Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15,** 568–573.
Yang, Z. (2002). Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* **12,** 688–694.
Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide sequences with variable rates over sites: Approximate methods. *Genetics* **141,** 1641–1650.
Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17,** 32–43.
Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19,** 908–917.
Yang, Z. R., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000). Codon-substitution models for heterogenous selection pressure at amino acid sites. *Genetics* **155,** 431–449.
Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-sites models. *Mol. Biol. Evol.* **21,** 1332–1339.
Zhang, J., and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **44,** S139–S146.
Zhang, J., Kumar, S., and Nei, M. (1997). Small-sample test of episodic adaptive evolution: A case study of primate lysozymes. *Mol. Biol. Evol.* **14,** 1335–1338.
Zhang, J., Rosenberg, H. F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95,** 3708–3713.
Zhang, J., and Rosenberg, H. F. (2002). Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc. Natl. Acad. Sci. USA* **99,** 5486–5491.

[38] Supertree Construction in the Genomic Age

By OLAF R. P. BININDA-EMONDS

Abstract

Supertree construction is the process whereby overlapping phylogenetic trees, and not character data, are combined to yield a larger, more comprehensive phylogeny. In this chapter, I review the logic and methodology behind supertree construction and argue that it holds a necessary place in phylogenetic inference. Much of the justification for supertrees is admittedly practical. As I show with an empirical example, most large groups have insufficient sequence data to build complete phylogenies for them. By being able to indirectly combine diverse forms of phylogenetic information, supertrees are the best method for constructing complete phylogenies of groups with hundreds of species. However, supertree construction can also be justified on theoretical grounds. As whole genomic data are obtained for increasing numbers of species, the theoretical and practical advantages of supertrees together will ensure that the method will

0076-6879/05 $35.00

play a necessary analytical role as part of a divide-and-conquer strategy to reconstructing the Tree of Life.

Introduction

Since the beginning of the molecular revolution in the 1960s, a progressive array of molecular data has been used to elucidate the phylogenetic relationships of the species in the world around us. These data types include amino acid and DNA sequences, immunology and serology, DNA–DNA hybridization, isozymes, chromosomal banding patterns and rearrangements, SINEs and LINEs, gene order data, gene composition data, and linkage map data. Many of these data types are described elsewhere in both this volume and its predecessor (Zimmer *et al*., 1993).

In the age of genomics, the prospect of whole genomic data for phylogenetic inference is an exciting possibility. In fact, as of March 2005, whole genomic data already exist for 179 diverse microbial species in TIGR's Comprehensive Microbial Resource (Peterson *et al*., 2001), with the sequencing of several additional microbial genomes due for completion in 2005. The situation is not as advanced for eukaryotic organisms, however, where the larger genome sizes mean that sequencing efforts are concentrated in a few model species. Thus, for the moment, building large, comprehensive phylogenies for many large clades involves the combination of existing phylogenetic data. Traditionally, the data that are combined are character data, which has come to be known as the *total evidence approach* (Kluge, 1989) or the *supermatrix approach* (Sanderson *et al*., 1998) to phylogenetic inference.

In this chapter, I review a different approach for building comprehensive phylogenies in which the data that are combined are overlapping phylogenetic trees rather than the primary character data underlying those trees. This supertree approach has been used increasingly to construct (virtually) complete phylogenetic trees of clades with several hundred species (Davies *et al*., 2004; Jones *et al*., 2002; Pisani *et al*., 2002; Salamin *et al*., 2002), which in many cases, represent the only complete phylogenies for the groups in question (at the taxonomic level of the study). I first briefly introduce the concept of supertrees before describing the desirable features of supertree construction that will prove necessary in our efforts to reconstruct the Tree of Life, even in an age of whole genomic data.

