Review TRENDS in Ecology and Evolution Vol.19 No.6 June 2004
Full text provided by www.sciencedirect.com

Phylogenetics series

The evolution of supertrees

Olaf R.P. Bininda-Emonds

Lehrstuhl für Tierzucht, Technical University of Munich, D–85354 Freising-Weihenstephan, Germany

Supertrees result from combining many smaller, overlapping phylogenetic trees into a single, more comprehensive tree. As such, supertree construction is probably as old as the field of systematics itself, and remains our only way of visualizing the Tree of Life as a whole. Over the past decade, supertree construction has gained a more formal, objective footing, and has become an area of active theoretical and practical research. Here, I review the history of the supertree approach, focusing mainly on its current implementation. The supertrees of today represent some of the largest, complete phylogenies available for many groups, but are not without their critics. I conclude by arguing that the ever-growing molecular revolution will result in supertree construction taking on a new role and implementation in the future for analyzing large DNA sequence matrices as part of a divide-and-conquer phylogenetic approach.

SUPERTREE (see Glossary) construction is a phylogenetic approach in which many overlapping source trees, rather than the character data used to derive those trees, are combined to produce a single, larger supertree. In the six years since Sanderson *et al.*'s review of supertrees in *TREE* [1], there has been substantial development in the field, such that supertrees are an even more timely, and more controversial, topic. Perhaps coincident with the recent focus on the Tree of Life, current interest in supertrees is extremely high, with both supertrees and new supertree methods being published in increasing numbers. Supertree construction is currently the only phylogenetic method that can build complete phylogenies of very large clades (i.e. with hundreds of species). It has also been identified as a possibly necessary solution for reconstructing the Tree of Life [2–4]. However, the supertree approach has also been increasingly criticized (e.g. [4–6]).

Here, I examine the changing methodology and roles of supertrees with respect to phylogenetic inference. I review past and present forms of supertree construction, and their utility and limitations. I also examine the future of supertree construction, and argue that its current implementation will eventually be superseded by the use of supertrees as an efficient and necessary tool for analyzing very large phylogenetic datasets.

Supertrees past

Although the term 'supertree' was only coined and formalized in 1986 [7], the concept behind supertrees existed informally long before this. Going possibly as far back as the field of systematics itself, hierarchically nested trees were simply pasted together in a form of TAXONOMIC SUBSTITUTION to yield a more encompassing tree (Figure 1a). Only through such informal supertrees did previous systematists have any picture of the Tree of Life as a whole. Informal supertrees continue to be constructed (e.g. [8,9]), and remain our only means of

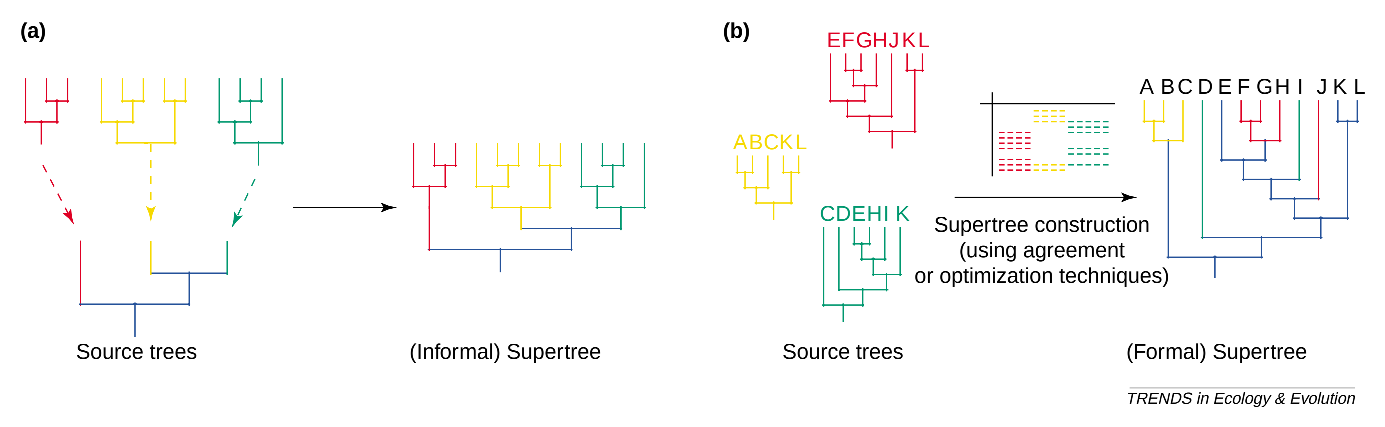


Figure 1. Supertree techniques past and present. **(a)** In the past, hierarchically nested source trees were grafted together to yield the supertree. Overlapping portions are shown in the same colour. **(b)** In the present, overlapping source trees are combined to yield the supertree. The supertree construction need not use a matrix representation of the source trees as shown. Portions of the supertree determined from a single source tree are displayed in the colour of that source tree.

Corresponding author: Olaf R.P. Bininda-Emonds (Olaf.Bininda@tierzucht.tum.de).

www.sciencedirect.com 0169-5347/$ - see front matter © 2004 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2004.03.015

Glossary

Adams consensus: a consensus method that preserves all nestings common to a set of source trees. Nestings are statements of the phylogenetic relationship of the form A is more closely related to B than either is to C. A, B and C need not be each other's closest relatives.

Additive binary coding: the coding of a complex composed of nested states (e.g. the nodes in a phylogenetic tree or certain biological characters of three or more states) as an equivalent series of nonindependent characters each with only two states.

Comparative method: the (comparative) study of the biology among a set of taxa in which the evolutionary relatedness of the organisms is taken into account to distinguish between similarities that arise for functional reasons (i.e., selection) and those that are due to common ancestry.

Compatibility: in mathematical terms, a set of trees is compatible if another tree exists that is consistent with (i.e. the same as or a less resolved version of) each tree in the set of trees.

