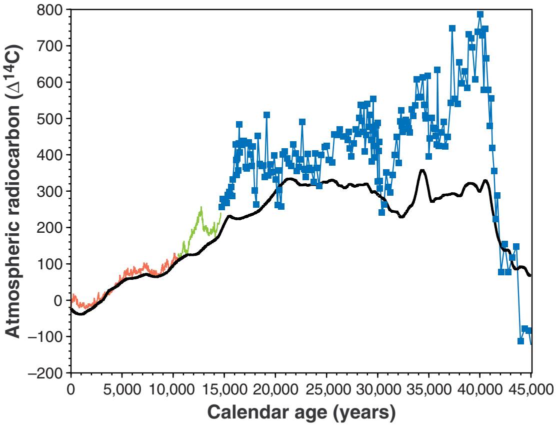


Atmospheric radiocarbon content over the past 45,000 years. Data from tree rings [red line (*11*)] and from foraminifera from the Cariaco Basin [green line (*12*) and blue points (*2*)] are compared to modeled production rate variations (*2*) (black line). Deviations between the data and the model reflect changes in the exchange of carbon between active reservoirs relative to today. The two most important processes for the largest deviations are changes in carbonate sedimentation and in the ocean overturning rate. The data suggest that, assuming constant carbonate flux, the glacial ocean overturned more slowly than it does today.

the deep ocean. Land biota and soils also take up radiocarbon, but export to the deep sea is by far the largest sink for modern atmospheric radiocarbon. If this sink were reduced, the atmospheric radiocarbon content would rise, and rise fast. If a new steady state were reached, and the production rate remained constant, the partitioning of radiocarbon between the atmosphere and deep ocean would shift toward higher atmospheric $^{14}C/^{12}C$ ($\Delta^{14}C$) ratios. This is exactly what Hughen *et al.* found for 11,500 to 13,000 years ago (see the second figure, green line) (*2*).

But there are several important other factors to consider. The production rate of radiocarbon is not constant. Radiocarbon production can be tracked through time by measuring the abundance of other radioactive isotopes produced by cosmic rays, such as ^{10}Be and ^{36}Cl, in ice cores (*3*); these isotopes are not affected by carbon sinks. It turns out that radiocarbon production was higher in the past (see the second figure, black line). But the measured $\Delta^{14}C$ at the Last Glacial Maximum was even higher (see the second figure, blue symbols). Hence, the ocean overturning must have been slower. However, Broecker *et al.* show with a simple calculation that the implied increase in the radiocarbon age of the deep ocean is much larger than their measured age difference from the foraminifera.

How can these disparate results be reconciled? One possibility is that the carbon cycle was fundamentally different at the Last Glacial Maximum. Changes in the burial rate of carbonate sediments can affect atmospheric radiocarbon, because these sediments remove radiocarbon from the system. Another option is that the waters below 2-km depth were much older than those measured by Broecker *et al.* There is not much data, but two papers have found very radiocarbon-depleted waters in the deepest parts of the ocean at the Last Glacial Maximum (*4*, *5*). Clearly we need water column profiles of radiocarbon for the Last Glacial Maximum from both the Atlantic and the Pacific.

Finally, it is possible that 19,000 years ago—the age of some of the data used by Broecker *et al.*—the system was not in a steady state. The atmospheric record in the figure shows that large swings in $\Delta^{14}C$ occurred regularly during the last glacial period. Such swings did not occur over the past 10,000 years of relatively constant climate, during which the radiocarbon cycle was in a steady state. Ice core records in Greenland show that no equivalent period occurred during the last glacial period.

Moreover, the small but growing number of deep radiocarbon values from the Last Glacial Maximum (*1*, *4*, *5*) provides insights into what drives the strength of the overturning circulation: Contrary to a widely held belief, the circulation rate is not driven by surface density gradients in the north-south direction. This conclusion agrees with recent theoretical arguments (*6*–*8*). For example, Huang has shown that a modeled increase in the density of high-latitude waters did not directly result in an increased rate of deep-water formation (*8*). In fact, a large meridional surface density gradient induces a strong vertical stratification, which inhibits the return of deep water to the surface and weakens circulation.

Observational data from the paleoclimate record support this theory. The ocean interior was more highly stratified at the Last Glacial Maximum than it is in the modern ocean (*9*), but the circulation was not much stronger, and was possibly slower, than it is today. More data are needed to determine whether the strength of the overturning circulation depends on winds and tidal energy (*7*, *8*, *10*), rather than on surface warming/cooling or evaporation/precipitation budgets at the surface. The growing paleoceanographic data set shows great promise to answer this question.

