Natural selection and phylogenetic analysis

Scott V. Edwards[1]
Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138

If Darwin were to survey the entirety of the biological sciences today, he would be pleased to observe how central phylogenies and "tree thinking" are to integrative research (1). Biologists of all stripes now realize that phylogenies are not exotic, but fundamental and routine tools for understanding not only history but mechanism, organization, and function of biological networks at all levels, from molecular and cellular to ecological. The last two decades have seen an explosion of sophisticated statistical methods for inferring phylogenetic trees (2), and these methods are remarkably robust to a variety of forces that can conceivably derail phylogenetic analysis and lead researchers to incorrect conclusions about phylogenetic relationships—forces such as vagaries of the molecular clock, changing base compositions of DNA sequences, even evolutionary convergence, whether driven by natural selection or simple biases of mutation. Yet some genes in some groups of species exhibit evolutionary convergence on such a vast scale that even the best phylogenetic methods fail and erroneous relationships result. The report by Castoe et al. in this issue of PNAS (3) documents an example of rampant convergence in the mitochondrial DNA of snakes, and it raises intriguing questions as to how widespread such convergence is in molecular data.

Convergence is the acquisition of similar phenotypic or genetic states in unrelated lineages, and is usually assumed to be driven by natural selection. Molecular data are by no means immune to convergence, but the type of convergence most often observed, called homoplasy, can be thought of as the product not of natural selection but of one of many kinds of biases—developmental, as observed in morphological traits (4), or mutational, as frequently observed in DNA sequence data (such as the bias for C–T transitions in animal mitochondrial DNA). Although ubiquitous, homoplasy usually occurs at a low enough rate, and at few enough sites in the DNA sequence data collected by researchers, that it generally does not pose a problem for phylogenetic analysis, and systematists have developed a number of ways to detect, quantify, and deal with it (2). By contrast, there have been relatively few cases in which adaptive convergence has shaped the evolution of particular genes to such an extent that it dominates their phylogenetic signal (5).

In analyzing DNA sequences from two new mitochondrial genomes from snakes, as well as additional mitochondrial genomes from squamates (lizards and snakes), Castoe et al. (3) have documented convergence in mitochondrial protein-coding genes on a scale hitherto unappreciated. They reach this conclusion by comparing their mitochondrial tree, in which agamid lizards and snakes form a clade to the exclusion of iguanas and chameleons, with the tree yielded by nuclear DNA sequences, in which the Iguania (consisting of iguanas, chameleons, and agamid lizards) is monophyletic. The tree implied by whole mitochondrial genomes thus contradicted the signal in much previous phylogenetic data, resulting in a lack of congruence, the ultimate arbiter of accuracy in phylogenetic analysis.

The data analyzed by Castoe et al. (3) are noteworthy in a number of ways. The signal in a relatively small number of sites in the mtDNA genomes appears to overwhelm the signal in the remainder of the mitochondrial genome. The authors make a good case that the patterns found in the mtDNA sequences are the result not of standard homoplasy at the nucleotide level but rather of selection-driven convergence at the amino acid level. They point out that the second positions of codons, which usually exhibit low levels of homoplasy in vertebrate data sets, nonetheless yielded a tree that linked agamids and snakes, as do amino acid sequences. The snake and agamid mtDNA sequences did not exhibit conspicuous base compositional patterns that would result in a misleading tree. A well-known signal that can mislead phylogenetic analysis is long-branch attraction, in which homoplasy can accumulate between unrelated lineages to such an extent that phylogenetic analysis

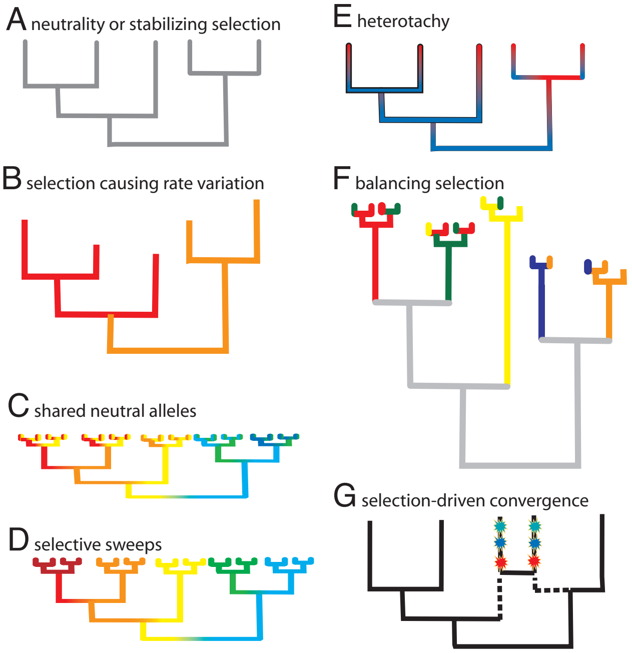
A neutrality or stabilizing selection
E heterotachy
B selection causing rate variation
F balancing selection
C shared neutral alleles
D selective sweeps
G selection-driven convergence

