Homoplasy: A New Term in the Lexicon of Phylogenetics

JOHN C. AVISE1 AND TERENCE J. ROBINSON2

1Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA; E-mail, javise@uci.edu
2Evolutionary Genomics Group, Department of Botany and Zoology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa

Homoplasy (trait similarity due to evolutionary convergence, parallelism, or character reversals) is a well-appreciated form of phylogenetic noise that systematists strive to identify and avoid when reconstructing species phylogenies. However, another source of phylogenetic “noise” is often neglected: the idiosyncratic sorting of gene-tree lineages into descendant taxa from character-state polymorphisms retained across successive nodes in a species tree. Here we introduce a term (hemiplasy) that formalizes a category of outcomes that can emerge from this evolutionary lineage-sorting phenomenon, and we make a case for why a wider recognition of hemiplasy (and attempts to ameliorate its complications) can play an important role in phylogenetics.

The word homoplasy, meaning shaped (-plasy) in the same (homo-) way, refers to any trait correspondence or similarity not due to common ancestry. A central challenge in phylogenetic reconstruction is thus to distinguish the phylogenetic noise of homoplasy from the phylogenetic signal of homology (similarity in biological features due directly to shared ancestry). However, homology itself bears a subtle relationship to phylogeny, as emphasized by Willi Hennig (1950) more than a half-century ago. Hennig introduced the critical distinction between shared ancestral homology (sympleisomorphic similarity) and shared derived homology (synapomorphic similarity), noting that only the latter is indicative of monophyly within an organismal phylogeny. Hennig’s cladistic insights fostered a fundamental revolution in phylogenetic principles and methodologies.

The molecular revolution in biology that began at about that same time added further nuances to the homology concept. For example, DNA sequence homology in a multigene family can be due either to paralogy (similarity tracing to a gene duplication event) or to orthology (similarity tracing to an allelic separation within a particular locus). Orthology and paralogy are both genuine forms of genetic homology, but a failure to distinguish them in comparisons of DNA sequences can lead to errors in phylogenetic reconstruction.

Phylogenetic jargon is already extensive but also important because words such as homoplasy, synapomorphy, and orthology capture and convey sophisticated evolutionary concepts that otherwise might remain opaque or underappreciated. In this spirit, here we formally define a new term—hemiplasy—for how the well-known phenomenon of idiosyncratic lineage sorting can lead to fundamental discordances between gene trees and organismal (species) trees. As will be described, hemiplasy is a bona fide form of homology (allelic orthology in this case) that nonetheless can give the illusion of homoplasy in an organismal tree. No other word or simple phrase currently exists to encapsulate the phenomenon that we will define under the suggested term.

CONCEPTUAL BACKGROUND

The nature of Mendelian heredity in sexually reproducing taxa ensures that alleles at unlinked loci transmit through an organismal pedigree via noncoincident genealogical pathways across multiple generations. Thus, both within and among related species, the true topologies of gene trees inevitably differ somewhat from locus to unlinked locus (Ball et al., 1990). Furthermore, gene genealogies can in principle differ in basic topology from the overall population tree or species tree of which they are a part, if for no other reason than stochastic lineage sorting across successive evolutionary nodes in an organismal phylogeny. These concepts and their corollaries have been available for more than two decades (Hudson, 1983; Tajima, 1983; Takahata and Nei, 1985; Neigel and Avise, 1986), and they are encapsulated...
today in the widely appreciated distinction between gene
trees and organismal trees (Doyle, 1992; Maddison, 1997;
Nichols, 2001). Nevertheless, as noted by Liu and Pearl
(2007), “the current molecular phylogenetic paradigm
still reconstructs gene trees to represent the species tree.”

Nei (1987:401–403) summarized the theoretical prob-
ability of qualitative discordance between a gene
tree and an organismal tree for the simple case of
three related species (for more complex situations, see
Rosenberg, 2002 and Degnan and Rosenberg, 2006).

For selectively neutral alleles, the probability of the
topological discordance illustrated in Fig. 1b is given by
\( \frac{2}{3}e^{-\frac{T}{2N}} \), where \( T \) is the number of generations be-
tween the first and second speciation events \( (T = t_1 -
 t_2) \) and \( N \) is the effective population size (Fig. 1). This
probability can also be interpreted as the percentage of
unlinked neutral loci expected to show topological dis-
agreement with the species tree. The formula supports
general intuition by showing that a gene tree is most
likely to “misrepresent” the topological structure of a
species phylogeny when internodal times are short rela-
tive to effective population sizes. For example, the discor-
dance probability is approximately 50% when \( T/2N =
0.3 \), but it is infinitesimally small when \( T/2N = 100 \).

