Perspective

Of Genes and Genomes

DAVID M. HILLIS AND J. J. BULL

VIRTUALLY ALL COMPARATIVE STUDIES OF BIOLOGICAL variation among species depend on a phylogenetic framework for interpretation (*1*). However, with few exceptions we have not directly observed any phylogeny. This fact means that, for the most part, phylogenetic methods have not been tested directly.

One method of testing phylogenetic methods indirectly is to simulate the evolution of a gene or set of genes on a computer and then test the ability of various methods to reproduce the simulated history. Such simulations are useful for understanding the consequences of various assumptions, but do not indicate which assumptions are biologically plausible. Worse, the assumptions of an algorithm used to simulate phylogeny can be matched exactly by the assumptions of a particular method of phylogenetic analysis, which may give a false impression that the particular method is "the best" for reconstructing phylogeny.

Another approach to testing phylogenetic methods is to examine real groups whose histories have been controlled by humans: domesticated and laboratory organisms. Although complete histories rarely exist, incomplete breeding records are available for some crop plants (*2*) and laboratory animals (*3*). In particular, inbred strains of laboratory mice have provided a useful system for comparing known phylogenies to inferred phylogenies (*3*). In this issue, Atchley and Fitch (*4*) continue these studies by examining a phylogeny of 24 inbred strains inferred from genetic data on 144 separate loci. Because a portion of the phylogeny is known from breeding records, they are able to examine the concordance between the trees inferred from various sets of genes and the known relationships among lineages.

In the nuclear genome of sexually reproducing organisms, the history of alleles at a single locus does not necessarily fit exactly with the phylogeny of species (*5*). Within any lineage, several alleles may exist at a given locus at any point in time. If a polymorphic lineage becomes divided (speciation), each of the daughter lineages may also be polymorphic. Subsequent differential fixation of alleles may not be correlated with the historical relationships of the lineages. The phylogenetic tree of alleles at this locus (the gene tree) would then differ from the phylogenetic tree of the taxa (the species tree).

The potential incongruence between gene trees and species trees suggests that inference of species trees should be based on analysis of multiple, independent loci. Deviations between gene trees and species trees should be uncorrelated among independent loci, so in analyses of several loci the historical signal should be apparent over the background noise of random fixations of polymorphisms. Atchley and Fitch demonstrate this principle for the mouse phylogeny: analyses of all 144 loci reproduce the known genealogical relationships. However, they also found that some subsets of loci were far more effective for reconstructing the phylogeny than were others. Analysis of the 70 protein-encoding loci alone also reproduced the correct relationships, but the relationships suggested by both the 41 immune-system loci and the 22 endogenous viral loci were at variance with historical records.

Why are some sets of loci less effective for reconstructing species trees? In the case of the immune-system genes, many of the loci are tightly linked. Atchley and Fitch note that this effectively reduces the number of independent loci. This not only reduces sample size, but may multiply the stochastic effects of single gene trees. For instance, the nine major histocompatibility loci that map to the same region of chromosome 12 are treated as nine independent characters. However, fixation of any one of these loci is likely to be closely tied to fixation at the other linked loci. Therefore, correlated fixation events (which are not randomly distributed among the lineages) can overwhelm the historical signal, thus producing a tree at variance with the recorded genealogy. The message is that analyses of multiple loci need to involve independent loci to overcome the stochastic effects of gene trees.

The endogenous viral loci may present a second problem. Although these viral loci are thought to be relatively stable, their retroviral origins suggest a simple mechanism for convergence among unrelated lineages (*6*). Therefore, although the alleles at viral loci may be homologous (similar because of common ancestry), the relevant common ancestors may be viruses rather than mice (that is, the genes may be xenologous rather than orthologous).

Although inbred mouse strains have provided a fruitful means for assessing phylogenetic methods, the system has some limitations. Evolution among the inbred lines occurred relatively slowly, and largely through fixation of alleles polymorphic in their ancestors rather than through fixation of new mutations. Although experimental replication of a given genealogical pattern is possible, the decades necessary to do so preclude its feasibility. Ancestral populations of mice cannot be maintained without further change, so comparisons of inferred to ancestral genotypes are not possible. Much of the history of the lineages is unrecorded, thereby limiting the comparisons between inferred and actual phylogenies. Finally, the crossing of inbred strains to produce new lines results in a genealogy that is fundamentally at odds with the bifurcating phylogeny model. To date, the methods necessary to infer such reticulations have not been developed, and the "correct" bifurcating tree is not always obvious for a system that includes reticulations. Nonetheless, the mice offer advantages over those systems that are more prone to manipulation, such as viruses (*7*). For instance, the complexities offered by major histocompatibility complex loci and integrated viruses are not feasibly studied in simpler systems, and represent examples of biological complexities that are rarely introduced into simulation studies. The current Atchley and Fitch study thus stands as one of the few empirical tests in an exciting field with ever-broadening applications.

Department of Zoology, The University of Texas, Austin, TX 78712.

REFERENCES

1. N. Eldredge and J. Cracraft, *Phylogenetic Patterns and the Evolutionary Process* (Columbia Univ. Press, New York, 1980); J. Felsenstein, *Am. Nat.* **125**, 1 (1985); D. R. Brooks and D. A. McLennan, *Phylogeny, Ecology, and Behavior* (Univ. of Chicago Press, Chicago, 1991); P. H. Harvey and M. D. Pagel, *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
2. B. R. Baum, in *Cladistics: Perspectives on the Reconstruction of Evolutionary History*, T. Duncan and T. F. Stuessy, Eds. (Columbia Univ. Press, New York, 1984), pp. 192–220.
3. W. M. Fitch and W. R. Atchley, *Science* **228**, 1169 (1985); in *Molecules and Morphology in Evolution: Conflict or Compromise?*, C. Patterson, Ed. (Cambridge Univ. Press, Cambridge, 1987), pp. 203–216.
4. W. R. Atchley and W. M. Fitch, *Science* **254**, 554 (1991).
5. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
6. R. Weiss, N. Teich, H. Varmus, J. Coffin, *RNA Tumor Viruses* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ed. 2, 1982); N. A. Jenkins, N. G. Copeland, B. A. Taylor, B. K. Lee, *J. Virol.* **43**, 26 (1982).
7. M. E. White, J. J. Bull, I. J. Molineux, D. M. Hillis, in *The Unity of Evolutionary Biology: Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology*, E. Dudley, Ed. (Dioscorides Press, Portland, OR, 1991), pp. 935–943.