References

1. W. Broecker *et al.*, *Science* **306**, 1169 (2004).
2. K. Hughen *et al.*, *Science* **303**, 202 (2004).
3. R. Muscheler *et al.*, *Earth Planet. Sci. Lett.* **219**, 325 (2004).
4. N. J. Shackleton *et al.*, *Nature* **335**, 708 (1988).
5. L. Keigwin, *Paleoceanography*, in press.
6. C. Wunsch, R. Ferrari, *Annu. Rev. Fluid Mech.* **36**, 281 (2004).
7. W. Munk, C. Wunsch, *Deep-Sea Res.* **45**, 1977 (1998).
8. R. X. Huang, *J. Phys. Oceanogr.* **29**, 727 (1999).
9. J. F. Adkins, K. McIntyre, D. P. Schrag, *Science* **298**, 1769 (2002).
10. J. R. Toggweiler, B. Samuels, *Deep-Sea Res.* **42**, 477 (1995).
11. M. Stuiver *et al.*, *Radiocarbon* **40**, 1041 (1998).
12. K. Hughen *et al.*, *Science* **290**, 1951 (2000).

EVOLUTION

Genomic Databases and the Tree of Life

Keith A. Crandall and Jennifer E. Buhay

Although we have not yet counted the total number of species on our planet, biologists in the field of systematics are eagerly assembling the Tree of Life (*1*, *2*). The Tree of Life aims to define the phylogenetic relationships of all organisms on Earth. On page 1172 of this issue, Driskell *et al.* (*3*) propose an intriguing computational method for assembling this phylogenetic tree.

These investigators probed the phylogenetic potential of ~300,000 protein sequences sampled from the GenBank and Swiss-Prot genetic databases. From these data, they generated "supermatrices" and then supertrees. Supermatrices are extremely large data sets of amino acid or nucleotide sequences (columns in the matrix) for many different taxa (rows in the matrix). Driskell *et al.* constructed a supermatrix of 185,000 protein sequences for more than 16,000 green plant taxa and one of 120,000 sequences for nearly 7500 metazoan taxa. This compares with a typical systematics study of, on a good day, four to six partial gene sequences for 100 or so taxa. Thus, the potential data enrichment that comes with carefully mining genetic databases is terrific. However, this enrichment comes at a cost. Traditional phylogenetic studies sequence

The authors are in the Department of Integrative Biology, and K. A. Crandall is also at the Monte L. Bean Life Science Museum and the Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA. E-mail: keith_crandall@byu.edu, crayfish@email.byu.edu

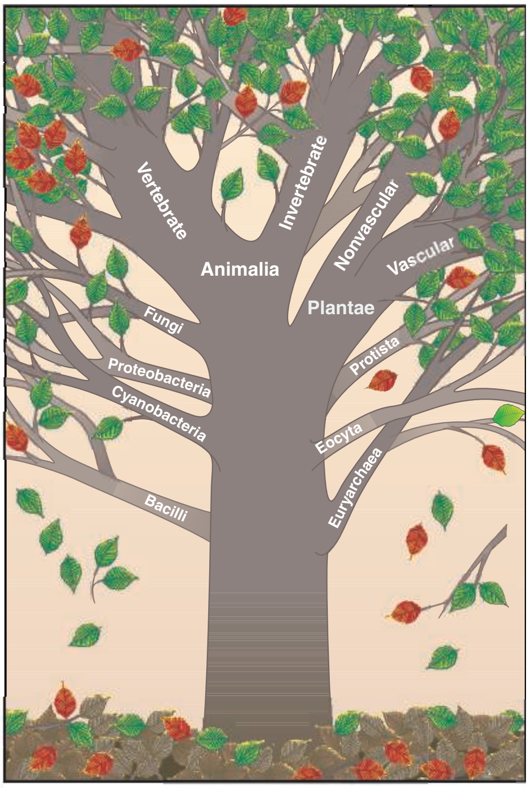
Published by AAAS

the same gene regions for all the taxa of interest while minimizing the overall amount of missing data. With the database supermatrix method, the data overlap is sparse, resulting in many empty cells in the supermatrix, but the total data set is massive.

To solve the problem of sparseness, the authors built a "supertree" (*4*). The supertree approach estimates phylogenies for subsets of data with good overlap, then combines these subtree estimates into a supertree. Driskell and colleagues took individual gene clusters and assembled them into subtrees, and then looked for sufficient taxonomic overlap to allow construction of a supertree. For example, using 254 genes (2777 sequences and 96,584 sites), the authors reduced the green plant supermatrix to 69 taxa from 16,000 taxa, with an average of 40 genes per taxon and 84% missing sequences! This represents one of the largest data sets for phylogeny estimation in terms of total nucleotide information; but it is the sparsest in terms of the percentage of overlapping data. Yet even with such sparseness, the authors are still able to estimate robust phylogenetic relationships that are congruent with those reported using more traditional methods. Computer simulation studies (*5*) recently showed that, contrary to the prevailing view, phylogenetic accuracy depends more on having sufficient characters (such as amino acids) than on whether data are missing. Clearly, building a supertree allows for an abundance of characters even though there are many missing entries in the resulting matrix.

