Prospects for building the tree of life from large sequence databases

Amy C. Driskell, Cécile Ané, J. Gordon Burleigh, Michelle M. McMahon, Brian C. O'Meara, Michael J. Sanderson

SUPPORTING ONLINE MATERIAL

MATERIALS AND METHODS

Sequence data. All records in GenBank with the taxonomic label "Viridiplantae" were extracted from the "gbpln.seq" flat files of release 137 (Aug. 15, 2003). Amino acid sequences were captured from the translation field of each "CDS" feature. The Swiss-Prot data was taken from the complete flat file of the full release 40 (2001) and the cumulative update 40.29 (Dec. 2002). Information on all sequences included in the analyses is available on our website (*S1*). Each sequence was "tagged" with its associated GenBank taxonomy ID number. Because subspecific taxa are sometimes treated ambiguously by users of GenBank's taxonomy, multiple taxa from the same species (e.g. *Pongo pygmaeus abelii* [taxon ID 9601] with *P. pygmaeus* [taxon ID 9600]) have different taxon IDs, and consequently our final supermatrices contained multiple representatives of some species.

Identification of clusters and orthologs. Sets of potentially homologous sequences ("clusters") were identified using the NCBI BLASTCLUST program (*S2*), which combines BLAST searches with single-linkage clustering. Many other methods have been developed to identify orthologs in databases for diverse purposes (*S3, S4*), but our goal is simply to put together sequences at roughly the appropriate level of divergence for phylogenetic inference (*S5*). The BLASTCLUST single-linkage cluster cutoff value was set at 60% identity and the default alignment length was used. Clustering by homology avoids mistakes that can arise by clustering based on annotations, but it occasionally separates divergent but homologous sequences into disjoint clusters if no members share at least 60% identity. These can be added later by either supermatrix or supertree methods, but they will not directly provide information about relationships among taxa in the two different clusters.

Clusters containing four or more taxa, termed “minimal phylogenetic clusters,” are sorted into two groups: those containing only orthologous sequences (“single-copy”) and those composed of a mix of paralogous and orthologous sequences. By default, clusters with only a single sequence per taxon were treated as orthologous. A phylogenetic test was used to classify the remaining clusters. First, sequences were aligned using default options in CLUSTALW (*S6*) and unconstrained and constrained maximum parsimony heuristic searches were conducted using a ‘‘protein parsimony’’ step matrix (*S7*) with the program PAUP* (*S8*). In the constrained search all sequences from a single taxon were forced to form a clade. If the constrained tree was significantly less parsimonious than the unconstrained tree (using a signed-rank test (*S7*)) the cluster was not classified as single-copy and was excluded from further analysis. An interesting exception to this procedure was the finding that the widely-used chloroplast atpB gene has a mitochondrial homolog, which caused rejection of this cluster (the fourth largest in the data set) by our orthology test. In this one instance, we included in further analysis the subset of the data in the cluster corresponding to the chloroplast sequences (>98% of the sequences in the cluster) in order to retain the great taxonomic diversity represented by this cluster. Although our orthology test performed well in detecting known single-copy organellar genes, very small clusters from nuclear gene families can mistakenly be inferred to be single-copy if paralogs are simply missing from the database.

Identification of Groves. A "grove" is a set of clusters (phylogenetic data sets) that has the potential to generate novel statements of phylogenetic relationships via supertree construction (*S9, S10*). This potential depends on the taxonomic overlap between clusters and may or may not be realized depending on the actual level of disagreement among the trees constructed from each cluster and the algorithms used to build the supertrees. To obtain lower bounds on the number of groves, we construct an intersection graph in which clusters are represented by nodes. Edges, weighted by the number of taxa shared, connect nodes that share at least one taxon in common. Distinct connected components in this graph do not share any taxa in common and therefore cannot be part of the same grove (no new information can be gleaned by piecing together input trees that do not share any taxa). This gives an estimate of the lower bound on the number of groves. In addition, different subgraphs of the same connected component cannot be in the same grove if they are separated by a bridge of only weight one (a "bridge" is an edge that, if removed, disconnects the graph). Since removing a bridge of weight one replaces a connected component by two subgraphs, we can obtain a lower bound on the number of groves by the sum of the number of connected components and the number of weight-one bridges. A lower bound on the size of the largest grove can be obtained in a different manner. Rooted trees that overlap by two taxa form a grove. Similarly, connected components in a graph in which edges of weight one are suppressed also form groves, but these may not be maximal. Nonetheless the largest of these groves represents a lower bound on the size of the largest grove possible.

