YMPEV 1360
DISK / 2/4/03 / Chitra(CE)/ Rajesh(TE)
ARTICLE IN PRESS
No. of pages: 11
DTD 4.3.1 / SPS

ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Molecular Phylogenetics and Evolution xxx (2003) xxx–xxx

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

Gene tree parsimony vs. uninode coding for phylogenetic reconstruction

James A. Cotton* and Roderic D.M. Page

Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Received 11 December 2002; revised 14 February 2003

Abstract

Simmons and Freudenstein (2002) have suggested that there are important weaknesses of gene tree parsimony in reconstructing phylogeny in the face of gene duplication, weaknesses that are addressed by Simmons et al.'s (2000) method of uninode coding. Here, we discuss Simmons and Freudenstein's criticisms and suggest a number of reasons why gene tree parsimony is preferable to uninode coding. During this discussion we introduce a number of recent developments of gene tree parsimony methods overlooked by Simmons and Freudenstein. Finally, we present a re-analysis of data from Page (2000) that produces a more reasonable phylogeny than that found by Simmons and Freudenstein, suggesting that gene tree parsimony outperforms uninode coding, at least on these data.

1. Introduction

Two very different methods of using paralogous genes for phylogenetic inference have been proposed: gene tree parsimony (Slowinski and Page, 1999) and uninode coding (Simmons et al., 2000). The first step in gene tree parsimony is to identify where gene duplications and gene losses have occurred on a gene family phylogeny, or set of gene phylogenies. This can only be done with some knowledge of the phylogenetic relationship of those taxa the genes are found in, or species tree. Gene tree parsimony (named by Slowinski et al., 1997) methods then propose that, if the species tree is unknown or uncertain, we should prefer the species tree that minimises the number of gene duplications, or duplications and losses, across a set of gene trees. This species tree is the most parsimonious tree in that it minimises the number of ad hoc assumptions of paralogy between sequences.

Uninode coding (Simmons et al., 2000) takes a rather different view—it circumvents the problem of including duplicate genes in a total-evidence analysis matrix by identifying clear orthology groups and coding them as separate columns in the matrix. This would leave a great deal of missing data, so a hypothetical ancestral sequence of all the duplicated copies—representing the sequence of the gene at the moment of duplications, reconstructed under maximum parsimony—is inserted into the matrix. Finally, a binary character representing the duplication event itself is added into the matrix. Fig. 1 shows the uninode coding scheme. Simmons and Freudenstein (2002) present a list of further rules for the implementation of uninode coding.

Here we discuss the 10 criticisms of gene tree parsimony suggested by Simmons and Freudenstein (2002), and suggest that many of them have little force, also apply to the uninode coding method, or hail from a particular perspective on phylogenetic methodology. Of the few remaining criticisms, most are reflections of a wider debate, that between consensus and "total-evidence" methods for using multiple sources of evidence in phylogenetic reconstruction. We revisit this debate briefly, to suggest that these criticisms are not decisive in deciding between gene tree parsimony and uninode coding methods. A further subset of the criticisms are aimed at only a particular implementation of the gene

* Corresponding author. Present address: Department of Zoology, Natural History Museum, London SW7 5BD, UK. Fax: +44-141-330-2792.

E-mail address: j.cotton@udcf.gla.ac.uk (J.A. Cotton).

1055-7903/03/$ - see front matter © 2003 Published by Elsevier Science (USA).
doi:10.1016/S1055-7903(03)00109-X

