Syst. Biol. 48(4):814–825, 1999

How Should Species Phylogenies Be Inferred from Sequence Data?

JOSEPH B. SLOWINSKI[1,3] AND RODERIC D. M. PAGE[2]

[1]*Department of Herpetology, California Academy of Sciences, Golden Gate Park, San Francisco, California 94118, USA; E-mail: jslowins@cas.calacademy.org*
[2]*Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK*

There are two levels of potential error in the inference of species phylogenies from molecular sequence data: (1) A gene tree (herein, we use the term gene for any contiguous block of nucleotides, regardless of whether it codes for a protein or not) for a set of molecular sequences will be incorrectly inferred if there is sufficient random or systematic error (Swofford et al., 1996), and (2) even if a gene tree is correctly inferred, the phenomena of deep coalescence, gene duplication, and horizontal gene transfer can produce a gene tree different from the true species tree (Goodman et al., 1979; Avise et al., 1983; Pamilo and Nei, 1988; Doyle, 1992; Maddison, 1996, 1997).

The second level of error would be quite worrisome if all nucleotides of genomes were historically linked (as is generally true with organellar genomes). In this situation, there would only be one gene tree, which might not be the same as the true species tree. But happily, because of intra- and interchromosomal recombination, the nuclear genome is composed of many historically linked sets of nucleotides with different histories. We call these *linkage partitions* (the c-genes of Doyle, 1992, 1997). Sequences sampled from several species and forming a single linkage partition are related by a unique, binary gene tree. Contrary to claims that natural data partitions do not exist (e.g., Kluge and Wolf, 1993; Siddall, 1997), linkage partitions are natural partitions of molecular sequence data and can be considered as independent estimators of the overlying species phylogeny.

This strongly suggests that the molecular phylogenetic analysis of species by using parsimony should be performed on two levels as follows: (1) Separate gene trees are inferred from each linkage partition, and (2) the species phylogeny is then inferred from the set of gene trees. A method (Maddison, 1997; Page and Charleston, 1997a, 1997b; Slowinski et al., 1997) termed *gene tree parsimony* by Slowinski et al. (1997) is the appropriate method for implementing the second step. Gene tree parsimony operates by finding the species tree or trees that minimizes the number of hypothesized gene tree/species tree conflict-producing events required to fit each gene tree to the species tree(s). Central to gene tree parsimony is the concept of fitting trees to other trees (Page, 1994a). Gene tree parsimony implements Doyle's (1992) insightful concept that nucleotides are characters of gene trees, whereas gene trees are characters of species trees. An important caveat relates to the serious question raised by Maddison (1997) of just what a species phylogeny is meant to represent. The search for a species phylogeny assumes that such a diagram has some meaning. This is a difficult issue that we leave to other workers.

Below, we briefly discuss the sources of conflict between gene and species trees, identify problems with previous methods for inferring species trees from molecular sequence data, define gene tree parsimony, and then illustrate the application of gene tree parsimony by using the computer program GeneTree (Page, 1998), which is free and available at http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html; it requires Mac OS 7.5 or later running on a PowerMac, or an Intel-based PC running Windows 95/NT 4.0 or later. The issues explored in this article bear directly on the controversial question of whether or not

[3]Address correspondence to this author.

data should be partitioned or combined for phylogenetic analyses. Some phylogeneticists (e.g., Kluge, 1989; Kluge and Wolf, 1993; Nixon and Carpenter, 1996) have argued that all available data should always be combined (the "simultaneous analysis" approach), whereas other phylogeneticists (e.g., Miyamoto and Fitch, 1995) have argued that data from different sources should never be combined (the "separate analysis" approach). Intermediate between these two extremes is the "prior agreement" method espoused by some (e.g., Bull et al., 1993), which argues that data sets should be combined only if not significantly heterogeneous. In the sense that we espouse combining all available independent gene trees when inferring the species phylogeny, gene tree parsimony can actually be considered simultaneous analysis. We have simply developed a more refined concept of the appropriate units for combination.

Sources of Conflict Between Gene and Species Trees

The phenomena of deep gene coalescence, gene duplication, and horizontal gene transfer (Fig. 1) can produce a gene tree different from the true species tree (Goodman et al., 1979; Avise et al., 1983; Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991; Doyle, 1992; Maddison, 1996, 1997). Gene duplication (Fig. 1a) produces conflict between gene and species trees when paralogous sequences are sampled. A deep coalescence (Fig. 1b) can produce conflict between a gene tree and the overlying species tree because there is a window of opportunity for a sequence from a less related species to coalesce with one of the descendant sequences of the deep coalescence. Deep coalescence produces conflict that is analogous to that produced by gene duplication because paralogous sequences are sequences that coalesced prior to the ancestor of the species from which they were sampled. Of course, the evolutionary dynamics of maintaining duplicated genes are different from those of maintaining multiple alleles at a locus, which are in direct competition. Nevertheless, without additional information (beyond the observed conflict between a gene and species tree), it

