

L. A. S. JOHNSON REVIEW No. 9

Construction and annotation of large phylogenetic trees

Michael J. Sanderson

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA.

Abstract. Broad availability of molecular sequence data allows construction of phylogenetic trees with 1000s or even 10 000s of taxa. This paper reviews methodological, technological and empirical issues raised in phylogenetic inference at this scale. Numerous algorithmic and computational challenges have been identified surrounding the core problem of reconstructing large trees accurately from sequence data, but many other obstacles, both upstream and downstream of this step, are less well understood. Before phylogenetic analysis, data must be generated *de novo* or extracted from existing databases, compiled into blocks of homologous data with controlled properties, aligned, examined for the presence of gene duplications or other kinds of complicating factors, and finally, combined with other evidence via supermatrix or supertree approaches. After phylogenetic analysis, confidence assessments are usually reported, along with other kinds of annotations, such as clade names, or annotations requiring additional inference procedures, such as trait evolution or divergence time estimates. Prospects for partial automation of large-tree construction are also discussed, as well as risks associated with ‘outsourcing’ phylogenetic inference beyond the systematics community.

Introduction

Phylogenies with more than 1000 species are becoming increasingly commonplace (Tehler *et al.* 2003; Ley *et al.* 2005) and botanists have been at the forefront of this trend. Not only are many phylogeneticists interested in constructing as grand a synthesis for the tree of life as possible, but other biologists are increasingly integrating the broad knowledge base of comparative biology into these increasingly complete trees (e.g. Moles *et al.* 2005; Chave *et al.* 2006). Chase *et al.*'s (1993) pioneering reconstruction of a phylogeny of 500 angiosperm species on the basis of plastid *rbcl* data had a profound impact. Not only did it ultimately encourage a revolution in angiosperm classification (Angiosperm Phylogeny Group 2003), but it also prompted sustained interest in the computational aspects of large-tree reconstruction (e.g. Rice *et al.* 1997) and instigated an intense and highly successful research program based on sequencing other loci for comparably large sets of plant taxa (Soltis *et al.* 1997, 1999; Qiu *et al.* 1999, 2005). Interestingly, however, after Källersjö *et al.*'s (1998) construction of a much larger *rbcl* tree a few years later, on the basis of 2500+ sequences, the plant systematics community did *not* invest much effort in iteratively building even larger trees from this one locus (although see Janssen and Bremer 2004). More than

20 000 *rbcl* sequences are now in GenBank, but as yet no one has made a phylogeny from this entire dataset. Perhaps this is because of technological obstacles of building alignments and trees at that scale, or more likely, because of lack of confidence in the power of a single gene to provide sufficient resolution. The ‘supergene’ approach has given way to that of the ‘supermatrix’ (combined from alignments of sequences) and ‘supertree’ (combined from trees), signalling recognition that data analysed at very large scales might have to be handled differently. This review addresses strategies for very large-scale phylogenetic inference. Note, however, that datasets for phylogenetics can get large in two directions; namely, in the number of taxa, N , or the number of characters or sites in sequence data, C . Generally, increasing N has much more serious computational consequences than increasing C . However, new tools, especially genomics technologies, are rapidly increasing C and may help leverage robust analyses at increasingly large values of N . For this reason, I include discussion of the genomics side of data analysis, even though it is only beginning to affect large-tree construction.¹

The amount of data already available for large-tree reconstruction is surprisingly large. As of January 2007, GenBank (Release 157) archived sequence data from 165 000

¹Reflecting on one of many raging arguments over phenetic systematics in the late 1960s, L.A.S. Johnson argued that problems of homology (‘matching’) would not all be whisked away by large oceans of data: ‘... even if we knew the entire nucleotide sequences over a set of organisms we should still have to make many decisions on matching...’ (Johnson 1970: p. 227, based on his presidential address for the Linnean Society of New South Wales in 1968). At the time the prospects for studying such complete genome sequences must have seemed remote. Now the data are here, and the newest genomics technologies (e.g. 454 Life Sciences’s FLX system) promise to deliver 100 million base pairs of sequence in an eight hour run (50 chloroplast genomes or one entire *Arabidopsis* genome...). However, the number of ‘decisions’ to be made regarding the analysis of such data has grown along with the quantity of information.

species, just under 10% of described global species diversity. For vascular plants, ~62% of the genera in Mabberley (1987) have at least one sequence in GenBank. Or, from a geographic perspective, almost half the species in the California flora, one of the world's best-studied biodiversity hotspots, have at least one sequence in GenBank (48%, counting as a match possibly different *infraspecific* taxa present within the same species). Even the fraction of the species in the flora of Gunung Palung (species list provided by C. O. Webb, pers. comm.) in a much less well inventoried part of Indonesia, with molecular data in GenBank, is surprisingly high, at 17%. Of course, these fractions are dominated by widespread species rather than narrow endemics. Nonetheless, they suggest that global biodiversity is indeed being captured, albeit in a biased fashion, by molecular sequencing efforts. TreeBASE, an online database of phylogenetic trees, includes more than 67 000 species among its 4000+ phylogenies.

The availability of *some* data for a large subset of global species diversity, of course, does not guarantee their *sufficiency* for large-scale tree inference. Indeed, there is reason to think that something like Pareto's rule applies to phylogenetic inference, so that one would not be surprised to find that, say, 80% of the tree of life can be reconstructed relatively easily but that the remaining 20% will be much harder to reconstruct. In any phylogeny, most of the nodes are near the tips and relatively few are ancient, and it is often the ancient ones that are recalcitrant. Assessing the quality of phylogenetic trees when they are large, or are assembled piecemeal, raises new difficulties and challenges.

Once trusted, large-scale phylogenies are reconstructed, they can be annotated. Perhaps the most common annotation is nomenclatural. The systematics community is presently wrestling with procedures for integrating phylogeny and nomenclature, and although no consensus is evident about the proper approach (e.g. Godfray and Knapp 2004; Laurin *et al.* 2005), there seems little doubt that phylogenies will both influence such work and themselves be annotated by its products. Other kinds of annotations require more quantitative analysis, and some may even emerge in parallel with tree-building itself. For example, there has been much interest in estimating divergence times of nodes in phylogenetic trees—and thus annotating trees with respect to geologic age. Other kinds of annotations include biogeographic history, ancestral-state reconstructions of traits, histories of gene or genome duplications or lateral transfers, inferences about diversification rates and their shifts across a tree, and even bioclimatic reconstructions or niche models of ancestors (Yesson and Culham 2006). Some of these are computationally challenging problems in their own right, even in the context of a *given* phylogenetic tree. Others are relatively straightforward but may benefit from integration with tree-building itself (such as attempts to simultaneously infer trees and divergence times with a relaxed molecular clock; Drummond *et al.* (2006)).

The present paper reviews some of the methodological and computational issues that are raised by recent attempts to scale phylogenetic inference up to the level that is allowed by existing data. Some of these problems are already well known and are merely exacerbated at this scale, others are novel, caused by heterogeneity of data, unprecedented scale of the trees or

the interaction of systematic methodology and sampling with the end products of a complex and as yet poorly understood evolutionary process.

Sources of evidence for large-tree reconstruction

Single loci

The largest phylogenies of plant taxa have been reconstructed by using one or a few highly conserved genes that have been sequenced with universal or nearly universal PCR primers. Higher-level phylogenetic analysis of plants has benefited tremendously from the slow rate of evolution of chloroplast protein-coding and nuclear rRNA genes. Concerted cooperative efforts among many investigators have effectively parallelised the workload and provided datasets of 1000–10 000s of taxa, most notably for plastid *rbcL*, *matK*, *ndhF* and 18S nuclear rDNA loci (Table 1). More rapidly evolving loci, such as internal transcribed spacers (ITS) of nuclear rDNA or various chloroplast spacer regions, have been sequenced even more widely in plants; however, difficulties of alignment have precluded their assembly into datasets as large as some of those assembled for protein-coding genes or rRNA genes (although see McMahon and Sanderson 2006 for one strategy).

Genomics

Although their impact on plant phylogenetic reconstruction has been limited, genomics data are becoming increasingly valuable for this purpose. The number of publicly available completely sequenced plant nuclear genomes is still small (*Arabidopsis thaliana*, *Oryza sativa* and *Populus trichocarpa*); however, another half dozen are nearing completion. These provide almost exhaustive coverage of the whole genomes of plants, limited only by the experimental design of the genomics pipeline and the quality of the assembly. The latter can be problematic in plant genomes for many reasons, including presence of highly repetitive DNA that is rampant in some large plant genomes. On the other hand, complete genomes of organelles have already proven very useful for phylogenetic efforts. Some 57 land-plant plastid genomes and 14 land-plant mitochondrial genomes have been fully sequenced. Analyses of the protein-coding genes of plastid genomes have formed the basis of numerous reconstructions of angiosperm phylogenies (Goremykin *et al.* 1997) and some interesting controversies. For example, the arguments in the literature about the position of *Amborella* as

Table 1. Numbers of sequences in GenBank (Release 157) for selected genes or non-coding regions widely used in plant phylogenetic reconstruction

Locus	Number of sequences
ITS	69 066
<i>trnL</i>	48 765
<i>rbcL</i>	25 721
18S rDNA	22 830
<i>matK</i>	19 562
<i>atpB</i>	9985
<i>ndhF</i>	8240
<i>psbA</i>	5588
<i>LEAFY</i>	5292
<i>rpl16</i>	4519

the sister group of the remaining angiosperms (Goremykin *et al.* 2003) have revealed a sensitivity to both the method of tree reconstruction and to taxon sampling, despite the presence in these datasets of 50 genes or more (Leebens-Mack *et al.* 2005). Incongruence and sensitivity to inference method may in fact be surprisingly common in phylogenomic analyses (Jeffroy *et al.* 2006). Non-coding regions in plastid genomes and the highly complex mitochondrial genomes have as yet been relatively untouched, at least at a scale of the whole genome, although specific genes, spacers and introns have been widely used. The relatively high rate of horizontal gene transfer observed between angiosperm mitochondrial genomes of distantly related plant species obviously limits the utility of some of these markers for species-level inference (Mower *et al.* 2004).

