should not be constructed from limited data, be they source trees or character data. Most supertrees have been constructed from many more source trees than under these limiting conditions, and no novel clades have been reported for any of the major supertree studies (see Bininda-Emonds, 2003), including that of Gatesy et al. (2004).

However, I would argue against the assertion of Gatesy et al. (2004) that novel clades in a supermatrix analysis are always justified because they derive from hidden support in character data. Consider the example they cited with approval regarding the novel clade associated with *Paratomistoma courtii*. In the supermatrix tree, this species clusters as the sister group to the remaining species of Gavialinae. However, this position conflicts with its placement deep within Tomistominae from the analysis of the morphological data set, the only one to specify the position of this fossil species. Although a subsignal within this data set does cluster *P. courtii* within gavialines (J. Gatesy, pers. comm.), it is unclear why hidden support is preventing this species in the supermatrix tree from remaining as sister to the clade comprising *Gavialosuchus eggenbergensisi*, *Tomistoma lusitanica*, and *Tomistoma schlegeli*, a clade present in both the morphological and supermatrix trees. Instead, the novel placement of *P. courtii* appears to be an artifact of missing data arising from this case of "taxon sampling." The lack of other, largely molecular, data for *P. courtii* means that its final position is determined largely by being optimized on the scaffold imposed on it by the much more numerous remaining data, where 13 of the 17 data sets favor a sister group relationship between gavialines and tomistomines (in contrast to the morphological data set). In a sense, *P. courtii* is being "left behind" while the extant species with more data sort themselves out. The end result is that its final position does not reflect the only data available for it. Such artifacts need not be limited to fossil species either, but to any poorly sampled species. Given the patchy distribution of molecular data (see Sanderson et al., 2003), it is likely that some novel clades arising because of hidden support in purely molecular data sets might be equally suspect, and should not be accepted uncritically.

DATA QUALITY

Gatesy et al. (2002, 2004) counter that the greater degree of taxonomic completeness that supertrees make possible comes at the cost of having to include what they hold to be source trees of poor quality (taxonomies in particular), whether as a result of "dubious data" or invalid analytical techniques. They therefore question the utility of several supertree studies as a framework under which to study evolutionary phenomena. The potential corollaries of this criticism to supertree construction are twofold: 1) that trees derived from poor data or analyses are necessarily inaccurate, and 2) that any inaccuracy is detrimental to the resulting supertree. However, evidence suggests otherwise in both cases.

The history of phylogenetic analysis is arguably one of broadly congruent results rather than widespread disagreement. Specific exceptions abound, of course, but I would suggest that the relative amount and degree of conflict has been overplayed. For example, I have shown elsewhere that estimates of phylogeny within the mammalian order Carnivora are indistinguishable statistically for the most part (Bininda-Emonds, 2000). This included phylogenies derived from good and poor data or analyses (including taxonomies and other "data-free" phylogenies). Naturally, this finding applies only to the Carnivora. However, it implies that source trees derived from "dubious" data or techniques should not be excluded automatically, but subjected to the same process of data assessment that Gatesy et al. (2002, 2004) advocate, and which is possible in a supertree framework (contrary to their claims).

An implicit assumption in phylogenetic analysis is that phylogenetic signal is coherent and will outweigh any non-phylogenetic "signals," which are random or at least less coherent. This is, in fact, the principle underlying signal enhancement and hidden support. However, it also explains why the inclusion of poor source trees need not be detrimental to a supertree analysis: any inaccurate information from such source trees should be outweighed by the coherent phylogenetic signal from the remaining source trees. Indirect support for this derives from the observation that heavily downweighting poor source trees had little appreciable effect in most supertree studies (e.g., Purvis, 1995; Bininda-Emonds et al., 1999; Jones et al., 2002; Stoner et al., 2003). Altogether, differential weighting schemes and other sensitivity analyses would seem to therefore be reasonable counterstrategies to ascertain any effects owing to the inclusion of poor data.

On a more practical note, Gatesy et al. (2004) argue that the lack of good data for a species should be taken as a sign to collect some. I agree wholeheartedly. But, what do we do in the meantime, especially in the face of the looming biodiversity crisis? In view of the increasing importance of phylogenetic studies to conservation biology (Purvis et al., in press), the stand of Gatesy et al. that poorly known species be excluded from phylogenetic (super)trees and the comparative analyses based on them has serious consequences. Most methods that can identify threatened, species-poor clades, and therefore possible factors correlating with this threat, rely on complete taxon sampling (see Gittleman et al., 2004). Yet, threatened species are precisely those for which good data are often lacking (see McKinney, 1999; Mace et al., 2003; Gittleman et al., 2004), and this will likely be true for some time to come, especially for less charismatic organisms (e.g., most invertebrate groups). Surely it is better for conservation biologists to draw inferences now, and ones that are based on phylogenies that only *might* be inaccurate as a result of using poor data, rather than to wait and start conservation efforts after enough good phylogenetic data has been amassed (when it might be too late). Gatesy et al. (2004) rightly note that their complete phylogeny of Crocodylia will have important implications for the conservation of this group, but complete phylogenies such as this, even for other equally small groups, remain very much the exception.

The Future: Global Congruence and Divide-and-Conquer

The recognition that the supertree and supermatrix approaches analyze different data using different assumptions and methods has an important consequence. Contrary to Gatesy et al. (2004), these approaches should be seen as being complementary, and not competing, strategies for phylogenetic reconstruction in a manner akin to the global congruence approach (sensu Lapointe et al., 1999). Where these different approaches both support the same set of relationships for a comparable set of studies, we can have increased confidence in the reality of those relationships. By contrast, relationships upon which the approaches disagree, especially poorly supported relationships, should be examined more closely for possible causes of this conflict (e.g., data used or assumptions made in either approach, or true conflict), and targeted for additional data collection and analysis.