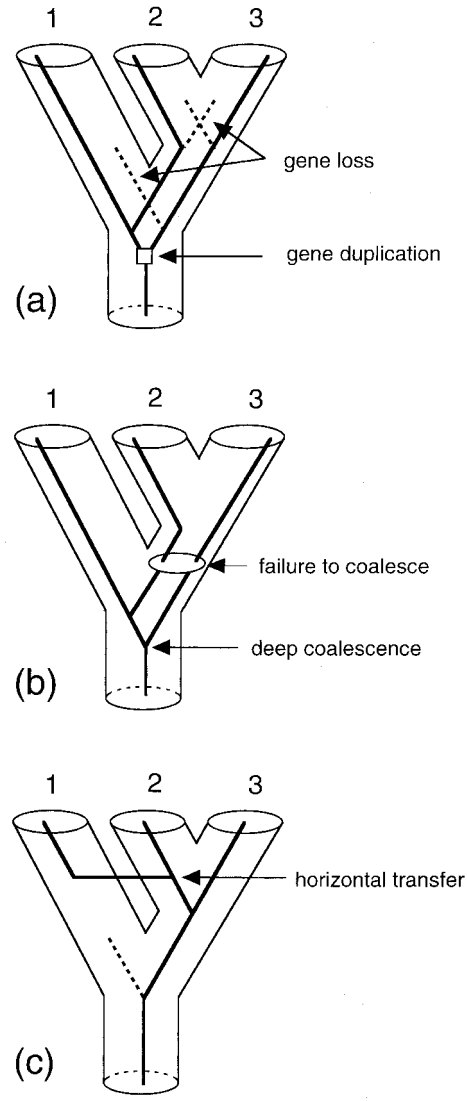

Figure 1. Examples of gene trees embedded in species trees, showing the sources of gene tree/species tree conflict. In each case, the gene tree groups species 1 and 2, which are not each other's closest relatives. (a) A basal gene duplication followed by subsequent gene loss results in a single copy of the gene in each of the three species. However, the genes in species 1 and 2 are paralogous with respect to the gene in species 3. (b) The allele in species 2 does not coalesce with the allele in species 3, but instead with the allele in species 1. This failure of the alleles in species 2 and 3 to coalesce in the immediate ancestor of species 2 and 3 results in a deep coalescence at the base of the species tree. (c) An example of lateral transfer, wherein a gene lineage with its ancestry in species 2 is transferred to species 1, leading to conflicting gene and species trees. This can occur in several ways, including hybridization between species and viral transfer of genes between hosts.

is impossible to distinguish between gene duplications and deep coalescences. Deep coalescences and gene duplications, however, are only part of the problem; the rest of the problem involves our failure to sample sequences that have or might have descended from the ancestral sequences. For example, the conflict between gene and species trees caused by sampling paralogous sequences from the two loci of a duplicated gene would disappear if sequences from both loci were sampled in all the species (Doyle, 1992). Horizontal gene transfer (Fig. 1c), which includes phenomena such as the transfer of genes between bacteria and hybridization between different species of bisexual eukaryotes, obscures species phylogeny because sequences from one species may introgress into another species. Many examples of lateral transfer are now known from viruses (e.g., Gibbs and Keese, 1995), prokaryotes (e.g., Lawrence and Ochman, 1998), and eukaryotes (e.g., Assali et al., 1990).

Some authors (e.g., Pamilo and Nei, 1988; Wu, 1991) have explored the conditions under which a gene tree can be expected to show congruence with the overlying or "containing" (Maddison, 1997) species phylogeny. Pamilo and Nei (1988) found that when internodes are narrow (in terms of the effective population size) and long (in terms of time), the probability of a gene tree being the same as the species tree is maximized. The worst-case scenario is obtained when every internode of a species phylogeny approaches infinite width and zero length. Under this situation, gene trees are distributed according to the Markovian distribution (Slowinski and Guyer, 1989).

Unfortunately, when a species phylogeny has short internodes, this will also adversely affect our our ability to resolve the underlying gene trees. Thus, the form of a species tree will affect phylogenetic analysis at both levels: the probability of correctly inferring gene trees, and the probability that gene trees will be congruent with the overlying species phylogeny.

Problems with Prior Approaches

As applied to sequence data, the simultaneous analysis or combined-data approach (Kluge, 1989; Kluge and Wolf, 1993; Nixon and Carpenter, 1996) concatenates all available nucleotides (or amino acids) for a set of taxa into a single matrix for analysis. In doing so, simultaneous analysis collapses what should be two levels of analysis into one. This results in three problems: First, and probably most important, the simultaneous analysis approach erroneously treats every nucleotide of all available genes as independent estimators of the overlying species phylogeny. To illustrate why this is a problem, consider constructing a combined matrix from three genes sampled from three species when the first gene has three times the number of nucleotides as the other two. Assume that because of recombination, the three genes have experienced independent phylogenetic histories. Further assume that the history of the first gene is (AC)B and is incongruent with the species phylogeny {(AB)C}, whereas the histories of the other two genes are congruent with the species phylogeny. In this situation, we are giving the first gene's history triple the weight of the other two histories solely because of the different numbers of nucleotides (Doyle, 1992). This is a serious problem because the first gene tree is incongruent with the species phylogeny, which is likely to result in the inference of the wrong species tree under the simultaneous analysis approach. On the other hand, if the three gene histories are treated as independent estimators of the species phylogeny, the true species phylogeny prevails. Under the assumption that each gene tree has been correctly inferred, there is simply no a priori justification for weighting one gene's history more than another's when inferring a species phylogeny. If the nucleotides of a given gene share the same history, then a gene phylogeny represents only a single character of the species phylogeny (Doyle, 1992), regardless of the number of nucleotides that gene comprises. Of course, it is generally unknowable whether a gene tree has been correctly inferred. Given this, it may be useful to weight gene trees by some function of their confidence. Confidence can vary according to different factors but may in some situations be positively correlated with the number of nucleotides, because longer genes generally suffer from less random error.

This simple example illustrates that character independence exists at least at two levels: the level of the gene tree, and the level of the species tree. Nucleotides of a single gene may form independent estimators of the gene tree, but they are not independent estimators of the species tree because collectively they track the history of a single gene. It is essential to identify sets of historically linked nucleotides and then to treat these as the fundamental units for phylogenetic analysis (Doyle, 1992; Page, 1996; Slowinski et al., 1997).

