Chapter 12

GARBAGE IN, GARBAGE OUT

Data issues in supertree construction

Olaf R. P. Bininda-Emonds, Kate E. Jones, Samantha A. Price, Marcel Cardillo, Richard Grenyer, and Andy Purvis

Abstract: As in conventional phylogenetic analyses, issues surrounding the source data are paramount in the supertree construction, but have received insufficient attention. In supertree construction, however, the source data represent phylogenetic trees rather than primary character data. This presents several supertree-specific problems. In this paper, we examine several key data issues for supertree construction, including data set non-independence, taxonomy of terminal taxa, and the question of what constitutes a valid source tree. Throughout, we present our suggested protocol for source tree collection and manipulation based on our experiences in building a supertree of mammals. Other protocols and decisions are naturally possible. What is important is that all collection protocols are presented explicitly and address minimally the issues that we have identified.

Keywords: character data; data non-independence; monophyly; paraphyly; source trees

1. Introduction

Supertree construction represents a class of techniques in which at least partially overlapping evolutionary trees are combined to produce a single (usually) more comprehensive tree, the supertree. It differs from most other forms of phylogenetic analysis in that the raw data comprise tree topologies rather than traditional morphological or molecular characters. Despite this important distinction, most of the theoretical research into supertrees has focused on assessing the performance of existing methods (both empirically and from first principles) and on developing new methods (for a review, see Bininda-Emonds *et al.*, 2002). Apart from a few papers (e.g., Springer and

Bininda-Emonds, O. R. P. (ed.) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, pp. 267–280. Computational Biology, volume 3 (Dress, A., series ed.).

de Jong, 2001; Gatesy *et al*., 2002; Gatesy and Springer, 2004), little attention has been paid to the raw data of a supertree analysis, namely the source trees that are combined into the supertree. This state of affairs parallels the situation in traditional, character-based phylogenetics, where discussions of the issues involving character selection and definition are comparatively rare (however, see Jenner, 2001).

Issues concerning source trees are of fundamental importance to supertree construction and transcend the different supertree methods (as well as applying to consensus techniques). Different supertree methods possess different properties (see Wilkinson *et al*., 2004) and have better or worse performance under certain circumstances, but they are all limited ultimately by characteristics of the source trees that are being combined. An especially important issue is that of source tree non-independence, whereby the same primary character data contributes to more than one source tree (Springer and de Jong, 2001; Gatesy *et al*., 2002).

In this paper, we examine issues related to the collection and manipulation of source trees as part of a supertree analysis. We also provide a suggested protocol based on our experiences in building a supertree of all extant species of mammal. Parts of this protocol will apply to all supertree methods, whereas others will be specific to matrix-based supertree methods such as matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992). Although the protocol does not resolve all issues involved in building supertrees, we believe it to be the best working protocol and one that is based on explicit procedures such that it can be applied easily by different researchers to generate the same end results.

2. Issues concerning source trees

2.1 Non-independence and duplication

It is common practice in phylogenetic systematics for characters (and often the character states) to be obtained from the literature and re-used in a novel analysis. This is true for both morphological (see Jenner, 2001) and molecular data. For example, in the early days of DNA-sequence analysis, individual research groups often published a series of papers, each of which in turn included newly sequenced species that were added to the base data set. As a result, the same basic piece of character information can contribute to more than one source tree, resulting in data duplication in a supertree analysis. In all cases of data duplication, the overlap of character data between source studies means that the associated source trees are not independent of one another, a key assumption of phylogenetic analysis.

Moreover, the degree of non-independence and duplication is often difficult to quantify between any two source trees and virtually impossible for even a modest set of interrelated source trees (but see Gatesy *et al.*, 2002).

Although most attention has focused on non-independence among trees from different papers (Springer and de Jong, 2001; Gatesy *et al.*, 2002; Gatesy and Springer, 2004), non-independence can also arise among trees presented within any single source study (between- and within-study non-independence, respectively). The latter form of non-independence arises because individual data sets are often analyzed using multiple optimization criteria and weighting schemes; using different subsets of the data (both characters and taxa); and, for molecular sequence data, using different alignments. Thus, a given study can present numerous potential source trees that are clearly not independent of one another.