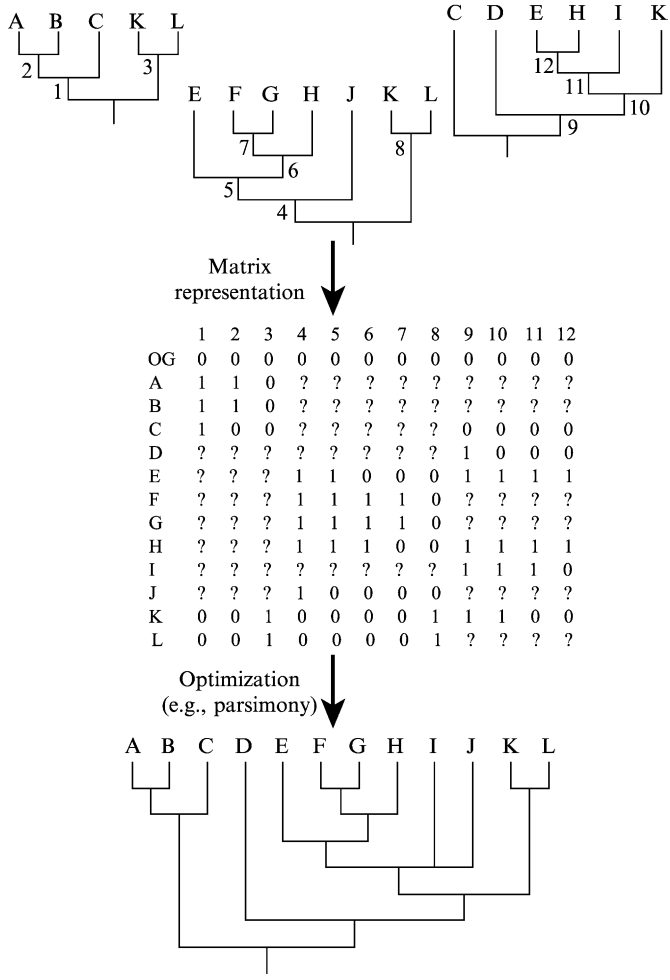
What Are Supertrees?

The idea underlying supertree construction is that a more comprehensive phylogeny can be constructed by combining two source trees that

overlap only partially in their taxon sets. In this way, statements of relationship can be made between two species that do not appear on the same tree (Sanderson *et al*., 1998) (Fig. 1). Because trees are combined in this approach, supertree construction has many obvious parallels with the more familiar field of consensus trees. However, a distinction is often made between the supertree and consensus settings (Bininda-Emonds *et al*., 2002), with the latter being a special case of the former where the source trees have identical taxon sets. Thus, although supertree methods will work in the consensus setting (i.e., to combine trees with identical taxon sets), the same is not true in reverse.

The principle of combining overlapping trees to yield a more comprehensive tree is probably as old as systematics itself, where trees were informally pasted together historically. As it is recognized currently, however, the field of supertree construction is only about a dozen years old, stemming from the independent description of the supertree method matrix representation with parsimony (MRP) by Baum (1992) and Ragan (1992). Although the first formal supertree method, and the term *supertree*, is attributable to Gordon (1986), it was the development of MRP with its numerous desirable properties that spurred the growth of supertrees. Among these properties is the ability to combine all possible statements of phylogenetic relatedness as long as they could be represented as a treelike structure, the use of the familiar and well-understood parsimony as an optimization criterion, and the ability to produce well-resolved trees. The potential for MRP (and supertree construction in general) to yield complete phylogenies of large clades was quickly realized by Purvis (1995), who used MRP to produce the first complete phylogeny for all 203 extant species of primates that was based on a rigorous objective methodology. The primate supertree has since gone on to be cited numerous times and used as a framework for understanding the biology of the entire order in an evolutionary perspective and at an unprecedented taxonomic scale.

Today, many supertree methods exist (Bininda-Emonds, 2004), all with slightly different properties. The basis for many of these methods is matrix representation (Fig. 1), whereby the topology of the source trees is coded into a matrix. This matrix is then optimized using any of a number of criteria (e.g., parsimony, compatibility, likelihood, least-squares, and Bayesian methods) to yield the supertree. Although the one-to-one correspondence between any single tree and its matrix representation is well founded in both graph and network theory, no such relationship exists between the joint set of matrix representations and the supertree (Baum and Ragan, 1993). Instead, the supertree must be viewed as the tree with the best fit to the set of source trees according to the given optimization criterion. However, each column (matrix element) in the combined matrix



	1	2	3	4	5	6	7	8	9	10	11	12
OG	0	0	0	0	0	0	0	0	0	0	0	0
A	1	1	0	?	?	?	?	?	?	?	?	?
B	1	1	0	?	?	?	?	?	?	?	?	?
C	1	0	0	?	?	?	?	?	0	0	0	0
D	?	?	?	?	?	?	?	?	1	0	0	0
E	?	?	?	1	1	0	0	0	1	1	1	1
F	?	?	?	1	1	1	1	0	?	?	?	?
G	?	?	?	1	1	1	1	0	?	?	?	?
H	?	?	?	1	1	1	0	0	1	1	1	1
I	?	?	?	?	?	?	?	?	1	1	1	0
J	?	?	?	1	0	0	0	0	?	?	?	?
K	0	0	1	0	0	0	0	1	1	1	0	0
L	0	0	1	0	0	0	0	1	?	?	?	?