Libraries of expressed sequence tags (ESTs) are much easier and cheaper to produce than whole nuclear genomes. They consist of small sequences derived from messenger RNA of protein-coding genes. ESTs must be assembled into larger consensus sequences to be useful for phylogenetic purposes. This assembly is complicated computationally by alternative splicing and by gene duplication, both of which can generate ambiguous assemblies (Dong *et al.* 2005). Moreover, the libraries are derived from the transcriptomes of specific tissues, and may be biased towards certain subsets of the genome. Relatively few studies have exploited these data; however, this may be changing (Schlueter *et al.* 2004; de la Torre *et al.* 2006; Sanderson and McMahon 2007). For example, by using seven large EST libraries for angiosperms (including pine as an outgroup) and culling through tens of thousands of EST sequences to find sets of homologues with phylogenetic signal, Sanderson and McMahon (2007) recovered the now well corroborated tree for these taxa. The potential is certainly great. Some 36 plant species have EST libraries with more than 50 000 sequences, and 132 species have significant EST libraries available (<http://www.ncbi.nlm.gov>). Moreover, although expensive, the cost is certainly much closer to the scale that molecular systematics has operated in the past, than is the cost of whole-genome sequencing.

Intermediate in cost and coverage between these two kinds of genomic resources are efforts based on BAC-end sequences (McCubbin and Roalson 2005). In these libraries the genome is randomly broken into relatively large pieces and inserted into numerous 'bacterial artificial chromosomes', and then just the ends of these sequences are sequenced. Given enough partially overlapping BACs, a significant fraction of a genome can be sequenced or primers can be designed to permit sequencing of homologues in close relatives. Large-scale sequencing efforts have generated in the order of 100 000 BAC-end sequences for rice and a dozen of its wild relatives (Ammiraju *et al.* 2006), which have been put to use to build a phylogeny of *Oryza* (B. Hurwitz, pers. comm).

Database mining

Few molecular phylogenetic studies rely exclusively on newly generated data. Molecular-sequence databases almost always contain relevant data for outgroup taxa at least. The taxonomic breadth of these databases provides a natural source of data for large-scale tree reconstruction. This taxonomic diversity is biased strongly towards model organisms and economically

important taxa, the usual targets of both traditional molecular biology and more recent genomics projects. However, the bulk of sequence data for most of the other species in the databases, probably deposited mainly by systematists and other evolutionary biologists, forms a long taxonomically enriched 'tail' in GenBank's distribution of sequences (Fig. 1). These diverse taxa all have relatively few sequences in the database, but collectively they form a potentially less biased resource than what a first glance at GenBank suggests. A new 'view' on the taxonomic diversity of GenBank is provided by our Phylota Browser database (<http://loco.biosci.arizona.edu/cgi-bin/pb.cgi>; Fig. 2), which assembles sets of homologous sequences suitable for use in phylogenetic analysis. For Release 157 of GenBank, the database consists of some 615 000 clusters, of which 57 000 are phylogenetically informative, meaning they contain sequences for at least four distinct taxa. Of course, these are not guaranteed to produce well supported phylogenetic trees, although they have that potential. It is worth noting that more than 1700 of these clusters are 'large', containing sequences for more than 100 distinct taxa, and many more clusters even larger than this remain undiscovered because the database limits assembly to clusters of 20 000 sequences as of this writing.

However, data-mining studies, such as the analysis of papilionoid legume diversity of McMahon and Sanderson (2006), confront several problems that are more serious than they would be if the investigators generated the data themselves. Mistaken annotations are a widely cited problem with GenBank (Vilgalys 2003; McMahon and Sanderson 2006), especially mistaken taxonomic identifications, which are among the most difficult to detect from the properties of the sequence data themselves (unlike, say, mistaken annotations about gene names or sequence features). The data of McMahon and Sanderson (2006) suggested a taxon identification error rate of as much as

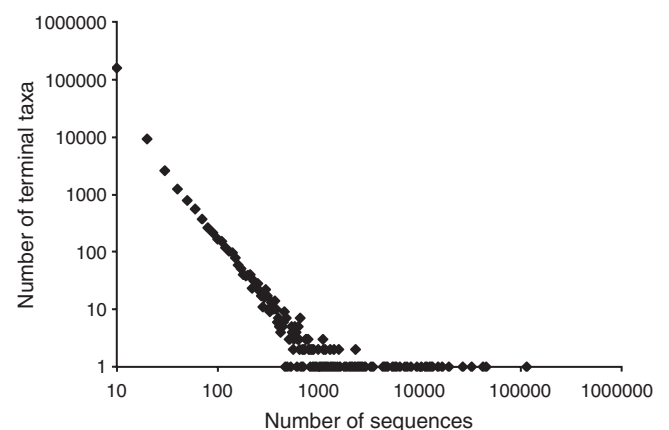


Fig. 1. The highly skewed distribution of sequences among taxa in GenBank (Release 157). Horizontal axis is the number of sequences per taxon (numbers are the upper boundary of bins that are 10 units wide). Vertical axis is the number of terminal taxa in the NCBI taxonomy tree having that many sequences per taxon. Log scale de-emphasises the dramatically skewed distribution but is necessary to plot the data. Data are from the taxonomically defined eukaryotic divisions of GenBank excluding primates and rodents, and also excluding high throughput, EST, and genome survey sequences, all of which are dominated by sequences from model organisms. In other words, the real skew across GenBank as a whole is even greater.

Sequence diversity and cluster set summaries

Sequence tallies include those from “model” organisms. To *exclude* model organisms, click [here](#)

NCBI taxon name ¹	Descendant species ²	Sequences (GIs)	Seq. clusters ³	Phylog. inform. seq. clusters ⁴
Mirbelieae up to Papilionoideae	276	811	<u>13</u>	<u>9</u>
<i>Almaleea</i>	1	3	<u>3</u>	0
<i>Aotus</i>	5	10	<u>4</u>	<u>1</u>
<i>Brachysema</i>	9	51	<u>6</u>	<u>6</u>
<i>Callistachys</i>	1	14	<u>8</u>	0
<i>Chorizema</i>	12	23	<u>4</u>	<u>2</u>
<i>Daviesia</i>	45	56	<u>5</u>	<u>2</u>
<i>Dillwynia</i>	2	5	<u>3</u>	0
<i>Erichsenia</i>	1	2	<u>2</u>	0
<i>Euchilopsis</i>	1	3	<u>3</u>	0
<i>Eutaxia</i>	1	2	<u>2</u>	0
<i>Gastrolobium</i>	51	165	<u>7</u>	<u>5</u>
<i>Gompholobium</i>	3	6	<u>3</u>	0
<i>Isotropis</i>	3	17	<u>8</u>	0
<i>Jacksonia</i>	3	12	<u>6</u>	0
<i>Jansonia</i>	1	6	<u>6</u>	0
<i>Latrobea</i>	3	11	<u>6</u>	0
<i>Leptosema</i>	2	4	<u>2</u>	0
<i>Mirbelia</i>	16	43	<u>6</u>	<u>2</u>
<i>Nemcia</i>	20	86	<u>7</u>	<u>5</u>
<i>Oxylobium</i>	7	35	<u>6</u>	<u>5</u>
<i>Phyllota</i>	1	5	<u>5</u>	0
<i>Podolobium</i>	6	32	<u>8</u>	<u>4</u>
<i>Pultenaea</i>	76	209	<u>7</u>	<u>3</u>
<i>Sphaerolobium</i>	3	6	<u>2</u>	0
<i>Stonesiella</i>	1	1	<u>1</u>	0
<i>Urodon</i>	1	2	<u>2</u>	0
<i>Viminaria</i>	1	2	<u>2</u>	0

¹Names refer to node and its subtree unless the term “node only” appears.

²Taxa at species level rank as annotated by NCBI.

³Italicized cluster totals refer to internal nodes having too many sequences to exhaustively cluster, so only a sample of sequences were clustered.

Non-italicized totals refer to clusters that were exhaustively built. For further information of about this sampling scheme for deep nodes see [here](#). A dash (-) means there were too many sequences to clusters or no sequences at all.

⁴Phylogenetically informative clusters have four or more taxa (not GIs) represented.

Fig. 2. Screen shot of one page from the Phylota Browser database of phylogenetically informative sequence clusters (<http://loco.biosci.arizona.edu/cgi-bin/pb.cgi>). Page is shown for the Australian legume tribe Mirbelieae (as circumscribed in the NCBI taxonomy). Clusters are sets of homologous sequences from different loci that could be aligned and submitted to phylogenetic analysis. Clusters were identified by an informatics pipeline consisting of all-by-all BLAST searches with stringent ‘Expect’ values, requiring 51% coverage on all hits, followed by single-linkage clustering of hit results. Phylogenetically informative clusters are those having at least four distinct taxon names among the sequences present. The database provides clusters across all eukaryotes.

7% for species of papilionoids, much lower than that cited for fungi (20%), but enough to cause significant problems in the data assembly and phylogenetic analysis. An increased incidence of sequence submissions that cite voucher specimens is a positive trend in GenBank in recent years.

