**References and Notes**

1. K. Hughen *et al.*, *Science* **303**, 202 (2004).
2. R. Muscheler *et al.*, *Earth Planet. Sci. Lett.* **219**, 325 (2004).
3. W. S. Broecker, E. Clark, I. Hajdas, G. Bonani, *Paleoceanography* **19**, doi: 2003PA000974 (2004).
4. L. Beaufort, T. de Garidel-Thoron, A. C. Mix, N. G. Pisias, *Science* **293**, 2440 (2001).
5. Alex Wiedenhoeft of the Forest Products Laboratory in Madison, Wisconsin, advised us that it was not possible to identify milligram-sized fragments of tropical woods; rather, a hand-sized piece of wood would be required. He also commented that, once waterlogged, any wood would be dense enough to sink in seawater.
6. T. P. Guilderson, D. P. Schrag, M. A. Cane, *J. Clim.* **17**, 1147 (2004).
7. T. Guilderson, personal communication, 2004.
8. Because conventional radiocarbon ages are calculated using the Libby half life of 5568 years, we have used this value to calculate all age differences in this study for consistency.
9. E. Bard, *Paleoceanography* **3**, 635 (1988).
10. W. S. Broecker, W. C. Patzert, J. R. Toggweiler, M. Stuiver, *J. Geophys. Res.* **91**, 14345 (1986).
11. R. M. Key *et al.*, *Radiocarbon* **44**, 239 (2002).
12. J. F. Adkins, E. A. Boyle, *Paleoceanography* **12**, 337 (1997).
13. J. F. Adkins, K. McIntyre, D. P. Schrag, *Science* **298**, 1769 (2002).
14. K. Matsumoto, T. Oba, J. Lynch-Stieglitz, H. Yamamoto, *Quat. Sci. Rev.* **21**, 1693 (2002).
15. W. S. Broecker, T.-H. Peng, S. Trumbore, G. Bonani, W. Wolfli, *Global Biogeochem. Cycles* **4**, 103 (1990).
16. L. D. Keigwin, *J. Oceanogr.* **58**, 421 (2002).
17. K. Ohkushi, M. Uchida, N. Ahagon, T. Mishima, T. Kanematsu, *Nucl. Instrum. Methods Phys. Res. Sec. B,* **406**, 223 (2004).
18. S. J. Goldstein, D. W. Lea, S. Chakraborty, M. Kashgarian, M. T. Murrell, *Earth Planet. Sci. Lett.* **193**, 167 (2001).
19. M. Andree, *Radiocarbon* **29**, 169 (1987).
20. W. S. Broecker, K. Matsumoto, E. Clark, I. Hajdas, G. Bonani, *Paleoceanography* **14**, 431 (1999).
21. E. L. Sikes, C. R. Samson, T. P. Guilderson, W. R. Howard, *Nature* **405**, 555 (2000).
22. Four benthic-planktic age differences for LGM samples from cores off New Zealand are reported in (*21*): 2.12 thousand years (ky) (depth 2.07 km, planktic $^{14}$C age 24.4 ky); 0.84 ky (depth 2.07 km, planktic $^{14}$C age 25.6 ky); 3.48 ky (depth 2.7 km, planktic $^{14}$C age 24.1 ky); and 0.75 ky (depth 1.3 km, planktic $^{14}$C age 24.8 ky). We choose not to include them because they represent a different age range from those that we studied.
23. We thank L. Beaufort for providing the Admiralty core sample and T. Guilderson for providing a survey of coral-derived prebomb $\Delta^{14}$C values. This paper is funded in part by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. Financial support was provided by NSF grant OCE 02-21979 and by NOAA Consortium on the Ocean's Role in Climate (CORC) grant UCSIO P.O. 10196097-003/NA17RJ1231. This study was funded in part by a Lamont-Doherty postdoctoral fellowship and a fellowship grant (award number CSEF CC3B) from the Comer Science and Education Foundation to S.B., Lamont-Doherty Earth Observatory contribution no. 6663.