Unless it is accounted for, data non-independence and the associated data duplication means that some data sets are effectively upweighted and might have more influence on the supertree analysis. For instance, both Springer and de Jong (2001) and Gatesy *et al*. (2002) point out examples of single data sources being replicated many times in the supertree analysis of Liu *et al*. (2001). Similar instances of data duplication occur undoubtedly in most of the published supertrees (exceptions include Daubin *et al.*, 2001, 2002; Kennedy and Page, 2002), despite steps taken to minimize them.

Data set non-independence is arguably the greatest problem facing supertree construction and all methods of combining trees. This problem usually cannot be eliminated entirely, and in fact will become more of an issue with the increasing number of total evidence studies and the re-use of data, particularly molecular data, facilitated by on-line archiving of data sets (e.g., GenBank, web pages of individual journals, or TreeBASE; Sanderson *et al.*, 1994; Piel *et al*., 2002). However, we feel that data set non-independence can be largely ameliorated using an appropriate source tree collection protocol. This was shown indirectly by Gatesy *et al*. (2002) in their re-analysis of the Liu *et al*. (2001) mammal supertree analysis. When Gatesy *et al*. pruned out source trees that they felt were redundant or poorly justified, they recovered a supertree containing Cetacea within a paraphyletic Artiodactyla (i.e., they recovered the clade Hippopotamidae + Cetacea), a result that they felt was more in accord with current systematic opinion. However, the reworked matrix was still held to contain redundant information (J. Gatesy, pers. comm.).

2.1.1 Identifying independent source trees

The guiding rule in our protocol, which is summarized in Figure 1, is to identify phylogenetic hypotheses that can be viewed reasonably as

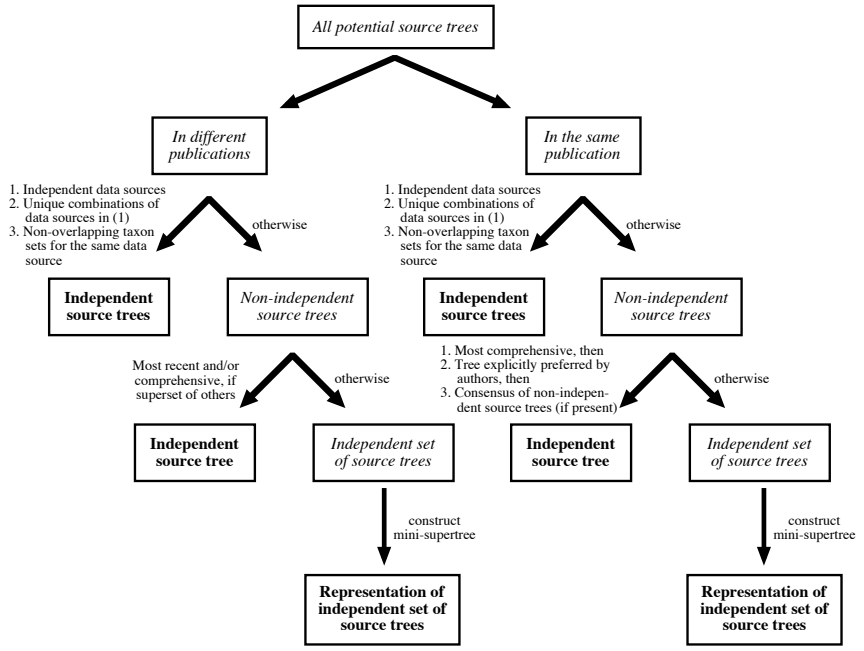

Figure 1. Decision tree summarizing our suggested protocol for source tree collection. Boxed entries in italics require further processing. Boxed entries in bold face represent reasonably independent units that can be included in a supertree analysis. Source trees can be rejected at any level (not shown).