Another problem is the heterogeneous sample of taxa and loci that is ultimately combined in a supermatrix or supertree

analysis. Since the data arrive in the database from many sources and many investigators, it is usually impossible to assemble a ‘complete’ matrix of species by loci—i.e. one with no missing sequences. This leads to datasets that are ‘fragmented’ into islands of data separated by seas of question marks (Sanderson *et al.* 2007; Table 2). The effects of this on phylogenetic analyses can range from benign to nearly catastrophic. For example, it is

Table 2. Hypothetical example of a combined alignment of gene fragments sampled from different collections of taxa, and for which no new phylogenetic conclusions are possible beyond those possible from the separate alignments

For example, there is no inference possible about a quartet relationship of any two species sampled from Gene 1 together with any two species from Gene 2 (e.g. Taxa A, B, E and F)

Taxon	Gene 1	Gene 2
A	acgtcccatgtatgt	???????????????
B	acgtcccatgaacgt	???????????????
C	acgtcccatgtatgt	???????????????
D	acgtcccatgaacgt	???????????????
E	acgtcccatgtatgt	ttaagctctccccc
F	???????????????	ttaagctctccccc
G	???????????????	ttaagctctccacc
H	???????????????	ttaagctctccaccg
I	???????????????	ttaagctctccaccg

possible to combine several sets of data in such a way that they shed absolutely no mutual phylogenetic light on one another—therefore wasting the computational effort required to handle the combined dataset (Table 2).

An alternative to downloading data for all the taxa in some clades with species in GenBank is to use formal algorithms to assemble such data, targeting certain properties of the final dataset. A simple example is to search for and enumerate maximal complete datasets; that is, the largest collections of taxa and genes that are fully sampled in the database (every taxon has every gene). Sanderson *et al.* (2003) and Driskell *et al.* (2004) explored these strategies (see these papers for visualisations). For sparse databases such as GenBank, it is often possible to quickly enumerate maximal complete datasets (they tend to be small), and then to use these as starting points for more ‘lacunose’ data matrices that permit some tolerable level of missing data (Yan *et al.* 2005).

Literature

An increasingly popular starting point for reconstructing large phylogenetic trees is a set of trees obtained from the literature, from databases of species trees such as TreeBASE or the Tree of Life Web Project, or other databases of gene trees, several of which now have phylogeny servers associated with them (e.g. the Pfam database). These studies use supertree methods (see below) to piece together the trees. Davies *et al.*'s (2004) 379-taxon angiosperm supertree is a recent example, but many others have been constructed for smaller clades, including pines (Grotkopp *et al.* 2004), temperate herbaceous legumes (Wojciechowski *et al.* 2000), grasses (Salamin *et al.* 2002) and so on. Empirical work has advanced somewhat slowly, in part because of relative inaccessibility of the data—the trees in the literature usually have to be manually extracted, but the increase in the rate of deposition of trees in the TreeBASE database promises to make these resources more widely available.

Tree inference

Single-gene datasets: scaling computation to large trees

Most of the plant phylogenetics community's experience with large phylogenetic trees has been with taxon-rich single-gene or

few-gene datasets, such as the large angiosperm trees, based on *rbcL* and a few other loci. Methods that rely on optimisation—searching among trees for one with a best ‘score’, such as parsimony or maximum likelihood—belong to a class of hard problems for which no algorithms are known that are guaranteed to find the best tree in anything less than a running time proportional to e^N , where N is the number of taxa, for worst-case datasets. Unfortunately, experience suggests that many phylogenetic datasets are close enough to ‘worst case’, so that algorithms that are guaranteed to find optimal trees, such as the ‘branch and bound’ method, simply never finish for datasets with more than 20–25 taxa.

Thus, the game in large-tree reconstruction is either to use approximate algorithms for optimisation methods (‘heuristics’), or non-optimisation methods, such as neighbour-joining. For many years, the number of heuristics available has been limited, but recently new heuristic strategies (e.g. Goloboff 1999, implemented in TNT; RaxML: Stamatakis *et al.* 2005; disk-covering: Huson *et al.* 1999) promise significant speed-ups. However, it is hard to imagine effective heuristics that do not require a running time that is at least proportional to N^2 , equivalent, for example, to building a tree by sequential addition with no branch-swapping, such that each taxon is provisionally added to the existing tree in every possible place before the next taxon is added. As trees grow to tens or even hundreds of thousands of sequences, even such very crude ‘polynomial-time’ heuristics begin to take a toll on computational resources (Bininda-Emonds *et al.* 2001). Nonetheless, there is a better chance that technology can keep up with algorithms that scale as N^2 than with those that scale exponentially.

Accuracy

The question of how accurate tree-inference methods are in general, and how that accuracy scales to large trees, has received considerable attention, both in simulation studies and more recently from theory. Early work focused on accuracy of different methods in small trees and led to the conclusion that a few methods (UPGMA, Lakes invariants) had poor performance, whereas other methods were largely indistinguishable in a large part of the reasonable set of evolutionary model conditions (Hillis *et al.* 1994). The main exception to this is the failure of parsimony methods in the so-called ‘Felsenstein zone’, a veritable black hole drawing in infinite amounts of data but never returning the correct answer (Felsenstein 1978). Later work showed analogous conditions under which maximum likelihood could be misled, particularly when the model of evolution used for inferring the tree did not closely match the model under which the characters evolved. One such case occurs when the true evolutionary process consists of a mixture of very different processes but the model used in inference does not (Chang 1996; Kolaczowski and Thornton 2004). In practice it has proven difficult to positively detect these pathological cases, although some studies have shown ‘long-branch-attraction’ effects through simulation and parametric bootstrap procedures (Huelsenbeck 1997; Sanderson *et al.* 2000).

A more relevant issue to this review is the question of how accurately large trees can be reconstructed, compared with smaller ones (Kim 1998). This question received considerable attention after Hillis (1996) described a large phylogeny

that could be reconstructed accurately with surprisingly few characters. Interestingly, this was based on the newly available large dataset of 18S sequences for angiosperms. Much debate followed about whether ‘adding taxa’ or ‘adding characters’ was the best way to improve accuracy (Graybeal 1998). However, this debate often conflated two separate questions. The first concerns how accurately a tree of a given size can be reconstructed by using data exhibiting certain properties. The second question concerns, given a set of data and a set of taxa *sampled from* a larger tree, what the best strategy is to improve accuracy (Kim 1998). In both questions it is important to distinguish the accuracy of some *specific* phylogenetic hypothesis about the tree (e.g. the presence of a specified quartet relationship between four specified taxa) from the average overall accuracy of the tree. As a tree gets larger (whether by more sampling or by expanding its depth), obviously it requires more information to specify all the details about its structure.

Some important recent theoretical results can at least place bounds on accuracy (reviewed in Mossel and Steel 2005). One way to state them is in terms of k_{\min} , the minimum number of characters needed to reconstruct the tree correctly with some specified probability. Given a lower bound on the rate of evolution across the tree, l_c (which can be thought of as the ‘length’ of the shortest branch), and depending on the shape of the tree, measured by the ‘diameter’ or length of the longest path in the tree, this number ranges from a worst case bound of

$$k_{\min} < cN^b \times \ln(N)/l_c^2, \quad (1)$$

which is polynomial in N , to

$$k_{\min} < c(\ln(N))^b/l_c^2, \quad (2)$$

which is polynomial in $\ln N$ and therefore grows more slowly (‘sublinearly’) than N . Here, b and c are constants that depend on the model and tree shape. Mossel and Steel (2005) conjectured that $b = 1$ (it can be proven to hold for completely balanced trees with a few other assumptions); however, this remains unproven. In either case, a sublinear dependence on N means it is not necessary to add characters proportional to the number of taxa, debunking an old rule of thumb in the literature that three characters per taxon are needed for a reliable tree (Sneath and Sokal 1973: p. 351). One piece of bad news is that in either case, k_{\min} will grow as the inverse *square* of the shortest, most difficult branch length, which may well explain the existence of so many polytomies that remain recalcitrant to phylogenetic inference.

Interestingly, if one is satisfied with inferring ‘most’ of the tree correctly, Mossel (2007) showed that it is possible to reconstruct the shallow parts of the tree, which after all include most of the nodes, with K_{\min} proportional to $\ln N$. This provides theoretical explanation of some simulation results we obtained earlier (Bininda-Emonds *et al.* 2001) that implied it could be easy to reconstruct most of the tree while simultaneously difficult to get a few deep nodes right.

To understand the second question it is important to note that a tree can ‘get’ larger in terms of number of taxa in the following two ways: by ‘infilling’ with more taxa sampled from the tree but keeping the most recent common ancestor the same; or ‘outfilling’ by making the root of the tree deeper. Both outfilling and infilling can increase the diameter or decrease the length of the shortest branch, depending on the distribution of branches on the tree. Only by making certain assumptions about

the tree, can a general trend in accuracy be predicted. Without these strong assumptions about branch lengths, it is impossible to know whether adding taxa will improve or degrade accuracy (Kim 1998).

Perhaps this can all be summed up by asserting that much of any clade will be reconstructed with a reasonable investment of characters, probably even a sublinear one; however, the remainder may require an unreasonable investment, making the consideration of cost per character an important factor.