FIG. 1. An example of matrix representation with parsimony (MRP) supertree construction. In the first step, the informative nodes (numbered) of all three source trees are coded such that taxa that are descended from that node are scored as 1; those that are not but are present on the tree are scored as 0; and those that are not present on the tree are scored as ? A hypothetical all-zero outgroup (OG) is added to the matrix to root the trees; it is later pruned from the supertree. In the second step, the combined matrix representations are optimized (here, using parsimony) to yield the supertree. Note that the supertree allows statements of relationship between taxa that do not co-occur on the any single source tree (e.g., taxon I with taxa F and G).

representations does maintain a one-to-one correspondence with a particular node in one source tree. This allows for the differential weighting of matrix elements according to the differential support of the nodes among the set of source trees. For instance, if the bootstrap frequencies are known for the nodes of a source tree, the matrix element representing each node can be weighted in proportion to the bootstrap frequency of that node. Such weighting has been shown to improve the fit of the matrix representation to the source tree (Ronquist, 1996) and to improve the performance of MRP in simulation (Bininda-Emonds and Sanderson, 2001).

As with consensus techniques, the choice of which supertree method to use is partly dependent on the question being asked. Strict (Gordon, 1986; Steel, 1992) and semistrict (Goloboff and Pol, 2002; Lanyon, 1993) supertree methods present the relationships that are common to or uncontradicted among, respectively, the set of source trees. As such, they provide a conservative summary of the information common to a set of source trees. Similarly, MinCutSupertree (Page, 2002; Semple and Steel, 2000) preserves nestings and, like Adams consensus, can be used to detect common statements of relationship among a set of source trees (e.g., A and B are more closely related than either is to C, where A, B, and C need not be each other's closest relatives). Gene tree parsimony (Slowinski and Page, 1999) yields a supertree that explains incongruence among the (molecular) source trees in terms of biological phenomena such as gene losses or duplications. Methods such as the average consensus (Lapointe and Cucumel, 1997) or RankedTree (Bryant *et al*., 2004) directly use branch-length information from the source trees, which is less easy to accommodate using matrix representation methods (although possible through weighting of the matrix elements; see earlier discussion). Finally, supertree methods derived from the Build algorithm of Aho *et al*. (1981) (e.g., strict, [modified] MinCutSupertree, RankedTree, AncestralBuild, and SemiLabeledBuild) run in polynomial time according to the number of taxa. Thus, they are much faster than the remaining supertrees methods (which are NP-complete and have no efficient solution, thereby requiring the use of less-desirable heuristics) and might be particularly well suited for very large supertree problems.

Why Use Supertrees? Supertrees vs. Supermatrices

Practical Considerations

Supertrees are often viewed as being an alternative to conventional, character-based phylogenetic analysis (Gatesy *et al*., 2004), with critics suggesting that supertrees have been justified largely on the basis of utility

and expediency (Gatesy and Springer, 2004; Gatesy *et al*., 2002). Much of the interest in supertrees does indeed derive from practical considerations. In particular, it is simply not possible to construct a complete phylogeny for most (large) groups of organisms because of a lack of data that can be analyzed using a single optimization criterion (i.e., compatible data). In contrast, combining trees as a supertree allows data of all forms to be combined indirectly (e.g., DNA hybridization, morphological, DNA or amino acid sequences, and immunological distances), thus potentially using the full phylogenetic dataset that exists. Of the complete supertrees that exist for many large clades of hundreds of species (Bininda-Emonds, 2004), probably none could be constructed using a supermatrix approach, although large character-based phylogenies do exist (Källersjö *et al*., 1998).