independent (following Purvis, 1995b; Bininda-Emonds *et al.*, 2003). Independence is a difficult concept about which to be precise in phylogenetics, where the aim is to recover the common hierarchical set of relationships that has underpinned the evolution of all characters. Hence, characters are not independent in the sense of being generated by independent processes. Given the acceptance of gene trees (versus species trees; Maddison, 1997), the recognition of pseudo-independent evolutionary "packets" (i.e., the genes, possibly down to their individual exons; and the trees derived from them) might be defensible, and accords with the original justification for MRP as a method to combine gene trees (Baum, and Ragan, 2004). Our protocol attempts to identify these packets and is based on an explicitly defined set of rules for deciding the precedence of one tree over another. These rules have been formulated according to the criteria of 1) data independence, 2) taxonomic inclusiveness, and 3) (informed) author preference.

We base decisions about independence on both the source of the character data and the taxon set, not on the publications in which they appear. Non-overlapping data sets (e.g., different genes) are considered to be

independent data sources, even if they appear on a single heritable unit like mitochondrial DNA. However, different portions of the same gene (and possibly each exon within a gene), because of their common evolutionary history, are not independent for an overlapping set of taxa, even if these gene portions do not overlap at all. Trees for non-overlapping taxon sets, even if they are derived from the same set of characters, are independent by practical necessity: there is no way to combine such data sets meaningfully other than to combine the primary character data.

We also hold unique combinations of genes to be independent sources and independent from data sets containing subsets of all the genes in the combination (i.e., a total evidence analysis; *sensu* Kluge, 1989). We base this decision on the phenomenon of "signal enhancement" (*sensu* de Queiroz *et al.*, 1995), whereby the combination of data sets can yield a novel solution that is not indicated by any of the constituent data sets (see Barrett *et al.*, 1991). In essence, we hold that signal enhancement causes total evidence solutions to constitute independent phylogenetic hypotheses, although they might be based on data used elsewhere. We would argue that different morphological data sets are equivalent to novel combinations of genes and so are considered to be independent of one another unless one data set is contained completely within another.

As mentioned, non-independence can exist between or within studies. When it is present between studies, we suggest using only the most recent and/or most comprehensive study (in terms of number of taxa), but only if this study is a superset of all other source trees. This is true whether the same or different research groups have published the studies. Where no single choice presents itself (e.g., no study exists that contains all the taxa found in the others), all equally suitable source trees should be collected for a subsequent, intermediate analysis (see below).

For within-study non-independence, the first step is to identify independent sets of trees based on data set independence. For each such set, the first choice of source tree is the most comprehensive one, based on both taxa and characters. Where this is not possible, the next choice is the phylogenetic hypothesis that is preferred explicitly by the authors of the study. If no single tree is preferred clearly, the (preferred) consensus of all the trees in the set should be used instead. Finally, should multiple equally suitable source trees remain, all should be collected.

2.1.2 Accommodating non-independent sets of source trees

Often it will be possible to select a single source tree to represent a set of non-independent source trees (e.g., the most comprehensive source tree or the one preferred explicitly by the author(s) of a study). When this is not

possible, the set of non-independent source trees should be treated as a single unit. However, these units can still possess a disproportionate influence on the analysis because they are not single trees, but sets of trees. Three solutions present themselves to correct for this. The first two apply only to matrix representation supertree methods.