2 July 2004; accepted 8 October 2004

# Prospects for Building the Tree of Life from Large Sequence Databases

**Amy C. Driskell,[1,2]* Cécile Ané,[1]†‡ J. Gordon Burleigh,[1]† Michelle M. McMahon,[1]† Brian C. O'Meara,[2]† Michael J. Sanderson[1]**

**We assess the phylogenetic potential of ~300,000 protein sequences sampled from Swiss-Prot and GenBank. Although only a small subset of these data was potentially phylogenetically informative, this subset retained a substantial fraction of the original taxonomic diversity. Sampling biases in the databases necessitate building phylogenetic data sets that have large numbers of missing entries. However, an analysis of two "supermatrices" suggests that even data sets with as much as 92% missing data can provide insights into broad sections of the tree of life.**

More than 100,000 species—about 6% of all those known to science—have at least one molecular sequence archived in public databases, but what fraction of these sequences is phylogenetically informative? Here, we examine two large samples of proteins and show how the answer depends on the pattern of homology among sequences and the distribution of sequences among taxa. We then parse these databases into phylogenetic supermatrices for metazoans and green plants. Although the databases have sampling biases that cause these matrices to be very sparse, they can still provide useful information for building the tree of life.

We examined the phylogenetic information content of the Swiss-Prot database of 120,000 sequences for nearly 7500 taxa and a "taxonomically enriched" subset of GenBank, which consisted of 185,000 amino acid sequences for more than 16,000 green plant taxa (*1*). Clusters of putative homologs were identified via $N \times N$ BLAST searches (*2*). Clustering procedures involve trade-offs among the reliability of homology assessment, the taxonomic breadth, and the accuracy of tree inference. The trade-offs are controlled by the stringency of homology searches and can be adjusted to maximize the phylogenetic utility of resulting clusters, on the basis of the depth and breadth of the phylogenetic question to be addressed (*3*, *4*). Clusters containing at least four taxa are termed "minimal phylogenetic clusters," because unrooted trees with fewer than four taxa contain no information about relations. Although minimal phylogenetic clusters were a small fraction of all clusters found [6.5% and 2.3% for Swiss-Prot and GenBank, respectively (Table 1)], they retained about one-third of the original sequences and a substantial fraction (74% and 95%) of the taxonomic diversity originally contained in the sample.

We screened minimal phylogenetic clusters for the presence of paralogs with a phylogenetic test of orthology (*5*). A species tree cannot be easily deduced from a cluster containing both orthologs and paralogs, although methods for this have been proposed (*6*–*8*). Screening reduced the candidate minimal clusters to smaller sets of orthologous "single-copy" clusters retaining only 24% and 21% of the original sequences, but still covering 59% and 89% of the original taxa in Swiss-Prot and GenBank, respectively (Table 1). These sequences are very sparsely distributed among taxa as measured by their "densities" (Table 1) (*9*).

Further assessment of the phylogenetic utility of these data requires consideration of how the data should be parsed for phylogenetic analyses. One approach is to build gene trees from individual clusters and to assemble these trees using supertree methods (*3*, *10*). Supertree methods require at least partial taxonomic overlap between trees. A set of trees (each inferred from a cluster) with enough taxonomic overlap to allow supertree construction is a "grove" (*1*). The minimum number of groves in a database is a lower bound on the number of supertrees required to encompass all its sequence data. The single-copy green plant proteins form at least 15 groves (Table 1). The largest of these groves minimally includes trees from 814 clusters and contains more than 14,000 taxa—87% of all the green plant taxa in the GenBank database. Swiss-Prot has at least eight times as many groves, which reflects its greater taxonomic breadth but higher fragmentation (Table 1). Both data sets also contain a small number of "orphans," clusters with no taxonomic overlap with other clusters.

[1]Section of Evolution and Ecology, [2]Center for Population Biology, University of California, One Shields Avenue, Davis, CA 95616, USA.

*To whom correspondence should be addressed. E-mail: acdriskell@ucdavis.edu

†These authors contributed equally to this work.