Diameter: an approximate measure of the evolutionary distance represented by a given group of taxa as given by the maximum DNA sequence distance between any pair of taxa in the group.

Homoplasy: the *ad hoc* explanation of incongruence among biological character data as represented by multiple evolutionary events. Specific forms of homoplasy are convergence and parallelism (independent origins of the same derived trait in different lineages) and reversal (the reacquisition of the primitive trait).

Majority rule consensus: a consensus method that preserves all relationships appearing in $>50\%$ of the source trees. In fully resolved majority rule consensus, relationships that appear in $<50\%$ of the source trees can appear in the consensus solution so long as they do not contradict relationships that occur more frequently.

Matrix representation: the process whereby a tree structure is converted into the form of a matrix using any one of several coding methods (e.g. additive binary coding). The tree structure and its matrix representation have a one-to-one correspondence and are equivalent structures.

Monophyletic group: all and only those organisms or taxa that are descended from a single common ancestor (and including the ancestor).

Objective function: the function that determines how good any solution is to a given optimization problem. Roughly equivalent to optimization criterion (e.g. parsimony or likelihood) in that the latter use an implicit objective function. For instance, the objective function for a phylogenetic parsimony analysis is the fewest number of character state changes on a tree.

Polytomies: nodes that give rise to three or more descendant lineages simultaneously. The lack of resolution occurs either because of insufficient or conflicting information regarding the order (a hard polytomy) or, much more commonly, because of insufficient or conflicting information regarding the order of the branching events (a soft polytomy).

Polynomial time: the case where the running time of a problem scales with the size of the problem (in phylogenetic systematics, the number of terminal taxa) according to a polynomial function. Polynomial time algorithms are considered to be computationally efficient.

Signal enhancement: the phenomenon whereby the analysis of two combined phylogenetic data sets can yield a novel solution not supported by the analysis of either data set separately. Arises because, when combined, the congruent subsignals in the two data sets outweigh the incongruent primary signals in each.

Supermatrix approach: a phylogenetic approach in which separate character data sets are concatenated and analyzed simultaneously to yield a phylogenetic tree.

Supertree approach: a phylogenetic approach in which phylogenetic trees are combined to yield another phylogenetic tree. Distinguished from classic consensus techniques in that the source trees need only have overlapping rather than identical taxon sets.

Taxon bipartition: the two sets of taxa on a phylogenetic tree that appear on opposite sides of a given internal branch of that tree.

Taxonomic substitution: the process of replacing a terminal taxon on one phylogenetic tree with a tree representing the internal relationships of that taxon.

Total evidence: the philosophical principle that the best hypothesis is the one derived from all the available data. In phylogenetic systematics, this principle has come to be equated with the supermatrix approach, whereby all available character information is combined and analyzed.

visualizing the Tree of Life (e.g. the Tree of Life Web Project; http://tolweb.org/tree/phylogeny.html).

One serious drawback of informal supertrees is that they cannot account readily for conflicting estimates of phylogeny because only one tree can be grafted onto the supertree for a given group. Although this single tree might derive from a very comprehensive analysis (e.g. a SUPERMATRIX analysis of several different genes), a subjective decision concerning the best phylogenetic estimate for that group must still be made. The nested groups are also often based on taxonomic categories (e.g. families or genera) that might not accurately reflect evolutionary history (i.e. they need not be MONOPHYLETIC). Ultimately, the inability of informal supertree construction to accommodate conflict means that most phylogenetic information for any given group will necessarily be ignored in favour of a single hypothesis. This limitation has led, at least indirectly, to supertree construction in its current form.

Supertrees present

The current incarnation of supertrees has improved upon the informal method to enable conflicting estimates of phylogeny to be combined. Thus, such formal supertree techniques have a more objective basis and yield phylogenetic inferences that are derived from the widest selection of evidence. Current supertree methods can be categorized into those indicating common or uncontested groupings among the set of source trees (agreement supertrees), or those yielding the supertree(s) that has the maximum fit to the set of source trees according to some OBJECTIVE FUNCTION (optimization supertrees) (Figure 1b).

The first formal supertree method was introduced by Gordon [7] and is analogous to strict consensus, outputting only those relationships common to all source trees. However, this strict supertree method is limited in that the source trees must be COMPATIBLE as a set (i.e. the supertree, beyond collapsing nodes into POLYTOMIES, cannot contradict the relationships in any source tree). If the source trees are not compatible, the method will not return a tree. The breakthrough for supertrees came when the optimization-based technique matrix representation using parsimony (MRP) was described independently by Baum [10] and Ragan [11] (also [12]). MRP represented a universally applicable method that could combine even incompatible sets of source trees using existing phylogenetic software.

At the time of Sanderson *et al.*'s review [1], strict and MRP supertrees represented the only recognized formal supertree methods. (Some methods predate Sanderson *et al.*'s review, but were generally unrecognized as supertree techniques at the time. In fact, the BUILD algorithm [13] that underlies many supertree techniques, including strict supertrees, pre-dates even Gordon's article. However, it was developed for other purposes and only applied subsequently to supertree construction by Steel [14].) Today, however, at least 16 methods and variants thereof exist (Table 1), all with different properties. Many methods have links to conventional consensus techniques. For example, strict and semi-strict supertrees are the analogues of their consensus namesakes, MINCUTSUPERTREE resembles ADAMS CONSENSUS, and MRP performs similarly to FULLY RESOLVED MAJORITY RULE CONSENSUS. Interestingly, Daniel and Semple's extensions [15] of the BUILD algorithm also provide a formal basis for informal supertrees that enable multiple, nested source trees to be combined. Similar to MRP, many current methods use

Table 1. Current formal supertree methods divided according to category

Agreement supertrees	Refs	Optimization supertrees	Refs
MinCutSupertree	[50]	Average consensus (matrix representation using distances, MRD)	[51]
Modified mincut supertree	[52]	Bayesian supertrees	[46]
RankedTree	[53]	Gene tree parsimony	[36]
Semi-Labelled- and AncestralBuild	[15]	Matrix representation using compatibility (MRC)	[38,54]
Semi-strict	[25,55]	Matrix representation using flipping (MRF; also known as MinFlip supertrees)	[26]
Strict	[7]	Matrix representation using parsimony (MRP) and variants	[10,11,24,54,56]
Strict consensus merger	[47]	Most similar supertree method (dfit)	[a]
		Quartet supertrees	[28,57]

[a] Chris Creevey *et al.*; http://bioinf.may.ie/software/clann/.

