Gene Tree Parsimony of Multilocus Snake Venom Protein Families Reveals Species Tree Conflict as a Result of Multiple Parallel Gene Loss

Nicholas R. Casewell, 1 Simon C. Wagstaff, 2 Robert A. Harrison, 2 and Wolfgang Wüster* 1

1 School of Biological Sciences, Environment Centre Wales, Bangor University, Bangor, United Kingdom
2 Alistair Reid Venom Research Unit, Liverpool School of Tropical Medicine, Liverpool, United Kingdom

*Corresponding author: E-mail: w.wuster@bangor.ac.uk.
Associate editor: Alexei Drummond

Abstract

The proliferation of gene data from multiple loci of large multigene families has been greatly facilitated by considerable recent advances in sequence generation. The evolution of such gene families, which often undergo complex histories and different rates of change, combined with increases in sequence data, pose complex problems for traditional phylogenetic analyses, and in particular, those that aim to successfully recover species relationships from gene trees. Here, we implement gene tree parsimony analyses on multilocus gene family data sets of snake venom proteins for two separate groups of taxa, incorporating Bayesian posterior distributions as a rigorous strategy to account for the uncertainty present in gene trees. Gene tree parsimony largely failed to infer species trees congruent with each other or with species phylogenies derived from mitochondrial and single-copy nuclear sequences. Analysis of four toxin gene families from a large expressed sequence tag data set from the viper genus Echis failed to produce a consistent topology, and reanalysis of a previously published gene tree parsimony data set, from the family Elapidae, suggested that species tree topologies were predominantly unsupported. We suggest that gene tree parsimony failure in the family Elapidae is likely the result of unequal and/or incomplete sampling of paralogous genes and demonstrate that multiple parallel gene losses are likely responsible for the significant species tree conflict observed in the genus Echis. These results highlight the potential for gene tree parsimony analyses to be undermined by rapidly evolving multilocus gene families under strong natural selection.

Key words: gene tree parsimony, multigene family, multiple copy genes, snake venom toxins, Bayesian posterior distributions, Serpentes: Viperidae: Echis.

Introduction

The key assumption of molecular systematics is that the generation of gene phylogenies provides information about the evolutionary relationship of the organisms from which the genes have been isolated (Cotton and Page 2002). It is often simply assumed that a gene phylogeny (gene tree) accurately represents the organismal phylogeny (species tree) of the species sampled (e.g., Okuda et al. 2001; Tsai et al. 2004, 2007). However, the suggestion that a species tree can be obtained simply by sampling a specific gene across a range of species is often erroneous (Page and Cotton 2000; Cotton and Page 2002), particularly if the gene is of multilocus origin rather than having a single chromosomal locus. Correctly inferred gene trees do not always correspond to species trees due to evolutionary processes such as duplication and loss, deep coalescence, and horizontal transfer (Goodman et al. 1979; Doyle 1992; Slowinski and Page 1999; Galtier and Daubin 2008). The combination of gene duplication and loss can produce conflicts with a species tree when paralogous sequences are sampled and treated as orthologous, a common occurrence in undersampled data sets (Slowinski et al. 1997; Page and Cotton 2000). Deep coalescence (or ancestral polymorphism) is an event at a single locus where a sequence from a less related species coalesces with one of the descendents of the deep coalescence (Slowinski et al. 1997; Slowinski and Page 1999). Deep coalescence can produce an analogous situation to duplication and loss because paralogous sequences are simply sequences that have coalesced prior to the ancestor of the species from which they were sampled (Slowinski et al. 1997; Slowinski and Page 1999). Sequencing both loci of a duplicated gene should resolve the discordance between species and gene trees due to paralogous sequences (Doyle 1992), highlighting the fundamental importance of substantial gene sampling. Horizontal transfer, including processes such as hybridization and gene transfer between species, is widely assumed to be more common in prokaryotes and of lesser importance in eukaryotic data sets (Syvanen 1994; Galtier and Daubin 2008).

Reconciliation of species and gene trees was first implemented by Goodman et al. (1979) and has subsequently been progressed by a number of different approaches over the years (e.g., Page 1994; Eulenstein et al. 1997; Ronquist 1997). An extension of tree reconciliation, gene tree parsimony, aims to identify the species tree that minimizes the assumptions of evolutionary events (duplications, losses, and/or deep coalescences) necessary to fit a given gene tree to the species tree (Slowinski et al. 1997; Slowinski and Page 1999).
a considerable challenge considering the frequency with which these events occur, particularly within rapidly diversifying gene families (Page and Cotton 2000). GeneTree (Page 1998) was the first program to implement this logical strategy by using simple, standard tree search heuristics to infer species trees from gene trees under three independent optimality criteria: duplications and losses, duplications-only, and deep coalescences. Subsequent programs and models have attempted to improve the biological realism of gene processes through time (e.g., Liu and Pearl 2007; Liu et al. 2010) or improve the implementation of gene tree parsimony (Sanderson and McMahon 2007; Oliver 2008; Wehe et al. 2008). Nevertheless, despite its simple search strategy, GeneTree remains the only widely available software that implements analyses for gene duplication and loss, gene duplications-only, and deep coalescence optimality criteria. Ideally, a strategy that uses heuristics to search for multiple gene processes simultaneously would be applied; however, such a method has yet to be implemented due to the fundamental problem of how to weight duplications, losses, and coalescences against each other. Despite this issue, gene tree parsimony has been reported to obtain results consistent with other analyses in snakes (Slowinski et al. 1997) and other vertebrates (Cotton and Page 2002) and performed well against a known species tree in an Angiosperm data set (Sanderson and McMahon 2007). Given the substantial increases in the generation of sequence data by recent genomic and transcriptomic studies, assessing the potential for gene tree parsimony to successfully recover species relationships from comprehensively sampled data sets has become a particularly timely exercise.