The second problem with the simultaneous analysis approach is that the distinction between homoplasy and gene tree/species tree conflict is ignored (Page and Charleston, 1997a; Slowinski et al., 1997). For example, if a true gene tree is (AB)C and the true species tree is (AC)B, then any substitutions occurring along the branch of descent leading to (AB) on the gene tree will be interpreted as homoplasy in the context of the species tree, even though they are not homoplasies at all.

The third problem with the simultaneous analysis approach concerns polymorphism at multiple loci. In this situation, it is not obvious how the sequences from the different loci can be incorporated into a combined matrix, unless the loci are physically linked on each sampled sequence. Any logical approach to the inference of species phylogenies from sequence data must be able to accommodate sequence polymorphism. Figure 2 shows the number of sequences plotted against number of species for vertebrate gene families in release 29 (17 March 1998) of the HOVERGEN (Duret et al., 1994) database. Note that usually each species has a single mitochondrial sequence for a given gene (hence the mitochondrial genes fall along the 1:1 line), whereas most nuclear genes are now known from multiple copies. Thus, multiple sequences are accumulating rapidly in genetic databases, making it therefore imperative that phylogenetic methods be able to accommodate these.

Note that we are not arguing that the combined approach is never appropriate, only that it is not applicable when genes have experienced different histories. To the extent that a set of contiguous nucleotides have shared the same history, it is desirable to combine the data because of the property of statistical consistency (Huelsenbeck et al., 1996), which ensures that the estimated tree converges on the true tree as ever-increasing numbers of characters are sampled. However, phylogenetic consistency requires certain assumptions of the evolutionary process that generated the data. Several tests have been proposed for the null hypothesis that two data sets represent character samples generated along the same phylogeny (reviewed by Huelsenbeck et al., 1996).

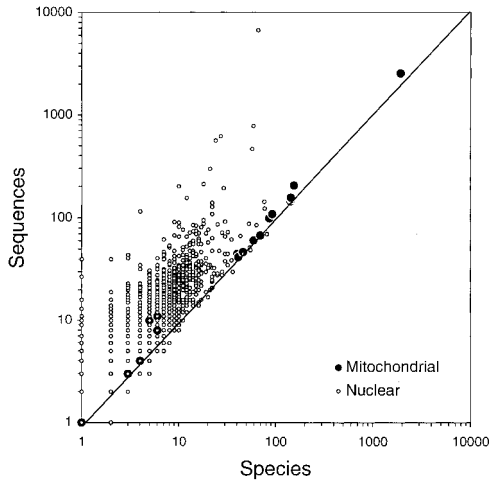

FIGURE 2. Number of sequences plotted against number of species for vertebrate gene families in release 29 (17 March 1998) of the HOVERGEN (Duret et al., 1994: no. 1182) database. Note that usually each species has a single mitochondrial sequence for a given gene (hence the mitochondrial genes fall along the 1:1 line), whereas most nuclear genes are present in multiple copies. Because of redundancy in species names (for example, use of "human" and "*Homo sapiens*" to describe the source of different genes in the same family), some gene families appear to have fewer sequences than species. This figure demonstrates the rapid accumulation of multiple sequences for species, underscoring the need for phylogenetic methods to be able to accommodate multiple sequences.

Because the nucleotides (or amino acids) from genes with different histories cannot be combined for phylogenetic analysis, several authors have proposed alternative methods for inferring species trees from gene trees, based on treating each gene tree as a character. Baum (1992), Doyle (1992), and Ragan (1992) suggested recoding each gene phylogeny into parsimony characters,

an approach that Ragan termed matrix representation of trees. The resulting series of characters can then be analyzed to find the minimum-length species phylogeny. This approach, however, is flawed because homoplasy in this context has no obvious biological meaning (Rodrigo, 1993; Page, 1994a). When an extra step occurs for a character (= gene tree), it is not clear just what that extra step means. Biological realism is imperative for any phylogenetic method.

de Queiroz (1993) advocated the use of consensus trees for inferring species trees in the face of gene tree/species tree conflict. This use of consensus analysis is inappropriate for two reasons (Mirkin et al., 1995): First, consensus methods cannot easily accommodate differing sets of terminal entities (Page, 1996; but see Sanderson et al., 1998), such as arise from polymorphism or simply from gene trees with sequences sampled from different species; second, many consensus methods are too conservative.

Linkage Partitions

Bull et al. (1993) argued that molecular data often are divided among partitions defined by different evolutionary processes or histories, and that data from different partitions should not be combined for phylogenetic analysis. Kluge and Wolf (1993), however, deny the existence of natural partitions in molecular data. We disagree with Kluge and Wolf. Clearly natural partitions of sequence data exist, and at several levels. Perhaps the most fundamental for inference of species phylogeny are partitions of nucleotides defined by their shared hierarchical descent, which we term *linkage partitions* (the c-genes of Doyle, 1992, 1995, 1997). A linkage partition is a set of nucleotide (or amino acid) sequences meeting the following criteria: Each member of the partition is a sequence of contiguous nucleotides (or the translated protein product); and the sequences are descended from ancestral sequences in a strictly hierarchical fashion (i.e., the relationships of the sequences can be represented by a tree).