MATRIX REPRESENTATION to represent the topologies of the source trees in a data matrix (Box 1).

MRP remains by far the most popular supertree method, owing to a combination of historical precedence coupled with partial software implementation (for the parsimony analyses), universal applicability (it can combine all source trees), a methodology and optimization criterion that were transparent and familiar to biologists, and good performance, in that it produces large, well resolved and apparently accurate supertrees (i.e. the supertrees do not significantly conflict with traditional hypotheses of phylogenetic relationships). It was shown only after MRP had gained general acceptance that the method does show good accuracy in simulation, performing about equally with analyses of the primary character data at reconstructing a known model tree [16].

Formal supertrees have been constructed for many groups of mammals, other vertebrates, and plants, and almost exclusively using MRP (Table 2). Some of these phylogenies are among the largest ever constructed (e.g. the bat supertree contains 916 species) and many represent the only complete phylogenetic estimates for their respective groups that are based on a rigorous methodology. Supertrees have also revolutionized the COMPARATIVE METHOD. Their unprecedented combination of large size and completeness has enabled biologists to test hypotheses on a larger scale and with more statistical power than is possible with conventional phylogenies. Supertrees have been used to address questions about evolutionary models, cladogenesis and species richness, evolutionary patterns and comparative biology, and biodiversity and conservation (see [17,18]).

Arguments for and against formal supertrees

Formal supertrees have been justified on both practical and theoretical grounds. As with comparative biology, their utility for descriptive systematics is clear: the large size of supertrees and their greater potential for taxonomic completeness enable systematic statements to be based on a larger amount of information. Moreover, supertrees can combine any phylogenetic estimates as long as they are representable as a tree-like structure. As such, supertrees can combine estimates obtained from incompatible data types (e.g. DNA sequences versus DNA–DNA hybridization data),

Box 1. Representing tree topologies and matrix representation

Combining tree topologies requires the hierarchical structure of a tree to be represented in some fashion. This is commonly done by listing, for each internal branch on the tree, those taxa that appear on one side of the branch as opposed to the other. Comparing these TAXON BIPARTITION sets between different trees forms the algorithmic basis for conventional consensus and most agreement supertree techniques.

Matrix representation forms the basis of most optimization supertree techniques (all except gene tree parsimony and quartet supertrees) and also the semi-strict agreement method (Table 1, main text). It also works with taxon bipartitions, and the most widely used form uses ADDITIVE BINARY CODING to represent the structure of a tree. In its most basic form, each informative node of the tree is coded in turn, with taxa that are descended from that node scored as 1 and taxa that are not scored as 0. The matrix so generated has a one-to-one correspondence with the tree and can be converted back into the tree using any of several optimization techniques (Figure I). The average consensus uses a variant of this coding (path-length distance matrices) to also encode branch length information.

In combining multiple trees, the matrix representations of each tree are concatenated into a single matrix. The coding is modified slightly such that taxa not present on a given source tree are scored as missing data (?) for the nodes on that source tree, and all trees are rooted with an all-zero outgroup. (In the average consensus, the missing values must be estimated from the known information). Optimization of the combined matrix representations then yields the supertree. Whereas the correspondence between a single tree and its matrix representation is well founded in graph and network theory, that between the supertree and the combined representations of the source trees must be viewed as a heuristic.

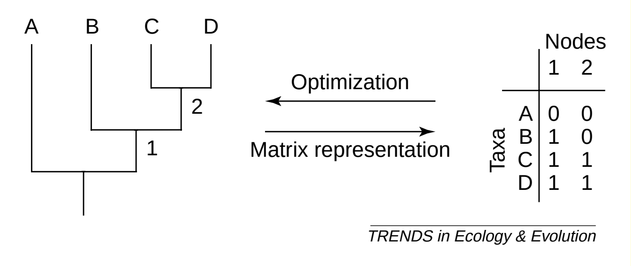


Figure I. The one-to-one correspondence between a single tree and its matrix representation.

The elements created by matrix representation (matrix elements or pseudocharacters) are statements of membership and, therefore, are only functionally equivalent to conventional characters [24]. However, this still yields numerous analytical advantages. In particular, matrix elements can be individually weighted to account for differential support or confidence in a source tree or the nodes of it. Such weighting can improve the fit between a source tree and its matrix representation [78] and the accuracy of MRP in simulation [16]. In many ways, the average consensus is akin to a weighted matrix representation in which the weights represent the branch lengths of the source tree, and these two techniques have been shown to be equivalent when all source trees have identical taxon sets [79].