First, the source trees in each unit can be downweighted such that the unit as a whole has the same weight as a single source tree. However, this solution will not be feasible usually because it is not clear in many cases what the corrected weight should be given that individual source trees will themselves have different weights (i.e., numbers of matrix elements) according to their size and resolution. Second, as suggested by Bininda-Emonds and Bryant (1998) for coding in a set of equally most parsimonious solutions, only each unique node in the set of source trees should be coded. By itself, this yields a solution identical with the strict consensus of the set of source trees. However, it also retains the conflicting clustering information that is otherwise subsumed in the strict consensus tree. This solution is possible only if unique nodes can be identified unambiguously. Thus, the set of source trees must have identical taxon sets and information regarding node support cannot be included. Third, one can produce a "mini-supertree" for the set of source trees, and then use this mini-supertree as the source tree in the main analysis (e.g., Bininda-Emonds *et al.*, 1999). This procedure can be used with all supertree algorithms and on sets of source trees that do not have identical taxon sets. However, there is a loss of information in that the mini-supertree usually will not display all the relationships present in the set of source trees. Moreover, the mini-supertree will be produced often under conditions where MRP methods at least have been demonstrated to possess reduced power (i.e., few source trees; Bininda-Emonds and Sanderson, 2001). Despite this, we recommend the use of mini-supertrees over the other two options because of its generality and ease of application.

2.2 Standardizing terminal taxa

In all studies where data are combined, be they primary character data or source trees, one needs to ensure that the terminal taxa, be they fossil or extant, are comparable throughout the source data (also Page, 2004). This is essential for terminal taxa that appear in more than one source data set. However, taxonomic differences between the data sets will often make the required assessments difficult. A useful solution, therefore, is to standardize the taxonomy of the terminal taxa. In so doing, all taxonomic information from the source study should be collected and retained. This will simplify the standardization process as well as facilitate the possible use of different taxonomic systems in future analyses.

Many of the ideas in this section have been implemented in the Perl script synonoTree.pl, available from http://www.tierzucht.tum.de/Bininda-Emonds or from the first author.

2.2.1 Combining trees at different taxonomic levels

The terminal taxa of different source trees can be species (or subspecies) or from higher taxonomic levels. The latter situation is more common in (older) morphological studies, where the states described usually do not refer to an actual species, but to the inferred ancestral (or groundplan) states of the higher taxon (e.g., Wyss and Flynn, 1993). The use of higher taxa in molecular studies is less common and derives usually from a given species being held to be an exemplar for (i.e., representative of all members of) a higher-level group.

It is possible to build supertrees with the terminal taxa representing different taxonomic levels. For example, the top-level supertree of the Carnivora (figure 1 of Bininda-Emonds *et al.*, 1999) contains both species (the red panda, *Ailurus fulgen*s and the walrus, *Odobenus rosmarus*) and numerous families. Similarly, the Liu *et al*. (2001) mammal supertree contains both orders (Carnivora and Primates) and families (remaining terminal taxa). This is valid in both cases because the terminal taxa are not nested hierarchically and so are independent of one another.

Where the terminal taxa are nested, the use of classification graphs (see Page, 2004) can allow the source trees to be combined, but only if they have labeled internal nodes (see also Daniel and Semple, 2004). A more universal solution is to standardize the names using either the higher- or lower-level names. The decision on which level to use depends obviously on the desired taxonomic level of the supertree. In both cases, the required decisions make important taxonomic assumptions and thus should be made according to a standard, well-recognized taxonomic reference containing sufficient information upon which to base them (e.g., synonymy lists and type localities).

If the higher-level name is to be used, then all constituent lower-level taxa should take on this name (i.e., essentially taken to be exemplars of the higher taxon), with all monophyletic clades being collapsed to a single terminal. This procedure can occasionally cause the higher taxon to be non-monophyletic in some source trees. Our solution to this problem appears in the following section. This procedure also makes a strong assumption regarding the monophyly of the higher-level taxon, which can cause erroneous results if it is wrong (Gatesy *et al.*, 2002; Malia *et al.*, 2003). It is also important to ensure that the higher-level taxon is comparable (i.e., that a consistent definition is used) across the source studies.