A major criticism of the original gene tree parsimony methodology was its failure to quantify confidence levels in the derived species tree by disregarding any uncertainty in the gene tree (Page and Cotton 2000; Sanderson and McMahon 2007). In order to account for gene tree uncertainty when inferring species trees, methodologies that incorporate the bootstrap have been implemented (Cotton and Page 2002; Sanderson and McMahon 2007) and the use of Bayesian posterior distributions has been advocated (Buckley et al. 2006; Oliver 2008). The use of Bayesian Markov Chain Monte Carlo analyses is particularly valuable, as they accommodate the inherent uncertainty present in gene genealogies, have increased sensitivity to phylogenetic signal, provide easy interpretation of results, and have computational advantages over other techniques (Larget and Simon 1999; Huelsenbeck and Ronquist 2001; Alfaro et al. 2003; Ronquist and Huelsenbeck 2003). Furthermore, it has been demonstrated that subjecting substantial numbers of gene tree Bayesian posterior distributions to multiple species tree searches prior to generating a majority-rule consensus tree can provide a rigorous assessment of node uncertainty within an inferred species tree (Buckley et al. 2006; Oliver 2008).

Snake venoms are complex mixtures of proteins and peptides; they exhibit a high level of biological activity and a diverse array of actions on both natural prey items and humans (Chippaux et al. 1991; Aird 2002). The majority of venom proteins (commonly referred to as toxins) appear to have been recruited into the venom gland from multigene protein families normally expressed in a variety of bodily tissues for ordinary physiological “housekeeping” purposes (Fry 2005). Following their recruitment, venom proteins evolve rapidly via a “birth and death” model of evolution, whereby frequent duplications of protein-encoding genes permit rapid functional and structural diversification alongside enhanced rates of sequence evolution (Nei et al. 1997; Kini and Chan 1999; Kordiš and Gubensˇek 2000; Župunski et al. 2003). Although some genes become deleted from the genome or degenerate into pseudogenes, others undergo neofunctionalization, resulting in the generation of a range of proteins that exhibit distinct functional diversification (Fry et al. 2003; Lynch 2007). These rapidly evolving gene families provide an ideal model to investigate whether species trees can be inferred from rapidly evolving multilocus gene families using gene tree parsimony.

Here, we use snake venom gland expressed sequence tags (ESTs) from four closely related species of saw-scaled vipers (Serpentes: Viperidae: Echis) (Casewell et al. 2009) to assess the validity of rapidly evolving multilocus gene families as species tree predictors. These identically generated multispecies EST data sets provide an unbiased, directly comparable, sampling resource and supply comprehensive multiple copy gene data for multiple gene families, whereas the rigorous generation of node support values using Bayesian posterior distributions provides a measure of confidence for species tree interpretation. Furthermore, the EST data, together with a quantitative measure of species tree support, permit a direct comparison between species trees inferred from nucleotide and translated nucleotides, thereby allowing us to investigate the relationship between gene trees and species trees in greater detail. We assess the capability of gene tree parsimony to resolve species relationships from multilocus genes by comparing inferred species trees derived from venom protein gene trees with a template species phylogeny of the genus Echis derived from multiple loci deep coalescence analysis (Maddison and Knowles 2006; Maddison WP and Maddison DR 2008; Oliver 2008) of four mitochondrial and three single-copy nuclear genes. Snake venom protein families have previously been analyzed using gene tree parsimony: Slowinski et al. (1997) inferred species relationships consistent with other analyses from phospholipase A<sub>2</sub> (PLA<sub>2</sub>) and short neurotoxin (NKS) venom proteins isolated from members of the Elapidae (Serpentes). However, this investigation did not take into account the substantial uncertainty observed in the gene trees; for this reason, we revisit this data set and apply a Bayesian gene tree parsimony methodology in order to interpret the inferred species trees alongside rigorously generated node support values.

**Materials and Methods**

**Mitochondrial and Single-Copy Nuclear Sequences**

In order to provide a “template” phylogeny for the four Echis taxa under consideration (E. ocellatus, E. coloratus, E. pyramidum and E. carinatus) and two outgroup taxa (Bitis arietans...
and Cerastes cerastes—Wüster et al. 2008; Pook et al. 2009), we used a combination of data from several single-copy loci. These included the data generated by Barlow et al. (2009) for four mitochondrial genes (cytochrome b, NADH dehydrogenase subunit 4, 12s rRNA, and 16s rRNA) and one single-copy nuclear gene (recombination activating gene 1 [RAG1]). Additionally, we generated novel sequences for two additional single-copy nuclear genes, prolactin receptor (PRLR) and ubiquitin 1 (UBN1) (Townsend et al. 2008). Nuclear gene DNA sequence alignments were generated using techniques previously described by Barlow et al. (2009) (details of polymerase chain reaction primers and reaction conditions are provided in the supplementary methods, Supplementary Material online). Five data sets were prepared for analysis by Bayesian inference: 1) all mitochondrial genes, 2) RAG1, 3) PRLR, 4) UBN1, and 5) a concatenated data set of all mitochondrial and nuclear genes.