Supermatrix approaches

With still very few exceptions (e.g. see *rbcl* and ITS in Table 1), the number of taxa sequenced for any given locus is still small with respect to the entire tree of life. This has prompted assembly of multiple loci into datasets that cover more taxa at the expense of having missing data for some taxa. The combination of several multiple-sequence alignments into a single giant alignment is referred to variously as a ‘supermatrix’ or ‘superalignment’. The basic structure of a supermatrix can be described by an auxiliary construct called a ‘data availability matrix’, which is also useful in a workflow setting for describing which data remain to be sequenced. It is simply a matrix of taxa by loci, in which a ‘1’ entered in a cell indicates the presence of a sequence for that locus and taxon, whereas a ‘0’ indicates absence. The ‘density’ of the supermatrix is then the fraction of cells in the data-availability matrix having a ‘1’ (Sanderson *et al.* 2007).

Although many phylogenetic projects are designed to build high-density many-locus datasets, these obviously require significant investment of resources if the goal is to build a large tree. Larger trees can be built more inexpensively by assembling data with lower densities. Indeed, almost all such mining studies tolerate some level of missing data, and this level tends to increase with the number of taxa included. For example, many studies of the order of a 100 taxa have been assembled with densities of 1/3 or higher (Baptiste *et al.* 2002); however, a study of McMahon and Sanderson (2006) of 2000+ papilionoid species had a density of only 5%.

Low-density supermatrices present several problems. Large numbers of question marks make life hard for heuristic search algorithms because so many alternative trees have the same score. Supermatrices can also become fragmented in ways that either limit the amount of new information that can be obtained (Sanderson *et al.* 2007; see Table 2) or conspire with other defects in the data to generate pathologies, such as rogue taxa that can roam the tree Flying-Dutchman-like (McMahon and Sanderson 2006). Combining alignments may even not be possible for non-coding regions such as ITS that form the workhorse of lower-level phylogenetic inference (in plants), forcing them to be kept separate and contributing thereby to decreases in overall density. Nonetheless, taxon-rich supermatrices can have a very low density and still generate relatively robust hypotheses that combine deep phylogenetic inferences with shallow ones (McMahon and Sanderson 2006).

Supertrees

Supertrees are trees assembled from other, smaller trees (reviewed in Bininda-Emonds 2004a, 2004b). Supertrees are distinguished from consensus trees, which are trees assembled from a collection of trees all having the *same* set of taxa. Sometimes supertrees are assembled more or less manually,

as with the protocols implemented in the phylomatic site (Webb and Donoghue 2005; <http://www.phylodiversity.net/phylomatic/phylomatic.html>) which integrate a 'trusted' backbone tree of plants with lower-level phylogenies of specific taxa. These have permitted comparative analyses of thousands of plant species (Moles *et al.* 2005; Chave *et al.* 2006). However, formal algorithms have also been developed, in which supertrees are built in a multi-step procedure where individual 'input' trees are built first and then these trees are combined. Supertrees have been constructed for many reasons, including conflict resolution, building phylogenies for a group that include all its species and building very large trees—or all of the above simultaneously. The methodology has evolved quickly in the last few years and attracted the renewed interest of computer scientists, who first developed the techniques as a tool in database queries (Aho *et al.* 1981). When the input trees are all compatible—meaning that there exists one or more supertrees that display all the input trees as subtrees—it is possible to build supertrees in polynomial time. See Bininda-Emonds *et al.* (2002) for an accessible description of one algorithm. In the real world, input trees tend not to agree and other algorithms are necessary. The most widely used algorithm is matrix representation with parsimony (MRP: Baum 1992), which first converts each input tree to a matrix of binary characters in which each column represents a clade on that tree, then combines the input matrices into a 'supermatrix' with whatever question marks are necessary to fill it in and finally reconstructs a supertree via parsimony analysis of that 'supermatrix'. The only homoplasy in this matrix arises from conflicts *between* the input trees, which may be much less than in a character (super)matrix derived from the original data used to build the input trees.

The strengths of the method are also its weaknesses. Because the inputs are trees, some of the complexity of the underlying data used to build those trees is removed before supertree analysis. This can translate into a much faster tree reconstruction than the corresponding supermatrix approach, but the loss of signal contained in the original data may mislead downstream analyses, especially if weak signals from individual datasets might have emerged when combined into a supermatrix. Criticisms that information about the reliability of the input trees is lost in supertree construction are less compelling. Reliability scores from the inputs can be incorporated into the supertree algorithms, and several supertree bootstrap procedures have now been developed (Burleigh *et al.* 2006; Moore *et al.* 2006).

Interestingly, supertree construction raises similar problems of fragmentation as does sampling from the sequence databases; the pattern of taxonomic overlap is important (Ané *et al.* 2006). For example, if two trees share only one species in common (as in Table 2), the set of supertrees that displays each of those input trees reveals nothing new about relationships between the taxa on the different trees. On the other hand, if two rooted input trees share *two* species then new information can be gleaned (Ané *et al.* 2006). However, this is a sufficient but not a necessary condition; *four* rooted trees that share only *one* taxon in common between every pair can lead to novel statements in the supertree (see Ané *et al.* 2006 for additional insights on these conditions).

An interesting recent advance which may help overcome some of these problems of partial taxonomic overlap of input trees is an algorithm that respects labelled internal nodes on

the input trees (Semple *et al.* 2004). Input trees may share no terminal taxa despite the fact that they each have representatives of some well known larger clade (e.g. taxa sampled from different clades of eudicots). If some knowledge of clade membership is assumed by annotating internal nodes, this information can help constrain the supertree analysis. As always, there will be risks in assuming knowledge of clade membership, but certain clades have achieved a level of credibility that might warrant this approach (legumes, eudicots, angiosperms, seed plants and so on). This is in some ways analogous to enforcing the monophyly of certain taxa in a conventional character-based (or supermatrix) analysis.

Supertrees from gene trees

A final topic that fits technically within any discussion of supertrees is the problem of constructing a species tree from one or more gene trees. Gene trees may be subject to a variety of processes making them incongruent or otherwise difficult to assemble into species trees, including lineage sorting, hybridisation, lateral transfer and, most challenging, gene or genome duplication and loss. The topic can be considered one of supertree construction because the basic unit of analysis is the (gene) tree, and the taxa included in the gene trees may well be fewer than those in the set of all gene trees together. What makes it stand on its own, historically as well as algorithmically, is the frequent presence of duplicate representatives of the same taxon, usually owing to gene duplication or multiple alleles.

When multiple copies of the same locus are present in taxa and these have arisen by duplication deep in the history of the collection of species (i.e. before the last speciation event), it is clearly not possible to directly infer species trees from the sequence data. The genes within a single species may be only distantly related by an ancient duplication event; the presence of homology between genes present in the same genome can promote recombinational processes that cloud the ancestry of the gene tree, and deletion of copies of the gene can paint a false picture of relationships among the surviving sequences. Goodman *et al.* (1979) proposed a method of inferring species by minimising the number of duplications that must be inferred in a collection of gene trees. The number will vary among all possible rooted species trees, providing a criterion for species-tree preference. Relative to conventional species-tree inference, fewer algorithms and software tools are available for building species trees in this way, but the topic is receiving renewed interest (Page 1998; Arvestad *et al.* 2003; Durand *et al.* 2005; Bansal *et al.* 2007). Good reviews of the approach can be found in Page and Charleston (1998). Sanderson and McMahon (2007) recently performed an exhaustive search across all species trees in a small number of taxa by using highly duplicated plant EST libraries and recovered what is generally regarded as the correct angiosperm tree. Until recently, however, it has been difficult to scale up this kind of analysis. Bansal *et al.* (2007) improved the running time of tree search algorithms for gene tree parsimony significantly, allowing construction of much larger trees, including that for plants shown in that paper. Arvestad *et al.* (2003) explored statistical Markov-chain Monte Carlo approaches to the core problem of reconciling gene trees and species trees, and there is some hope that this will lead to fast species-tree inference procedures; however, as yet, none is available in this model-based inference framework.

Incremental tree construction

As trees get larger it becomes increasingly attractive to simply add new data to old analyses. One of the most complete explorations of this approach to large-tree inference is in the ARB project (Ludwig *et al.* 2004), which originally was focused on constructing very large bacterial and archeal trees from tens of thousands of 16S rDNA sequences. ARB's parsimony insertion feature tries to insert a new sequence into an existing tree in the most parsimonious way possible, with little or no rearrangement of the existing tree. This requires only about $O(N)$ running time, as opposed to the $O(N^2)$ running time of building the whole base tree from scratch, with sequential addition heuristics. Limited branch swapping can be invoked following the optimal placement of a new sequence to look for global optimality, but of course this has no guarantee of success. This general approach is finding new applications in DNA barcoding projects, such as the Witness for the Whales Project (Ross *et al.* 2003), which attaches user-submitted sequences to the appropriate part of an already constructed reference tree (among other options).

Confidence limits

Bootstrap methods for assessing confidence in clades are fairly sensitive to the size of the tree; bootstrap values for clades tend to decline as larger samples of that clade are taken ('infilling'—see above; Mort *et al.* 2000; Sanderson and Wojciechowski 2000; Salamin *et al.* 2003). For example, when we bootstrapped the entire dataset of 2228 species of papilionoid legumes, few clades were supported by high bootstrap values, despite agreement in the literature about high support for certain familiar clades (McMahon and Sanderson 2006). However, when taxa were pruned from the bootstrap trees (*after*, not before, bootstrapping), the fraction of trees exhibiting these familiar clades was much higher. This same phenomenon was documented in detail for *Astragalus*, a species-rich clade of legumes (Sanderson and Wojciechowski 2000). Bayesian methods build a different kind of confidence assessment directly into their search procedures and thus offer an attractive alternative to bootstrapping; however, the widespread perception that Bayesian posteriors for clades are also biased, in this case upward, must be resolved satisfactorily (Alfaro and Holder 2006). For either bootstrapping or posterior calculations, the length of time needed to converge either to something near the optimal tree, or to stationarity, respectively, must grow with the size of the tree and thereby may have an impact on the reliability of these assessments (see e.g. Sanderson and Wojciechowski 2000, for bootstrapping). One cannot escape the suspicion that none of the existing tools for confidence assessments is ready for scaling to very large phylogenies.