Table 2. Examples of supertrees constructed using formal methods

Group	Taxonomic level	No. terminal taxa[a]	Method[b]	No. source trees	Refs
Non-mammalian vertebrates					
Caenophidia (snakes)	Species	63	MRP	15	[58]
Crocodylia (crocodiles and relatives)	Species	**22 extant** + 53 fossil	MRP	21	[59]
Dinosauria (dinosaurs)	Genus	**277**	MRP	134	[60]
‘Global avian fauna’	Genus and species	Not given	MRP/MRD/informal	90	[61]
Procellariiformes (seabirds)	Species	**122**	MRP	7	[34]
Mammals					
Artiodactyla (excl. whales) (even-toed ungulates)	Species	171	MRP	48	[62]
Carnivora (carnivores)	Species	**271**	MRP	177	[39]
Chiroptera (bats)	Species	**916**	MRP	105	[63]
Lipotyphla (insectivores)	Species	181	MRP	47	[64]
Lagomorpha (rabbits and pikas)	Species	**80**	MRP	146	[65]
Mammalia (mammals)	Order/Family	**90**	MRP	430	[30]
Marsupialia (marsupials)	Species	267	MRP	158	[66]
Primates (primates)	Species	**203**	MRP	112	[19,67]
Plants					
Angiosperms (flowering plants)	∼Order	128	MRP	7	[68]
Angiosperms (flowering plants)	Family	379	MRP	46	[69]
Apiales (umbelliferous plants)	∼Family	212	MRP	11	[68]
Cortaderia + outgroups (grasses)	Species	59	MRP	2	[70]
Hologalegina (legumes)	Species	571	MRP	43	[71]
Lithocarpus (tanbark oaks)	Species	22	MRP	5	[72]
Pinus (pines)	Species	**99**	MRP	14	[73]
Poaceae (grasses)	Genus	403	MRP	55	[74]
Other					
Bacteria	Phylum	9	MRD analogue	15	[75]
Bacteria	Species	37	MRP	130–196	[32]
Bacteria	Species	45	MRP	730	[33]
Diptera (true flies)	Family	**151**	MRP	12	[c]
Metazoa (animals)	‘Class’	102	MRP	156	[76]
Schistosoma (blood flukes)	Species	14	MRP	8	[77]

[a] Entries in bold face are complete at the given taxonomic level for the clade in question.

[b] MRP, matrix representation using parsimony; MRD, matrix representation using distances; informal, informal supertree construction.

[c] David Yeates *et al.*; http://www.inhs.uiuc.edu/cee/therevid/supertree.html.

or even those lacking any underlying data (if so desired). Combining the primary character data, by contrast, requires that these data can be analyzed using a single optimization criterion. As such, not all data can be included in a supermatrix analysis and supertrees have been justified on the principle of TOTAL EVIDENCE (e.g. [1,17,19]): the best hypothesis is the one that makes use of all the available information, or is derived from the most independent lines of evidence.

However, this desirable feature of supertrees also forms the basis for the strongest criticisms of the approach. Because supertree construction is one step removed from primary character data, critics argue that supertree construction entails a loss of valuable information (e.g. [4–6]). The effects of this loss are suggested by some to make supertree construction less desirable than are analyses of the primary character data or even simply invalid for three reasons: (i) accounting for differential signal strength within datasets; (ii) pseudoreplication among source trees; and (iii) the validity of supertrees as phylogenetic hypotheses.

Differential signal strength and signal enhancement

Barrett *et al.* ([20], also [21]) demonstrated that the simultaneous analysis of two datasets yielding conflicting phylogenetic trees can produce a novel tree when the congruent subsignals in each dataset outweigh the individual conflicting primary signals. This phenomenon of SIGNAL ENHANCEMENT cannot occur in supertree construction, which, by combining trees, cannot account easily for subsignals in the original datasets [22]. Although most optimization supertree methods can yield relationships that are not present or implied in the set of source trees [23,24], these novel or unsupported clades have no support in the raw data of a supertree analysis (i.e. the source trees), such that some researchers argue that they should be regarded as spurious (e.g. [22]). The inability of all supertree methods to account for signal enhancement and the potential for optimization supertree techniques to create spurious novel clades have been strongly criticized (e.g. [4–6,22,25]).

Yet, many supertree methods show good performance in simulation [16,26–28], such that they are often as accurate as analyses of the combined primary character data (the supermatrix approach) and produce few, if any, novel clades [23]. This indicates that the inherent loss of information is not detrimental in practice. Moreover, many supertree methods can account for differential signal strength in the primary data through differential weighting (Box 1). Such weighting improves performance in simulation to the point that weighted MRP analyses show slightly greater accuracy at reconstructing a known model tree than do supermatrix analyses [16]. Allowing for signal enhancement is clearly desirable from a theoretical perspective. However, both the frequency with which novel clades result from signal enhancement and the degree to which the primary signals differ from both one another and the congruent subsignals (‘severity’) in empirical data are not yet adequately documented (but see [29]). The examples from Barrett *et al.* [20] and Chippendale and Wiens [21] are perhaps unnaturally severe. Should such strong conflict prove to be the rule rather than the exception, phylogenetic estimates from single datasets must then be viewed with extreme caution. However, the generally good agreement among phylogenetic estimates from different sources and within different groups of organisms speaks against strong signal enhancement being widespread.

Pseudoreplication of data

The loss of contact with the primary character data means that supertree construction is susceptible to pseudoreplication, in that the same character data can contribute to more than one source tree. For instance, Springer and de Jong [5] showed that the same transferrin immunology dataset for bats formed part of five source trees in the supertree analysis of Liu *et al.* [30] (see [6] for additional examples of pseudoreplication in the same study). Replicated data such as these violate the key assumption in phylogenetic analysis of data independence. Such data are effectively upweighted in a supertree analysis and will influence the outcome more strongly than they ordinarily should. Given the continual recycling of phylogenetic information, and especially the increasing frequency of phylogenies derived from combined datasets, increasingly fewer source trees will be independent of one another at the level of the primary character data.

Data duplication can affect all supertree techniques, but can be largely mitigated through careful data collection following formalized protocols (e.g. [31]). Most published supertree analyses have attempted to minimize data duplication, admittedly with varying effectiveness. Supertrees also exist that are free of data duplication (e.g. [32–34]). Finally, it has been argued elsewhere that source trees, as phylogenetic hypotheses, have an emergent property such that they comprise more than just the character data underlying them [31,35]. As such, judgements concerning data duplication should be made at the level that is appropriate to the analysis [35]. Given the arguably emergent nature of source trees as phylogenetic hypotheses, and that supertrees combine tree topologies and not primary character data, data duplication at the character level might not be as problematic in a supertree context as some workers would argue.