Venom Toxin DNA and Amino Acid Sequences
Venom gland cDNA libraries were constructed using procedures previously outlined (Wagstaff and Harrison 2006; Casewell et al. 2009). Briefly, multiple cDNA libraries were constructed from ten wild-caught specimens each of E. ocelatus (Nigeria), E. coloratus (Egypt), E. pyрамидum leakeyi (Kenya), and E. carinatus sochureki (Sharjah, UAE); ~1,000 random clones per species were picked for sequencing using M13 forward primers. ESTs were bioinformatically processed using the PartiGene pipeline (Parkinson et al. 2004) with high stringency cluster on the basis of Blast similarity (CLOBB) clustering (Parkinson et al. 2002; Wagstaff and Harrison 2006) and Blast annotation against multiple databases (see Casewell et al. 2009). ESTs exhibiting significant (>1 × 10^-5) Blast annotation to the four most heavily represented venom proteins present in the venom gland transcriptomes, the snake venom metalloproteinase (SVMP), C-type lectin (CTL), PLA2, and serine protease (SP) protein families (Casewell et al. 2009), were identified before alignment in ClustalW (Thompson et al. 1994). Full-length sequencing of PLA2 and CTL clones were obtained during the initial round of sequencing, whereas reverse sequencing, using M13 reverse primers, was carried out on all SP clones to generate full-length sequences. Full-length SVMP sequence information was gained via primer walking a nonredundant set of SVMP clones (see supplementary methods, Supplementary Material online) that demonstrated sequence similarity to the catalytic site (H-box) of the metalloproteinase domain (Fox and Serrano 2005). Outgroup sequences for each gene family were identified by sequence similarity searches against a number of non-Serpentes databases. The data sets were trimmed to the open reading frame of the translated proteins; identical sequences and those containing truncations or frameshifts as the result of insertions or deletions were excluded in MEGA v4.0.2 (Tamura et al. 2007). The alignment of full-length variants with ClustalW (Thompson et al. 1994) preceded additional manual adjustments. The finalized DNA data sets were then translated into amino acids (AAs) and realigned before the exclusion of any remaining identical sequences.

The Elapidae PLA2 (59 sequences from 25 species) and NXS data sets (42 sequences from 27 species) analyzed by Slowinski et al. (1997) were retrieved from the protein database SWISS-PROT using the NCBI browser. Signal sequences were removed in MEGA v4.0.2 prior to alignment in ClustalW and subsequent manual adjustments. All alignment files have been submitted to TreeBASE and can be accessed via the following URL (http://purl.org/phylo/treebase/phylows/study/TB2:S10791).

Gene Tree Generation
DNA gene trees from mitochondrial and single-copy nuclear genes (for the generation of a template species tree) and DNA and AA gene trees (from multilocus toxin gene families) were produced using optimized models of sequence evolution combined with Bayesian inference. Given that complex models of sequence evolution have been demonstrated to extract additional phylogenetic signal from data (Castoe et al. 2005; Castoe and Parkinson 2006), we subjected all DNA data sets to analysis in MrModeltest v2.3 (Nylander 2004) and the AA data sets in ModelGenerator v0.85 (Keane et al. 2006). Prior to analysis of the DNA data sets, sequences were partitioned into first, second, and third codon partitions to incorporate any differences in patterns of sequence evolution. The model favored under the Akaike Information Criterion (Posada and Buckley 2004) was selected for all partitions. Bayesian inference analyses were undertaken in MrBayes v3.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) on the freely available bioinformatic platform Bioportal (www.bioportal.uio.no). Each data set was run in duplicate using four chains simultaneously (three heated and one cold) for 5 × 10^6 generations, sampling every 500th cycle from the chain and using default settings in regard to priors. We used the program Tracer v1.4 (Drummond and Rambaut 2007) to estimate effective sample sizes for all parameters and to construct plots of ln(L) against generation to verify the point of convergence (burnin); trees generated prior to the completion of burnin were discarded.

Gene Tree Parsimony
To infer species trees from both DNA and AA venom protein gene trees, we implemented a gene tree parsimony strategy similar to that described by Buckley et al. (2006) and Oliver (2008) using a novel bioinformatic pipeline consisting of GeneTree v1.0 (Page 1998) and PAUP* v4.0b10 (Swofford 2002). The topologies of all postburnin trees (36004) generated for each Bayesian data set were extracted in PAUP* using the savetrees command, while removing branch lengths and internal node labels. Tree topologies were edited to GeneTree input specifications before each of the trees was subjected to heuristic species tree searches in GeneTree using the steepest ascent option. Each analysis was run using 50 heuristic searches in order to undertake a comprehensive search of the tree space and to account for extraneous random starting trees, whereas
branch swapping was carried out using the most effective option (ALT), which alternates between nearest-neighbor interchanges and subtree pruning and regrafting (Page and Charleston 1997). The individual species trees inferred from each of the postburnin gene trees were subsequently summarized into a single majority-rule consensus species tree. The frequency of each node inferred from the 36004 species trees thus represents a measure of the uncertainty for the relationships present in the consensus species tree. For each venom protein family, the heuristic searches in GeneTree were implemented for three separate optimality criteria for both DNA and AA data sets: 1) duplications and losses, 2) duplications-only, and 3) deep coalescences.

The *Echis* mitochondrial and single-copy nuclear gene data set and the *Echis* and Elapidae venom toxin data sets comprise of independent nonhomologous gene families (Takahata 1989; Maddison 1997). It was therefore essential to undertake analyses that act to simultaneously minimize deep coalescences in the different gene families that comprise each of these three data sets. Analyses minimizing deep coalescences were undertaken for: 1) the four *Echis* DNA mitochondrial and single-copy nuclear genes (to simultaneously consider all genes used for species tree reconstruction) and 2) the *Echis* DNA SVMP, CTL, PLA₂, and SP genes (to reconstruct the phylogeny of *Echis* from all multilocus toxin gene families together), and 3) elapid AA PLA₂ and NXS genes (to simultaneously consider toxin genes used in gene tree parsimony). These analyses incorporated postburnin Bayesian posterior distributions into the multiple loci deep coalescence method of AUGIST in the Mesquite software system (Maddison and Knowles 2006; Maddison WP and Maddison DR 2008; Oliver 2008). Randomly sampled trees from multiple posterior distribution sources were rearranged using the subtree pruning and regrafting (SPR) option of the deep coalescence multiple loci analysis, using 500 replicates and the “contained polytomies auto-resolve” option rejected. The multiple resolved trees resulting from each analysis were summarized into a single majority-rule consensus species tree, as previously described, to provide estimates of node support for the inferred species relationships. In light of differing patterns of gene duplication and loss in the *Echis* toxin genes (see below), we repeated the analysis for this data set but excluding SP genes.