When using the lower-level name, the first step is to identify the actual species of the higher taxon that were examined in the source study. This is the most desirable, direct, and least assumption-laden option. Two solutions present themselves if this is not possible. First, one can assume the monophyly of the higher taxon and create an extra node consisting of all the species (or relevant taxonomic units) in the higher taxon. However, this procedure will elevate the support for the monophyly of the higher taxon artificially according to the membership of the taxonomic reference. As such, the support derives from an appeal to authority (*sensu* Gatesy *et al.*, 2002) and not hard evidence. Moreover, the current content of the higher-level group might differ from that recognized at the time of the source study as a result of changes in phylogenetic opinion or simply the use of a different taxonomic system. Instead, we suggest that the type species of the higher taxon be identified and used in its place (following Jones *et al.*, 2002). Although this procedure is also assumption-laden, it makes fewer assumptions of monophyly and thus influences the topology of the supertree to a lesser degree.

2.2.2 Non-monophyletic species

The ever-changing and contentious nature of taxonomy and species assignment means that source studies are often not comparable in the species that they recognize (i.e., they use different synonyms or even synonyms that are no longer valid). It is therefore desirable to standardize the species taxonomy using a single, recognized reference. However, doing so can result in the creation of non-monophyletic species (Figure 2). For example, a given source study might recognize X and Y as being distinct species, and ones that are not each other's closest relatives. By contrast, the reference taxonomy recognizes X and Y as being the same species, Z, rendering Z non-monophyletic in that source study. For clarity, we refer to X and Y as the "source species" and Z as the "reference species" in the following.

Two solutions present themselves. First, as for coding higher-level taxa, the source species that represents the type species in the reference taxonomy should be recognized as the reference species. Second, if this is not possible (e.g., the type species is unknown or neither source species can be equated with it), then the position of the reference species should be considered as being uncertain in that source tree. In essence, the single source tree represents multiple source trees, one for each source species (now representing the reference species) with the remaining source species being pruned (Figure 2b). These multiple non-independent source trees can be handled in the normal fashion. Either the source trees can be used to form a

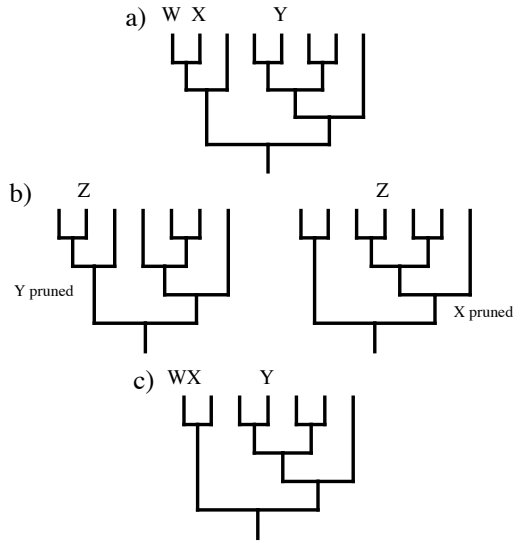

Figure 2. The problem of non-monophyletic species. The reference taxonomy holds the source species X and Y to be the same species, Z, rendering Z polyphyletic on the tree in (a). One possible solution is to prune each of the source species in turn to yield a set of source trees reflecting the uncertain placement of Z (b). If two or more source species of Z form a monophyletic clade (e.g., W and X in (a)), this clade can be collapsed to a single terminal (c).

mini-supertree or, for matrix representation methods, each unique node among the set of source trees can be coded.

It is also possible that a source study will have a reference species formed from a combination of monophyletic and non-monophyletic source species. For example, in Figure 2a, consider that the sister species of X, W, is also a source species of Z. In this case, W and X can be collapsed into the single source species WX (Figure 2c). However, both W and X should be considered separately in WX when attempting to determine if either is the type species of Z.

2.3 Valid versus invalid source trees

A multitude of phylogenies derived from a variety of data sources and methodologies exist in the systematic literature. Older source trees in particular are often not based on any explicit data source or methodology. Obviously, all these potential source trees can differ widely in quality and potential utility.