Supertrees as phylogenetic hypotheses

Finally, it has been argued that supertrees, as summaries of summaries, are not valid phylogenetic hypotheses and, therefore, should not be used as a basis for comparative biology or systematics (e.g. [4–6]). MRP supertrees have come under the strongest criticism here, with critics arguing further that the phenomena of convergence, parallelism and reversal (together, HOMOPLASY) are not biologically meaningful for the raw data of an MRP analysis (e.g. [4,36–38]). In fact, several supertree techniques, such as semi-strict supertrees, gene tree parsimony and matrix representation using compatibility or flipping (Table 1), have been promoted on the basis that this criticism of MRP supertrees does not apply to them. However, the principle of parsimony (a.k.a. Ockham's Razor: prefer the simplest explanation that explains the observations) makes no explicit mention of homoplasy. Instead, homoplasy is an *ad hoc* explanation by biologists of incongruence in character data. Given that supertree construction does not combine character data, incongruence in an MRP supertree analysis need not be interpreted as homoplasy, but rather as simply incongruence among source trees [35].

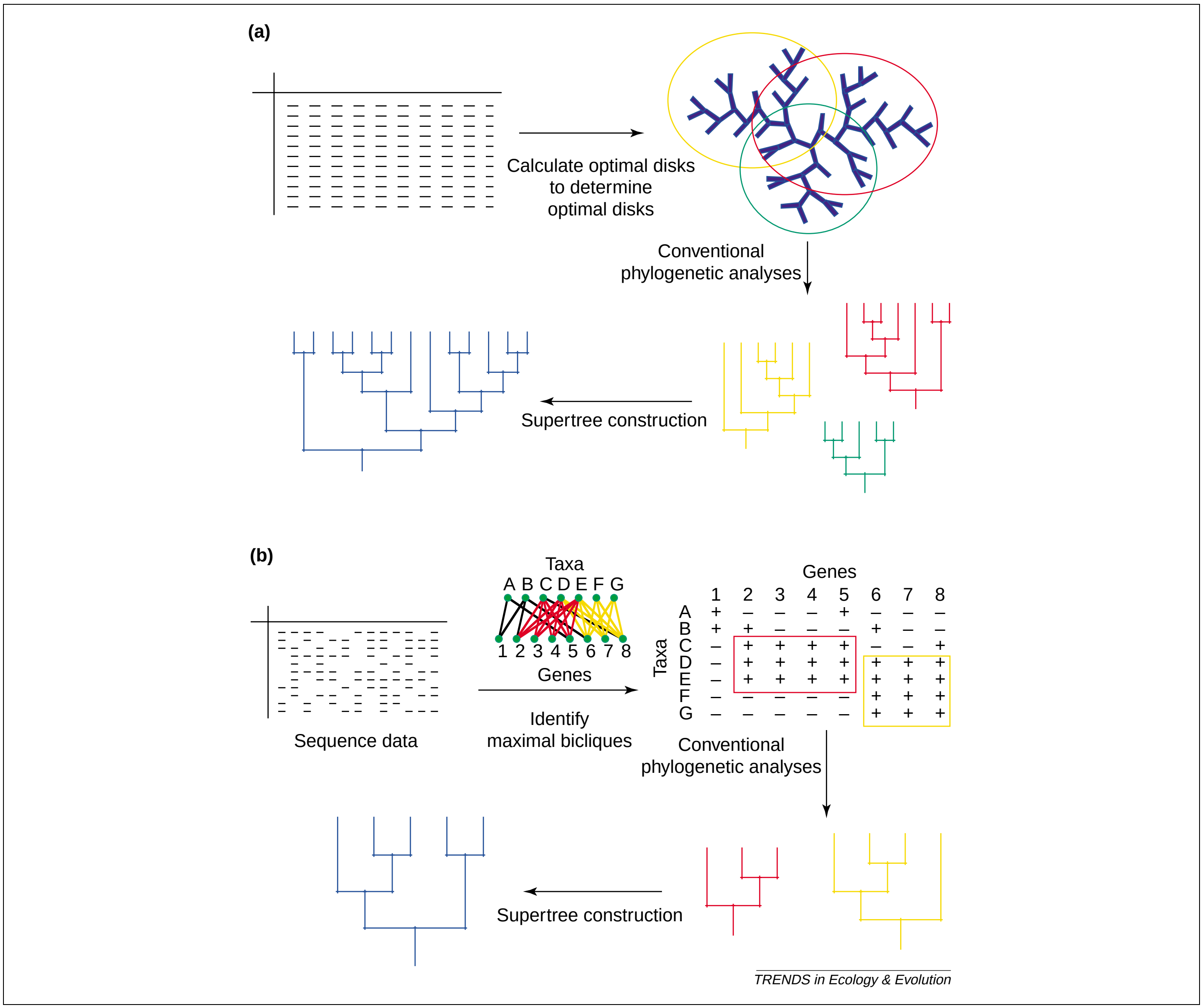
In the end, supertrees propose a hypothesis of statements of relationship among taxa, one that can be evaluated in a similar way to any other such phylogenetic hypothesis. Although most conventional phylogenetic support measures are invalid in a supertree context, several supertree-specific measures now exist (see [35]). As argued elsewhere [35], discrepancies between comparable supertree and supermatrix analyses should be treated the same as conflicts between conventional phylogenetic analyses. Are any differences the result of a poor analysis (e.g. as determined by Gatesy *et al.* [6] of the Liu *et al.* [30] supertree analysis) or are the differences restricted to weakly supported regions in one or both trees, thereby signalling the need for additional data collection and analysis?

Supertrees future

To date, supertrees have been predominantly used to combine source trees independently of the source data. The primary justification here is a lack of compatible (molecular) data to enable a comparable supermatrix analysis to be produced (e.g. [1,19,30,34,39]). Large character-based phylogenies do exist (e.g. [40]), but none are as taxonomically complete or based on as much diverse information as many of the large supertrees. However, the ever-increasing pace of the molecular revolution means that sufficient data for many groups will be available in the foreseeable future, although those for less charismatic groups might lag behind for some time. Coupled with a probable decreased emphasis on the collection and phylogenetic analysis of morphological data in the future [41], the current implementation of supertrees will become largely obsolete eventually.