**Results**

**Echis** Species Tree

A total of 578 bp of PRLR (*n = 6*) (GenBank: HM623444–HM623449) and 470 bp of UBN1 (*n = 6*) (GenBank: HM623450–HM623455) nuclear gene sequence data was aligned alongside the 3,145 bp (*n = 6*) generated by Barlow et al. (2009). For Bayesian inference, models of sequence evolution for the novel data partitions (PRLR and UBN1) were identified in MrModelTest v2.3 (supplementary table 1, Supplementary Material online). The species tree generated by Bayesian analysis of the concatenated mitochondrial and nuclear genes of Barlow et al. (2009) combined with PRLR and UBN1 revealed a topology (fig. 1A) partially incongruent with previous analyses of the genus *Echis* (Barlow et al. 2009; Pook et al. 2009). However, the node responsible for this incongruence, the monophyly of *E. pyramidum* and *E. ocellatus*, is not significantly supported (<0.95). Moreover, the individual species trees generated by Bayesian analysis of the four single gene loci (mitochondrial, RAG1, PRLR, and UBN1) produced contrasting topologies (supplementary fig. 1A–D, Supplementary Material online), with different nuclear genes significantly supporting nodes that are significantly contradicted by others; this strongly implies that, as unlinked loci, they have undergone different histories (Takahata 1989; Maddison 1997). We therefore adopted a multiple loci deep coalescence approach, incorporating Bayesian posterior distributions generated for the four mitochondrial and single-copy nuclear loci, to resolve the evolutionary history of the genus *Echis*. Deep coalescence analysis produced a species tree with the same topology as that determined by Barlow et al. (2009) (fig. 1B), although the monophyly of
Gene Tree Parsimony of Venom Proteins

E. coloratus and E. p. leakeyi was not significantly supported. Nevertheless, evidence of contrasting gene histories (supplementary fig. 1, Supplementary Material online) and congruence with a previous analysis (Barlow et al. 2009) (B—in parentheses) are shown for relevant nodes. Outgroup taxa are Cerastes cerastes and Bitis arietans.

Venom Protein Gene Trees

The Echis venom protein data sets comprised of a total of 2,004 bp of SVMP (n = 209) (GenBank: GU012123–GU012315 and AM039691–AM039701), 780 bp of SP (n = 32) (GenBank: GU012092–GU012122), 519 bp of CTL (n = 130) and 444 bp of PLA2 (n = 42) aligned sequence data (a list of CTL and PLA2 GenBank accession numbers can be found in supplementary table 2, Supplementary Material online). The aligned Echis AA data sets represented 667 AAs of SVMP (n = 194), 260 AAs of SP (n = 27), 173 AAs of CTL (n = 116), and 144 AAs of PLA2 (n = 33) sequence data. The Elapidae data sets implemented by Slowinski et al. (1997) were aligned into 126 AA of PLA2 (n = 59) and 65 AA of NXS (n = 42) sequence data. Models of evolution assigned by MrModelTest v2.3 for DNA partitions and ModelGenerator v0.85 for AA data sets are displayed in supplementary table 1 (Supplementary Material online) Tracer v1.4 (Drummond and Rambaut 2007) revealed the burn-in period for each Bayesian analysis had occurred prior to the first 1 × 10^5 generations for all parameters, but we conservatively discarded the first 5 × 10^5 generations and calculated the consensus trees from the remaining posterior distributions. All parameters of the Tracer analyses had effective sample sizes greater than 250, in most cases by a large margin. All consensus gene trees generated by Bayesian inference are available on TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S10791).

Gene Tree Parsimony in the Genus Echis

The majority-rule consensus trees generated by gene tree parsimony analyses for duplication and loss, duplications-only, and deep coalescences are shown in figure 2 (DNA) and figure 3 (AA). Notably, considerable variation was observed in the species trees inferred from the different venom protein families; no less than nine differing species tree topologies (out of 26 possible) resulted from the 24 analyses. Furthermore, only two fully resolved species trees, generated using the SVMP AA gene trees under the duplication and loss and duplications-only optimality criteria, matched the baseline phylogeny for this genus derived from deep coalescence analysis of multiple mitochondrial and nuclear gene loci (fig. 1B). Furthermore, only one node, supporting the monophyly of E. coloratus and E. p. leakeyi, was strongly supported (>95%) in both of these trees. Although this node represented the most frequent node observed in the Echis species trees, it was not ubiquitous and only strongly supported in 25% (DNA) and 58% (AA) of the inferred trees. However, many other toxin-derived species trees contained nodes incongruent both with the baseline species phylogeny (fig. 1B) and with species trees inferred from other toxin gene trees, although only one of these was strongly supported (monophyly of E. ocellatus and E. p. leakeyi in the case of the SPs under the duplication and loss and deep coalescence criteria, for both DNA and AA sequence data). A number of other nodes were unresolved or weakly supported, highlighting the lack of topological consistency observed throughout the inferred species trees.