‡Present address: Department of Statistics, University of Wisconsin, Medical Science Center, 1300 University Avenue, Madison, WI 53706, USA.

A second and more widely used strategy is to concatenate clusters into multigene (protein) data matrices. Increasing the sequence data per taxon should improve accuracy (*11*, *12*), a theoretical result supported by several recent phylogenomic studies (*13*–*15*). However, as with supertree construction, the taxonomic structure of the set of clusters limits the size of data matrices that can be assembled from them. To explore these limits, we used an exact algorithm derived in the context of a well-known graph problem on "bicliques" (*5*, *16*) to enumerate all possible concatenated matrices that are both maximal (not contained in larger matrices) and complete (no missing sequences). The largest complete multiprotein Swiss-Prot and GenBank matrices have either many taxa and few proteins or the reverse; none has large numbers of proteins and taxa simultaneously (Fig. 1), and these matrices retain only a small fraction of the original taxonomic diversity in the databases (Table 1). The numbers of taxa and genes in a matrix can be greatly increased by allowing missing sequences, or "holes," in the matrix (*17*, *18*). Such "supermatrices" are also affected by the grove structure of the database, because the missing entries induce patterns of partial overlap among taxa.

Statistics on clusters, groves, and bicliques offer a glimpse of the potential phylogenetic information content of the sequence databases. To explore the feasibility of realizing this potential, we assembled supermatrices spanning a small (but broad) taxonomic sample from each database. We identified the largest grove and constructed a supermatrix from complete matrices having at least 10 clusters and four taxa (*1*). The green plant matrix retained representatives across the group, but Swiss-Prot was culled to metazoans (plus fungal outgroups) to avoid conflicting gene histories of nuclear, mitochondrial, and chloroplast genomes. The resulting two supermatrices had comparable numbers of taxa and shared no sequences. The Swiss-Prot supermatrix had 70 taxa × 1131 genes (6623 sequences and 469,497 characters), an average of 95 genes per taxon. The green plant supermatrix contained 69 taxa × 254 genes (2777 sequences and 96,698 characters), an average of 40 genes per taxon. These are among the largest supermatrices yet analyzed for phylogenetic inference, as well as the sparsest, with 92% and 84% missing entries, respectively (Fig. 2). Yet, although sparse, they are 46 and 75 times as dense, respectively, as the single-copy minimal sequence collections from which they were derived (Table 1).

Trees from these matrices (*1*) broadly agree with conventional views on phylogenetic relations, and many nodes, particularly in the Swiss-Prot tree, are well supported by bootstrap values (figs. S1 and S2). However, some of the "backbone" of the green plant tree is weakly supported, and some unconventional and likely incorrect relations are depicted in both topologies. Many of these unconventional relations have been seen in previous molecular studies (*19*–*22*), including even the nonmonophyly of monocots (*23*). To characterize the sources of these signals, we compared (Fig. 3) every clade in each individual protein tree against the final trees (figs. S1 and S2). Compared with the green plant supermatrix, fewer of the proteins in the Swiss-Prot data conflict with the relations in the final tree. The level of conflict in green plants is remarkably high: More individual protein trees conflict with any given final clade than support it. Nonetheless, many nodes on this tree are still well supported by bootstrap values. Surprisingly little relation was seen between the number of proteins supporting a node and its bootstrap score. For example, placement of the lepidopteran *Spodoptera frugiperda* within the Diptera receives 96% bootstrap support, but only one protein cluster of the 1131 sampled contains sequence for both *Spodoptera* and other dipterans (none contain