Conflict resolution

Regardless of methodology, constructing large datasets for phylogenetics raises the possibility that different subsets of the dataset may contain different signals about phylogenetic history. This might arise from any number of causes ranging from analytical artefacts (some genes suffer from long-branch attraction; others do not) to real differences in evolutionary history owing to lineage sorting, recombination, hybridisation, introgression, lateral gene transfer and gene duplication. Data

heterogeneity has been one of the best-studied phylogenetic issues of the last decade. (de Queiroz *et al.* 1995; Cunningham 1997; Rokas *et al.* 2003; Vogl *et al.* 2003).

Perhaps the first stage in any large-scale analysis of phylogenetic data is an assessment about the presence or absence of paralogous copies of sequence in the dataset. As discussed earlier, special phylogenetic algorithms are needed to reconstruct species trees when gene duplications are present, and most workers have opted therefore to either focus on single-copy loci from the get-go, or develop methods for excluding paralogs from the dataset (e.g. Storm and Sonnhammer 2002; Sanderson *et al.* 2003; Robbertse *et al.* 2006). Even if paralogs have been successfully excluded, conflict may still arise from a variety of sources. Thus, assembly of large supermatrices or supertrees would therefore ideally be followed by careful analysis of whether partitions of the data (or subsets of the input trees in supertree analyses) track different histories or different models of evolution. Unfortunately, the technology to do so efficiently is not yet in place. In supermatrix analysis, tests for data heterogeneity can be undertaken with established tools such as the ILD test (Farris *et al.* 1994, 1995; although note numerous statistical criticisms, e.g. Shimodaira 2002). These kinds of tests must now be extended to tens or hundreds of loci at a time. Enumerating all the subsets of loci that might each obey the same probabilistic model is computationally prohibitive, irrespective of the statistical approach chosen, and the alternative of simply allowing every partition to have its own very complex collection of parameters seems guaranteed to eventually run afoul of the bias–variance tradeoff in model selection (Burnham and Anderson 1998)—that is, increasing inaccuracy of tree estimation as the model complexity increases.

Thus, not surprisingly, most attempts to build large trees from numerous loci or input trees have just sidestepped the question of detecting data heterogeneity and taken an implicit 'total evidence' approach—probably without much deference to the extensive discussions in the systematics literature about the philosophical issues involved (e.g. Kluge 1989). It would not be surprising if the data heterogeneity problem reared its head yet again in the context of large-scale phylogenetics.

Annotation of large trees

General comments

Although the construction of large phylogenies, with an eye towards the ultimate goal of building the tree of life, has its own appeal to phylogeneticists, most biologists are interested in the information that can be conveyed directly by such trees (who's related to whom) and can be attached to it (names, branch lengths, times, rates, areas and so on). Indeed, the topological structure alone is something perhaps only a phylogenetic systematist could love. Scaling *annotation* to large trees raises some new problems for certain kinds of data, whereas other data are less affected. New software tools are available, aimed explicitly at easing the process of manual annotation and providing queries on the basis of those annotations (e.g. TreeDyn: Chevenet *et al.* 2006; and the phyloXML project: <http://www.phyloxml.org>). However, other kinds of annotations are expected to be more automated or driven by quantitative computational procedures themselves.

One generic issue that arises irrespective of the kind of annotation, however, and which does raise scaling problems, is the choice of whether to annotate a point estimate (a single tree) or a confidence interval (credible set) of trees. As the techniques of phylogenetics have become increasingly influenced by statistical model-based inference methods, the annotation of trees with non-phylogenetic information has had to confront a basic issue; namely, whether to annotate a single tree or a spectrum of the reasonable trees implied by the data. The former is easiest to communicate and use in downstream analyses; the latter is a more faithful portrayal of phylogenetic uncertainty. As trees become larger, however, the inclusion of a spectrum of trees in subsequent analysis becomes increasingly computationally expensive.

Nomenclature

Many phylogeneticists annotate trees with taxonomic names, almost always the names of clades. Generally, the choice of which clades to name has been rather arbitrarily based on considerations such as congruence with existing classifications, agreement with reconstructions of key character innovations, and so on. This is one setting in which phylogenetic uncertainty expressed in a Bayesian or other framework is not easily communicated. Ideally, names are added only if phylogenetic results can be presumed correct—the uncertainty having been set aside. Hibbett *et al.* (2005) arrived at an interesting answer to this problem by developing an automated system for applying phylogenetically precise definitions of taxa to large phylogenies of homobasidiomycetes that would change as the trees change. This could serve as an interesting model for nomenclatural annotation of large phylogenies. The systematist constructs a set of node-based definitions of taxa (lists of terminal taxa that together uniquely identify a most recent common ancestor and the clade descended from it), and the system automatically annotates the tree correctly on the basis of this definition, even as the tree increases in size or changes shape because of incorporation of new data. The system is limited only by the small computational problem of finding the most recent common ancestor of sets of taxa repeatedly, which can be done in constant time if the tree is pre-processed correctly (Bender and Farach-Colton 2000; Hibbett describes a somewhat slower algorithm). This particular form of nomenclatural annotation shares the schema expressed in the Phylocode, but it seems that any reasonable alternative schema might have similar properties.

Divergence times

Rivalling nomenclatural annotations of phylogenies in frequency in the last few years are estimates of divergence times. With the advent of methods that attempt to account for variable (non-clocklike) rates of molecular evolution, the rate of publication of time-calibrated phylogenetic trees has soared, despite numerous methodological hurdles that remain (see especially Britton 2005, reviewed in e.g. Rutschmann 2006). In the best cases, in which evolution is nearly clocklike, the estimation of divergence times scales as the square of the number of taxa for many likelihood-based methods (such as ML in PAUP or LF in r8s, assuming optimal quadratic convergence of the continuous function hill-climbing routines), and as linear in the

number of taxa for fast non-model based methods such as the mean path length method of Britton *et al.* (2002). It is likely that for trees in which rates of substitution vary substantially across the tree, rate-smoothing methods will not converge as fast as when a clock is present (nor as accurately!), making scaling divergence times to large trees somewhat problematic. One thing that should help is the inclusion of proportionally more fossil calibration points, to impose additional constraints on the inferences (Yang and Rannala 2006). However, this is obviously not feasible for many clades, with a paucity of such fossils to begin with. For example, Lavin *et al.* (2005), in their large analysis of divergence times in Leguminosae, a clade with 19 000 species and as many unknown internal node ages, scoured the literature to find even a small number of reliable fossils. The only saving grace for the problem of adding divergence times to large trees is the possibility that for some large clades the tempo of the rate change is sufficiently slow so that occasional calibration points added to the tree as needed will be sufficient to bound the divergence-time estimates. However, this is an open and entirely empirical question.

Until recently, divergence-time estimation was done *post hoc*, on either a single given tree or a spectrum of trees. However, recent work (Drummond *et al.* 2006) has shown how trees can be inferred simultaneously with divergence times, without making the strong and usually unwarranted assumption of a molecular clock.

Trait data

Many comparative biologists want to display and analyse trait information on large phylogenetic trees. For example, a recent analysis examined the correlates of seed-size variation across a tree of 13 000 species of seed plants (Moles *et al.* 2005). Software such as MacClade (Maddison and Maddison 2000) and Mesquite (Maddison and Maddison 2007) are well established and provide estimates of the evolutionary history of traits, both discrete and continuous, under various sets of assumptions, as well as nice visualisations. More specialised procedures may be necessary for features more loosely associated with the species' biology, such as their geographic range or climatic tolerances (Hardy and Linder 2005) or their membership in a specific community (Webb *et al.* 2006). Most of these reconstruction algorithms run very efficiently, usually in linear time, and thus scale well to large trees; however, the problem is increasingly one of archiving and visualisation rather than algorithm engineering. It is barely practical to annotate a large tree with some nomenclatural information and divergence times (if the tree is drawn as a chronogram), and adding the reconstructed evolutionary history of a set of traits is simply not feasible. This implies that online databases of results, perhaps an extended annotation-aware version of TreeBASE, will be necessary to communicate information of this sort. Alternatively, because reconstructions are indeed quite fast, perhaps it will suffice to have real-time tools available so that users can simply input a tree and the particular trait data of interest and have the reconstructions done on the fly. This should become more feasible when existing tools such as Mesquite become more integrated into other databases and web services (cf. the CIPRES project: www.phylo.org).

Technological issues

Computational complexity and harnessing computing resources

Program running times for both tree building and tree annotation have been discussed above at length. The basic conundrum of computational phylogenetics is that most of the relevant problems, including multiple-sequence alignment and tree inference, are 'intractable' in that no exact solution is known with better than exponential running times in the number of taxa. Because present 'large-scale' phylogenetic efforts hover around perhaps 1000 species, and there exists great interest in extending this input 1000-fold to 1 million or more species (the order of the tree of life), this presents daunting computational challenges (i.e. e^{1000} times harder, at best). All the computing resources of the planet could be harnessed and still no known exact algorithm (e.g. branch and bound) would finish running for large datasets that have already been published. Barring someone discovering the solution to these difficult problems, the phylogenetics community individually and collectively has to develop strategies to balance both their requests for massive computing infrastructure and their need for good heuristic solutions that require long (but not exponential) running times.