For example, consider sampling the entire mitochondrial genome from a series of species. Because mitochondrial genomes are (usually) free from recombination, the nucleotides form a series of historically linked characters, defining a single linkage partition. Now consider a nuclear gene sampled from the same series of species that has not itself experienced recombination within the history spanned by the sampled species. Regardless of whether these two sets of sequences have tracked the same or different histories, they define two linkage partitions and should be considered independent estimators of the species phylogeny. The boundaries between linkage partitions are created by recombination, whether intrachromosomal, interchromosomal, or nonreciprocal.

We suggest that a major goal of molecular phylogenetics should be the identification and demarcation of linkage partitions. How can this be done? Consider sampling a set of sequences representing two different loci for a series of species, where "loci" is used very loosely to refer to any two disjoint segments of DNA and each locus by itself is assumed to define a linkage partition. There are several possible situations: (1) The loci fall on different chromosomes. In this situation, the loci have experienced independent histories and correspond to separate linkage partitions, assuming that there has not been concerted evolution between the loci. (2) The loci fall on the same chromosome and the loci are physically linked on each sampled sequence. In this situation, we should be able to test whether the loci correspond to the same or separate linkage partitions by using any of several methods devised to detect heterogeneity between genes. These methods can be divided between those methods developed for the specific goal of detecting recombination (e.g., Stephens, 1985; Sawyer, 1989; Dykhuizen and Green, 1991; Weiller, 1998) and those developed for the more general goal of detecting different evolutionary processes in phylogenetic data (reviewed by Huelsenbeck et al., 1996). In the present context, the null hypothesis is simply that the present-day covalent linkages between the loci on the sampled sequences have been historically maintained through the span of time covered by the sampled sequences. If recombination in the past has broken the linkages and concatenated DNA

segments with different histories, then hopefully application of the heterogeneity tests will lead to rejection of the null hypothesis, impelling one to treat the loci as separate linkage partitions. (3) The two loci fall on the same chromosome, but because of the sampling strategy, some or all of the sampled sequences represent partial DNA fragments on which only one of the loci is represented. Unfortunately, in this situation, testing for linkage partitions using heterogeneity tests is not straightforward because the null hypothesis is not defined (i.e., we do not know which, if any, of the partial sequences were physically linked in vivo). In this difficult situation, it seems best to simply consider the loci as forming two linkage partitions.

Gene Tree Parsimony

Maddison (1997), Page and Charleston (1997a, 1997b), and Slowinski et al. (1997) described a procedure—termed *gene tree parsimony* by Slowinski et al. (1997)—for finding the species tree that minimizes a weighted sum of deep coalescences, gene duplications and losses (here we use the term loss to refer to either a real loss or simply a failure to sample a sequence), and horizontal gene transfers necessary to fit each gene tree to the species tree. Gene tree parsimony is based on the very important concept of fitting a tree to a containing tree (Fig. 3). Examples of this include fitting species trees to area trees, fitting gene trees to species trees, and fitting parasite trees to host trees (Page, 1993a). Fitting trees to other trees was originally investigated by Goodman et al. (1979) in the context of fitting a gene tree to a species tree in such a way as to minimize the number of gene duplications and losses necessary to reconcile the two trees. This procedure (Fig. 3) was later formalized by Page (1994a) as tree reconciliation, which is implemented in the program COMPONENT 2.0 (Page, 1993b). Fitting trees to containing trees is analogous to optimizing nucleotide characters to gene trees and could also be considered an optimization procedure, albeit at a higher level.

Page (1994a) suggested that tree reconciliation could be used as an optimality criterion for phylogenetic analysis by selecting the species tree that allows the number of duplications and losses to be minimized, an idea further developed by Mirkin et al. (1995) and Guigo et al. (1996). Gene tree parsimony as described in this article generalizes this method by including the additional optimality criteria of minimizing deep coalescence and lateral transfer. Optimization procedures for deep coalescence are discussed in Maddison (1997) and Slowinski et al. (1997) and illustrated by Figure 3. Optimization procedures for lateral transfer are much more difficult to implement (Page, 1994b; Page and Charleston, 1997a, 1997b; Charleston, 1998) and the program GeneTree (Page, 1998) does not yet incorporate lateral transfer as an optimality criterion. Perhaps for many taxa, lateral transfer occurs so rarely as to be safely ignored; however, the phenomenon is known to occur widely (e.g., Assali et al., 1990; Gibbs and Keese, 1995; Lawrence and Ochman, 1998).

Gene tree parsimony makes several assumptions about gene trees: (1) that each is based on sequences that have not experienced recombination, i.e. based on a single linkage partition; (2) that each has been correctly inferred; and (3) that they are independent. Assumption 1 requires that linkage partitions be correctly identified, which is almost certainly difficult, but methods for doing so now exist (e.g., Stephens, 1985; Sawyer, 1989; Dykhuizen and Green, 1991; Huelsenbeck et al., 1996; Weiller, 1998). With regard to assumption 2, the conditions under which a gene tree will be correctly inferred have been extensively explored in the systematics literature going back to Felsenstein (1978) and will not be discussed here except to point out the following potential problem: If recombination has been extensive, it is possible that linkage partitions will be so small as to make accurate inference of gene trees impossible because of random error. Assumption 3 is problematic because linkage partitions will often not be truly independent according to the strict definition of that word. This will especially be true for gene regions close together on a chromosome.