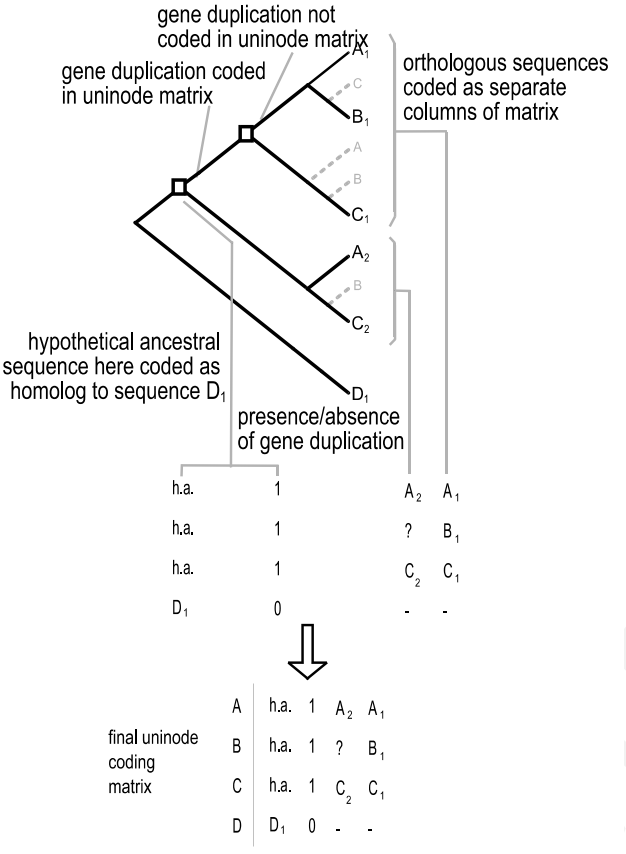

Fig. 1. The uninode coding scheme for a gene tree for genes from species A–D (A_1, etc. are gene copies). If we assume a species tree (((A, (B, C)), D), reconciled tree methods would recognise two gene duplications and four gene losses. Only one of these duplications is recognised by uninode coding, as sequences are only present for one copy of the more recent duplication in any species. The uninode coding matrix for this gene tree is shown below—A_1, etc. represent the aligned sequences of the respective genes, ?, is missing data; -, is inapplicable data and h.a. represents the hypothetical ancestral sequence of A_1, A_2, B_1, C_1, and C_2.

tree parsimony method—that of the program GeneTree (Page, 1998), and overlook a number of recent algorithmic developments.

Simmons and Fruedenstein's 10 criticisms of gene tree parsimony are listed in Table 1. They appear in this table in the order they appear in the original manuscript—the titles given here are not from the original, but

Table 1
Simmons and Freudenstein's criticisms of gene tree parsimony

1.	Problematic selection among variants
2.	Non-independence of duplication events
3.	Incomplete sampling of gene copies
4.	Weighting of nucleotide/amino-acid characters
5.	Partitioned data
6.	Slow searching in GeneTree
7.	Requires resolved gene trees
8.	Assumes correct gene trees
9.	Conflict between gene trees given equal weight
10.	No branch support values

(hopefully faithfully) paraphrase the main point made by Simmons and Freudenstein. These criticisms are valuable in drawing attention to certain features of the gene tree parsimony method, and in highlighting the value of certain new developments in gene tree parsimony techniques, but we disagree with Simmons and Freudenstein's conclusion that these criticisms imply that "uninode coding be used instead of gene tree parsimony for phylogenetic inference from paralogous genes."

2. Different algorithms and new techniques

GeneTree is a single implementation of reconciled tree methods to infer phylogeny from gene families, and it is important no to confuse the limitations of the GeneTree program with the conceptual limitations of the reconciled tree methods themselves. This is particularly clear in the case of criticism #6—about the slowness of GeneTree's heuristic searches—GeneTree currently implements the algorithm of Eulenstein (1997), a development of the original mapping algorithm (Page, 1994), and then uses heuristic searches through tree space to find the optimal species tree. More efficient search strategies are available—Hallett and Lagergren (2000) present a fixed-parameter tractable algorithm for finding the optimal species tree for a set of gene trees under the duplication-and-loss criterion without the need for this heuristic search. This has been implemented in the interpreted language Darwin (Gonnet et al., 2000; Hallett and Lagergren, 2000), and is likely to be implemented in a future version of GeneTree, and certainly demonstrates that slowness is not a property of the gene tree parsimony method itself. Simmons and Freudenstein's use of Page and Charleston's (1997) search strategy to claim that "GeneTree is too slow to thoroughly search the tree space" is particularly misleading given that Hallett and Lagergren (2000) demonstrate that Page and Charleston do indeed identify species trees with the globally best cost for Guigo et al.'s data (Guigo et al., 1996).