This type of qualitative discordance between the
branching topology of a gene tree and a species tree can
also be interpreted to reflect the retention of a polymor-
phism across successive nodes in a species tree, followed
by lineage sorting and idiosyncratic fixation of alterna-
tive character states in the descendant species (Fig. 2).
Note that allele “b” in Fig. 2 is a derived character state
and that it is shared by two descendant taxa (B and
C) that nonetheless do not constitute a clade at the or-
ganismal level. In other words, character state “b” is
a clade-defining synapomorphy, but the monophyletic
assemblage that it earmarks is within the gene tree per-
se rather than at the composite species level.

The probability of topological discordance between a
gene tree and a species tree can also reflect additional fac-
tors that impact the ratio of \( T/2N \). For example, balanc-
ning selection can maintain a genetic polymorphism for
long periods of time, in effect making \( T/2N \) smaller and
thereby increasing the probability of an eventual discor-
dance between a species tree and the particular gene tree
whose alleles are under selection. Conversely, a genetic
polymorphism experiencing underdominant selection,
or one whose alleles undergo positive selective sweeps,
tends to be transient in a species and thereby is less likely
to eventuate in a gene-tree/species-tree incongruence.

Disparities between the topologies of gene trees and
species trees due simply to idiosyncratic lineage sorting
can also characterize taxa that separated anciently but
whose speciation events were close in evolutionary time
(Fig. 3). In such cases, the lineages from the polymorphic
ancestral gene pool that happen to have reached fixation
in distant descendants are those that produced the orig-
inal gene-tree/species-tree disharmony (Takahata, 1989;
Wu, 1991). For example, with respect to organismal phy-
logeny, taxa D and E in Fig. 3 are members of a clade to
the exclusion of F, whereas with respect to the gene tree
in Fig. 3, taxa E and F are members of a clade to the exclu-
sion of D. Thus, in principle the gene-tree/species-tree
“problem” is not confined to recently separated taxa.

**Example**

The overall phylogeny for human, chimpanzee, and
gorilla appears to be a near trichotomy with the most
Another depiction of how alleles at a gene (or alternative states of any polymorphic trait) can be misleading with respect to the tree topology for the species in which the alleles are housed. Shown is a polymorphism that traversed successive speciation nodes only to sort idiosyncratically and later become fixed in the descendant species in a pattern that at face value would appear to be discordant with the species phylogeny.

Diagrammatic representation of how an ancient discordance between a gene tree and a species tree can be perpetuated indefinitely and thereby retained as a permanent incongruity between the gene tree and the species tree of descendant taxa (after Avise, 2000).

Likely resolution being sister-taxa status for *Homo* and *Pan* to the exclusion of *Gorilla* (e.g., Stanyon et al., 2006). Not all homoplasy-free gene trees or sets of character states are expected to match this composite species topology, however, if some polymorphisms happen to have traversed the adjacent evolutionary nodes before sorting idiosyncratically into various pairs of the descendant taxa (Takahata et al., 1995). For example, Chen and Li (2001) reported that whereas DNA sequences from each of 31 independent loci support the *Homo-Pan* clade, 12 appear to support a *Pan-Gorilla* clade and 10 appear to support a *Homo-Gorilla* clade; and in a more extensive recent analysis, Ebersberger et al. (2007) reported that about 23% of 23,210 DNA sequence alignments in the great apes implied at face value that chimpanzees are not the closest genetic relatives of humans (see Patterson et al., 2006, for comparable findings based on 20 million base pairs of aligned human and chimpanzee sequence). In the human-chimpanzee case, the causes of these discrepancies are not fully understood (and may include postspeciation introgression; Patterson et al., 2006). However, at least in principle, each gene tree could be correct in the sense of providing valid genealogical signal (i.e., without homoplasy) for the specific portion of the genome that it represents.
DEFINITION

A formal term seems desirable to encapsulate the essence of the phylogenetic processes described above that can lead to genuine discordances between particular gene trees (components of the genome) and a composite or overall species phylogeny. We suggest the word hemiplasy, because the responsible lineage sorting processes have homoplasy-like consequences despite the fact that the character states themselves are genuinely homologous and apomorphic. So, hemiplasy is somewhat like homoplasy in terms of its face-value phylogenetic consequences, yet its evolutionary etiology is fully distinct from homoplasy. We suggest the following formal definition of hemiplasy: the topological discordance between a gene tree and a species tree attributable to lineage sorting of genetic polymorphisms that were retained across successive nodes in a species tree. A set of hemiplasious alleles, genes, or other character states would thus be those that contribute to hemiplasy in a phylogenetic data set.