An "orphan" is a cluster that has no edges to any other clusters in the graph.

Identification of bicliques. A multi-gene (multi-protein) matrix can be represented by a bipartite graph in which one set of nodes corresponds to proteins and the other to taxa, with an edge connecting a protein node to a taxon node if the protein sequence exists for the taxon in the database. A data matrix is "complete" if all proteins are sampled for all taxa—that is, if there are no missing entries. A complete matrix is "maximal" if no larger matrices contain it. In the graph this corresponds to a "maximal biclique," which is a completely connected subgraph of the bipartite graph that cannot be extended. We call matrices containing only one protein as "trivial" maximal bicliques, as these are equivalent to the original clusters or to subsets thereof. Identifying all maximal complete matrices in a sequence database or maximal bicliques in a graph is an NP-complete problem However, exact algorithms are fast for sparse matrices (*S11, S12*), and these were implemented to identify all maximal bicliques from the set of single-copy minimal phylogenetic clusters for the two databases.

Supermatrix assembly. A "supermatrix" is a (usually) *incomplete* concatenated matrix assembled from more than one gene (here protein). Incomplete matrices correspond to "quasi-bicliques," which are graphs that can be turned into bicliques with the addition of some edges. Few formal methods are available for quasi-biclique construction (*S13*, *S14*), so we imposed minimal conditions on each taxon and protein and let the structure of the databases determine which taxa and proteins would be included in the resulting supermatrices. In particular, we discarded all bicliques with fewer than 10 proteins and 4 taxa in the biclique collections for each database (Table 1). The Swiss-Prot data set was culled to include only metazoans and fungal outgroups to avoid conflicting gene signals arising from the non-parallel histories of the nuclear, mitochondrial, and chloroplast genomes in photosynthetic eukaryotes. We next identified the largest grove in each collection of 10 × 4 bicliques to ensure sufficient taxonomic overlap among the genes. Finally, we constructed the supermatrix by compiling the sequences from all the bicliques contained within the largest grove and any additional sequences for the taxa in the grove not in the set of maximal bicliques. See the paper for statistics on these matrices and our website for the matrices themselves (*S1)*.

Tree reconstruction and evaluation. Sequence alignments were obtained from single cluster alignments as described above. We performed maximum parsimony analyses on the green plant and metazoan supermatrices using a heuristic tree search algorithm in PAUP* 4.0b10 (*S8*), and searches of the green plant supermatrix were performed using a "protein parsimony" step matrix (*S7*). The search strategy consisted of 100 random addition sequence replicates (96 for the Swiss-Prot matrix) and TBR branch swapping. Each addition sequence replicate was time-limited to 12 hours and retained a maximum of 10,000 equally parsimonious trees. Trees from each replicate were combined and a strict consensus topology was computed. In addition, we performed 100 replicates of nonparametric bootstrapping. Each bootstrap replicate employed four random addition sequence starting tree replicates and TBR branch-swapping, was time-limited to six hours of branch-swapping per replicate, and retained a maximum of 10,000 trees. Finally, parsimony ratchet searches (*S15*) were implemented for both data sets using PAUP* and a custom Perl script, with five runs for each data set with 11 re-weighting/flat-weighting cycles (each limited to one hour), and five runs with 35 cycles (each limited to twenty minutes). The ratchet runs did not find any trees not already recovered from the other heuristic searches.