The same algorithms also answer criticism #7—both Eulenstein, and Hallett and Lagergren suggest that their algorithms can be easily extended to cases where gene trees contain polytomies. One easy way to include polytomies, which we have implemented in a version of GeneTree, is to allow a set of gene trees to be input, and to minimise duplications or duplications and losses across this set of trees. If a polytomy is considered to be a "soft" polytomy (Maddison, 1989), it represents uncertainty between a number of different possible bifurcating trees, differing in the order of branching above this node. A set of trees could thus include all the possible dichotomous resolutions of any polytomies in the input gene tree, but equally could be a set of most-

parsimonious trees from a parsimony analysis, or some similar representation of the uncertainty in the gene tree estimate.

Simmons and Freudenstein state that branch support values are not provided for reconciled trees (criticism #10), but a number of ways to present such measures have been proposed. One way to incorporate branch support values is to use a bootstrap profile of gene trees as an input to the gene tree parsimony step, generating a set of species trees. The proportion of these trees containing a clade of interest would then be a direct analogue of standard bootstrap proportions, as suggested by Page and Cotton (2000), and recently used in Cotton and Page (2002). In fact, this method also helps answer criticism #8—using a set of bootstrap trees effectively provides a confidence interval around the best estimate of each gene tree, relaxing the requirement for correct, fully resolved gene family trees. This should also improve inferences about the patterns of gene duplications and losses. In fact, we need not use a set of bootstrap trees—using a Bayesian credible set of trees might give a more statistically rigorous confidence interval (Huelsenbeck et al., 2000).

3. Selection among variants of gene tree parsimony

The choice between different analysis methods is not unusual in scientific methods, and is hardly a substantive criticism—in parsimony methods generally (including analysis of uninode coded data) we must choose between different weighting schemes (e.g., weighting transitions higher than transversions) and we frequently have to make choices between methods of phylogenetic reconstruction. Beyond this, a number of methods of phylogenetic analysis are available when faced with a sequence alignment. Flexibility in analytical method only seems a problem under the view that there is only a single "true" method of phylogenetic inference, a philosophy not shared by all systematic biologists. We see the availability of a range of analytical tools is a positive thing, not a negative one.

In any case, the fact that Simmons and Freudenstein (2002) and Simmons et al. (2000) only suggest a single uninode coding method does not imply that other variants cannot be proposed. For example, Simmons et al. (2000) make no defence as to why the binary gene duplication characters need to be included in the matrix at all—uninode coding would still be logically consistent without these characters, or with these characters weighted twice, or three times or any number at all. This problem was recognised more than 20 years ago (Fitch, 1979)—there is no logical way to decide how to weight a duplication character relative to a nucleotide or amino-acid substitution. Uninode coding methods suffer from the same 'problem' of multiple variants as gene tree parsimony methods.

A final point is that it seems that duplication-and-loss and duplication-only scores will always give compatible results, but that the duplication-and-loss result will be better resolved. Using the duplication-only criterion is, in this case, merely more conservative, avoiding the risk of grouping some taxa together by sampling failure. This is a corrollary of conjecture 3 of Page and Charleston (1997, p. 63), which is still formally unproven. Even if this conjecture is shown to be formally false, there is certainly a close relationship between the different cost functions used in gene tree parsimony—both duplication-only and duplication-and-loss scores will be highly correlated with the deep coalescence cost (as Zhang, 2000, has shown for a slightly different cost to that implemented in GeneTree).