Gene tree parsimony has an undeniable logic: If a gene tree differs from the true species tree, the discrepancy is due to some

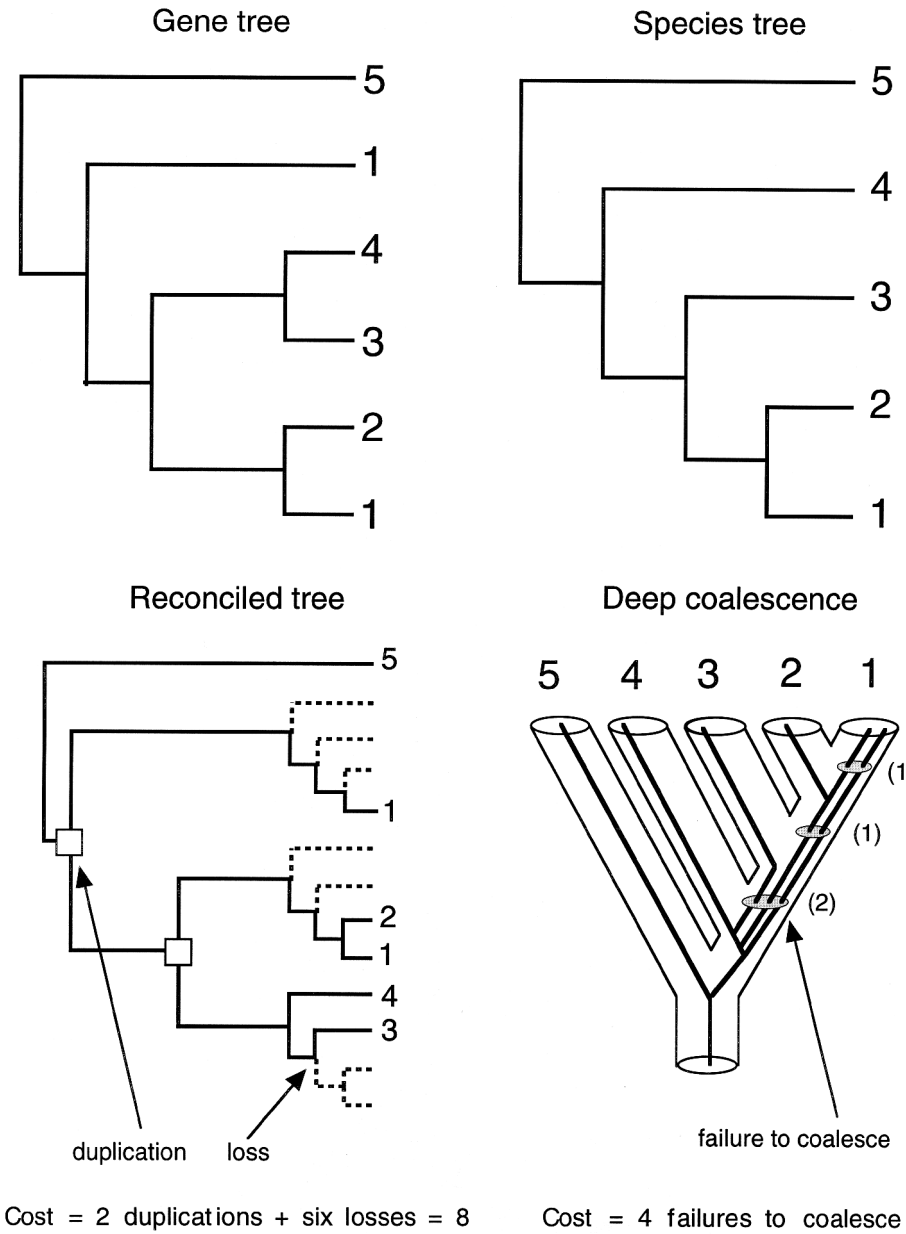

FIGURE 3. Optimization methods for implementing gene tree parsimony. Given the gene and species trees shown, embedding the gene tree in the species tree (under the assumption that the observed gene tree/species tree conflict is due to gene duplication coupled with unsampled or extinct sequences) results in a reconciled tree that requires two duplications and six losses for a total of eight events. Under the assumption that gene tree/species tree conflict is due to deep coalescence, one can use the same procedure to fit the gene tree into the species tree; however, instead of counting duplications and losses, one counts the number of gene lineages that do not coalesce on each branch of the species tree. The two alleles in species 1 fail to coalesce until the most recent common ancestor (MRCA) of species 1–4, which counts as three failures to coalesce (once in species 1, once in the MRCA of species 1 and 2, and lastly in the MRCA of species 1–3). The allele in species 3 and the MRCA of alleles 1 and 2 fail to coalesce in the ancestor of these three species; instead they meet in the MRCA of species 1–4, bringing the total cost to four events.

combination of deep coalescences, gene duplications, and lateral transfers. Therefore, it is biologically realistic to consider these as the fundamental events to minimize under a parsimony criterion for inferring species phylogeny. A practical feature of gene tree parsimony is that the individual gene trees need not be based on the same sets of species.

Parsimony is not the only possible method for finding a species phylogeny based on reconstructing deep coalescent, gene duplication, and lateral transfer events. If one is willing to assign probabilities to these events, then a maximum likelihood approach is also possible (Maddison, 1997). Many papers have modeled the probabilities of deep coalescence under simple stochastic models (e.g., Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991).

Under the restrictive assumptions that all gene trees have been correctly inferred and deep coalescence is the only source of discord between gene and species trees, we believe that gene tree parsimony will possess the desirable quality of being consistent (sensu Felsenstein, 1978). Pamilo and Nei (1988) explored the conditions under which a gene tree can be expected to show congruence with the overlying species phylogeny when deep coalescence is the only source of discord between gene and species trees. They found that the narrower (in terms of effective population size) and longer (in terms of time) the internodes are, the greater the probability that a gene tree will match the species tree. In the worst-case scenario, every internode of the species phylogeny approaches infinite width and zero length, in which case each gene tree has a probability according to the Markovian model (Slowinski and Guyer, 1989). The fact that most internodes can be expected to have a finite width and length greater than zero means that, under the assumptions above, the probability of a gene tree being identical to the species tree will exceed the probabilities of the other possible gene trees. Thus, given enough gene trees, gene tree parsimony should in theory reconstruct the correct species tree. But more work is needed to test this supposition.