Fortunately, extensive work is underway. Explicitly parallelisable methods, such as the parsimony heuristics of TNT, the Bayesian MCMC runs of MrBayes, neighbour-joining (pNJTree: Du and Lin 2006), maximum likelihood (Minh *et al.* 2005, piQPNNI; Till *et al.* 2004; Keane *et al.* 2006), DCM (Du *et al.* 2005), quartet puzzling methods (Schmidt *et al.* 2002) and genetic algorithms (Zwickl 2006), or trivially parallelisable bootstrap analyses (Stamatakis 2006), can be run on compute clusters that are becoming increasingly common or on 'grid' computing systems, i.e. collections of under-utilised networks of machines in classrooms, office buildings, and so on, distributed broadly over the internet (Myers and Cummings 2003; Walters *et al.* 2005). Parmentier *et al.* (2006) described promising algorithms for parallelising simultaneous alignment and tree building, following in a line of work on parallel versions of multiple-sequence alignment in recent years. Exotic solutions such as porting RaxML to the graphics processor on desktop computers have even been explored (Charalambous *et al.* 2005). One reason for optimism is the new enthusiasm for these problems among computer scientists.

Automation?

Can the construction of large phylogenetic trees be automated? For genomic analysis such work is well underway (see e.g. Nilsson *et al.* 2004 for tree-based EST assembly). Efforts aimed at species-tree reconstruction are also moving forward. Ciccarelli *et al.* (2006) built a computational pipeline to assemble orthologues from whole-genome sequences across the tree of life and build a phylogeny of 191 taxa. Driskell *et al.* (2004) and then McMahon and Sanderson (2006) constructed a pipeline to download sequences from GenBank, assemble and align them into large supermatrices and build phylogenetic trees of up to a few thousand species. Robbertse *et al.* (2006) described a similar set of protocols to mine genome-sequence data across ascomycetes. These pipelines generally

still all require non-trivial human intervention. Hibbett *et al.* (2005) developed the Mor pipeline to automatically extract rDNA sequences for homobasidiomycetes and build very large phylogenies from them in a continually updated fashion (the only such system existing, I believe). General computing tools to develop these pipelines further may be helpful in the future, such as the Kepler (<http://kepler-project.org/>) and Pise projects (<http://www.pasteur.fr/recherche/unites/sis/Pise/>), aided by extensive libraries of modular programs, such as BioPERL.

In their work, Driskell *et al.* (2004), McMahon and Sanderson (2006) and Sanderson and McMahon (2007) found several important obstacles to automating such pipelines—in other words, steps that required human intervention to prevent introduction of egregious but avoidable errors. Perhaps the most important of these was alignment. No alignment program has been found that can reliably deal with a combination of non-coding sequence data, deep divergence and heterogeneous sequence length. Hundreds of hours of manual sequence examination were required to minimise introduction of misleading phylogenetically significant errors at the alignment step. Work to develop good quantitative indicators of bad alignments is relatively experimental and heuristic (Castresana 2000; Lassmann and Sonnhammer 2002). Another obstacle is mistaken annotations. Similar issues face anyone relying on data-based information, and it may be possible to develop automated tools that send alarms when consistency checks fail (e.g. two angiosperm species with a gene labelled *rbcl* do not BLAST). Finally, the patterns of fragmentation of data that can arise if multiple loci or deep divergences are included can introduce pathological results that can be difficult to explain without detailed *post hoc* analysis of the steps in the pipeline.

Visualisation of large trees

A surprisingly obstinate problem that has emerged as phylogenies have grown in size and information content is visualisation. Trees are odd structures that can be displayed in many apparently distinct ways (e.g. rotating sister groups around a node does not change anything but the tree's visual appearance). As their size increases, it becomes difficult for the eye to capture grouping information, the content of clades, in a way that is meaningful.

In large trees, a more basic problem arises; namely, how to see the terminal taxon or clade *names* and place them in context with their phylogenetic neighbours and more inclusive clades? One approach is to 'hide' part of the tree. This is the strategy of tools that collapse large clades to nodes or icons (ARB: Ludwig *et al.* 2004) and to web browser-based tools such as the Tree of Life Web Project (Maddison and Schulz 1996–2007) which uses hyperlinks to move up and down in the phylogenetic hierarchy. Instead of hiding the rest of the tree completely, other tools distort the space that the tree is found in, magnifying the local neighbourhood and shrinking the more distant parts of the tree. The first innovations in this direction embedded large trees in a 3-D 'hyperbolic space' (Munzner 1998; Hughes *et al.* 2004), in which the user, by navigating through this space, encountered highly magnified nearby parts of the tree, with distant parts vanishing. A more recent approach, taken in the



Fig. 3. View of the 2228-taxon tree of McMahon and Sanderson (2006) using the program TreeJuxtaposer (Munzner *et al.* 2003). The program permits the user to interactively magnify clades as shown here for a predominantly Australian collection of papilionoid legumes. Taxon names are sampled from the terminal nodes according to a density parameter specified by the user.

program TreeJuxtaposer (Munzner *et al.* 2003), lays out the tree in 2-D space, but lets the user interactively magnify one part of the tree while shrinking the rest of the tree in a visually appealing fashion (Fig. 3). Paloverde (Sanderson 2006) is not nearly so geometrically innovative. It merely lays out a large circular tree in a 3-D world, providing rapid mouse-driven methods for zooming in on local areas of interest (a sort of GoogleEarth for trees), while maintaining correct undistorted perspective on what lies in the distance.

The success or failure of these visualisation efforts will probably lie in the ability of these programs to communicate phylogenetically relevant information and annotations. For example, in Paloverde one feature will collapse all species within genera to nodes if the taxon names are given as binomials. TreeJuxtaposer has a nice feature to highlight similarities and differences between trees laid out side by side. Other automated ways to convey information about the higher-order structure of large trees are easy to imagine, such as providing node-based definitions of clades to the program, as suggested by the structure of the Mor project described above.

Conclusions: who will assemble the tree of life?

Anyone on the planet with an internet connection and a reasonably new desktop computer can download sequence data and build large phylogenetic trees. This is both exciting and sobering. On the one hand, a clear lesson of the bioinformatics revolution is that open access to data and tools can accelerate discovery and innovation. The many costs incurred by individual investigators forced to deposit their sequence data into GenBank, for example, have probably been more than compensated for on a community-wide basis. On the other hand, although phylogenetics—like bioinformatics—has a relatively recent origin, phylogenetics differs in that it is also intimately tied to systematics, a discipline with a much longer history, a vast literature and a complex set of intellectual and philosophical issues with important implications for all of biology (e.g. taxonomy and nomenclature). The computational tools and data are in place for many workers with little training in systematics to build large phylogenetic trees. This is not necessarily a recipe for disaster, but it does beg many questions. Would such efforts be undertaken in full cognisance of existing hypotheses of relationships and with a view towards integrating new work with less accessible forms of data (e.g. morphology, fossils, geography)? Would it proceed with an appreciation of its taxonomic and nomenclatural implications; the downstream effects on the very practical impacts that classifications derived from phylogenies will have on the biological sciences? Finally, will it be tempered by the same level of native scepticism and desire to examine results for congruence in the most synthetic fashion possible that has traditionally characterised systematically driven phylogenetic work?

Because sequence data are so readily available for such a diversity of organisms, building large-scale trees from it is likely to be too tempting to resist for computationally savvy biologists of many stripes. If phylogenetic systematists do not take this opportunity, it is likely that others will. Most of the efforts along these lines have been undertaken by workers with a largely phylogenetic research program, but this might change

if the phylogenetics community loses interest in advances in computational biology.

Furthermore, since it is much easier and cheaper to download data than it is to collect new plant material or even sequence DNA obtained from herbarium specimens, there is a risk that data mining may come to dominate the phylogenetics culture. If so, the fraction of described biodiversity now in the databases will not continue to grow at its present rate, which has seen a roughly linear increase of ~17 000 new species a year across the last 5 years (or ~1% of described species diversity per year) (note the accumulation of *sequences* is exponential, doubling roughly every 2 years). The interest and efforts of many focused on mining the databases should not occur at the expense of the extraordinary efforts, driven almost entirely by the systematics community, to populate those databases with information drawn from the global biota.

References

- Aho AV, Sagiv Y, Szymanski TG, Ullman JD (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing* **10**, 405–421. doi: 10.1137/0210030
- Alfaro ME, Holder MT (2006) The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology Evolution and Systematics* **37**, 19–42. doi: 10.1146/annurev.ecolsys.37.091305.110021
- Ammiraju JSS, Luo MZ, Goicoechea JL, Wang W, Kudrna D, Mueller C, Talag J, Kim HR, Sisneros NB, Blackmon B, Fang E, Tomkins JB, Brar D, MacKill D, McCouch S, Kurata N, Lambert G, Galbraith DW, Arumuganathan K, Rao K, Walling JG, Gill N, Yu1 Y, SanMiguel P, Soderlund C, Jackson S, Jackson S, Wang RA (2006) The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Research* **16**, 140–147. doi: 10.1101/gr.3766306
- Ané C, Eulenstein O, Piaggio-Talice R, Sanderson MJ (2006) Groves of phylogenetic trees. Technical Report. University of Wisconsin, Madison, WI.
- Angiosperm Phylogeny Group (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* **141**, 399–436. doi: 10.1046/j.1095-8339.2003.t01-1-00158.x
- Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **19**(suppl. 1), i7–i15. doi: 10.1093/bioinformatics/btg1000
- Bansal M, Burleigh JG, Eulenstein O, Wehe A (2007) Heuristics for the gene-duplication problem: an W(N) speed-up for the local search. *RECOMB 2007*.
- Baptiste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruffé L, Gaasterland T, Lopez P, Müller M, Hervé P (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 1414–1419. doi: 10.1073/pnas.032662799
- Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**, 3–10. doi: 10.2307/1222480
- Bender MA, Farach-Colton M (2000) The LCA problem revisited. *Lecture Notes in Computer Science* **1776**, 88–94.
- Bininda-Emonds ORP (2004a) The evolution of supertrees. *Trends in Ecology & Evolution* **19**, 315–322. doi: 10.1016/j.tree.2004.03.015
- Bininda-Emonds ORP (Ed.) (2004b) 'Phylogenetic supertrees.' (Kluwer: Boston)