However, supertree construction, rather than disappearing, will transform again into an important tool for the analysis of very large datasets. Tree space grows superexponentially with the number of species in the analysis [42], which, even with faster computers (and computer clusters) and algorithms, will ultimately constrain the size of our analyses. Instead, a more profitable strategy is the divide-and-conquer approach, whereby a very large phylogenetic problem is decomposed into many smaller subproblems, the solutions to which are combined (as a supertree) to derive the global answer. The subproblems are more computationally tractable because they are smaller and often of smaller DIAMETER (i.e. the phylogenetic distance between the taxa is smaller). Therefore, they are both faster to analyze and possibly more accurate than the larger overall problem [43].

The divide-and-conquer approach is not new to phylogenetics. It underlies strategies such as compartmentalization [44] and quartet puzzling [45], among others. However, despite being suggested by Gordon [7], the use of supertrees as part of a divide-and-conquer strategy is relatively novel. Because supertree construction will



	1	2	3	4	5	6	7	8
A	+	–	–	–	+	–	–	–
B	+	+	–	–	–	+	–	–
C	–	+	+	+	+	–	–	+
D	–	+	+	+	+	+	+	+
E	–	+	+	+	+	+	+	+
F	–	–	–	–	–	+	+	+
G	–	–	–	–	–	+	+	+

TRENDS in Ecology & Evolution

Figure 2. Supertree techniques future, showing two implementations of the divide-and-conquer approach. **(a)** Disk-covering methods decompose a data matrix into overlapping disks. The matrix need not be complete as shown. **(b)** Biclique methods determine those regions of a data matrix that contain no missing data. For both methods, the disks or regions identified are analyzed independently, with the results combined to yield the supertree.

represent a technique for analyzing very large datasets rather than purely combining source trees under this scenario, many of the criticisms above no longer apply. Instead, a supertree approach will show many desirable properties and can make use of natural partitions in the data (e.g. individual genes). (In fact, MRP was proposed originally as a method to combine gene trees [10,12].) The different partitions can be individually analyzed under the most appropriate model of evolution and optimization criterion to obtain the most robust estimate possible [17,32]. A few supertrees have been constructed already in such a context (e.g. [32,33]), and a supertree approach seems well suited to identify the shared components among the gene trees that should reflect the overall species tree. By contrast, the validity of a supermatrix approach to analyze genes of different evolutionary histories simultaneously seems questionable [32,33]. Partitions can also derive from practical considerations. For instance, aligning many genes across all life (or even large diameter problems) will not be possible, precluding a single supermatrix-style analysis. Finally, some supertree methods [e.g. those based on the BUILD algorithm, such as strict and (modified) MinCut supertrees] obtain results in POLYNOMIAL TIME and are therefore vastly more computationally efficient than are optimization-based character (or supertree) analyses. Bayesian supertree analyses are also faster than comparable character-based Bayesian analyses [46].

I conclude by highlighting two new divide-and-conquer approaches that show the way for supertrees into the future: disk-covering methods (DCMs; [47]; Figure 2a) and the biclique method (37,48]; Figure 2b). Each approach is one solution for the divide step of the divide-and-conquer strategy to yield subproblems to be analyzed using conventional phylogenetic techniques. Supertree construction solves the conquer step to combine the subproblems to obtain the global solution.

DCMs are a suite of methods that differ in the size and diameter of, and the overlap between, the subproblems that they yield, with different combinations of these parameters being more suited to different optimization criteria (e.g. parsimony, likelihood or phenetic algorithms). The subproblems are determined based on the pairwise distances between the taxa in the global dataset [43] and are, therefore, analytically determined from the data. They differ in their taxon sets only (i.e. subproblems are partitioned according to taxa, not characters). As shown by Roshan *et al.* [43], divide-and-conquer approaches based on DCMs coupled with the strict consensus merger supertree method can show improved accuracy and speed compared with a conventional phylogenetic analysis of large molecular datasets (although performance gains were not achieved in all cases).

Biclique methods attempt to identify partitions of the data that are ‘worth’ analyzing in that they will yield a well resolved solution in a reasonable amount of time. Current sequence repositories, such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html), have extremely patchy distributions of data [48]: sequence data are concentrated among relatively few model organisms and ‘model genes’ (e.g. 18S rDNA, or cyt *b* in mammals). Thus, supermatrix approaches must contend with high amounts of missing data, which can increase analysis times and decrease the decisiveness of the final solution [49]. The biclique method [37,48] delineates data partitions by identifying the largest regions (bicliques) of sequence databases that are data rich in terms of both species and genes. Ideally, these are largest regions for which all genes are present for all species. However, this criterion might be overly stringent, at least currently, in that the bicliques will be very small (i.e. few taxa and/or few genes). Therefore, biclique methods employing relaxed criteria (e.g. 75% completeness) are also being developed [37]. A logical extension of the biclique method would be to identify the maximal set of bicliques with sufficient taxonomic overlap to be combined as a supertree: in essence, bicliques of bicliques.

Summary

Although a vast improvement over informal supertree techniques, the present incarnation of supertree construction has attracted increasing criticism. Many criticisms have documented important shortcomings in the application of supertree methods, rather than of the approach as a whole. In turn, supertree analyses have become increasingly refined to address the identified deficiencies. Although the supertree approach as currently employed is still not ideal, it remains the only current way to build complete and reasonably accurate phylogenies of large clades. As such, the supertrees produced represent valuable phylogenetic hypotheses that can be used now, and are subject to the same process of refinement as any other hypothesis. They should not be used uncritically, however. Just like any other phylogeny, supertrees should be judged according to the data and analyses that are used to produce them. The continued evolution of supertrees will provide them with an increasingly rigorous basis, and ensure that supertree construction has a valid and important place in phylogenetic inference into the future.