Duplication-and-loss results can be misleading in certain circumstances. If the sampling of genes is incomplete, the absence of a gene copy from the sequence database could be for two different reasons—because the gene copy does not exist in the species' genome or because it has not been sequenced. Duplication-and-loss costs risk conflating these two costs, and so supporting relationships on the basis of the uneven sampling of molecular biologists. In some studies, such as that of Martin and Burg (2002, p. 584), where sampling is known to be fairly complete, duplication-and-loss costs are appropriate. However, studies using only a small selection of sequences taken from the public sequence databases, and including taxa that are not fully sequenced (e.g., Cotton and Page, 2002), such as the data used here, are likely to produce biased results under this criterion.

4. Consensus methods vs combined analysis

The debate over whether to combine data from multiple different sources of evidence in a single data matrix for phylogenetic analysis has been on-going for over a decade (for reviews see de Queiroz et al., 1995; Huelsenbeck and Bull, 1996). Three different opinions have been reflected in the literature—taxonomic congruence, which supports separate analysis and the use of consensus methods to investigate similarities between them (Miyamato and Fitch, 1995; Swofford, 1991), "total evidence" or combined analysis, which supports combining separate datasets before analysis (Barrett et al., 1991; Kluge, 1989) and an intermediate position, which advises combining data when statistical tests suggest they are compatible (Bull et al., 1993; Huelsenbeck and Bull, 1996). There has been a long debate between proponents of these methods for dealing with multiple data sources in systematics.

We believe that, in the context of this debate, a number of Simmons and Freudenstein's criticisims of gene tree parsimony merely reflect differences between these positions. These criticisms have thus been addressed in previous discussions, and are, in any case, not decisive criticisms of the gene tree parsimony method. Simmons and Freudenstein suggest that both reconciled trees and uninode coding are "total-evidence" or "simultaneous-analysis" approaches, in the sense of Kluge (1989). However, Kluge uses "total-evidence" to apply to methods that seek to find the hypothesis that maximises total "character congruence" rather than "taxonomic congruence"—by including all possible evidence in analysis of a single data matrix. Gene tree parsimony is not a total-evidence method in the sense of maximising congruence between all sequence characters in this way—as Page (2000, p. 99), explicitly states "It should be emphasized that the topology of this species tree depends entirely on the topology of the nine gene trees (and the constraint tree); no reference is made to the underlying sequence data."

In fact, gene tree parsimony methods have something in common with both consensus methods and total evidence approaches. Gene tree parsimony is a total-evidence method in the sense that it seeks the best explanation for all the available data, but the data it uses are the phylogenies for the gene families rather than the sequence alignments themselves—effectively applying total evidence under the parsimony criterion to higher-level characters, namely gene trees. On the other hand, if we use the terminology of de Queiroz et al. (1995), gene tree parsimony is clearly a consensus method, in that 'characters in two (or more) data sets are not allowed to interact directly with one another in a single analysis, but instead interact only through the trees derived from them.' Gene tree parsimony is not a traditional consensus method, however, in that rather than seeking to summarise the a set of source trees, it seeks to find a tree best representing the evolution of a set of gene trees in a biologically meaningful way.

Traditional consensus methods are likely to be a poor choice for studying historically associated lineages such as genes and their species, as discussed by Page (1996), and acknowledged by authors on both sides of the debate (e.g., Cognato and Vogler, 2001). Consensus methods seek to represent incongruence between source trees, whereas reconciled tree methods attempt to resolve this incongruence by explaining it in terms of evolutionary events such as gene duplication and gene loss—effectively taking this incongruence 'at face value' as needing a biological explanation. The uninode coding method simply makes the minimum variation to simple combined analysis needed to incorporate multiple gene copies—any incongruence is treated as statistical error, to be submerged by the weight of combined data from multiple loci. By relaxing the requirement of gene tree topologies to be exactly correct (e.g., by using a bootstrap profile or Bayesian credible set of trees, as discussed above), we effectively allow gene tree parsimony methods to only find evolutionary explanations for significant incongruence. In fact, the difference between combined analysis and methods relying only on the reconstructed phylogeny (such as consensus methods and gene tree parsimony) reflects a statistical trade-off between reducing bias (by combining all data) and correctly estimating variance in the estimate of phylogeny (by partitioning data)—a trade-off widely accepted in the statistical literature (Holmes, 2003). In the sense that one uses the sequence data directly and the other considers tree from the separate data partitions, gene tree parsimony and uninode coding represent alternative sides of the debate over combined analysis vs. consensus methods. Simmons and Freudenstein's criticisms #4 and #5 reflect this debate—a debate that is still active (Levausser and Lapointe, 2001) and can hardly be considered a decisive criticism of gene tree parsimony.