Several questions remain, however, about this strategy. First, the supertree strategy depends fundamentally on our ability to distinguish between orthologous (derived from a speciation event) and paralogous (derived from a duplication event) gene sequences (*6*). The methods to draw this distinction are in their infancy. Little work has been done to compare such methods in terms of their accuracy and their robustness with respect to data that do not fit underlying assumptions (such as neutral evolution). The distinction between the two types of gene sequences typically relies on a well-populated database. Second, supertree approaches themselves are controversial, in part because the methodology results in a degree of disconnect between the underlying genetic data and the final tree produced. Moreover, this strategy has yet to be validated by computer simulation or well-established phylogenetic methods. Third, the supertree approach makes a fundamental assumption: that a bifurcating tree topology represents the genomic evolutionary history of species. This assumption has been called into question because of the reality of genetic exchange across species boundaries through mechanisms such as horizontal gene transfer and hybridization. Depicting genealogical relationships as networks might better represent the true underlying biology (*7*, *8*).

Nonetheless, the ability of Driskell *et al.* to estimate apparently robust phylogenetic estimates from an impressively large and equally impressively sparse data set—all collected from existing databases—has important implications for future work on the Tree of Life. This and other studies demonstrate the importance of computational approaches compared with brute-force gene sequencing. In addition, Driskell *et al.*'s approach allows for a more comprehensive sampling of existing data, which enables characterization of evolution across the greatest possible diversity of life. Driskell *et al.* report that more than 100,000 species have at least one molecular sequence archived in public databases. They fail, however, to mention the extreme sampling bias in these databases toward vertebrate animals and green plants. They suggest that the number of these favored species is 6% of those known to science, which is roughly equivalent to the National Science Foundation's estimate of 1.7 million species. However, estimates of the total number of species on Earth (not just those known to science) range from 4 million to 100 million. The computational methodology of Driskell and colleagues may be well suited to future efforts both in terms of taxon sampling and gene sampling for maximizing coverage of the Tree of Life. This represents an extreme departure from the notion of "barcoding" all of life, which emphasizes sequencing one gene for all species (*9*). Both approaches rely heavily on having a well-populated database, but the supertree strategy does not have the constraint that these data must be from the same gene. It may turn out that two different databases are needed for these two distinct purposes (that is, establishing relationships versus diagnosing species).



Building the Tree of Life. A current view of the Tree of Life (*7*). Information is biased toward vertebrate animals and vascular plants (the thick branches); lesser-known groups such as bacteria, fungi, and protists are largely underrepresented. Also shown are species known to science (green leaves), extinct species (leaf litter, brown), endangered species (falling leaves), and species for which "barcode" information is available (red leaves).

If the supertree approach establishes the trunk and thick branches of the Tree of Life, then perhaps the barcoding approach is more appropriate for discerning the twigs and leaves of the tree (see the figure). Currently, most attention has been focused on the trunk at the expense of the leaves. However, the leaves are dropping quickly. We are losing 27,000 species each year while only describing 18,000 new species (*10*). The Driskell *et al.* study provides hope for combining diverse and sparse data sets collected from both leaf and trunk areas of the Tree of Life to provide a robust estimate of this tree, but this should in no way undermine efforts to characterize as many leaves as possible before they hit the ground. Future applications of Driskell *et al.*'s computational method and verification of its performance compared with computer-simulated known phylogenetic histories and empirically "known" histories will provide further insights into the generality of mining our genetic databases to assemble the Tree of Life.

References and Notes

1. M. Pagel, *Nature* **401**, 877 (1999).
2. A new NSF program funds computational approaches for "assembling the Tree of Life" (AToL). Total AToL program funding is $13 million for fiscal year 2004. NSF, *Assembling the Tree of Life: Program Solicitation NSF 04-526* (www.nsf.gov/pubs/2004/ nsf04526/nsf04526.pdf).
3. A. C. Driskell *et al.*, *Science* **306**, 1172 (2004).
4. M. J. Sanderson *et al.*, *Trends Ecol. Evol.* **13**, 105 (1998).
5. J. Wiens, *Syst. Biol.* **52**, 528 (2003).
6. J. W. Thornton, R. DeSalle, *Annu. Rev. Genomics Hum. Genet.* **1**, 41 (2000).
7. M. C. Rivera, J. A. Lake, *Nature* **431**, 152 (2004).
8. W. Doolittle, *Science* **284**, 2124 (1999).
9. P. D. N. Herbert *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14812 (2004).
10. E. O. Wilson, *The Diversity of Life* (Norton, New York, 1992).