In the sense that we espouse combining all available gene trees (based on different linkage partitions) when inferring the species phylogeny, gene tree parsimony can actually be considered simultaneous analysis. But instead of advocating the simultaneous analysis of nucleotides from different genes, we advocate the simultaneous analysis of all gene trees based on different linkage partitions. We have simply developed a more refined concept of the appropriate units for combination.

An Example from the Snake Family Elapidae

We illustrate gene tree parsimony with an empirical example (Slowinski et al., 1997) from the snake family Elapidae. Slowinski et al. inferred separate gene trees (Fig. 4) from amino acid sequences for the venom proteins phospholipase A_2 (PLA_2; Fig. 4a) and the short-chain neurotoxin (NXS; Fig. 4b) from various elapids. The PLA_2 tree was rooted with outgroup sequences from *Pseudonaja textilis* and *Oxyuranus scutellatus;* the NXS tree was rooted with outgroup sequences from *Dendroaspis.* This example is especially useful because of the high degree of sequence polymorphism. Many of the polymorphic sequences sampled do not mutually cluster by species. Any method that purports to infer species phylogenies from sequence data must be able to accommodate polymorphism.

To implement gene tree parsimony, Slowinski et al. (1997) used GeneTree (Page, 1998), which will locate species trees under the optimality criteria of minimizing the number of deep coalescences or gene duplications plus losses (as mentioned above, GeneTree does not incorporate horizontal transfer). At present, GeneTree minimizes either the number of deep coalescences or the number of gene duplications plus losses during a run, but not both simultaneously. Mixed analyses are probably more realistic, but this involves the complex issue of how deep coalescence and duplication events should be weighted relative to each other. When no sequences for a particular gene have been sampled from a species, GeneTree treats this as missing data; that is, the missing sequences are not counted as losses under the duplication criterion. This has the desirable effect of eliminating spurious