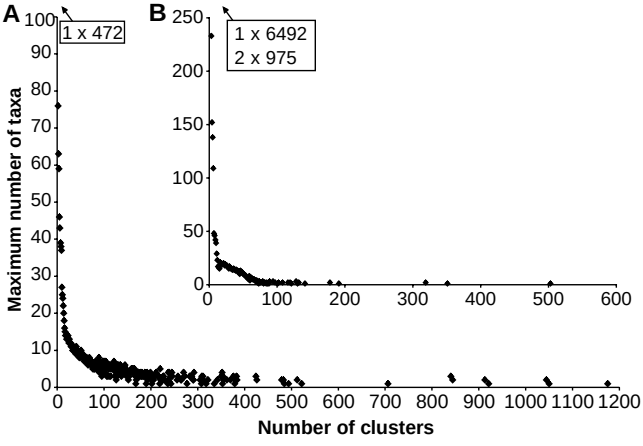


**Fig. 1.** Size distribution of the maximal complete supermatrices (maximal bicliques) for two data sets. Each point represents the biclique with the largest number of taxa for a particular number of clusters. (**A**) Swiss-Prot. (**B**) Green plants from GenBank. Bicliques on the tails are not shown, but their sizes are indicated in the inset boxes.

**Table 1.** Summary statistics on the phylogenetic information content of proteins from Swiss-Prot and GenBank (green plant only) databases. "Minimal phylogenetic clusters" contain at least four taxa. "Nontrivial" bicliques contain at least two genes and four taxa. Density = (number of cells in the taxon × cluster matrix containing ≥ 1 sequence)/(number of taxa × number of clusters) (*1*). See text for other definitions.

| Database statistics | Swiss-Prot release 40.29 | GenBank release 137 (green plant) |
|---|---|---|
| Summary statistics | | |
| Number of sequences in release | 121,218 | 185,418 |
| Total number of sequences clustered | 121,218 | 185,089 |
| Total number of taxa | 7449 | 16,348 |
| Total number of clusters | 64,712 | 59,144 |
| Minimal phylogenetic clusters | | |
| Number of clusters | 4214 (6.5%) | 1365 (2.3%) |
| Sequence coverage | 41,812 (34%) | 65,113 (35%) |
| Taxon coverage | 5538 (74%) | 15,599 (95%) |
| Single-copy clusters | | |
| Number of clusters | 3592 (5.6%) | 853 (1.4%) |
| Sequence coverage | 28,742 (24%) | 39,443 (21%) |
| Taxon coverage | 4404 (59%) | 14,502 (89%) |
| Density | 0.0018 | 0.0021 |
| Groves | | |
| Minimum number of groves | 123 | 15 |
| Number of orphan clusters | 67 | 7 |
| Minimum number of clusters in largest grove | 3183 | 814 |
| Minimum number of sequences in largest grove | 25,272 (21%) | 38,700 (49%) |
| Minimum number of taxa in largest grove | 2695 (36%) | 14,169 (87%) |
| Bicliques | | |
| Number of nontrivial maximal bicliques | 43,576 | 5587 |
| Sequence coverage | 23,855 (20%) | 15,092 (8.2%) |
| Taxon coverage | 1449 (19%) | 4230 (26%) |
| Number of clusters in biclique set | 3187 (4.9%) | 645 (1.1%) |
| Largest biclique (in terms of taxa) | 2 × 76 | 2 × 975 |
| Largest biclique (in terms of clusters) | 352 × 4 | 70 × 4 |

sequence from *Spodoptera* and both *Anopheles* species) and can therefore shed light on the final topology. Other relations, such as the placement of *Spathiphyllum* outside of the main clade of monocots, are not supported by a single protein and must therefore be an emergent property of weak signals buried within several proteins.

These supermatrices differ from other recent phylogenomic analyses of diverse taxa (*13*–*15*, *18*) in two important respects. First, the taxa and proteins in our supermatrices were determined largely by the structure of the databases rather than by decisions of the investigators. Second, they contain more taxa (two or more times as many), but one-fourth or fewer proteins per taxon, with a substantially lower overall density. In two studies with <15 taxa, bootstrap support was uniformly very high across the tree (*14*, *15*) and could be obtained even with a subset of sequences (*14*). In two other studies (*13*, *18*) with larger taxon sets (30 to 36 taxa), many clades were strongly supported but some were not—an expected consequence of increased taxon sampling (*24*). Our results support the general conclusion of these other studies that combining many genes can provide strong support for nodes in a large and complex tree, with the added twist that genes need not be sampled evenly across taxa and a surprising amount of missing data is tolerable (*18*, *25*). Nevertheless, even our threshold of 10 proteins per taxon was not enough to prevent recovering some unconventional results, as our surprising findings regarding monocots document. Such results, if they stem from genomewide long branch–attraction artifacts (*26*), will also occur in phylogenomic studies unless problematic taxa are deliberately or accidentally excluded and will only be overcome by additional taxon sampling or intensive analytical efforts. Therefore, we conclude that exploitation of existing databases, taking into account the inherent sample biases of the data, provides a cost-effective complement to intensive genomewide sequencing efforts, especially if we wish to include large numbers of taxa or remote corners of the tree of life.