The decision as to what constitutes a valid (or usable) source tree has differed in past supertree analyses. These decisions have been guided by the

often-conflicting issues of data quality, achieving sufficient taxonomic coverage, and the goal of the supertree analysis. For example, Pisani *et al*. (2002) explicitly collected source trees of the Dinosauria that were obtained using cladistic methods, and therefore postdate 1966. By contrast, Purvis (1995a) included all phylogenetic hypotheses available, even those that were never intended as such, to obtain complete taxonomic coverage of the Primates. However, he acknowledged the difference in source tree quality by downweighting those trees that were not derived using a robust methodology heavily. Finally, in trying to summarize the prevalent systematic opinion for a given historical period, Bininda-Emonds (in press) included all available source trees and weighted them equally.

When the goal of the supertree study is to derive the best possible estimate of the phylogeny of a given group, only source trees of the highest available quality should be used (Gatesy *et al.*, 2002; Bininda-Emonds *et al.*, 2003), within the constraints of being able to assess this quality *a priori* and achieving the desired taxonomic coverage. The deleterious effects of including poor quality data manifest themselves probably only if wrong statements from these trees act in concert and, more importantly, if there are few good data. Thus, data quality control is particularly important when there are relatively few source trees. With many source trees, the inclusion of lesser quality data should have little effect on the result. However, one should be mindful that there is likely to be a trade-off between source tree number and quality. Implicit in this discussion is that poor quality data are misleading. In one case study, however, poor and good quality data produced estimates of phylogeny that were indistinguishable from one another statistically (Bininda-Emonds, 2000). This finding obviously need not be representative of all taxonomic groups or analyses, however.

Our suggestion is that only source trees based on original analyses should be collected. Secondary representations of a source tree from another study should be deferred to the tree in the original study. It is debatable whether or not taxonomies or even other supertrees constitute original analyses. Both represent secondary manipulations of original data, but also potentially novel phylogenetic hypotheses. In our work, we have included the former (unless the phylogeny that it is based on is clear), but not the latter. Instead, we collected the source trees from which the supertree was derived.

On a more practical note, we also recommend that only trees from published sources, or minimally ones that are "in press", be collected. Doing so increases the accountability (*sensu* Gatesy *et al.*, 2002) of the supertree analysis with respect to its source data. Unpublished data are liable to change before they are published, might be published in another format or venue and therefore be untraceable, or might never be published. We would include web journals and, so long as the results have not been published elsewhere,

graduate dissertations as valid published sources. Although the latter are not peer reviewed in the strict sense, the key issues here are data accessibility and accountability. The inferred quality of the source trees is an important, but separate consideration.

In the end, the researcher must make the final decision as to what represents a valid source tree. However, it should be made according to explicit guidelines, based on, but not limited to, issues such as data quality, the methodology used, and assumptions made in the study (e.g., appeals to authority or other assumptions of monophyly). Moreover, we urge the use of differential weighting to explore the effect of source tree quality on the supertree analysis (e.g., Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Jones *et al.*, 2002; Stoner *et al.*, 2003).

2.4 Source-tree collection

In all cases, we recommend that suitable source trees be collected exactly as they appear in the source study, together with all important taxonomic information for the taxa that appear on them. Copies of the source trees can then be modified to suit the particular supertree analysis, providing maximum flexibility. For example, the source trees can be pruned to include only the set of taxa of interest (e.g., pruning fossil species) or the taxon names can be standardized to accord with a given taxonomic reference or the desired taxonomic level of the supertree.

3. Conclusion

The protocol that we outline above represents a refinement and formalization of the different procedures used previously in supertree construction (Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002; Stoner *et al.*, 2003) and has been useful in our efforts to build a complete species-level mammal supertree. The protocol should not be taken to be the absolute solution to supertree construction, but instead be interpreted as some guidelines to a set of defined problems. Different solutions can, and probably will, be necessary for supertrees of other groups and depending on the goal of the supertree study. What is important for every supertree analysis is that the rules used to select and manipulate the source trees are made explicit to allow reconstruction of the results.