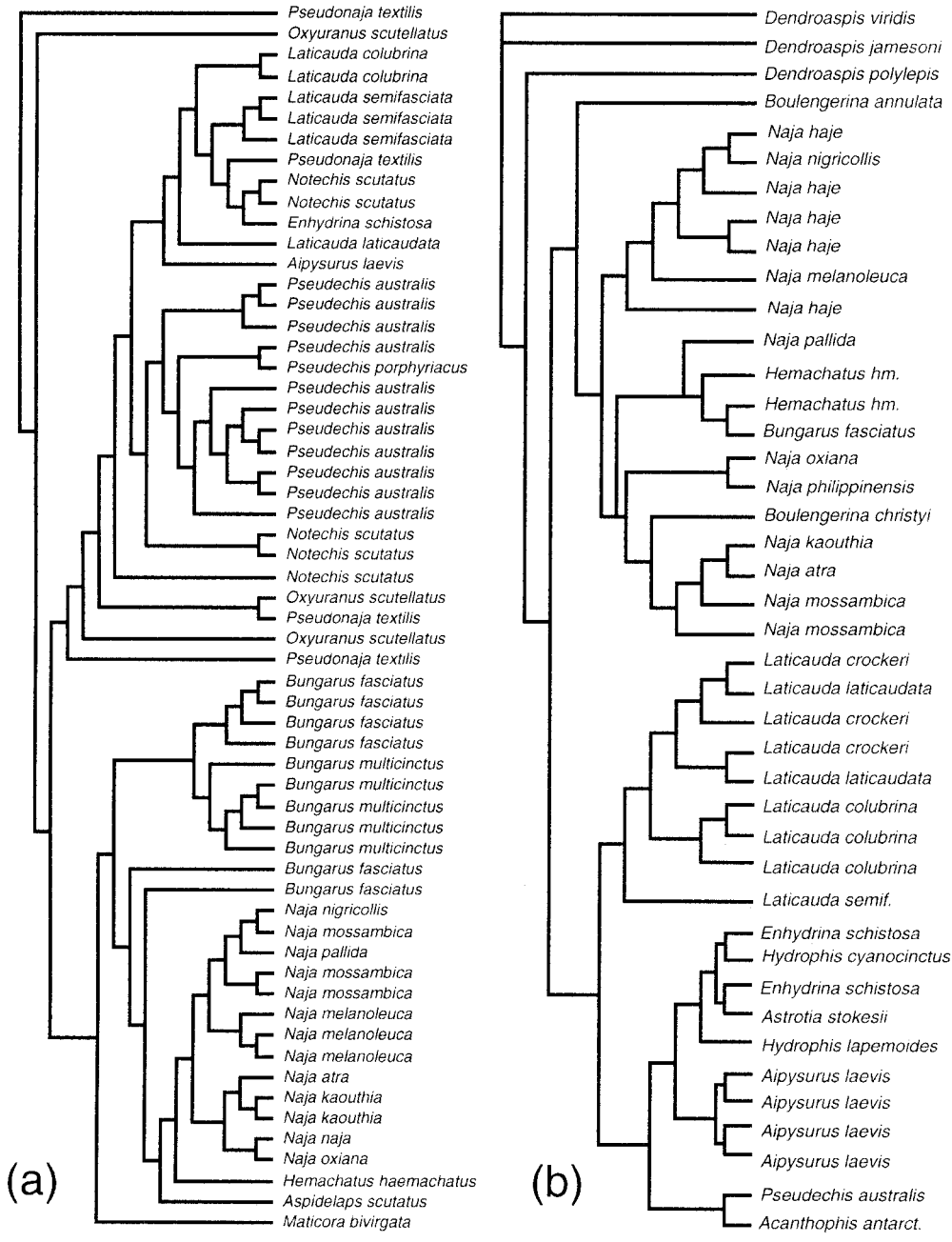

FIGURE 4. Shortest trees resulting from parsimony analyses of 59 elapid PLA_2 (a) and 42 NXS (b) amino acid sequences (see Slowinski et al. {1997} for details). These gene trees are the data for a gene tree parsimony analysis to find an optimal species phylogeny. Abbreviations: *Laticauda semif.* = *L. semifasciata; Hemachatus hm.* = *H. haemachatus; Acanthophis antarct.* = *A. antarcticus.*

clades formed by the shared absence of sequences for a gene. Prior to the GeneTree analyses, the outgroup sequences were pruned from the gene trees.

Under the criterion of minimizing duplications plus losses, GeneTree found >99 shortest species trees (the version of GeneTree used by Slowinski et al. {1997} stored only up to 99 trees; currently, GeneTree can store up to 999 trees) (Fig. 5) from the PLA_2 and NXS trees with a cost of 122 duplications plus losses (deep coalescence cost ranged from 58 to 62 failed coalescences). Under the criterion of minimizing deep coalescences, GeneTree found >99 shortest trees (Fig. 6) with a cost of 54 (duplication cost ranged from 131 to 135 failed coalescences).

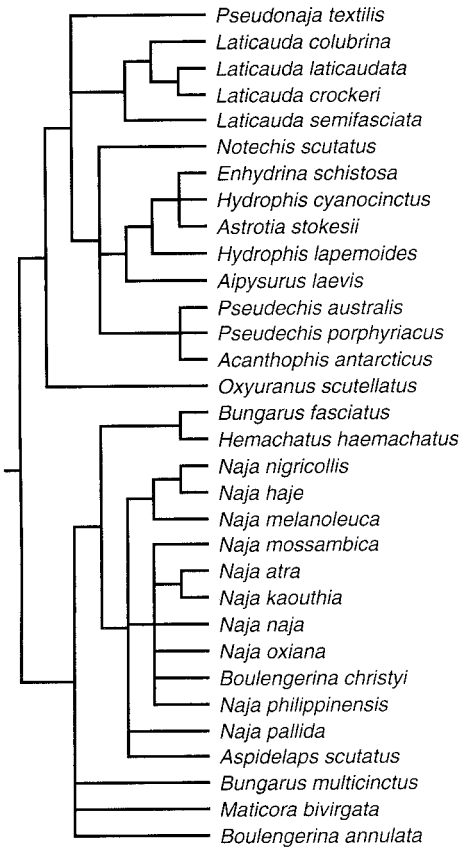

FIGURE 5. The strict consensus tree of the 99 shortest species trees resulting from analysis of the PLA_2 and NXS gene trees by using gene tree parsimony implemented with GeneTree, minimizing the number of duplications plus losses (cost = 122 duplications plus losses).

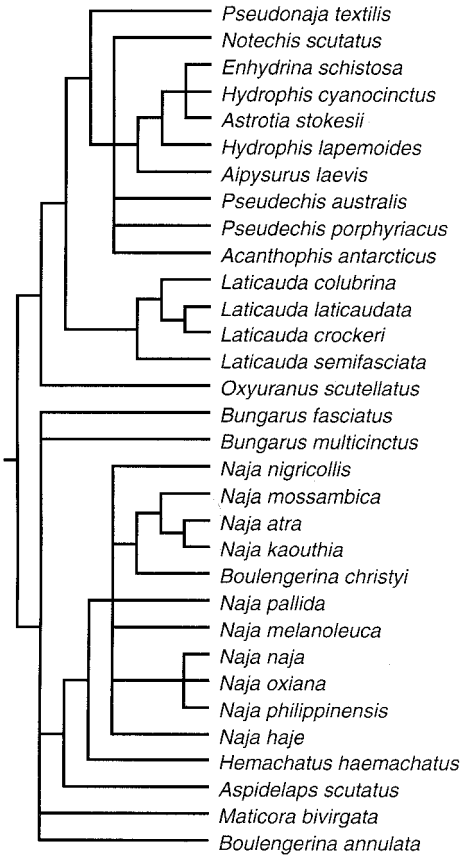

FIGURE 6. The strict consensus tree of the 99 shortest species trees resulting from analysis of the PLA_2 and NXS gene trees by using gene tree parsimony implemented with GeneTree, minimizing deep coalescences (cost = 54 failed coalescences).

The results (Figs. 5, 6) illustrate the utility of gene tree parsimony along several lines:

1. Gene tree parsimony is able to reduce a series of gene trees with extensive polymorphism to an easily interpretable species phylogeny. For example, from the elapid gene trees (Fig. 4a and 4b), it is very difficult to interpret the relationships among the species of the cobra genus *Naja*, both because of the polymorphism as well as the interpolation within that group of species in other genera. But gene tree parsimony reduces this complex picture to the straightforward hypothesis of relationships for *Naja* shown on the species phylogenies

(Figs. 5, 6), wherein all *Naja* are brought together with *Boulengerina*, a result corroborated by mtDNA data (Slowinski and Keogh, unpubl.). This ability to deal effectively with polymorphism is a major asset of gene tree parsimony, not just because future molecular work will continue to accumulate polymorphic variants of genes but also because polymorphic gene trees are likely to provide more accurate estimators of species phylogeny than are nonpolymorphic gene trees (Takahata, 1989).

