Minireview

The imbalanced supertree of flowering-plant phylogeny

Sean W Graham and Quentin CB Cronk

Address: UBC Botanical Garden and Centre for Plant Research, MacMillan Building, 2357 Main Mall, The University of British Columbia, Vancouver BC, V6T 1Z4, Canada.

Correspondence: Quentin CB Cronk. E-mail: quentin.cronk@ubc.ca

Published: 15 July 2004

Genome ***Biology*** 2004, **5:**236

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/8/236

Abstract

Two contrasting approaches have been used to construct the overall tree of life from molecular data: one involves the analysis of single large datasets, while the other involves joining many independent smaller analyses into a supertree. A recent study uses the latter approach to produce the most complete phylogeny yet of flowering plant families.

Comparative biology and the tree of life

Many questions in biology cannot be fully examined without a phylogenetic framework. Examples are developmental questions, such as the nature and origin of leaves; character-evolution questions, such as how many times leaves have evolved; and ecological questions, such as the correlation between function and morphology during leaf evolution. Thus, when large, relatively reliable phylogenies first became available for flowering plants - fueled by the same technical advances in computational and molecular biology that promoted the rise of the genomic sciences (see page 136 of [1]) - an explosion of new biology resulted. Answering the most interesting questions, particularly those concerning the origin of particular traits, require the phylogenies to be as large and as complete as possible. We are still far from completeness, however: millions of species of organism (many still uncollected) are thought to be on our planet, and only a fraction have been subject to comparative gene-sequence analysis for phylogenetic studies.

There are two basic approaches to adding species to the tree of life. One is to perform ever larger single analyses; this is, in theory at least, the most advantageous approach, as all species are analyzed using comparable data. But the analysis of even a few hundred taxa can pose serious computing challenges, although these may be overcome in the future by using gridded computer power [2]. Another approach to generating a complete tree of life is to use existing data, which often take the form of numerous small independent analyses containing some overlap of species. In a type of 'meta-analysis', these independent analyses can be 'stitched together' into supertrees using various algorithms. This supertree approach may have the shortcoming that it is a composite of disparate analyses, but its main advantages are that it mirrors how molecular systematics is being done in practice, and that it can use datasets that already exist.

Methods for constructing supertrees were developed in the early 1990s [3,4] and most commonly use matrix representation with parsimony (MRP). In the MRP approach, each tree is represented as a matrix, the matrices are combined, analyzed using parsimony, and the most parsimonious tree that fits all the matrix information is selected. The matrix representation may take different forms to accommodate various theoretical considerations [5,6] and may be weighted to allow for differences in the reliability of the data. The MRP method, as implemented in a recent software program [7], has been used by Davies *et al.* [8] to build the most complete evolutionary tree of the families of flowering plants to date. The authors then use this tree to answer a comparative biology question: why have some lineages led to groups of very high diversity while other lineages of equal age have produced groups of very low diversity?

Unbalanced evolution

The 20th century biologist John Haldane is said to have mused about God's inordinate fondness for beetles. A similar predilection could be construed for the flowering plants (angiosperms), which number in the hundreds of thousands of species. The earliest discernible branch point in the lineage of flowering plants yields two branches; one seems to comprise almost all of the living species, but the other has only a single modern survivor, *Amborella trichopoda* [9]. The discovery that this hitherto obscure South Pacific shrub represents a major branch arising from the deepest point of angiosperm phylogeny resulted in a flurry of exciting new research on its biology [10], and a massive re-evaluation of our understanding of the early evolution of modern angiosperms. Similar remarkable numerical disparities are also found scattered throughout the angiosperm portion of the tree of life. There are about 10,000 species of grass, for example, making Poaceae one of the largest families - it is also, of course, one of the most economically and ecologically important plant groups. The closest relatives of the grasses are the relatively obscure families Ecdeiocoleaceae (tussocky cord rush) and Joinvilleaceae (joinvillea), consisting of two species apiece.