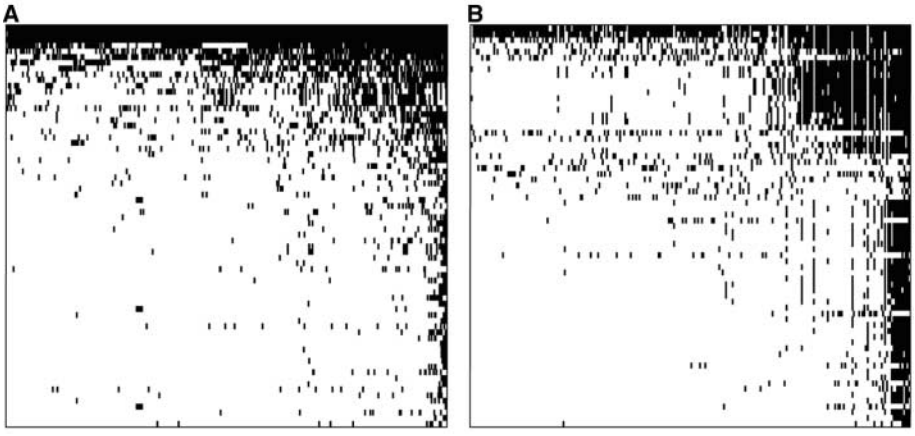


**Fig. 2.** Visualization of the distribution of sequences among taxa for the supermatrices. (**A**) SwissProt metazoan supermatrix. (**B**) GenBank green plant supermatrix. Columns correspond to clusters (proteins); rows to taxa, ordered so that density increases to upper right. Black indicates a sequence is present in that cluster for that taxon.

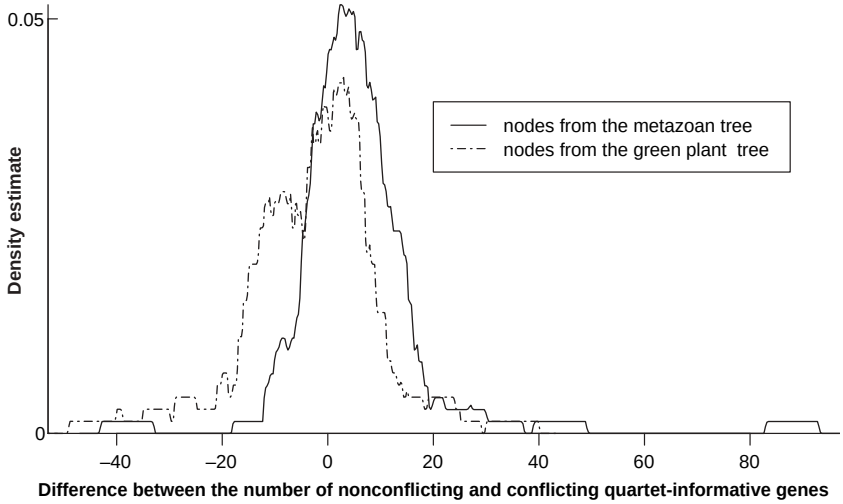


**Fig. 3.** Protein-by-protein distribution of support for clades in supermatrix trees (figs. S1 and S2). A protein tree is "quartet informative" for a given branch (bipartition) of the supermatrix tree if its data set had sequence for at least one taxon in each of the four groups attached to this branch. It is considered "nonconflicting" if it has at least one most-parsimonious (MP) tree displaying the bipartition (pruned to the taxon set for the protein) or if all of its MP trees display a polytomy in place of the branch. Otherwise, the protein is classified as "conflicting." Plots are smoothed histograms (density estimates) of the number of quartet-informative proteins that are nonconflicting minus the number that conflict. Values to the right of the origin are therefore branches in which the majority of proteins do not conflict. Density estimation was done with *R* statistical software (*27*), using a rectangular kernel and bandwidth of 3 for both trees.