Acknowledgements

We thank Harold Bryant and John Gatesy for their helpful comments. This work was conducted as part of the "Phylogeny and Conservation" Working Group supported by the National Center for Ecological Analysis and Synthesis, a center funded by NSF (grant DEB–94–21535), the University of California at Santa Barbara, and the State of California. Additional support was through the German research program BMBF (OBE), a NERC studentship GT04 1999/TS/0140 (RG), NERC grant NER/A/S/2001/000581 (MC and AP) and NSF grant DEB/0129009 (KEJ and SAP). This work was completed while KEJ was at the Department of Biology, University of Virginia and RG was at the Department of Biological Sciences, Imperial College.

References

BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 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 RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.

BININDA-EMONDS, O. R. P. 2000. Factors influencing phylogenetic inference: a case study using the mammalian carnivores. *Molecular Phylogenetics and Evolution* 16:113–126.

BININDA-EMONDS, O. R. P. In press. The phylogenetic position of the giant panda (*Ailuropoda melanoleuca*): a historical consensus through supertree analysis. In D. G. Lindburg and K. Baragona (eds), *Pandas: Biology and Conservation*. University of California Press, Berkeley.

BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.

BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.

BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.

BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., GRENYER, R., CARDILLO, M., HABIB, M., PURVIS, A., AND GITTLEMAN, J. L. 2003. Supertrees are a necessary not-so-evil: a comment on Gatesy *et al*. *Systematic Biology* 52:724–729.

BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony supertree construction. *Systematic Biology* 50:565–579.

DANIEL, P. AND SEMPLE, C. 2004. A supertree algorithm for nested taxa. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 151–171. Kluwer Academic, Dordrecht, the Netherlands.

DAUBIN, V., GOUY, M., AND PERRIÈRE, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics* 12:155–164.

DAUBIN, V., GOUY, M., AND PERRIÈRE, G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* 12:1080–1090.

DE QUEIROZ, A., DONOGHUE, M. J., AND KIM, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* 26:657–681.

GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.

GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.

JENNER, R. A. 2001. Bilaterian phylogeny and uncritical recycling of morphological data sets. *Systematic Biology* 50:730–742.

JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.

KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology* 38:7–25.

LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.

MADDISON, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.

MALIA, M. J., JR., LIPSCOMB, D. L., AND ALLARD, M. W. 2003. The misleading effects of composite taxa in supermatrices. *Molecular Phylogenetics and Evolution* 27:522–527.

PAGE, R. D. M. 2004. Taxonomy, supertrees, and the Tree of Life. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 247–265. Kluwer Academic, Dordrecht, the Netherlands.

PIEL, W. H., DONOGHUE, M. J., AND SANDERSON, M. J. 2002. TreeBASE: a database of phylogenetic knowledge. In K. Shimura, K. L. Wilson, and D. Gordon (eds), *To the Interoperable Catalogue of Life with Partners — Species 2000 Asia Oceania. Proceedings of 2nd International Workshop of Species 2000*, pp. 41–47. National Institute of Environmental Studies (Research Report R–171–2002), Tsukuba, Japan. (http://www.nies.go.jp/kanko/kenkyu/pdf/r-171-2002.pdf)

PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.

PURVIS, A. 1995a. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.

PURVIS, A. 1995b. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.

RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.

SANDERSON, M. J., DONOGHUE, M. J., PIEL, W., AND ERIKSSON, T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany* 81:183.

SPRINGER, M. S. AND DE JONG, W. W. 2001. Phylogenetics. Which mammalian supertree to bark up? *Science* 291:1709–1711.

STONER, C. J., BININDA-EMONDS, O. R. P., AND CARO, T. M. 2003. The adaptive significance of coloration in lagomorphs. *Biological Journal of the Linnean Society* 79:309–328.

WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.

WYSS, A. R. AND FLYNN, J. J. 1993. A phylogenetic analysis and definition of the Carnivora. In F. S. Szalay, M. J. Novacek, and M. C. McKenna (eds), *Mammalian Phylogeny: Placental*s, pp. 32–52. Springer-Verlag, New York.