In addition to the conventional bootstrap measure of support, we examined the distribution of support from individual proteins. We analyzed each protein data set separately using maximum parsimony and compared the resulting protein trees to the supermatrix trees. Branches present in the supermatrix tree can be missing from a protein tree for at least two reasons: (1) the single protein data set, by itself, does not support the relationship, or (2) as each protein data set can have a different set of taxa, the protein tree does not contain the relevant taxa. The results of the single protein analyses are summarized in four numbers above each branch in the supermatrix trees (Figs. S1 and S2). The upper pair of numbers reports on protein trees that were "bipartition-informative," and the lower pair reports on protein trees that were "quartet-informative." For a protein tree to be bipartition-informative, it must have at least two taxa on each side of the branch of interest. For a protein tree to be quartet-informative, it must be bipartition-informative and it must have at least one taxon from each of the four groups (or more, in the case of polytomies) connected to the branch of interest. The first number of each pair gives the number of "non-conflicting" protein trees, and the second gives the number of "conflicting" protein trees. The single protein data set analyses frequently resulted in multiple most parsimonious (MP) trees. We considered a protein tree to be non-conflicting if at least one of the single protein's MP trees displayed the branch, or if all of the MP trees displayed a polytomy in place of the branch. Otherwise the protein tree was considered to be conflicting.

Fig. S1. The consensus tree from the Swiss-Prot metazoan supermatrix. The tree is a strict consensus of 163,145 equally parsimonious trees of length 218,960. Numbers before the taxon names are NCBI taxon IDs. Numbers below the nodes in bold italic are bootstrap proportions. Numbers above the nodes are as follows: the first line indicates the number of bipartition-informative genes in the matrix that are not conflicting/are conflicting; the second line indicates the number of quartet-informative genes that are not conflicting/are conflicting. More detail on these statistics is in Fig. 3. The tree was rooted with *Dictyostelium*. Asterisks indicate non-monophyletic taxa.