Simultaneous deep coalescence analysis of the four Echis venom protein families produced a tree topology (fig. 4A) entirely incongruent with the template species phylogeny (fig. 1B). Reanalysis of this data with the exclusion of the SP toxin family recovered one node congruent (monophyly of E. p. leakeyi and E. coloratus) with the template phylogeny (fig. 4B).

Gene Tree Parsimony in the Family Elapidae

The species trees inferred from gene tree parsimony analysis of the PLA2 family are displayed in figure 5; despite the differences in the optimality criterion employed by gene tree parsimony, the inferred species trees display similar topologies. The NXS data (fig. 6) produced largely unresolved species topologies, except when implementing the duplications-only criterion in DupTree.

Fig. 1. Phylogeny of the major Echis species groups inferred by (A) Bayesian analysis of four mitochondrial genes and three nuclear genes and (B) deep coalescence analysis of posterior distributions generated by Bayesian analysis of the multiple mitochondrial and nuclear gene loci. Bayesian posterior probabilities (A) and those inferred by AUGUST analysis (B) and Barlow et al. (2009) (B—in parentheses) are shown for relevant nodes. Outgroup taxa are Cerastes cerastes and Bitis arietans.

1161
coalescence analysis, simultaneously incorporating both venom protein families, revealed a largely unresolved and unsupported species topology of the elapids (fig. 7).

**Discussion**

**Gene Tree Parsimony in the Genus *Echis***

Evolutionary analyses of mitochondrial and single-copy nuclear genes were undertaken to generate a template species phylogeny for comparison with species trees inferred by gene tree parsimony analyses of venom protein families. The species trees generated by Bayesian inference of single gene loci (mitochondrial, RAG1, PRLR, and UBN1) produced contrasting topologies (supplementary fig. 1, Supplementary Material online), suggesting that these genes have different histories. Evidence of deep coalescence in the genus *Echis* is not surprising, given that the divergence of the four *Echis* species groups has likely occurred in quick succession (Pook et al. 2009). We subsequently undertook multiple loci deep coalescence analysis of posterior distributions of the mitochondrial and nuclear gene trees to account for this source of discordance, producing a species tree topology (fig. 1B) congruent with previous phylogenetic analyses (Barlow et al. 2009; Pook et al. 2009). We therefore used this topology as the template *Echis* species phylogeny for comparison with species trees inferred by gene tree parsimony.

A total of nine distinct species trees were generated from gene tree parsimony analyses of four *Echis* venom protein families, simultaneously incorporating both venom protein families, revealed a largely unresolved and unsupported species topology of the elapids (fig. 7).

**Fig. 2.** Majority-rule consensus trees for four DNA data sets of *Echis* venom protein families using gene tree parsimony. Separate analyses were implemented to minimize duplications and loss, duplications-only and deep coalescences. Circles indicate nodes congruent and crosses indicate nodes incongruent with the species phylogeny determined by multiple loci deep coalescence analysis. Black circles and crosses indicate that a node is robustly supported (>95%), gray symbols indicate insignificant node support.
protein families. The incongruence among species trees derived from these data sets is highlighted by the fact that the most frequent species tree topology represents only 25% of the total number of inferred species trees. Furthermore, the most commonly observed species tree is only partially resolved. This incongruence is particularly surprising; the SVMP and CTL DNA duplication and loss analyses were alone in producing fully resolved identical topologies, although neither tree exhibited strong support (>95%) for every node. The lack of consistency between species trees is important, as it highlights the absence of any strong signal opposing that of the template species tree inferred from mitochondrial and single-copy nuclear gene sequences (fig. 1B): consistent conflict between species trees inferred from gene tree parsimony and the template species phylogeny might suggest that the latter tree is in error; however, we did not uncover any consistent pattern of conflict. Moreover, the two nodes present in the inferred species trees that are congruent with the template phylogeny are only strongly supported in 42% (monophyly of *E. coloratus* and *E. p. leakeyi*) and 13% (monophyly of *E. coloratus*, *E. p. leakeyi*, and *E. ocellatus*) of the total derived trees. Interestingly, the SPs were the only venom protein family that strongly contradicts the monophyly of *E. p. leakeyi* and *E. coloratus*; instead, the monophyly of *E. p. leakeyi* and *E. ocellatus* is strongly supported, except in the duplications-only analysis, which fails to resolve the relationships among *E. coloratus*, *E. p. leakeyi*, and *E. ocellatus*.

**Fig. 3.** Majority-rule consensus trees for four AA data sets of *Echis* venom protein families using gene tree parsimony. Separate analyses were implemented to minimize duplications and loss, duplications-only, and deep coalescences. Circles indicate nodes congruent and crosses indicate nodes incongruent with the species phylogeny determined by multiple loci deep coalescence analysis. Black circles and crosses indicate that a node is robustly supported (>95%), gray symbols indicate insignificant node support.
Species trees derived from multiple loci deep coalescence analyses of Echis venom protein families. (A) SVMP, CTL, PLA2, and SP loci and (B) SVMP, CTL, and PLA2 loci. Circles indicate nodes congruent and crosses indicate nodes incongruent with the species phylogeny determined by multiple loci deep coalescence analysis.

The majority of GeneTree analyses of Bayesian posterior distributions generated from the Echis DNA data sets inferred fully resolved species trees that are typically supported by higher node values than their AA counterparts. This is not unexpected given the increase in the number of phylogenetically informative characters used to resolve the DNA gene trees and indicates, in this case, that nucleotide data sets are preferable for gene tree parsimony analyses of snake venom toxins. Although many AA species trees display unresolved nodes, the majority of the resolved clades display topologies identical to those inferred by the corresponding DNA data sets. The main exception is that the AA duplication and loss and duplications-only SVMP species trees both exhibit different tree topologies from their DNA counterparts, with greater node support. Coincidentally, these species trees are unique in inferring the topology of the genus Echis as predicted by the deep coalescence analysis of mitochondrial and multiple single-locus nuclear genes.