The molecular revolution, however, has done much to close this gap by yielding compatible data in great quantities for many species. Even so, data collection remains patchy and incomplete for many groups (Sanderson *et al*., 2003). Consider, in particular, the Carnivora, a well-sampled order in a well-sampled class (Mammalia). When I completed the supertree for all 271 extant carnivore species (Bininda-Emonds *et al*., 1999) in January 1996, there was no possibility of producing a molecular phylogeny on the same scale: GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) contained only 677 sequences for 48 species (Fig. 2). In the meantime, however, molecular sampling for the order has increased tremendously, so there were 1,984,623 sequences for 197 species as of March 21, 2004. A molecular phylogeny for the order seems within reach.

However, these raw numbers are somewhat deceptive. Of the nearly two million carnivore sequences, 99.6% and 0.2% are for the domestic dog and domestic cat, respectively (Table I), two carnivore species with active genome projects. Although this still leaves an average of 3900 sequences for each of the remaining 195 species, or 20 sequences per species on average, many species have been sampled repeatedly for the same gene. For example, 191 of the 219 sequences for *Martes americana* are for cytochrome *b*. Thus, many species are represented by very few genes and sequences (Fig. 3), and a complete molecular phylogeny for the Carnivora based on a wide variety of sequence data might be a more distant possibility than it might at first glance seem. The situation for less charismatic groups that have attracted less attention will naturally be even worse (Bininda-Emonds *et al*., 2002).

Thus, the supertrees of today are providing complete phylogenetic hypotheses for many groups that could not otherwise be achieved. These phylogenies have proven valuable for understanding the biology of the groups in question, with their large size and completeness giving unprecedented statistical power and scope to studies of descriptive systematics,

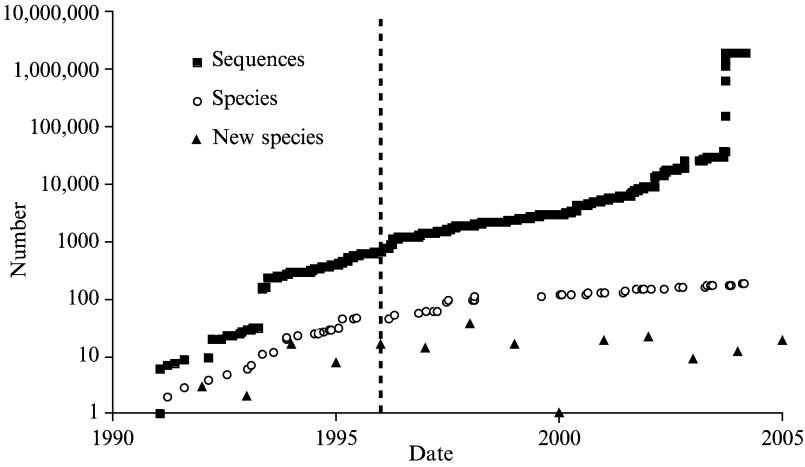


FIG. 2. The number of Carnivora sequences in GenBank and the number of extant species and new species per year represented by those sequences as a function of time. The sequence data are complete to March 12, 2004. The dashed line indicates the completion of the carnivore supertree in January 1996.

TABLE I
THE 10 MOST SEQUENCED CARNIVORA SPECIES IN GENBANK AS OF MARCH 12, 2004

Species	Common name	No. of sequences
Canis lupus	Gray wolf (includes domestic dog)	1,976,358
Felis silvestris	Wild cat (includes domestic cat)	4365
Leopardus pardalis	Ocelot	295
Ursus arctos	Brown bear	253
Martes americana	American marten	219
Panthera onca	Jaguar	173
Ursus americanus	American black bear	168
Mustela vison	American mink	160
Leopardus wiedii	Margay	151
Meles meles	Eurasian badger	141

Note: In total, 1,984,623 sequences were available for 197 of the 271 extant Carnivora species. Taxonomy follows Wozencraft (1993).

evolutionary models, cladogenesis and species richness, evolutionary patterns and comparative biology, and biodiversity and conservation (Gittleman *et al.*, 2004). However, given that more data are becoming available daily, is there a future for supertrees?