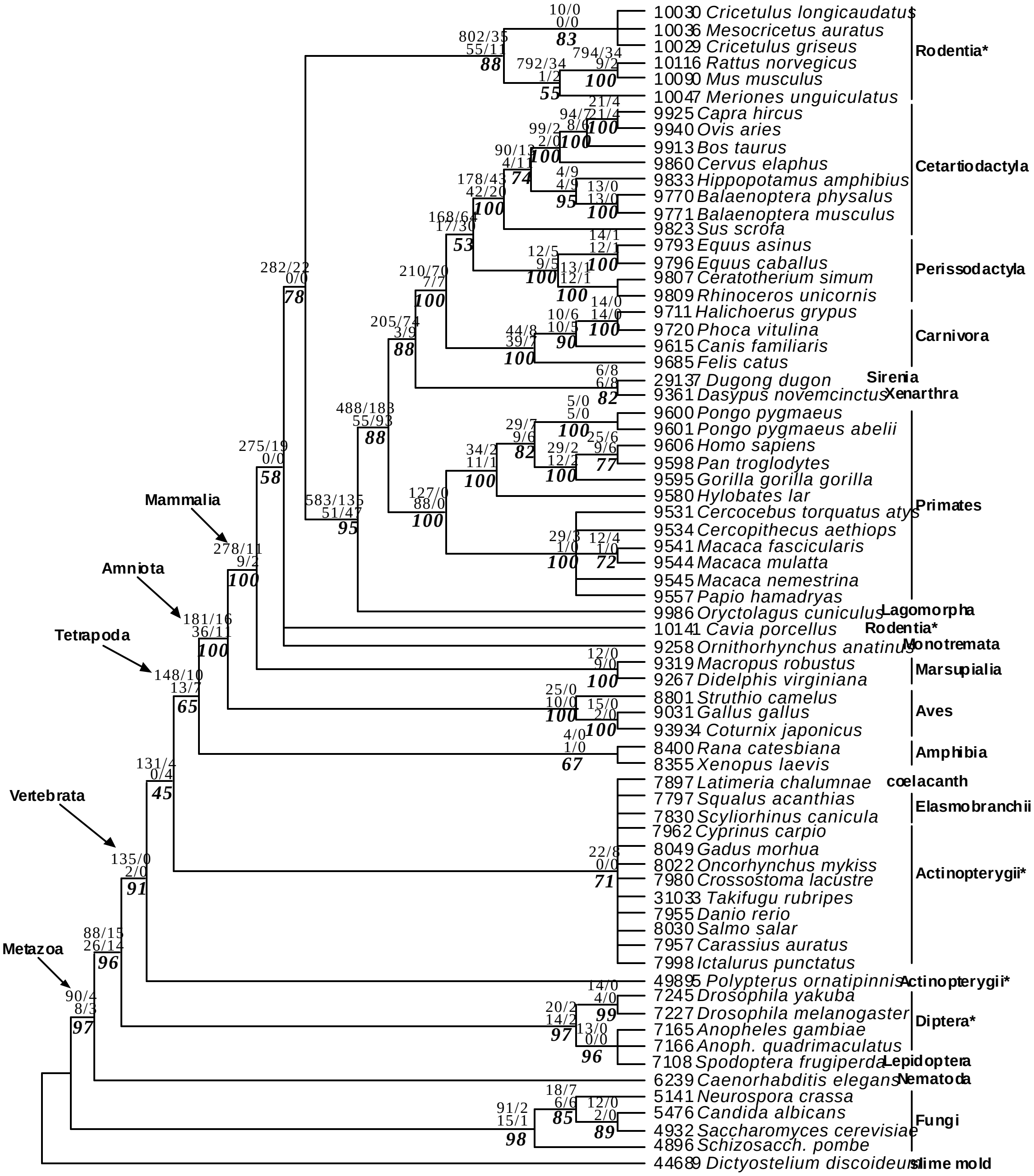
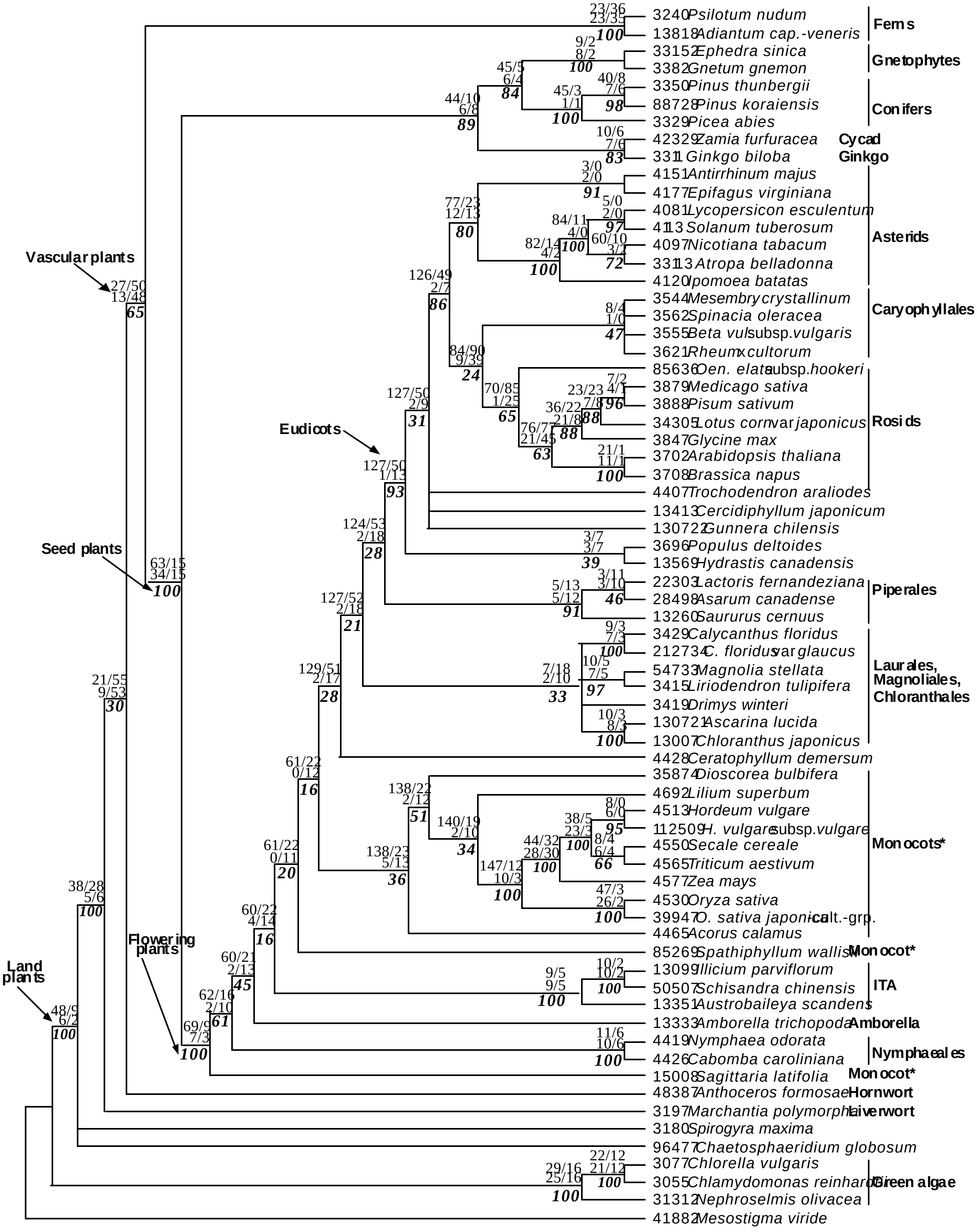