In fact, for practical purposes, the debate over consensus methods vs. total evidence is probably not of crucial importance. Simmons and Freudenstein, in common with other advocates of total evidence methods, suggest that total evidence methods may be more successful in that they allow "hidden support" for certain nodes to emerge from the combined matrix (Gatesy et al., 1999; Nixon and Carpenter, 1996). Hidden support refers to support across data partitions for relationships that are not evident in the most-parsimonious tree for the partitions analysed separately. While a number of studies have identified hidden support, they do not demonstrate that the hidden support is truly hidden in the sense of not being evident in a number of the trees from a bootstrap profile, or being excluded from the credible interval of trees in a Bayesian framework. Relaxing the dependence of gene tree parsimony on a single estimate of the gene trees would be expected to identify most significant hidden support.

5. Non-independence of gene duplications

The potential non-independence of gene duplications on trees has been recognized by a number of authors—some of the earliest theoretical work presented a method for identifying larger-scale genome duplications on a tree (Guigo et al., 1996). Most authors have followed Guigo et al. in considering independence of gene duplications as a valid simplifying hypothesis which can later be tested by comparing the distributions of duplications under this assumption and under the assumptions that the individual duplications are clustered into the minimum number of larger-scale episodes (Page and Cotton, 2002). This parallels a common assumption of phylogenetic methods, where nucleotide substitutions

are considered independent because modelling dependencies between substitutions at different sites would be intractable except in simple cases where this dependency is clear, such as in the stems of RNA molecules (Jow et al., 2002). In particular, uninode coding also makes the same assumption—the "gene duplication characters" are duplications coded as independent characters, as (Simmons and Freudenstein, 2002) admit. Another, pragmatic reason that we do not attempt to find the species tree minimising the number of gene duplication episodes is that this is demonstrably NP-hard (Fellows et al., 1998).

6. Hidden paralogy

The main criticism we have of the uninode coding method is that it ignores the possibility of hidden paralogy—paralogy that is not obvious due to the presence of both gene copies existing in extant genomes (Fig. 2).

How frequent hidden paralogy will be depends upon rates of gene duplication and loss—as gene families evolve under a birth-and-death process (Nei et al., 2000). This process may be even more common, as duplicate genes are complementary, so one copy will rapidly go extinct if a mutation renders one of the copies non-functional—there is no selective pressure to retain both copies of the gene (Lynch and Conery, 2000). If a speciation event occurs during this process, then different paralogous copies could easily go extinct in each lineage—in the simple case in which the two lineages have an equal chance of survival this will occur 50% of the time. Where gene duplications are frequent, and gene silencing and subsequent loss relatively slow, hidden paralogy will be very common. Apparent hidden paralogy could also pose a problem for the uninode coding method—even where multiple gene copies from a duplication exist in the genomes of some species, there will be situations in which no species shows both gene copies because of the incomplete sampling of genomes.

Uninode coding also ignores the possibility that the gene duplications present on the most-parsimonious gene tree (in stage 1) are incorrect—these duplications will be incorporated into the uninode coding matrix. This matrix pseudo-replicates some of the data by incorporating hypothetical ancestral sequences many times into the matrix, which are entirely dependent on the sequences they are calculated from. This pseudo-replication has two effects—it makes it very unlikely that the phylogenetic groups supported by gene duplications on the original parsimony trees will not be present in the final parsimony trees, particularly for duplications ancestral to large numbers of species, and it makes bootstrap values for these nodes very difficult to interpret.