Altering the optimality criteria implemented in GeneTree also resulted in alterations to the inferred species tree topologies: Trees minimizing deep coalescences are often incongruent with and display lower node support values than those inferred by minimizing duplications and losses and duplications-only. In general, major changes in species tree topologies are not observed between the duplication and loss and duplications-only analyses, consistent with the assumption that losses are informative in the EST data sets as a consequence of comprehensive gene sampling. However, in contrast to the duplications and loss analyses, none of the species trees derived from the duplications-only criteria exhibit strongly supported nodes that conflict with the template species tree, implying that the inclusion of loss events may be partially responsible for gene tree parsimony failure in this data set. Notably, the duplications-only analyses for the SP venom protein family reveal a consistent change in tree topology from fully resolved to partially resolved, the collapsed node being the grouping of E. ocellatus and E. p. leakeyi that is inconsistent with the template species tree (figs. 2 and 3). Given the comprehensive sampling methodology for the four venom protein families, this observation implies that gene loss has a greater influence on the outcome of species tree reconstruction from the SPs than from the other venom protein families.

Considering the lack of congruence between Echis species trees inferred from venom protein gene trees and the template species phylogeny, gene tree parsimony was subsequently undertaken by simultaneously considering posterior distributions derived from the multiple toxin loci (SVMP, CTL, PLA2, and SP) using the deep coalescences multiple loci analysis of AUGUST in Mesquite (Maddison and Knowles 2006; Maddison WP and Maddison DR 2008; Oliver 2008). This approach also failed to infer a species tree congruent with the template species phylogeny determined by deep coalescence analysis of multiple mitochondrial and nuclear loci (fig. 1B) and included the monophyly of E. p. leakeyi and E. ocellatus (fig. 4A). Considering different venom proteins represent independent nonhomologous gene families, these results are not unexpected; differing gene families likely exhibit different gene histories and rates of change (Takahata 1989; Maddison 1997). However, it is notable that when excluding sequence data responsible for the generation of strongly supported nodes incongruent with the template phylogeny (i.e., the SP toxin family) from the deep coalescence multiple loci analysis, the monophyly of E. p. leakeyi and E. coloratus (fig. 4B), a node congruent with our template species tree (fig. 1B) and the analyses of Barlow et al. (2009) and Pook et al. (2009), is recovered. Although none of the nodes in this species tree are significantly supported, these results suggest that patterns of gene loss in the SP gene family may have an important effect on the results of gene tree parsimony analysis in the genus Echis.

Gene Tree Parsimony in the Family Elapidae

The inferred species trees generated from the Elapidae PLA2 gene family provided support for the monophyly of the Australian and marine elapid radiation throughout the varying gene optimality criteria (fig. 5). The analysis minimizing duplications-only was alone in resolving the relationships within this clade with any significant support, although a number of inferred nodes are strongly contradicted by recent molecular phylogenetic studies (Slowinski and Keogh 2000; Sanders et al. 2008). The monophyly of marine and terrestrial Australian species was also inferred by Slowinski et al. (1997), with their analysis suggesting a fully resolved topology incongruent with our analyses, likely reflecting an unsupported species relationship. In the context of the African and Asian elapids, the monophyly of Aaspidelaps, Hemachatus, and Naja established by our consensus trees matched those of Slowinski et al. (1997); despite this consistency, support values are insufficient to exclude the possibility of an alternate topology. Furthermore, Slowinski et al.’s (1997) placement of Bungarus as outgroup to Aaspidelaps, Hemachatus, and Naja is unsupported by our PLA2 analyses, with different gene optimality criteria producing contrasting topologies. However, within the genus Naja, our consensus trees display a number of nodes consistent with both Slowinski et al. (1997) and recent mitochondrial phylogenies (Wüster
and Broadley 2007; Wüster et al. 2007): for example, monophilies of the African spitting cobras (N. mossambica, N. pallida, and N. nigricollis) and the Asian cobras (N. kaouthia, N. atra, N. naja, and N. oxiana).

The NXS consensus trees produced largely unresolved species topologies, except when implementing the duplications-only criterion in DupTree (fig. 6). The observed unresolved topologies and corresponding low node support values are
perhaps unsurprising given that the NXS data set contains less sequence data (65 AAs) and thus fewer characters than the other toxin families, due to the short length of the NXS genes. All three gene analyses provided strong support for Laticauda as the sister taxon of all other Elapidae, conflicting with the analyses of Slowinski et al. (1997) and a recent multigene phylogeny (Sanders et al. 2008). The significant support values associated with our placement of Laticauda suggest that the topology obtained by Slowinski et al. (1997) may not have been strongly supported, despite its consistency with Sanders et al. (2008). On the other hand, in keeping with other recent studies (Sanders et al. 2008).
et al. 2008), the monophyly of the marine and terrestrial Australian snakes, excluding *Laticauda*, is supported in each consensus tree (all nodes >90%), displaying a topology similar to that described previously (Slowinski et al. 1997), although inconsistencies with recent molecular phylogenies exist within the clade (Lukoschek and Keogh 2006; Sanders et al. 2008). The NXS consensus trees also fail to resolve the species relationships of the African and Asian elapids, except when minimizing duplications-only. Only three nodes exhibit significant support values, of which one, the monophyly of *N. mossambica*, *N. kaouthia*, and *N. atra*, is refuted by a recent mitochondrial phylogeny (Wüster et al. 2007). In contrast, Slowinski et al. (1997) obtained a largely resolved topology for the African and Asian elapids; this incongruence suggests again that the topology provided by Slowinski et al. (1997) is largely unsupported.