Fig. S2. The consensus tree from the GenBank green plant supermatrix. The tree is a strict consensus of 180 equally parsimonious trees of length 86,876 steps. Numbers at the nodes are as in Fig. S1. The unicellular green alga *Mesostigma* was used to root the tree (*S16*). Asterisks indicate non-monophyletic taxa.



References and Notes

S1. http://ginger.ucdavis.edu

S2. I. Dondoshansky, *BLASTCLUST vers. 6.1,* (NCBI, Bethesda, MD, 2002).

S3. R. L. Tatusov *et al.*, *Nucleic Acids Res.* **29**, 22 (2001).

S4. A. Krause, J. Stoye, M. Vingron, *Nucleic Acids Res.* **28**, 270 (2000).

S5. Z. Yang, *Syst. Biol.* **47**, 125 (1998).

S6. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).

S7. D. L. Swofford, G. J. Olsen, P. J. Waddell, D. M. Hillis, in *Molecular Systematics*, D. M. Hillis, C. Moritz, B. K. Mable, Eds. (Sinauer, Sunderland, MA, 1996) pp. 407-514.

S8. D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) vers. 4.0,* (Sinauer, Sunderland, MA, 2003).

S9. M. J. Sanderson, A. Purvis, C. Henze, *Trends Ecol. Evol.* **13**, 105 (1998).

S10. C. Ané, O. Eulenstein, R. Piaggio, M. J. Sanderson, in preparation.

S11. M. J. Sanderson, A. C. Driskell, R. H. Ree, O. Eulenstein, S. Langley, *Mol. Biol. Evol.* **20**, 1036 (2003).

S12. G. Alexe *et al.*, *DIMACS Technical Report 2002-52* (2002).

S13. N. Mishra, D. Ron, R. Swaminathan, *Machine Learning J.* **56**, 115 (2004).

S14. C. Yan, J. G. Burleigh, O. Eulenstein, in preparation.

S15. K. C. Nixon, *Cladistics* **15**, 407 (1999).

S16. M. Turmel, C. Otis, C. Lemieux, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11275 (2002).