Are these and other imbalances in the tree of life mere accidents of history? And why do some groups prosper, while others fade, or persist as 'living fossils' - faint echoes, perhaps, of what might have been? Davies *et al.* [8] address the puzzle of differential disparity by bringing to the table a new global estimate of angiosperm phylogeny. Since the first broad phylogenetic study of the gene encoding the large subunit of the Rubisco protein (*rbc*L) in 1993 [11], numerous broad-level angiosperm phylogenies have accumulated. Using the MRP method, Davies *et al.* [8] constructed an angiosperm supertree by stitching together a patchwork of approximately 50 overlapping phylogenetic studies - based on a variety of different gene combinations and morphological characters - into a single (nearly) complete family-level angiosperm supertree with almost 400 terminal 'twigs'.

Two measurements of tree imbalance [12,13] and diversification rate over the supertree were used by Davies *et al.* [8] to demonstrate a significant disparity in diversity. They pinpointed particular nodes on the supertree where there is a change in the rate of change of species diversification. In addition to demonstrating a substantial lability in the rate of diversification, a 'top ten' list was drawn up of nodes associated with the most extreme imbalances in diversification rate (Table 1); the grass case mentioned above is on the list, for example. The 'hotspot' nodes can be related back to possible biotic and abiotic triggers of diversification-rate changes. The authors assessed a number of characters as potential triggers such as biotic versus abiotic pollination - insect pollination is often highly specific and might conceivably drive speciation by promoting genetic isolation. A major conclusion of their article is that no such correlations were found, although convincingly ruling out the importance of character triggers (sometimes called 'key innovations') would require more formal reconstructions of character evolution and the assessment of many more characters. Davies *et al.* [8] do

Table 1

Sister taxa at the top ten most imbalanced nodes of flowering-plant phylogeny

	Diverse clade		Less diverse clade	
Node	Clade name	Geographical distribution	Clade name	Geographical distribution
1	Lamiales I (mints)	Cosmopolitan	Plocospermataceae	Central America
2	Poaceae (grasses)	Cosmopolitan	Ecdeiocoleaceae (tussocky cord rush)	Australia
3	Monocots	Cosmopolitan	Acoraceae (sweet flag)	Old World and North America
4	Asparagales (asparagus)	Cosmopolitan	Xeronemataceae	New Zealand and New Caledonia
5	Lamiales II (mints)	Cosmopolitan	Tetrachondraceae	New Zealand and Patagonia
6	Fabaceae (legumes)	Cosmopolitan	Surianaceae	Pan subtropical to tropical
7	Caryophyllales I (carnations)	Cosmopolitan	Asteropeiaceae and Physenaceae	Madagascar
8	Caryophyllales II (carnations)	Cosmopolitan	Stegnospermaceae	North and Central America
9	Ranunculales (buttercups)	Cosmopolitan	Eupteleaceae	East Asia
10	Cyperaceae and Juncaceae (sedges and rushes)	Cosmopolitan	Thurniaceae	North and South America

The common clade name or an example of a representative species is given in brackets after the formal name. Most of the diverse clades refer to a subset of the group noted (for example, monocots refers to all monocots except Acoraceae); clades annotated I or II refer to different subsets of the same larger clade (for more details see [8]).

demonstrate a weak tendency for the diversification rate itself to be inherited along the tree, which is consistent with the idea that these rates may be based on inherited organismal traits. One potential problem with the analyses of diversification rates, which is addressed by Davies *et al*. [8], is that the sizes of clades are not independent, as a result of the hierarchical nesting of phylogenies [14]. Given that exceptionally big or small families occur, the larger clades that contain these families will also tend to be bigger or smaller, respectively. To solve this problem of non-independence, the authors devised a novel heuristic method that, for the purposes of subsequent calculations, adjusts species counts in clades shown to have a change in diversification rate to match those seen in their sister clade.