Fig. 1. Ways in which natural selection can influence phylogenetic reconstruction. Colors of branches correspond to different species from which sequences are sampled, except in *E*, wherein colors indicate different rates of evolution at a site. (*A*) A gene tree of five species whose evolution is largely neutral or dominated by stabilizing selection. (*B*) Violations of the molecular clock caused by directional selection along lineages. (*C* and *D*) Contrast between shared polymorphisms commonly observed between closely related species at neutral loci (*C*) versus reciprocal monophyly of alleles between closely related species driven by selective sweeps (*D*). (*E*) Heterotachy, the change in rate of sites over time, may or may not be driven by natural selection. (*F*) Balancing selection can create patterns of "transspecies evolution," such as observed at genes of the major histocompatibility complex. (*G*) Selection-driven convergence of amino acid substitutions (starbursts) in unrelated lineages causes misleading phylogenies, drawing lineages together that are in fact unrelated (true relationships indicated by dotted lines).

Author contributions: S.V.E. wrote the paper.

The author declares no conflict of interest.

See companion article on page 8986.

[1] E-mail: sedwards@fas.harvard.edu.

groups them together (6). However, long branch attraction is predicted to yield a pattern in which the sites with the highest evolutionary rate show the greatest signal favoring the wrong tree (7); this was not the case in their data. A careful process of elimination drove the authors to the conclusion that pervasive adaptation at the level of amino acids is providing the misleading signal in these reptile mitochondrial genomes.

The Castoe et al. paper (3) raises an important question: Is natural selection a universal hindrance to phylogenetic analysis? (Fig. 1). The question has not often been tackled head on; usually challenges to phylogenetic analysis are framed not by the evolutionary forces themselves but by the consequences of those forces for changing the rates and patterns of substitution within and between lineages over time. A review of various kinds of forces suggests that natural selection need not be a problem for phylogenetic analysis (Fig. 1). For example, stabilizing selection, probably the most common type of selection on proteins, simply lowers the overall rate of evolution (8) (Fig. 1*A*). Directional selection resulting in novel substitutions along a lineage might violate the molecular clock only moderately (Fig. 1*B*), a situation that is dealt with well by many phylogenetic methods (2, 9). When several alleles per species are sampled, directional selection can "clean up" phylogenies such that species appear in discrete clusters in gene trees even when those same species do not form discrete clusters at genomic loci evolving neutrally (10) (Fig. 1 *C* and *D*). By contrast, some kinds of natural selection, such as balancing selection (frequency-dependent selection or heterozygous advantage) can produce bizarre phylogenetic trees. By continually rescuing rare alleles from extinction by genetic drift, balancing selection prolongs the lifespan of alleles such that allelic lineages can persist through many speciation events, sometimes spanning tens of millions of years, resulting in trees that appear scrambled with respect to species boundaries even if the gene tree itself is reconstructed accurately (Fig. 1*F*). Phylogenetic trees of major histocompatibility complex (MHC) genes fall into this category (11).

Other aberrant patterns of molecular evolution, such as heterotachy (when the rate of evolution of sites changes over time) have recently emerged as a potentially serious problem for phylogenetic analysis (12–14) (Fig. 1*E*). However, neither heterotachy nor deviations from a clock need be explained by natural selection; one might first look to changes in generation time to explain heterotachy, and aberrant clocks are routinely accounted for in this way or by fluctuations in the neutral space of alleles or fixation of slightly deleterious mutations (15). The type of selection-driven convergence identified by Castoe et al. (3), especially when spread throughout the gene(s) being used for phylogenetic analysis, is perhaps the most insidious, and there are no sure-fire ways for phylogenetic analysis to deal with it (Fig. 1*G*).