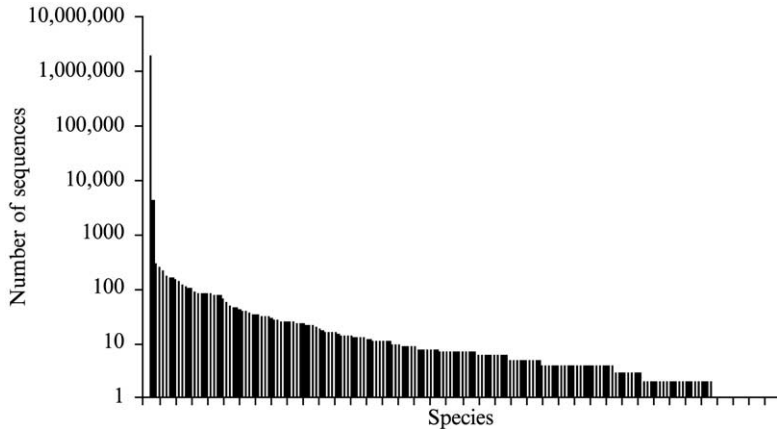


Fig. 3. The number of sequences for each of the 197 extant Carnivora species represented in GenBank. Species are presented in decreasing order according to the number of sequences. The sequence data are complete to March 12, 2004.

Theoretical Considerations

What is less appreciated is that there are also good theoretical arguments for using supertrees (Bininda-Emonds *et al*., 2002) and ones that will give an increasingly important role to supertree construction in the age of complete genomic information. In so doing, however, the role of supertree construction will change from its current form, from being used mostly to combine phylogenetic estimates derived from the literature, to being an important analytical technique.

The primary hindrance to reconstructing the Tree of Life is its sheer size. It has long been appreciated that the number of possible phylogenetic trees increases superexponentially with the number of taxa being examined (Felsenstein, 1978). Thus, the larger the phylogenetic problem, the greater the number of algorithmic shortcuts that must be taken to derive a solution in a reasonable amount of time, and the greater the probability that the globally optimal solution will not be found. Furthermore, the breadth of the Tree of Life makes deriving a globally informative dataset problematic. For example, a major stumbling block in deriving a morphological phylogeny of the metazoan phyla has been the difficulty in identifying homologous features among the phyla. Molecular data in the form of a few highly conserved genes such as 18S ribosomal DNA (rDNA) have helped to address this problem. Even so, suitable genes at this level (or beyond) are comparatively rare. More importantly, Sanderson *et al*. (1998) have suggested that aligning such genes at such levels (i.e., to identify

homologous features) will prove difficult, a situation parallel to that for morphological data.

Instead, any attempt to reconstruct large portions of the Tree of Life will require the use of supertree construction as part of a divide-and-conquer strategy to phylogenetic reconstruction. The principle underlying the divide-and-conquer approach is to break a large phylogenetic problem into numerous smaller subproblems, each of which is solved using conventional analyses. The results of the subproblems are then recombined (as a supertree) to derive the global answer. The reduced sizes of the subproblems make them computationally easier to solve and possibly more accurate because they are both smaller (fewer species) and of reduced breadth, allowing more data to be used (Roshan *et al*., 2004).

The value of supertrees as part of a divide-and-conquer strategy has been demonstrated nicely by Daubin *et al*. (2001, 2002). In the latter study in particular, Daubin *et al*. (2002) derived a phylogenetic estimate of 45 bacterial species using whole genomic data. Instead of analyzing all 730 orthologous genes that they were able to identify simultaneously, Daubin *et al*. (2002) analyzed each separately and so were able to analyze each according to the most appropriate model of evolution for it. [Although mixed-model analyses are possible in a supermatrix setting, especially in a Bayesian framework, they are more intense computationally. Therefore, many supermatrix studies use a parsimony criterion (Gatesy *et al*., 2002), which cannot account for models of molecular evolution as fully.] Each gene tree could also be pared to only those species for which data were present, thereby avoiding the adverse effects of including a large amount of missing data in the analysis (Wilkinson, 1995). The final tree was obtained by forming an MRP supertree of the individual gene trees and is regarded as one of the most robust estimates for the phylogenetic relationships of the species it contains.

Although Daubin *et al*. (2002) partitioned their supermatrix into orthologous genes, other possible strategies for the divide step are possible. Two especially promising approaches include disk-covering methods (Huson *et al*., 1999a,b) and bicliques (Burleigh *et al*., 2004; Sanderson *et al*., 2003) [for more detail, see Bininda-Emonds (2004)]. Bicliques in particular can identify portions of a supermatrix that are data rich in terms of both species and characters, thereby again avoiding the inclusion of large amounts of missing data.