- Bininda-Emonds ORP, Brady SG, Kim J, Sanderson MJ (2001) Scaling of accuracy in extremely large phylogenetic trees. *Pacific Symposium on Biocomputing* **6**, 547–558.
- Bininda-Emonds ORP, Gittleman JL, Steel MA (2002) The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* **33**, 265–290.
- Britton T (2005) Estimating divergence times in phylogenetic trees without a molecular clock. *Systematic Biology* **54**, 500–507. doi: 10.1080/10635150590947311
- Britton T, Oxelman B, Vinnersten A, Bremer K (2002) Phylogenetic dating with confidence intervals using mean path lengths. *Molecular Phylogenetics and Evolution* **24**, 58–65. doi: 10.1016/S1055-7903(02)00268-3
- Burleigh JG, Driskell AC, Sanderson MJ (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Systematic Biology* **55**, 426–440. doi: 10.1080/10635150500541722
- Burnham KP, Anderson DR (1998) 'Model selection and inference.' (Springer: New York)
- Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**, 540–552.
- Chang JT (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences* **134**, 189–215. doi: 10.1016/0025-5564(95)00172-7
- Charalambous M, Trancoso P, Stamatakis A (2005) Initial experiences porting a bioinformatics application to a graphics processor. *Lecture Notes in Computer Science* **3746**, 415–425.
- Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedrén M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang Q-Y, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH Jr, Graham SW, Barrett SCH, Dayanandan S, Albert VA (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* **80**, 528–580. doi: 10.2307/2399846
- Chave J, Muller-Landau HC, Baker TR, Easdale TA, Ter Steege H, Webb CO (2006) Regional and phylogenetic variation of wood density across 2456 neotropical tree species. *Ecological Applications* **16**, 2356–2367. doi: 10.1890/1051-0761(2006)016[2356:RAPVOW]2.0.CO;2
- Chevenet F, Brun C, Banuls AL, Jacq B, Christen R (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, doi: 10.1186/1471-2105-7-439
- Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287. doi: 10.1126/science.1123061
- Cunningham CW (1997) Can three incongruence tests predict when data should be combined? *Molecular Biology and Evolution* **14**, 733–740.
- Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V (2004) Darwin's abominable mystery: insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences, USA* **101**, 1904–1909. doi: 10.1073/pnas.0308127100
- Dong QF, Kroiss L, Oakley FD, Wang BB, Brendel V (2005) Comparative EST analyses in plant systems. *Methods in Enzymology* **395**, 400–419. doi: 10.1016/S0076-6879(05)95022-2
- Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara B, Sanderson MJ (2004) Prospects for building the tree of life from large sequence databases. *Science* **306**, 1172–1174. doi: 10.1126/science.1102036
- Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, e88. doi: 10.1371/journal.pbio.0040088
- Du ZH, Lin F (2006) pNJTree: a parallel program for reconstruction of neighbor-joining tree and its application in ClustalW. *Parallel Computing* **32**, 441–446. doi: 10.1016/j.parco.2006.05.001
- Du ZH, Lin F, Roshan UW (2005) Reconstruction of large phylogenetic trees: a parallel approach. *Computational Biology and Chemistry* **29**, 273–280. doi: 10.1016/j.compbiolchem.2005.06.003
- Durand D, Halldorsson BV, Vernot B (2005) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Lecture Notes in Computer Science* **3500**, 250–264.
- Farris JS, Källersjö M, Kluge AG, Bult C (1994) Testing significance of incongruence. *Cladistics* **10**, 315–319. doi: 10.1111/j.1096-0031.1994.tb00181.x
- Farris JS, Källersjö M, Kluge AG, Bult C (1995) Constructing a significance test for incongruence. *Systematic Biology* **44**, 570–572. doi: 10.2307/2413663
- Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**, 401–410. doi: 10.2307/2412923
- Godfray HCJ, Knapp S (2004) Taxonomy for the twenty-first century—Introduction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **359**, 559–569. doi: 10.1098/rstb.2003.1457
- Goloboff PA (1999) Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15**, 415–428. doi: 10.1111/j.1096-0031.1999.tb00278.x
- Goodman M, Czelusniak J, Moore GW, Romeroherrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology* **28**, 132–163. doi: 10.2307/2412519
- Goremykin VV, Hansmann S, Martin WF (1997) Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. *Plant Systematics and Evolution* **206**, 337–351. doi: 10.1007/BF00987956
- Goremykin V, Hirsch-Ernst K, Wolf S, Hellwig F (2003) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Molecular Biology and Evolution* **20**, 1499–1505. doi: 10.1093/molbev/msg159
- Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problems? *Systematic Biology* **47**, 9–17. doi: 10.1080/106351598260996
- Grotkopp E, Rejmanek M, Sanderson MJ, Rost TL (2004) Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. *Evolution* **58**, 1705–1729.
- Hardy CR, Linder HP (2005) Intraspecific variability and timing in ancestral ecology reconstruction: a test case from the Cape flora. *Systematic Biology* **54**, 299–316. doi: 10.1080/10635150590923317
- Hibbett D, Nilsson R, Snyder M, Fonseca M, Costanzo J, Shonfeld M (2005) Automated phylogenetic taxonomy: an example in the homobasidiomycetes (mushroom-forming fungi). *Systematic Biology* **54**, 660–668. doi: 10.1080/10635150590947104
- Hillis DM (1996) Inferring complex phylogenies. *Nature* **383**, 130–131. doi: 10.1038/383130a0
- Hillis DM, Huelsenbeck JP, Cunningham CW (1994) Application and accuracy of molecular phylogenies. *Science* **264**, 671–677. doi: 10.1126/science.8171318
- Huelsenbeck JP (1997) Is the Felsenstein zone a fly trap? *Systematic Biology* **46**, 69–74. doi: 10.2307/2413636
- Hughes T, Hyun Y, Liberles DA (2004) Visualizing very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinformatics* **5**, 48–53. doi: 10.1186/1471-2105-5-48
- Huson DH, Nettles SM, Warnow TJ (1999) Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology* **6**, 369–386. doi: 10.1089/106652799318337