Castoe et al. (3) pose a crucial unanswered question that begs for experimental analysis: What has caused this widespread molecular convergence? The amino acid substitutions found to be shared between agamid lizards and snakes may facilitate the extreme shifts in metabolic rate and high metabolic efficiency exhibited by these groups and may have fundamentally altered the reducing and coupling functions of the mitochondrial proton pump. Perhaps mitochondrial proteins act as such a tightly coupled integrated unit that physiological adaptations require concomitant changes throughout the 13 proteins of the genome. Castoe et al. raise the possibility that nuclear genes may also be subject to such rampant convergence. Although this remains a possibility, it is less likely that such convergence could occur on such a wide scale, across so many genes, that it would mislead phylogenetic analysis. Where such phenomena might be found is when the base composition of an entire gene or genome has shifted from that of its close relatives and has come to resemble an unrelated lineage, as was recently documented for the mammalian *RAG1* gene (16). As whole-genome sequencing accelerates, cases of widespread aberrant signal in the nuclear genome will no doubt crop up.

Because of its ease of amplification and sequencing, the mitochondrial genome became a workhorse of phylogenetics near the species level (phylogeography) during the 1990s (17), and in recent years whole-mitochondrial genome sequencing has been used to understand the phylogenetic relationships of many groups, especially vertebrates, for which there are now hundreds of complete genomes. Its rapid evolution clearly makes it a boon for analysis among close relatives, but some have questioned its utility as a phylogenetic marker among higher taxa: its evolutionary rate is rapid enough that high-frequency changes such as transitions often need to be masked so that phylogenetic noise does not swamp out signal (18). Indeed, given the increasing appreciation that phylogenies represent trees of species and lineages, each of which comprise many independently segregating genes whose gene trees inevitably vary at least slightly from one another, systematists today would question the sole use of a mitochondrial gene trees as a simple proxy for the relationships of the *species* in which that gene tree is embedded (19). Methods for estimating species trees—the trees of species and lineages in which gene trees percolate through history—are increasingly available and derive their power not from the accumulation of many sites within single genetic loci such as mtDNA, but via the signal in many loci, each of which exhibits phylogenetic signals that are correlated across loci because of their shared history, namely the species tree. For this reason, the motivation for mitogenomic studies (3, 18) is not phylogenetics per se, but a deeper understanding of mitochondrial genome evolution, a goal that would make Darwin and his intellectual descendants justly proud.

1. O'Hara RJ (1988) Homage to Clio, or, toward an historical philosophy for evolutionary biology. *Syst Zool* 37:142–155.
2. Felsenstein J (2003) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
3. Castoe TA, et al. (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci USA* 106:8986–8991.
4. Wake DB (1991) Homoplasy - the result of natural selection, or evidence of design limitations? *Am Nat* 138:543–567.
5. Stewart CB, Schilling JW, Wilson AC (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* 330:401–404.
6. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:401–410.
7. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54:743–757.
8. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
9. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. *Molecular Systematics*, eds. Hillis DM, Moritz C, Mable BK (Sinauer, Sunderland, MA), pp 407–514.
10. Ting CT, Tsaur SC, Wu CI (2000) *Proc Natl Acad Sci USA* 97:5313–5316.
11. Klein J, Satta Y, O'hUigin C, Takahata N (1993) The molecular descent of the major histocompatibility complex. *Annu Rev Immunol* 11:269–295.
12. Kolaczkowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980–984.
13. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* 5:50.
14. Baele G, Raes J, Van de Peer Y, Vansteelandt S (2006) An improved statistical method for detecting heterotachy in nucleotide sequences. *Mol Biol Evol* 23:1397–1405.
15. Takahata N (2007) Molecular clock: An anti-neo-Darwinian legacy. *Genetics* 176:1–6.
16. Gruber KF, Voss RS, Jansa SA (2007) Base-compositional heterogeneity in the RAG1 locus among didelphid marsupials: Implications for phylogenetic inference and the evolution of GC content. *Syst Biol* 56:83–96.
17. Avise JC (2000) *Phylogeography: The History and Formation of Species* (Harvard Univ Press, Cambridge, MA).
18. Pratt RC, et al. (2009) Toward resolving deep Neoaves phylogeny: Data, signal enhancement, and priors. *Mol Biol Evol* 26:313–326.
19. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19.