2. Gene tree parsimony is synthetic. Gene tree parsimony can build a species phylogeny from gene trees based on sequences sampled from different sets of species, as is the case with Figures 4a and 4b.
3. Gene tree parsimony produces species phylogenies corroborated by previous studies. Figures 5 and 6 show a basal division between the Australian/marine species (= sea snakes) and the African/Asian species. Further, the laticaudine (*Laticauda*) and hydrophiine (*Aipysurus*, *Astrotia*, *Enhydrina*, *Hydrophis*) sea snakes are shown to have separate origins. These results are very strongly corroborated by previous studies, both morphological and molecular (reviewed by Slowinski et al., 1997). The ability of a phylogenetic method to provide corroboration with previous hypotheses is probably the best arbiter of that method's reliability (Penny et al., 1982; Miyamoto and Cracraft, 1991; Slowinski, 1993).

Acknowledgments

We thank J. Doyle and W. Maddison for useful comments on the manuscript.

References

Assali, N. E., R. Mache, and S. L. De Goer. 1990. Evidence for a composite phylogenetic origin of the plastid genome of the brown alga *Pylaiella littoralis* (L.). Kjellm. Plant Mol. Biol. 15:307–315.

Avise, J. C., J. F. Shapiro, S. W. Daniel, C. F. Aquadro, and R. A. Lansman. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. Mol. Biol. Evol. 1:38–56.

Baum, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42:384–397.

Charleston, M. A. 1998. Jungles: A new solution to the host/parasite phylogeny reconciliation problem. Math. Biosci. 149:191–223.

de Queiroz, A. 1993. For consensus (sometimes). Syst. Biol. 42:368–372.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

Doyle, J. J. 1995. The irrelevance of allele tree topologies for species delimitation, and a nontopological alternative. Syst. Bot. 20:574–588.

Doyle, J. J. 1997. Trees within trees: Genes and species, molecules and morphology. Syst. Biol. 46:537–553.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN: A database of homologous vertebrate genes. Nucleic Acids Res. 22:2360–2365.

Dykhuizen, D. E., and L. Green. 1991. Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. 173:7257–7268.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Gibbs, A. J., and P. K. Keese. 1995. In search of origins of viral genes. Pages 76–91 *in* Molecular basis of virus evolution (A. J. Gibbs, C. H. Callisher, and F. Garcia-Arenal, eds.). Cambridge Univ. Press, Cambridge, UK.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.

Guigo, R., I. Muchnik, and T. F. Smith. 1996. Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 6:189–213.

Huelsenbeck, J. P., J. J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. 38:7–25.

Kluge, A. G., and A. J. Wolf. 1993. Cladistics: What's in a word. Cladistics 9:183–199.

Lawrence, J. G., and H. Ochman. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA 95:9413–9417.

Maddison, W. P. 1996. Molecular approaches and the growth of phylogenetic biology. Pages 47–63 *in* Molecular zoology: Advances, strategies, and protocols (J. D. Ferraris and S. R. Palumbi, eds.). Wiley-Liss, New York.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mirkin, B., I. Muchnik, and T. F. Smith. 1995. A biologically consistent model for comparing molecular phylogenies. J. Comp. Biol. 2:493–507.

Miyamoto, M. M., and J. Cracraft. 1991. Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. Pages 3–17 *in* Phylogenetic analysis of DNA sequences (M. M.

Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44:64–76.

Nixon, K. C., and J. M. Carpenter. 1996. On simultaneous analysis. Cladistics 12:221–241.

Page, R. D. M. 1993a. Genes, organisms, and areas: The problem of multiple lineages. Syst. Biol. 42: 77–84.

Page, R. D. M. 1993b. COMPONENT, version 2.0. The Natural History Museum, London.

Page, R. D. M. 1994a. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43:58–77.

Page, R. D. M. 1994b. Parallel phylogenies: Reconstructing the history of host–parasite assemblages. Cladistics 10:155–173.

Page, R. D. M. 1996. On consensus, confidence, and "total evidence." Cladistics 12:83–92.

Page, R. D. M. 1998. GeneTree: Comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819–820.

Page, R. D. M., and M. A. Charleston. 1997a. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. 7:231–240.

Page, R. D. M., and M. A. Charleston. 1997b. Reconciled trees and incongruent gene and species trees. Pages 57–70 *in* Mathematical hierarchies in biology, volume 37 (B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds.). American Mathematical Society.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Penny, D., L. R. Foulds, and M. D. Hendy. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297:197–200.

Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53–58.

Rodrigo, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. Taxon 42:631–636.

Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13:105–109.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

Siddall, M. E. 1997. Prior agreement: Arbitration or arbitrary. Syst. Biol. 46:765–769.

Slowinski, J. B. 1993. "Unordered" versus "ordered" characters. Syst. Biol. 42:155–165.

Slowinski, J. B., and C. Guyer. 1989. Testing the stochasticity of patterns of organismal diversity: An improved null model. Am. Nat. 134:907–921.

Slowinski, J. B., A. Knight, and A. P. Rooney. 1997. Inferring species trees from gene trees: A phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 8:349–362.

Stephens, J. C. 1985. Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. Mol. Biol. Evol. 2:539–556.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogeny inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.

Weiller, G. F. 1998. Phylogenetic profiles: A graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. 15:326–335.

Wu, C.-I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127:429–435.

Received 9 October 1998; accepted 10 March 1999
Associate Editor: R. Olmstead