As part of a divide-and-conquer strategy, supertree construction shows two desirable properties. The first is that simulation studies have shown that several supertree methods show good accuracy at reconstructing a known model tree under such circumstances (Bininda-Emonds and Sanderson, 2001; Chen *et al*., 2003; Levasseur and Lapointe, 2003;

Piaggio-Talice *et al*., 2004). In most cases, this accuracy was usually as good as that achieved by a simultaneous analysis of the combined character data. However, when the differential support within individual source trees was accounted for using weighting, MRP and the average consensus, at least, slightly outperformed the analogous supermatrix analysis (Bininda-Emonds and Sanderson, 2001; Levasseur and Lapointe, 2003). Thus, the inherent loss of information in combining trees as opposed to the primary character data does not appear to be detrimental in practice. Instead, the use of supertrees potentially allows for the inclusion of more information than a supermatrix approach in the form of the use of appropriate models of evolution for each partition (see earlier discussion; Bininda-Emonds *et al*., 2003).

The second advantage of a supertree-based divide-and-conquer search strategy of genomic data is the promise of decreased analysis time compared to a traditional supermatrix approach. The suitability of such a strategy for parallel processing is immediately clear, with each subproblem forming an independent analysis. As mentioned earlier, the smaller subproblems are also computationally easier to solve. Furthermore, as mentioned earlier, many supertree methods achieve results in polynomial time and are, therefore, much faster than character-based optimization criteria such as maximum likelihood, maximum parsimony, or neighbor joining. Even Bayesian supertrees display a speed advantage over comparable Bayesian analyses of molecular sequence data because of the special properties of Bayesian supertrees that allow a more efficient sampling strategy (Ronquist *et al*., 2004).

Thus, a divide-and-conquer strategy promises to show gains in both accuracy and speed compared to a conventional phylogenetic analysis. Evidence in support of this was provided by Roshan *et al*. (2004), who showed the postulated performance gains in the analysis of some, but not all, large molecular datasets that they examined. This is clearly an area of great promise and one that needs to be researched in more detail.

Summary: Future of Supertree Construction

Instead of being viewed as an alternative to the supermatrix approach (Gatesy *et al*., 2004), supertree construction should be viewed as being a complementary approach, both now and in the future. The basis for this is the realization that the two approaches analyze different datasets. The supermatrix approach uses the primary character data, whereas supertree construction analyzes phylogenetic hypotheses in the form of trees. These hypotheses not only derive from the primary character data, but also incorporate the many auxiliary assumptions made in the analysis of them (Bininda-Emonds *et al*., 2003). Thus, the supermatrix and supertree

approaches form important components of a global congruence framework (Lapointe *et al*., 1999), whereby well-supported relationships are those common to both sets of analyses. In contrast, conflicting sets of relationships indicate the need to identify possible sources of the conflict, be they inadequate analyses, insufficient data, or true conflict. The true complementarity of the supertree and supermatrix approaches, however, will be seen in the future, when their respective strengths contribute to a divide-and-conquer search strategy that probably represents our best opportunity to reconstruct larger portions of the Tree of Life.

References

Aho, A. V., Sagiv, Y., Szymanski, T. G., and Ullman, J. D. (1981). Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.* **10,** 405–421.

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. (1993). Reply to A.G. Rodrigo's ''A comment on Baum's method for combining phylogenetic trees.'' *Taxon.* **42,** 637–640.

Bininda-Emonds, O. R. P., Gittleman, J. L., and Purvis, A. (1999). Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). *Biol. Rev.* **74,** 143–175.

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

Bininda-Emonds, O. R. P., Gittleman, J. L., and Steel, M. A. (2002). The (super)tree of life: Procedures, problems, and prospects. *Annu. Rev. Ecol. Syst.* **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*. *Syst. Biol.* **52,** 724–729.

Bininda-Emonds, O. R. P. (2004). The evolution of supertrees. *Trends Ecol. Evol.* **6,** 315–322.

Bryant, D., Semple, C., and Steel, M. (2004). *In* ''Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 129–150. Kluwer Academic, Dordrecht, The Netherlands.

Burleigh, J. G., Eulenstein, O., Fernández-Baca, D., and Sanderson, M. J. (2004). Supertree methods for ancestral divergence dates and other applications. MRF Supertrees. *In* ''Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 65–85. Kluwer Academic, Dordrecht, The Netherlands.