Acknowledgements

This paper was the result of a presentation at the Workshop on New Trends in Phylogenetics and Genomics (University of Tübingen, July 1–4, 2003), for which I thank Daniel Huson, Kay Nieselt-Struwe and Mike Steel for the invitation to attend. I thank Julia Gockel, Andy Purvis, Charles Semple, Mike Steel and four anonymous reviewers for comments that helped to improve this article. Financial support was provided through the research program Bioinformatics for the Functional Analysis of Mammalian Genomes (BFAM), a project funded by the BMBF (Germany).

References

1 Sanderson, M.J. *et al.* (1998) Phylogenetic supertrees: assembling the trees of life. *Trends Ecol. Evol.* 13, 105–109

2 Soltis, P.S. and Soltis, D.E. (2001) Molecular systematics: assembling and using the Tree of Life. *Taxon* 50, 663–677

3 Pennisi, E. (2003) Modernizing the Tree of Life. *Science* 300, 1692–1697

4 Gatesy, J. and Springer, M.S. (2004) A critique of matrix representation with parsimony supertrees. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 369–388, Kluwer Academic

5 Springer, M.S. and de Jong, W.W. (2001) Phylogenetics. Which mammalian supertree to bark up? *Science* 291, 1709–1711

6 Gatesy, J. *et al.* (2002) Resolution of a supertree/supermatrix paradox. *Syst. Biol.* 51, 652–664

7 Gordon, A.D. (1986) Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *J. Classif.* 3, 31–39

8 Ortolani, A. (1999) Spots, stripes, tail tips and dark eyes: predicting the function of carnivore colour patterns using the comparative method. *Biol. J. Linn. Soc.* 67, 433–476

9 Cardillo, M. and Bromham, L. (2001) Body size and risk of extinction in Australian mammals. *Conserv. Biol.* 15, 1435–1440

10 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

11 Ragan, M.A. (1992) Phylogenetic inference based on matrix representation of trees. *Mol. Phylo. Evol.* 1, 53–58

12 Doyle, J.J. (1992) Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* 17, 144–163

13 Aho, A.V. *et al.* (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.* 10, 405–421

14 Steel, M. (1992) The complexity of reconstructing trees from qualitative characters and subtrees. *J. Classif.* 9, 91–116

15 Daniel, P. and Semple, C. (2004) A supertree algorithm for nested taxa. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 151–171, Kluwer Academic

16 Bininda-Emonds, O.R.P. and Sanderson, M.J. (2001) Assessment of the accuracy of matrix representation with parsimony supertree construction. *Syst. Biol.* 50, 565–579

17 Bininda-Emonds, O.R.P. *et al.* (2002) The (super)tree of life: procedures, problems, and prospects. *Annu. Rev. Ecol. Syst.* 33, 265–289

18 Gittleman, J.L. *et al.* (2004) Supertrees: using complete phylogenies in comparative biology. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 439–460, Kluwer Academic

19 Purvis, A. (1995) A composite estimate of primate phylogeny. *Philos. Trans. R. Soc. Lond. Ser. B* 348, 405–421

20 Barrett, M. *et al.* (1991) Against consensus. *Syst. Zool.* 40, 486–493

21 Chippindale, P.T. and Wiens, J.J. (1994) Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst. Biol.* 43, 278–287

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

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

24 Bininda-Emonds, O.R.P. and Bryant, H.N. (1998) Properties of matrix representation with parsimony analyses. *Syst. Biol.* 47, 497–508

25 Goloboff, P.A. and Pol, D. (2002) Semi-strict supertrees. *Cladistics* 18, 514–525

26 Chen, D. *et al.* (2003) Flipping: a supertree construction method. In *Bioconsensus* (Vol. 61) (Janowitz, M.F. *et al.*, eds), pp. 135–160, American Mathematical Society

27 Levasseur, C. and Lapointe, F-J. (2003) Increasing phylogenetic accuracy with global congruence. In *Bioconsensus* (Vol. 61) (Janowitz, M.F. *et al.*, eds), pp. 221–230, American Mathematical Society

28 Piaggio-Talice, R. *et al.* (2004) Quartet supertrees. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 173–191, Kluwer Academic

29 Cognato, A.I. and Vogler, A.P. (2001) Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Syst. Biol.* 50, 758–780

30 Liu, F-G.R. *et al.* (2001) Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291, 1786–1789

31 Bininda-Emonds, O.R.P. *et al.* (2004) Garbage in, garbage out: data issues in supertree construction. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 267–280, Kluwer Academic

32 Daubin, V. *et al.* (2001) Bacterial molecular phylogeny using supertree approach. *Genome Inform.* 12, 155–164

33 Daubin, V. *et al.* (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* 12, 1080–1090

34 Kennedy, M. and Page, R.D.M. (2002) Seabird supertrees: combining partial estimates of procellariiform phylogeny. *Auk* 119, 88–108

35 Bininda-Emonds, O.R.P. Trees versus characters and the supertree/supermatrix 'paradox'. *Syst. Biol.* (in press)

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

37 Burleigh, J.G. *et al.* (2004) MRF supertrees. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 65–85, Kluwer Academic

38 Ross, H.A. and Rodrigo, A.G. (2004) An assessment of matrix representation with compatibility in supertree construction. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 35–63, Kluwer Academic

39 Bininda-Emonds, O.R.P. *et al.* (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biol. Rev. Camb. Philos. Soc.* 74, 143–175

40 Källersjö, M. *et al.* (1998) Simultaneous parsimony jackknife analysis of 2538 *rbc* L DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Syst. Evol.* 213, 259–287

41 Scotland, R.W. *et al.* (2003) Phylogeny reconstruction: the role of morphology. *Syst. Biol.* 52, 539–548

42 Felsenstein, J. (1978) The number of evolutionary trees. *Syst. Zool.* 27, 27–33

43 Roshan, U. *et al.* (2004) Performance of supertree methods on various data set decompositions. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 301–328, Kluwer Academic