Deep coalescence analysis of the two venom protein family loci in AUGIST (Maddison and Knowles 2006; Maddison WP and Maddison DR 2008; Oliver 2008) also failed to resolve the species relationships of the elapids, producing an almost entirely unresolved tree topology, with only one node exhibiting significant support values (fig. 7). The simultaneous incorporation of multiple genes isolated from unlinked loci reduced support for the elapid species relationships in comparison with independent analyses of PLA₂ and NXS venom protein families (figs. 5 and 6). These outcomes are likely the result of using only two genes with highly incomplete species sampling, combined with unequal sampling of paralogous genes within each gene family, thereby resulting in a reduction in the resolution of the inferred species tree.

The Basis of Unsuccessful Tree Reconciliation

Our gene tree parsimony analyses were largely unsuccessful at reconstructing species trees from venom protein families that are congruent with species phylogenies derived from mitochondrial and single-copy nuclear genes. Despite the previous apparent success of venom protein gene tree parsimony (Slowinski et al. 1997), our results show that significant changes in these inferred Elapidae tree topologies occur when incorporating gene tree uncertainty. Furthermore, a number of relationships obtained by Slowinski et al. (1997) are not significantly supported in our species trees; these inconsistencies emphasize the importance of assessing node support in species trees obtained through gene tree parsimony. Despite partial species tree congruence between the analysis of Slowinski et al. (1997) and more recent molecular studies, little consensus can be derived from the inferred species trees, with different venom protein families inferring different evolutionary histories within the family Elapidae, whereas simultaneous analysis of both venom protein families produced an unresolved species topology. It is highly plausible that the failure of gene tree parsimony in the elapid data sets is a result of unequal and/or highly incomplete sampling of paralogous genes (mean number of sequences per species = 1.6 [NXS] and 2.0 [PLA₂]) preventing the correct species tree being extracted. Previous analyses of these venom protein families demonstrate the presence of multiple copies of NXS (Fry et al. 2003) and PLA₂ (Lynch 2007) genes in numerous elapid species, highlighting the likelihood that paralogous gene sampling has occurred. The observed conflict between species trees obtained from different toxin families, and between them and those from single locus genes, is the inevitable result of using sequence data collected non-systematically during the course of diverse toxinological studies.

However, gene tree parsimony was also unsuccessful at inferring the species relationship in the genus *Echis*; despite using substantially greater numbers of sequences, base pairs, venom protein families, and fewer species, only two derived species trees correctly inferred the topology determined by deep coalescence analysis of multiple mitochondrial and nuclear gene loci. In addition, the unbiased
EST sampling method incorporated for the Echis data set more likely reflects the multiple copy nature of these venom protein families and has been demonstrated to be representative of proteomic venom expression (Wagstaff et al. 2009). Notably, nodes throughout the Echis species trees are typically weakly supported, whether they are consistent or inconsistent with the template species phylogeny (fig. 1B). Conflict arising from species tree reconciliation is likely to be a result of this weak signal; the majority of nodes responsible for causing species tree incongruence with the species phylogeny are unsupported (<95%). The consistent exception to this observation occurs in the SP venom protein family, where the duplication and loss inferred species trees produced a strongly supported (>95%) topology incongruent with the species phylogeny, both for AA and DNA-derived gene trees.

Recombination was excluded as a factor confounding species tree reconciliation; analyses undertaken in the Recombination Detection Program v3.34 (RDP3) (Heath et al. 2006) (see supplementary methods, Supplementary Material online) revealed only false positive results in the CTL, PLA2, and SP data sets (data not shown) that

---

**Fig. 8.** Selective and parallel loss events preventing correct species tree reconciliation. Numbers refer to alleles and letters A–D refer to species. Circles indicate duplication events and crosses indicate loss events. (A) Repeated selective loss in species D leads to gene tree parsimony inferring the incorrect species tree if duplications and losses are taken into account. (B) Multiple parallel gene loss in species B and C leads to gene tree parsimony inferring the incorrect species tree if duplications and losses are taken into account. In both cases, the number of events required to derive the correct species tree is four (two duplications and two losses), whereas the most parsimonious explanation infers an incorrect species tree with only three gene events (two duplications and one loss). Note also that, in both cases, gene tree parsimony will underestimate the number of gene losses.
exhibited significance scores similar to those obtained from a vertebrate mitochondrial cytochrome b data set (supplementary table 3, Supplementary Material online) devoid of recombination and much lower than a snake venom protein data set that had previously been demonstrated to contain recombinants (Zha et al. 2006). Although the RDP3 analysis revealed four SVMP sequences (out of 209) containing evidence of apparent recombination (GenBank: GU012190, GU012203, GU012213, GU012261), all but one of these recombinants are nested within species-specific toxin clades and therefore would not have influenced the reconstruction of the species tree. Furthermore, as all the recombinant sequences are from *E. coloratus* and *E. p. leakeyi*, yet the relationship between these two species is correctly inferred in five of the six SVMP gene tree parsimony analyses, we exclude recombination as a factor responsible for confounding gene tree parsimony. Venom protein families may also be subjected to additional evolutionary phenomenon such as accelerated segment switches in exons to alter targeting (ASSET), where exons are radically changed to unrelated sequences leading to rapid functional evolution (Doley et al. 2008, 2009). Recent analyses demonstrated that ASSET may play a significant role in the evolution of certain venom protein families, including SVMPs, PLA2s, and SPs (Doley et al. 2009). In order to exclude the potential role of ASSET confounding gene tree parsimony, we repeated the analyses for the venom protein families described previously excluding the regions of DNA and corresponding AA sequence demonstrated to be under the influence of ASSET (Doley et al. 2009 and see supplementary methods, Supplementary Material online). All of these analyses produced inferred species tree topologies consistent with the original analyses (data not shown).