Gatesy and colleagues have done much to improve the phylogenetic database by generating and collating a tremendous amount of character data on two very different taxonomic groups. However, they have also now produced two sets of twin supertree-supermatrix analyses that could be profitably compared to help elucidate the phylogenetic relationships of the respective groups, especially outstanding areas of uncertainty. Together, the global solution provided by the supertree and supermatrix approaches is stronger than the solution from either analysis alone.

Even so, the full promise of this complementarity has yet to be realized. As part of a divide-and-conquer strategy—whereby a large phylogenetic problem is broken down into many smaller, computationally easier ones, the results of which are later combined—supertree construction could play a vital role in the analysis of very large supermatrices. Preliminary results indicate this to be the case (Sanderson et al., 2003; Roshan et al., 2004). Therefore, the complementary nature of the supertree and supermatrix approaches will become increasingly important as we tackle ever-larger portions of the Tree of Life for analysis.

Acknowledgments

I thank John Gatesy, John Gittleman, Kate Jones, Andy Purvis, and David Williams for their helpful comments, and the German research program BMBF, through the "Bioinformatics for the Functional Analysis of Mammalian Genomes" (BFAM) project, for financial support.

References

Bininda-Emonds, O. R. P. 2000. Factors influencing phylogenetic inference: A case study using the mammalian carnivores. Mol. Phylogenet. Evol. 16:113–126.

Bininda-Emonds, O. R. P. 2003. Novel versus unsupported clades: Assessing the qualitative support for clades in MRP supertrees. Syst. Biol. 52:839–848.

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., K. E. Jones, S. A. Price, M. Cardillo, R. Grenyer, and A. Purvis. 2004. Garbage in, garbage out: Data issues in supertree construction. Pages 267–280 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, the Netherlands.

Bininda-Emonds, O. R. P., K. E. Jones, S. A. Price, R. Grenyer, M. Cardillo, M. Habib, A. Purvis, and J. L. Gittleman. 2003. Supertrees are a necessary not-so-evil: A comment on Gatesy et al. Syst. Biol. 52:724–729.

Gatesy, J., R. H. Baker, and C. Hayashi. 2004. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. Syst. Biol. 53:342–355.

Gatesy, J., C. Matthee, R. DeSalle, and C. Hayashi. 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51:652–664.

Gatesy, J., and M. S. Springer. 2004. A critique of matrix representation with parsimony supertrees. Pages 369–388 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, the Netherlands.

Gittleman, J. L., K. E. Jones, and S. A. Price. 2004. Supertrees: Using complete phylogenies in comparative biology. Pages 439–460 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, the Netherlands.

Jones, K. E., A. Purvis, A. MacLarnon, O. R. P. Bininda-Emonds, and N. B. Simmons. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). Biol. Rev. 77:223–259.

Lapointe, F.-J., J. A. W. Kirsch, and J. M. Hutcheon. 1999. Total evidence, consensus, and bat phylogeny: A distance based approach. Mol. Phylogenet. Evol. 11:55–66.

Mace, G. M., J. L. Gittleman, and A. Purvis. 2003. Preserving the Tree of Life. Science 300:1707–1709.

Maddison, D. R., M. D. Baker, and K. A. Ober. 1999. Phylogeny of carabid beetles as inferred from 18S ribosomal DNA (Coleoptera: Carabidae). Syst. Entomol. 24:103–138.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.

Malia, M. J., Jr., D. L. Lipscomb, and M. W. Allard. 2003. The misleading effects of composite taxa in supermatrices. Mol. Phylogenet. Evol. 27:522–527.

McKinney, M. L. 1999. High rates of extinction and threat in poorly studied taxa. Conserv. Biol. 13:1273–1281.

Page, R. D. M. 2002. Modified mincut supertrees. Pages 537–552 *in* Algorithms in Bioinformatics, Second International Workshop, WABI, 2002, Rome, Italy, September 17–21, 2002, Proceedings (R. Guigó and D. Gusfield, eds.). Springer, Berlin.

Piaggio-Talice, R., J. G. Burleigh, and O. Eulenstein. 2004. Quartet supertrees. Pages 173–191 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, the Netherlands.

Pisani, D., and M. Wilkinson. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Syst. Biol. 51:151–155.

Purvis, A. 1995. A composite estimate of primate phylogeny. Philos. Trans. R. Soc. Lond. B 348:405–421.

Purvis, A., J. L. Gittleman, and T. M. Brooks. In press. Phylogeny and conservation. Cambridge University Press, Cambridge.

Rodrigo, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. Taxon 42:631–636.

Rodrigo, A. G. 1996. On combining cladograms. Taxon 45:267–274.

Roshan, U., B. M. E. Moret, T. L. Williams, and T. Warnow. 2004. Performance of supertree methods on various data set decompositions. Pages 301–328 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, the Netherlands.

Sanderson, M. J., A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20:1036–1042.

Slowinski, J. B., and R. D. M. Page. 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48:814–825.

Stoner, C. J., O. R. P. Bininda-Emonds, and T. M. Caro. 2003. The adaptive significance of coloration in lagomorphs. Biol. J. Linn. Soc. 79:309–328.

First submitted 31 December 2003; final acceptance 29 January 2004
Associate Editor: Mike Steel