Other evolutionary processes are also capable of producing homoplasy-free discordances between gene trees and species phylogenies. For example, particular alleles can leak across species boundaries via hybridization and introgression, and pieces of DNA sometimes move between species via true horizontal transfer (viral-mediated, for example). In such cases, the transferred DNA is a bona fide part of the genetic history of the species in question, but the gene tree would differ dramatically from the majority phylogeny for the remainder of the genome. We acknowledge that such outcomes may be difficult to distinguish from genuine hemiplasy in some particular empirical instances. Nevertheless, for epistemological clarity we recommend that the term hemiplasy not include these additional (and well appreciated) generators of phylogenetic discordance between gene trees and species trees but instead be confined to discordances that arise from idiosyncratic lineage sorting per se.

The importance of the hemiplasy concept is further evidenced by the fact that several recently introduced phylogenetic approaches in effect acknowledge and attempt to accommodate the phenomenon (Carstens and Knowles, 2007; Edwards et al., 2007; Liu and Pearl, 2007; Maddison and Knowles, 2006). Additional cutting-edge approaches underdominant selection (such as some types of chromosomal markers). Markers that should be relatively immune to hemiplasy might include, for example, those with smaller effective population sizes (such as cytoplasmic loci compared to autosomal nuclear genes) or those that are likely to experience underdominant selection (such as some types of chromosomal markers).

In conclusion, adoption of the word hemiplasy should contribute to the injection of oft-neglected "population thinking" into phylogenetic assessments and thereby provide another incremental step toward unifying the traditionally disparate fields of phylogenetic biology and population genetics.

REFERENCES

Species Names in the PhyloCode: The Approach Adopted by the International Society for Phylogenetic Nomenclature

BENOÎT DAYRAT,1 PHILIP D. CANTINO,2 JULIA A. CLARKE,3,4 AND KEVIN DE QUEIROZ5

1School of Natural Sciences, University of California, P.O. Box 2039, Merced, CA 95344, USA; E-mail: bdayrat@ucmerced.edu
2Department of Environmental and Plant Biology, Ohio University, Athens, OH 45701, USA; E-mail: cantino@ohiou.edu
3Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, Campus Box 8208, Raleigh, NC 27695-8208, USA; E-mail: julia.clarke@ncsu.edu
4Department of Paleontology, North Carolina Museum of Natural Sciences, 11 West Jones Street, Raleigh, NC 27601-1029, USA
5Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA; E-mail: dequeiroz@si.edu

On May 24, 2007, the Committee on Phylogenetic Nomenclature (CPN), which consists of 12 elected members from the International Society for Phylogenetic Nomenclature (ISPN), adopted a new article in the International Code of Phylogenetic Nomenclature (ICPN or PhyloCode; Cantino and de Queiroz, 2007) addressing the naming of species in the context of phylogenetic nomenclature. This vote, which took place after more than 10 years of discussion on how to handle species names in phylogenetic nomenclature, represents a major step in the development of the PhyloCode. Until now, the successive drafts of the PhyloCode have only dealt with clade names, although the application of phylogenetic nomenclature to species has been heavily debated at workshops and symposia on phylogenetic nomenclature (e.g., Cantino et al., 1999a), in the literature (e.g., de Queiroz and Gauthier, 1992; Graybeal, 1995; Schander and Thollesson, 1995; Cantino, 1998; Cantino et al., 1999a, 1999b; Freshefsky, 1999; Pleijel, 1999; Pleijel and Rouse, 2000a, 2000b, 2003; Artois, 2001; Hillis et al., 2001; Lee, 2002; Spangler, 2003; Dayrat et al., 2004; Dayrat, 2005; Dayrat and Gosliner, 2005; Fisher, 2006; Wolsan, 2007a, 2007b), and at the two meetings of the ISPN (Laurin and Cantino, 2004, 2006; see the Preface to the PhyloCode for additional information). The article on species names (Article 21) that the CPN recently adopted was prepared by the four of us. Here, we wish to explain its rationale and advantages.

POINTS OF INCOMPATIBILITY OF LINNAEAN BINOMINAL SPECIES NAMES WITH PHYLOGENETIC NOMENCLATURE

The Linnaean binomial nomenclature used for species names in the rank-based codes (Bacterial Code [BC], International Code of Botanical Nomenclature [ICBN], International Code of Zoological Nomenclature [ICZN]) is not fully compatible with phylogenetic nomenclature because their requirements regarding ranks differ. Under the rank-based codes, the name of a species is a combination of two words, i.e., a binomen (or binomial), the first part of which is a generic name and the second