The composition of snake venom proteins is under strong natural selection for adaptation toward specific diets (e.g., Daltry et al. 1996; Gibbs and Rossiter 2008; Barlow et al. 2009; Gibbs and Mackessy 2009). Consequently, we cannot rule out the effects of selection on patterns of gene duplication and loss as a factor confounding gene tree parsimony by influencing gene events within lineages with divergent diets. Because members of the genus *Echis* exhibit considerable variation in prey preference (Barlow et al. 2009), we propose that adaptive selection pressures are responsible for generating the strongly supported SP species trees that are incongruent with the template *Echis* phylogeny derived by deep coalescence analysis of multiple mitochondrial and nuclear loci (fig. 18). The presence of repeated selective loss in one lineage (fig. 8A) or multiple parallel loss in multiple lineages (fig. 8B) can confound gene tree parsimony; in both cases, the most parsimonious explanation for the species relationship requires fewer gene events than that of the true species tree. In the case of the SPs, the gene trees (supplementary fig. 2, Supplementary Material online) exhibit minimal representation of clades containing *E. ocellatus* and *E. p. leakeyi* SPs, suggesting that multiple parallel gene loss may have occurred in these two species. Consequently, any gene tree parsimony analyses seeking to minimize the required number of assumptions of gene loss would result in a species tree grouping these taxa together (see fig. 8). This hypothesis was tested by analyzing the AA SP gene data and the template *Echis* phylogeny (fig. 18) in GeneTree by implementing the reconciliation option. Reconciling the gene tree within the template species tree suggests multiple parallel loss events are occurring in each lineage, with the exclusion of *E. coloratus* (fig. 9). We therefore infer that gene tree parsimony is failing to produce a species tree topology congruent with the template species tree as a result of multiple parallel gene loss events; the incongruent monophyly of *E. ocellatus* and *E. p. leakeyi* occurs as parsimony minimizes the number of gene events required to reconcile the gene tree to the species tree (see fig. 8B). These results explain the gene processes that are responsible for the presence of strongly supported incongruent nodes in the trees derived from the *Echis* SPs and highlight the method by which gene tree parsimony can be undermined by non-random gene events occurring in multigene families evolving under strong selection pressures.

**Conclusions**

We have demonstrated the importance of rigorously assessing node support for inferred species trees generated by gene tree parsimony. The implementation of Bayesian posterior distributions for multiple venom protein families allowed inferred species trees to be interpreted with confidence and highlighted a lack of support for a number of previously reconstructed evolutionary relationships in our two data sets. In this case, gene tree parsimony largely failed to infer strongly supported species trees from a comprehensive data set of four multigene venom protein families isolated from four closely related members of the genus *Echis* and from a smaller data set of two venom protein families from members of the family Elapidae. It is notable, yet not unexpected, that when incorporating gene tree uncertainty for estimates of species tree inference, the estimates of species relationships appear to reflect more uncertainty. Here, we suggest the failure of gene tree parsimony to consistently resolve the elapid species relationship is the result of unequal and/or highly incomplete sampling of paralogous genes. Weak signal,
evident by low node support values, undermines species tree derivation in the *Echis* data sets, although we suggest that strongly supported conflict observed in the SP gene family is the result of nonrandom patterns of parallel gene loss and we have described how such gene processes may confound gene tree parsimony. Given that the relationship between venom protein gene trees and inferred species trees has been demonstrated to be complex, we suggest that utmost caution should be employed when interpreting gene tree data generated from rapidly evolving multilocus gene families likely to be under natural selection.

**Supplementary Material**

Supplementary figures 1 and 2, tables 1–3, and supplementary methods are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors wish to thank Paul Rowley (Liverpool School of Tropical Medicine) for expert herpetological assistance, Axel Barlow (University of Bangor) for design and optimization of the UBN1 primers, Damien Egan and Paul Ver-cammen (Breeding Centre for Endangered Arabian Wildlife, United Arab Emirates) for providing specimens of *E. c. sochureki*, Ann Hedley and Mark Blaxter (NERC Molecular Genetics Facility, University of Edinburgh) and Tim Booth, Bela Tiwari, and Jorge Soares (NERC Environmental Bioinformatics Centre, Oxford) for providing bioinformat-ic advice and Wayne Maddison (University of British Columbia) and Jeffrey Oliver (Yale University) for assistance with Mesquite and AUGIST. We also thank three anonymous reviewers for their insightful comments on various aspects of the manuscript. This work was funded by Research Studentship NER/S/A/2006/14086 from the Natural Environmental Research Council (NERC) to N.R.C., access to the NERC Molecular Genetics Facility at the University of Edinburgh (ref MGF 150) to W.W., the Leverhulme Trust (Grant F/00 174/l) to W.W. and R.H. and the Biotechnology and Biological Sciences Research Council (BBSRC) to R.H. and S.C.W. (BB/ F012675/1).

**References**


Nylander JAA. 2004. MrModeltest v2. Program distributed by the author, Evolutionary Biology Centre, Uppsala University.


Page RDM. 2004. MrModeltest v2. Program distributed by the author, Evolutionary Biology Centre, Uppsala University.


Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW. 2008. Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. Mol Phylogenet Evol. 47:129–142.