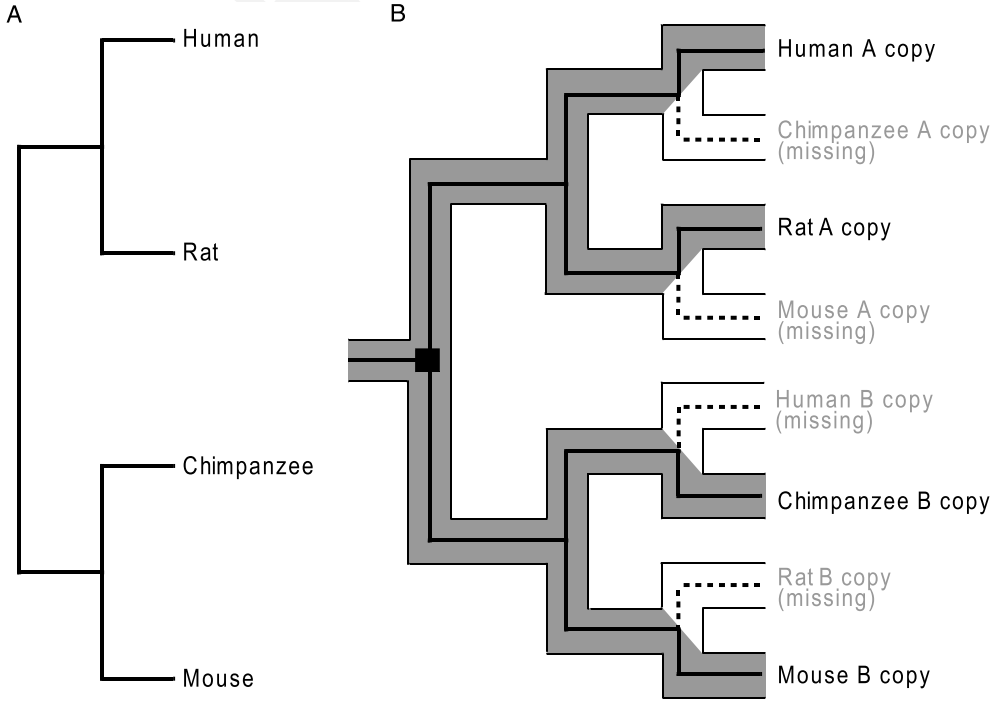

Fig. 2. Hidden paralogy. The gene tree: (A) shows no duplicated genes that would be coded as such in the uninode coding method, but any reasonably assumption about the relationships between these four species would suggest that the true pattern of evolution in this gene family is as seen in the reconciled tree (B). (B) Shows a duplication at the base of the gene tree, followed by four losses (or failure to sample four of the genes), suggesting that the rat and human genes are orthologues, and are paralogous to the mouse and chimpanzee orthologues.

7. An empirical example

Simmons and Freudenstein present a re-analysis of data from Page (2000) using uninode coding, and find a substantially different result. We use these data again to demonstrate some of the more recently developed methods discussed above. Page originally used the neighbour-joining method to generate gene family trees for the 9 genes used, while Simmons and Freudenstein use parsimony trees to infer the locations of gene duplications in stage 1 of the uninode coding process. To investigate how much the differences between the results of these two studies was due to the use of parsimony rather than neighbour-joining, and to demonstrate how multiple most-parsimious trees can be used in gene tree parsimony, we also use parsimony gene trees here.