### References and Notes

1. Materials and methods are available as supporting material on *Science* Online.
2. I. Dondoshansky, *BLASTCLUST vers. 6.1*, (National Center for Biotechnology Information, Bethesda, MD, 2002).
3. O. R. P. Bininda-Emonds, S. G. Brady, J. Kim, M. J. Sanderson, *Pac. Symp. Biocomp.* **6**, 547 (2001).
4. Z. Yang, *Syst. Biol.* **47**, 125 (1998).
5. M. J. Sanderson, A. C. Driskell, R. H. Ree, O. Eulenstein, S. Langley, *Mol. Biol. Evol.* **20**, 1036 (2003).
6. R. D. M. Page, *Mol. Phylogenet. Evol.* **14**, 89 (2000).
7. J. Kim, B. Salisbury, *Pac. Symp. Biocomp.* **6**, 571 (2001).
8. L. Arvestad, A.-C. Berglund, J. Lagergren, B. Sennblad, *Bioinformatics* **19** (suppl. 1), i7 (2003).
9. M. J. Sanderson, A. C. Driskell, *Trends Plant Sci.* **8**, 374 (2003).
10. M. J. Sanderson, A. Purvis, C. Henze, *Trends Ecol. Evol.* **13**, 105 (1998).
11. P. L. Erdös, M. A. Steel, L. A. Székely, T. J. Warnow, *Rand. Struct. Algorithms* **14**, 153 (1999).
12. D. M. Hillis, *Nature* **383**, 130 (1996).
13. E. Bapteste *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 1414 (2002).
14. A. Rokas, B. Williams, N. King, S. Carroll, *Nature* **425**, 798 (2003).
15. E. Lerat, V. Daubin, N. Moran, *PLoS Biol.* **1**, 101 (2003).
16. G. Alexe *et al.*, *DIMACS Tech. Rep. 2002-52* (2002).
17. C. Yan, J. G. Burleigh, O. Eulenstein, in preparation.
18. H. Philippe *et al.*, *Mol. Biol. Evol.* **21**, 1740 (2004).
19. A. M. D'Erchia, C. Gissi, G. Pesole, C. Saccone, U. Arnason, *Nature* **381**, 597 (1996).
20. M. J. Phillips, D. Penny, *Mol. Phylogenet. Evol.* **28**, 171 (2003).
21. R. Zardoya, Y. Cao, M. Hasegawa, A. Meyer, *Mol. Biol. Evol.* **15**, 506 (1998).
22. M. R. Duvall, A. B. Ervin, *Mol. Phylogenet. Evol.* **30**, 97 (2004).
23. D. E. Soltis *et al.*, *Ann. Mo. Bot. Gard.* **84**, 1 (1997).
24. J. Kim, *Syst. Biol.* **45**, 363 (1996).
25. J. J. Wiens, *Syst. Biol.* **52**, 528 (2003).
26. J. Felsenstein, *Biol. J. Linn. Soc.* **16**, 183 (1978).
27. See www.r-project.org/.
28. This research was supported by the National Science Foundation. Thanks to R. Piaggio and O. Eulenstein for insights on the properties of groves and Wen-Chieh Chang for biclique code. We also thank D. Fernández-Baca and J. Kim.

**Supporting Online Material**

www.sciencemag.org/cgi/content/full/306/5699/1172/DC1
Materials and Methods
Figs. S1 and S2

28 June 2004; accepted 14 September 2004