- Janssen T, Bremer K (2004) The age of major monocot groups inferred from 800+ *rbcL* sequences. *Botanical Journal of the Linnean Society* **146**, 385–398. doi: 10.1111/j.1095-8339.2004.00345.x
- Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? *Trends in Genetics* **22**, 225–231. doi: 10.1016/j.tig.2006.02.003
- Johnson LAS (1970) Rainbow's end: the quest for an optimal taxonomy. *Systematic Zoology* **19**, 203–239.
- Källersjö M, Farris JS, Chase MW, Bremer B, Fay MF, Humphries CJ, Petersen G, Seberg O, Bremer K (1998) Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Systematics and Evolution* **213**, 259–287. doi: 10.1007/BF00985205
- Keane TM, Page AJ, Naughton TJ, Travers SAA, McInerney JO (2006) Building large phylogenetic trees on coarse-grained parallel machines. *Algorithmica* **45**, 285–300. doi: 10.1007/s00453-006-1215-0
- Kim J (1998) Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Systematic Biology* **47**, 43–60. doi: 10.1080/106351598261021
- Kluge AG (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology* **38**, 7–25. doi: 10.2307/2992432
- Kolaczowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980–984. doi: 10.1038/nature02917
- Lassmann T, Sonnhammer ELL (2002) Quality assessment of multiple alignment programs. *FEBS Letters* **529**, 126–130. doi: 10.1016/S0014-5793(02)03189-7
- Laurin M, de Queiroz K, Cantino P, Cellinese N, Olmstead R (2005) The PhyloCode, types, ranks and monophyly: a response to Pickett. *Cladistics* **21**, 605–607. doi: 10.1111/j.1096-0031.2005.00090.x
- Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology* **54**, 575–594. doi: 10.1080/10635150590947131
- Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, de Pamphilis CW (2005) Identifying the basal angiosperms node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution* **22**, 1948–1963. doi: 10.1093/molbev/msi191
- Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences, USA* **102**, 11 070–11 075. doi: 10.1073/pnas.0504978102
- Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüßmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H (2004) ARB: a software environment for sequence data. *Nucleic Acids Research* **32**, 1363–1371. doi: 10.1093/nar/gkh293
- Mabberley DJ (1987) 'The plant book.' (Cambridge University Press: Cambridge, UK)
- Maddison DR, Schulz K-S (1996–2007) 'The tree of life web project.' 2006 <http://tolweb.org> [verified 17 July 2007].
- Maddison WP, Maddison DR (2000) 'MacClade 4: analysis of phylogeny and character evolution.' (Sinauer: Sunderland, MA)
- Maddison WP, Maddison DR (2007) Mesquite: a modular system for evolutionary analysis. <http://mesquiteproject.org/mesquite/mesquite.html> [verified 17 July 2007].
- McCubbin AG, Roalson EH (2005) Construction of bacterial artificial chromosome libraries for use in phylogenetic studies. *Methods in Enzymology* **395**, 384–400. doi: 10.1016/S0076-6879(05)95021-0
- McMahon MM, Sanderson MJ (2006) Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Systematic Biology* **55**, 818–836. doi: 10.1080/10635150600999150
- Minh BQ, Vinh LS, von Haeseler A, Schmidt HA (2005) pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics (Oxford, England)* **21**, 3794–3796. doi: 10.1093/bioinformatics/bti594
- Moles A, Ackerly D, Webb C, Tweddle J, Dickie J, Westoby M (2005) A brief history of seed size. *Science* **307**, 576–580. doi: 10.1126/science.1104863
- Moore B, Smith S, Donoghue MJ (2006) Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Systematic Biology* **55**, 662–676. doi: 10.1080/10635150600920693
- Mort ME, Soltis PS, Soltis DE, Mabry ML (2000) Comparison of three methods for estimating internal support on phylogenetic trees. *Systematic Biology* **49**, 160–171. doi: 10.1080/10635150050207456
- Mossel E (2007) Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **4**, 108–116. doi: 10.1109/TCBB.2007.1010
- Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In 'Mathematics of evolution and phylogeny'. (Eds O Gascuel, M Steel) pp. 384–412. (Oxford University Press: New York)
- Mower JP, Stefanovic S, Young GJ, Palmer JD (2004) Plant genetics—Gene transfer from parasitic to host plants. *Nature* **432**, 165–166. doi: 10.1038/432165b
- Munzner T (1998) Exploring large graphs in 3D hyperbolic space. *IEEE Computer Graphics and Applications* **18**, 18–23. doi: 10.1109/38.689657
- Munzner T, Guimbretiere F, Tasiran S, Zhang L, Zhou YH (2003) TreeJuxtaposer: scalable tree comparison using Focus + Context with guaranteed visibility. *ACM Transactions on Graphics* **22**, 453–462. doi: 10.1145/882262.882291
- Myers DS, Cummings MP (2003) Necessity is the mother of invention: a simple grid computing system using commodity tools. *Journal of Parallel and Distributed Computing* **63**, 578–589. doi: 10.1016/S0743-7315(03)00004-2
- Nilsson RH, Rajashekar B, Larsson KH, Ursing BM (2004) GalaxieEST: addressing EST identity through automated phylogenetic analysis. *BMC Bioinformatics* **5**, doi: 10.1186/1471-2105-5-87
- Page RDM (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* **14**, 819–820. doi: 10.1093/bioinformatics/14.9.819
- Page RDM, Charleston MA (1998) Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution* **13**, 356–359. doi: 10.1016/S0169-5347(98)01438-4
- Parmentier G, Trystram D, Zola J (2006) Large scale multiple sequence alignment with simultaneous phylogeny inference. *Journal of Parallel and Distributed Computing* **66**, 1534–1545.
- Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404–407. doi: 10.1038/46536
- Qiu YL, Dombrowska O, Lee J, Li L, Whitlock BA, Bernasconi-Quadroni F, Rest JS, Davis CC, Borsch T, Hilu KW, Renner SS, Soltis DE, Soltis PS, Zanis MJ, Cannone JJ, Gutell RR, Powell M, Savolainen V, Chatrou LW, Chase MW (2005) Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. *International Journal of Plant Sciences* **166**, 815–842. doi: 10.1086/431800
- de Queiroz A, Donoghue MJ, Kim J (1995) Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* **26**, 657–681. doi: 10.1146/annurev.es.26.110195.003301
- Rice KA, Donoghue MJ, Olmstead RG (1997) Analyzing large data sets: *rbcL* 500 revisited. *Systematic Biology* **46**, 554–563. doi: 10.2307/2413696

- Robbertse B, Reeves JB, Schoch CL, Spatafora JW (2006) A phylogenomic analysis of the Ascomycota. *Fungal Genetics and Biology* **43**, 715–725. doi: 10.1016/j.fgb.2006.05.001
- Rokas A, Williams B, King N, Carroll S (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804. doi: 10.1038/nature02053
- Ross HA, Lento GM, Dalebout ML, Goode M, Ewing G, McLaren P, Rodrigo AG, Lavery S, Baker CS (2003) DNA Surveillance: web-based molecular identification of whales, dolphins and porpoises. *Journal of Heredity* **94**, 111–114. doi: 10.1093/jhered/esg027
- Rutschmann F (2006) Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. *Diversity & Distributions* **12**, 35–48. doi: 10.1111/j.1366-9516.2006.00210.x
- Salamín N, Hodkinson TR, Savolainen V (2002) Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* **51**, 136–150. doi: 10.1080/106351502753475916
- Salamín N, Chase MW, Hodkinson TR, Savolainen V (2003) Assessing internal support with large phylogenetic DNA matrices. *Molecular Phylogenetics and Evolution* **27**, 528–539. doi: 10.1016/S1055-7903(03)00011-3
- Sanderson MJ (2006) Paloverde: an OpenGL 3D phylogeny browser. *Bioinformatics* **22**, 1004–1006.
- Sanderson MJ, McMahon MM (2007) Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evolutionary Biology* **7**(Suppl. 1), S3. doi: 10.1186/1471-2148-7-S1-S3
- Sanderson MJ, Wojciechowski MF (2000) Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Systematic Biology* **49**, 671–685. doi: 10.1080/106351500750049761
- Sanderson MJ, Wojciechowski MF, Hu JM, Khan TS, Brady SG (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution* **17**, 782–797.
- Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S (2003) Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Molecular Biology and Evolution* **20**, 1036–1042.
- Sanderson MJ, Ané C, Eulenstein O, Fernandez-Baca D, Kim J, McMahon MM, Piaggio-Talice R (2007) Fragmentation of large data sets in phylogenetic analysis. In 'Mathematics of evolution and phylogeny II'. (Eds O Gascuel, M Steel) (Oxford University Press: Oxford)
- Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**, 868–876. doi: 10.1139/g04-047
- Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504. doi: 10.1093/bioinformatics/18.3.502
- Semple C, Daniel P, Hordijk W, Page RDM, Steel M (2004) Supertree algorithms for ancestral divergence dates and nested taxa. *Bioinformatics* **20**, 2355–2360. doi: 10.1093/bioinformatics/bth246
- Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51**, 492–508. doi: 10.1080/10635150290069913
- Sneath P, Sokal R (1973) 'Numerical taxonomy.' (WH Freeman and Co.: San Francisco)
- Soltis DE, Soltis PS, Nickrent DL, Johnson LA, Hahn WJ, Hoot SB, Sweere JA (1997) Angiosperm phylogeny inferred from 18S ribosomal sequences. *Annals of the Missouri Botanical Garden* **84**, 1–49. doi: 10.2307/2399952
- Soltis PS, Soltis DE, Wolf PG, Nickrent DL, Chaw S-M, Chapman RL (1999) The phylogeny of land plants inferred from 18S rDNA sequences: pushing the limits of rDNA signal? *Molecular Biology and Evolution* **16**, 1774–1784.
- Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. doi: 10.1093/bioinformatics/btl446
- Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463. doi: 10.1093/bioinformatics/bti191
- Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18**, 92–99. doi: 10.1093/bioinformatics/18.1.92
- Tehler A, Little DP, Farris JS (2003) The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi. *Mycological Research* **107**, 901–916. doi: 10.1017/S0953756203008128
- Till M, Zhou BB, Zomaya A, Jermini LS (2004) Phylogenetic analysis using maximum likelihood methods in homogeneous parallel environments. *Lecture Notes in Computer Science* **3320**, 274–279.
- de la Torre J, Egan M, Katari M, Brenner E, Stevenson D, Coruzzi G, Desalle R (2006) ESTimating plant phylogeny: lessons from partitioning. *BMC Evolutionary Biology* **6**, 48. doi: 10.1186/1471-2148-6-48
- Vilgalys R (2003) Taxonomic misidentification in public DNA databases. *New Phytologist* **160**, 4–5. doi: 10.1046/j.1469-8137.2003.00894.x
- Vogl C, Badger J, Kearney P, Li M, Clegg M, Jian T (2003) Probabilistic analysis indicates discordant gene trees in chloroplast evolution. *Journal of Molecular Evolution* **56**, 330–340. doi: 10.1007/s00239-002-2404-3
- Walters JD, Casavant TL, Robinson JP, Bair TB, Braun TA, Scheetz TE (2005) XenoCluster: a grid computing approach to finding ancient evolutionary genetic anomalies. *Lecture Notes in Computer Science* **3606**, 355–366.
- Webb CO, Donoghue MJ (2005) Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes* **5**, 181–183. doi: 10.1111/j.1471-8286.2004.00829.x
- Webb CO, Losos JB, Agrawal AA (2006) Integrating phylogenies into community ecology. *Ecology* **87**, S1–S2. doi: 10.1890/0012-9658(2006)87[1:IPICE]2.0.CO;2
- Wojciechowski MF, Sanderson MJ, Steel KP, Liston A (2000) Molecular phylogeny of the 'temperate herbaceous tribes' of papilionoid legumes: a supertree approach. In 'Advances in legume systematics'. (Eds PS Herendeen, A Bruneau) pp. 277–298. (Royal Botanic Gardens, Kew: London)
- Yan CH, Burleigh JG, Eulenstein O (2005) Identifying optimal incomplete phylogenetic data sets from sequence databases. *Molecular Phylogenetics and Evolution* **35**, 528–535. doi: 10.1016/j.ympev.2005.02.008
- Yang ZH, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* **23**, 212–226. doi: 10.1093/molbev/msj024
- Yesson C, Culham A (2006) A phylogenetic study of cyclamen. *BMC Evolutionary Biology* **6**, 72.
- Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Dissertation, University of Texas at Austin, Austin, TX.

Manuscript received 28 February 2007, accepted 22 May 2007