Chen, D., Diao, L., Eulenstein, O., Fernández-Baca, D., and Sanderson, M. J. (2003). Flipping: A supertree construction method. *In* ''Bioconsensus'' (M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.), Vol. 61, pp. 135–160. American Mathematical Society, Providence, RI.

Daubin, V., Gouy, M., and Perrière, G. (2001). Bacterial molecular phylogeny using supertree approach. *Genome Inform.* **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 Res.* **12,** 1080–1090.

Davies, T. J., Barraclough, T. G., Chase, M. W., Soltis, P. S., Soltis, D. E., and Savolainen, V. (2004). Darwin's abominable mystery: Insights from a supertree of angiosperms. *Proc. Natl. Acad. Sci. USA* **101,** 1904–1909.

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

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

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'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 369–388. Kluwer Academic, Dordrecht, The Netherlands.

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

Gittleman, J. L., Jones, K. E., and Price, S. A. (2004). Supertrees: Using complete phylogenies in comparative biology. *In* ''Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.

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

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

Huson, D. H., Nettles, S. M., and Warnow, T. J. (1999a). Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J. Comput. Biol.* **6,** 369–386.

Huson, D. H., Vawter, L., and Warnow, T. J. (1999b). Solving large scale phylogenetic problems using DCM2. *In* ''Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology'' (T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, and R. Zimmer, eds.), Vol. 7, pp. 118–129. AAAI Press, Menlo Park, CA.

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). *Biol. Rev.* **77,** 223–259.

Källersjö, M., Farris, J. S., Chase, M. L., Bremer, B., Fay, M. F., Humphries, C. J., Petersen, G., Seberg, O., and Bremer, K. (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. *Pl. Syst. Evol.* **213,** 259–287.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* **38,** 7–25.

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

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.

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

Levasseur, C., and Lapointe, F.-J. (2003). Increasing phylogenetic accuracy with global congruence. *In* ''Bioconsensus'' (M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.), Vol. 61, pp. 221–230. American Mathematical Society, Providence, RI.

Page, R. D. M. (2002). Modified minifut supertrees. *In* ''Algorithms in Bioinformatics, Second International Workshop, WABI, 2002, Rome, Italy, September 17–21, 2002, Proceedings'' (R. Guigó and D. Gusfield, eds.), Vol. 2452, pp. 537–552. Springer, Berlin.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. (2001). The comprehensive microbial resource. *Nucleic Acids Res.* **29,** 123–125.

Piaggio-Talice, R., Burleigh, J. G., and Eulenstein, O. (2004). Quartet supertrees. *In* ''Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 173–191. Kluwer Academic, Dordrecht, the Netherlands.

Pisani, D., Yates, A. M., Langer, M. C., and Benton, M. J. (2002). A genus-level supertree of the Dinosauria. *Proc. R. Soc. Lond. B* **269,** 915–921.

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

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

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

Ronquist, F., Huelsenbeck, J. P., and Britton, T. (2004). Bayesian supertrees. *In* ''Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 193–224. Kluwer Academic, Dordrecht, the Netherlands.

Roshan, U., Moret, B. M. E., Williams, T. L., and Warnow, T. (2004). Performance of supertree mathods on various data set decompositions. *In* ''Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life'' (O. R. P. Bininda-Emonds, ed.), Vol. 3, pp. 301–328. Kluwer Academic, Dordrecht, the Netherlands.

Salamin, N., Hodkinson, T. R., and Savolainen, V. (2002). Building supertrees: An empirical assessment using the grass family (Poaceae). *Syst. Biol.* **51,** 136–150.

Sanderson, M. J., Purvis, A., and Henze, C. (1998). Phylogenetic supertrees: Assembling the trees of life. *Trends Ecol. Evol.* **13,** 105–109.

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

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

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

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

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

Zimmer, E. A., White, T. J., Cann, R. L. and Wilson, A. C. (eds.) (1993). ''Molecular Evolution: Producing the Biochemical Data.'' *Methods Enzymol.* **224**, 3–725.

[39] Maximum-Likelihood Methods for Phylogeny Estimation

By JACK SULLIVAN

Abstract

Maximum-likelihood (ML) estimation of phylogenies has reached a rather high level of sophistication because of algorithmic advances, improvements in models of sequence evolution, and improvements in statistical approaches and application of cluster computing. Here, I provide a brief basic background in application of the general principle of ML estimation to phylogenetics and provide an example of selecting among

 0076-6879/05 $35.00