A tree of all genomes

The supertree constructed by Davies *et al*. [8] can be viewed as the first major family-level treatment of the angiosperm portion of the tree of life - something of a landmark event. But it is rather a coarse approximation; the 'pixels' of resolution are entire families of flowering plants, rather than individual species. Improving the resolution and accuracy of angiosperm phylogeny remains a major goal. A further goal is a robust species-level tree of all organisms, but this is a challenge substantially greater in scope than most genome projects, because of the number of species involved, the desperate need for taxonomic work to define what the units (species) are and the need to better characterize the degree to which the tree of life metaphor breaks down among closely related species as a result of lateral gene transfer and related processes. Addressing the latter question will ultimately require the fusion of two disparate fields: comparative genomics and tree of life studies. A 'tree of all genomes' would provide the most fundamental insights into the kinds of molecular evolutionary processes and patterns that underpin all of biology. Such a tree would be complex, however, as organellar and nuclear genomes from the same organism may have different histories, and the nuclear genome is a composite of elements that, to a greater or lesser extent, also have independent histories. In this context, supertree reconstruction, although a pragmatic option, will always be more problematic to interpret than large primary analyses in which the data are consistent across the tree.

More information is needed on the relative contributions that speciation and extinction make to species-diversification rates. Currently these two distinct processes are conflated in a single measure of diversity; teasing them apart will require substantial new evidence from the fossil record. More realistic short-term tasks include the use of large-scale phylogenies for explicit reconstructions of character evolution in order to assess better the circumstances under which differential diversification rates may occur. A recent study [15] that used a less complete phylogenetic framework of the angiosperms has demonstrated that a particular floral characteristic (bilateral symmetry) can play a key role in angiosperm diversification rates. This is in contrast to the absence of correlations found among the set of characters examined by Davies *et al*. [8]. With the advent of large-scale trees and supertrees, addressing whether there is a detectable correlation between parameters of interest is now becoming a more tractable problem at the level of angiosperm phylogeny as a whole.

References

1. Felsenstein J: *Inferring Phylogenies*. Sunderland MA: Sinauer Associates; 2004.
2. **Cyber Infrastructure for Phylogenetic Research (CIPRes)** [http://landscape.sdsc.edu:8080/CIPRes]
3. Baum BR: **Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees.** *Taxon* 1992, **41:**3-10.
4. Ragan MA: **Phylogenetic inference based on matrix representation of trees.** *Mol Phylogenet Evol* 1992, **1:**53-58.
5. Purvis A: **A composite estimate of primate phylogeny.** *Philos Trans R Soc Lond B Bio Sci* 1995, **348:**405-421.
6. Bininda-Emonds ORP, Sanderson MJ: **Assessment of the accuracy of matrix representation with parsimony analysis supertree construction.** *Syst Biol* 2001, **50:**565-579.
7. Salamin S, Hodkinson TR, Savolainen V: **Building supertrees: an empirical assessment using the grass family (Poaceae).** *Syst Biol* 2002, **51:**136-150.
8. Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V: **Darwin's abominable mystery: insights from a supertree of the angiosperms.** *Proc Natl Acad Sci USA* 2004, **101:**1904-1909.
9. Qiu Y-L, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW: **The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes.** *Nature* 1999, **402:**404-407.
10. Thien LB, Sage TL, Jaffre T, Bernhardt P, Pontieri V, Weston PH, Malloch D, Azuma H, Graham SW, McPherson MA, *et al.*: **The population structure and floral biology of *Amborella trichopoda* (Amborellaceae).** *Ann Mo Bot Gard* 2003, **90:**466-490.
11. Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, PriceRA, Hills HG, Qiu YL, *et al.*: **Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L.** *Ann Mo Bot Gard* 1993, **80:**528-580.
12. Slowinski JB, Guyer C: **Adaptive radiations and the topology of large phylogenies.** *Am Nat* 1993, **142:**1019-1024.
13. Purvis A, Katzourakis A, Agapow PM: **Evaluating phylogenetic tree shape: two modifications to Fusco and Cronk's method.** *J Theor Biol* 2002, **214:**99-103.
14. Sanderson MJ, Donoghue MJ: **Shifts in diversification rate with the origin of angiosperms.** *Science* 1994, **264:**1590-1593.
15. Sargent RD: **Floral symmetry affects speciation rates in angiosperms.** *Proc R Soc Lond B Biol Sci* 2004, **271:**603-608.