44 Mishler, B.D. (1994) Cladistic analysis of molecular and morphological data. *Am. J. Phys. Anthropol.* 94, 143–156

45 Strimmer, K. and von Haeseler, A. (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13, 964–969

46 Ronquist, F. *et al.* (2004) Bayesian supertrees. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 193–224, Kluwer Academic

47 Huson, D.H. *et al.* (1999) Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J. Comput. Biol.* 6, 369–386

48 Sanderson, M.J. *et al.* (2003) Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol. Biol. Evol.* 20, 1036–1042

49 Wilkinson, M. (1995) Coping with abundant missing entries in phylogenetic inference using parsimony. *Syst. Biol.* 44, 501–514

50 Semple, C. and Steel, M. (2000) A supertree method for rooted trees. *Discrete Appl. Math.* 105, 147–158

51 Lapointe, F-J. and Cucumel, G. (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Syst. Biol.* 46, 306–312

52 Page, R.D.M. (2002) Modified mincut supertrees. In *Algorithms in Bioinformatics, Second International Workshop, WABI, 2002, Rome, Italy, September 17-21, 2002, Proceedings* (Vol. 2452) (Guigó, R. *et al.*, eds), pp. 537–552, Springer

53 Bryant, D. *et al.* (2004) Supertree methods for ancestral divergence dates and other applications. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 129–150, Kluwer Academic

54 Purvis, A. (1995) A modification to Baum and Ragan's method for combining phylogenetic trees. *Syst. Biol.* 44, 251–255

55 Lanyon, S.M. (1993) Phylogenetic frameworks: towards a firmer foundation for the comparative approach. *Biol. J. Linn. Soc.* 49, 45–61

56 Semple, C. and Steel, M. (2002) Tree reconstruction from multi-state characters. *Adv. Appl. Math.* 28, 169–184

57 Thorley, J.L. and Page, R.D. (2000) RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16, 486–487

58 Kelly, C.M. *et al.* (2003) Phylogenetics of advanced snakes (Caenophidia) based on four mitochondrial genes. *Syst. Biol.* 52, 439–459

59 Gatesy, J. *et al.* Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Syst. Biol.* (in press)

60 Pisani, D. *et al.* (2002) A genus-level supertree of the Dinosauria. *Proc. R. Soc. Lond. Ser. B* 269, 915–921

61 Barker, G.M. (2002) Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biol. J. Linn. Soc.* 76, 165–194

62 Mahon, A.S. (2004) A molecular supertree of the Artiodactyla. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (Vol. 3) (Bininda-Emonds, O.R.P., ed.), pp. 411–437, Kluwer Academic

63 Jones, K.E. *et al.* (2002) A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biol. Rev. Camb. Philos. Soc.* 77, 223–259

64 Grenyer, R. and Purvis, A. (2003) A composite species-level phylogeny of the 'Insectivora' (Mammalia: Order Lipotyphla Haeckel, 1866). *J. Zool.* 260, 245–257

65 Stoner, C.J. *et al.* (2003) The adaptive significance of coloration in lagomorphs. *Biol. J. Linn. Soc.* 79, 309–328

66 Cardillo, M. *et al.* A species-level phylogenetic supertree of marsupials. *J. Zool.* (in press)

67 Purvis, A. and Webster, A.J. (1999) Phylogenetically independent comparisons and primate phylogeny. In *Comparative Primate Socioecology* (Lee, P.C., ed.), pp. 44–70, Cambridge University Press

68 Plunkett, G.M. (2001) Relationship of the order Apiales to subclass Asteridae: a re-evaluation of morphological characters based on insights from molecular data. *Edinb. J. Bot.* 58, 183–200

69 Davies, T.J. *et al.* (2004) Darwin's abominable mystery: insights from a supertree of angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1904–1909

70 Barker, N.P. *et al.* (2003) The paraphyly of Cortaderia (Danthonioideae; Poaceae): evidence from morphology and chloroplast and nuclear DNA sequence data. *Ann. Miss. Bot. Gard.* 90, 1–24

71 Wojciechowski, M.F. *et al.* (2000) Molecular phylogeny of the "temperate herbaceous tribes" of papilionoid legumes: a supertree approach. In *Advances in Legume Systematics* (Vol. 9) (Herendeen, P. *et al.*, eds), pp. 277–298, Royal Botanic Gardens

72 Cannon, C.H. and Manos, P.S. (2001) Combining and comparing morphometric shape descriptors with a molecular phylogeny: the case of fruit type evolution in Bornean *Lithocarpus* (Fagaceae). *Syst. Biol.* 50, 860–880

73 Schwilk, D.W. and Ackerly, D.D. (2001) Flammability and serotiny as strategies: correlated evolution in pines. *Oikos* 94, 326–336

74 Salamin, N. *et al.* (2002) Building supertrees: an empirical assessment using the grass family (Poaceae). *Syst. Biol.* 51, 136–150

75 Galtier, N. and Gouy, M. (1994) Molecular phylogeny of Eubacteria: a new multiple tree analysis method applied to 15 sequence data sets questions the monophyly of Gram-positive bacteria. *Res. Microbiol.* 145, 531–541

76 Zrzavy, J. and Rican, O. (2003) Morphological and molecular supertrees for metazoan phyla: conflict and accord in diverse comparative. *Cladistics* 19, 162–163

77 Morand, S. and Muller-Graf, C.D.M. (2000) Muscles or testes? Comparative evidence for sexual competition among dioecious blood parasites (Schistosomatidae) of vertebrates. *Parasitology* 120, 45–56

78 Ronquist, F. (1996) Matrix representation of trees, redundancy, and weighting. *Syst. Biol.* 45, 247–253

79 Lapointe, F-J. *et al.* (2003) Matrix representations with parsimony or with distances: two sides of the same coin? *Syst. Biol.* 52, 865–868