7.1. Methods

The gene trees for this analysis were generated from the ClustalX alignments used by Page (2000) and available from http://taxonomy.zoology.gla.ac.uk/rod/data/vertebrates/. These alignments were converted to the NEXUS format and then analysed using PAUP 4b10 (Swofford, 1998) under the parsimony criterion, with 50 random addition-sequence replicates and TBR branch-swapping to completion, keeping multiple trees. All most-parsimonious trees found were incorporated into a GeneTree format NEXUS file and analysed using a specially written version of the GeneTree program, which treats multiple gene trees as equally parsimonious gene trees, searching for the species tree that minimises the cost across the set of trees, by, for each iteration of branch-swapping during the heuristic search, reconciling the species tree with each gene tree in turn, and recording as the correct cost the minimum cost across all the trees for that gene family. As discussed by Page, constrained searches are needed for this data to address the limited taxonomic coverage of most gene families, and the same constraints as used by Page (and Simmons and Freudenstein) were used in all analyses shown here.

Because of the complexity of searching across the profiles of most-parsimonious trees for each gene family, for every postulated species tree during the heuristic search, the searches for these data were very slow. The inclusion of multiple MPTs for each gene family also greatly increased the numbers of equal-cost trees found, so a two-step search strategy was employed. For the first step, a large number of starting tree replicates were used, but branch-swapping was performed on only a single tree during the search, thus preventing the searches becoming trapped on plateaus of equally parsimonious trees. The shortest trees from these searches were then swapped on to sample more widely from the island of trees identified during the first stage. This two-stage procedure gives us a reasonable chance of locating the shortest trees, and ensures that we sample adequately from the island (or islands) of trees found.

For both duplication-only and duplication-and-loss criteria, 100 searches starting from random addition-sequence replicate trees were performed. Under the duplication-only criterion, seven of these searches found the lowest cost of 92 duplications, finding seven different species trees. Several additional searches, holding multiple trees, were also run under this criterion, which were not run to completion but found over 15,000 trees of this cost without finding any lower-cost solutions. Under the duplication-and-loss criterion, 21 searches found trees with the lowest cost, of 383 duplications and losses. All seven of the duplication-only optimal trees found in these searches, and a randomly chosen sample of 10 duplication-and-loss optimal trees were used as starting points for searches swapping on multiple trees. Each of these searches was run until at least 1000 trees had been found, and in many cases were left for much longer, with none of the searches finding shorter trees than were identified in the first stage searches. The Adam's and strict consensus for each of the seven sets of duplication-and-loss results and each of the 10 duplication-only sets of trees were identical, or differed only in the resolution of a single node within the reptiles (for the duplication-and-loss data), confirming that each search had successfully sampled from across the island of minimal trees. The strict consensus trees of the 7000 duplication-and-loss trees and the 10,000 duplication-only trees found in these searches are shown in Fig. 3.

As pointed out by Simmons and Freudenstein, the standard gene tree parsimony analyses described above use only a single fully resolved phylogeny for each gene family, and so can take no account of weaknesses in the gene family trees. For example, many gene families may be unable to resolve particular relationships or show only limited support for a particular resolution. To incorporate this information, we have adopted a gene tree bootstrapping protocol (Cotton and Page, 2002; Page and Cotton, 2000). A set of 100 bootstrap trees for each gene family in the dataset, using the fast heuristic bootstrapping method of Paup 4b10. The species tree minimising the number of gene duplications were then found for successive trees from the bootstrap profile of each gene family, producing 100 sets of species trees. A single, complete search from a single random starting tree, keeping multiple solutions, was performed for each replicate, with multiple equal solutions down-weighted appropriately in the final calculation of support values. Support values analogous to standard bootstrap values could then be calculated as the number of times nodes appeared in these 100 species tree.

7.2. Results and discussion

The full phylogenetic results of the analyses described here are shown in Figs. 3 and 4. A summary of these results, showing relationships between the major vertebrate groups and comparing these results with the results of Page (2000) and Simmons and Freudenstein (2002) is shown in Fig. 5. We restrict this discussion to relationships between major vertebrate groups, all of which are unconstrained in the analyses discussed, and for which there is a clear idea of what the expected relationships are.

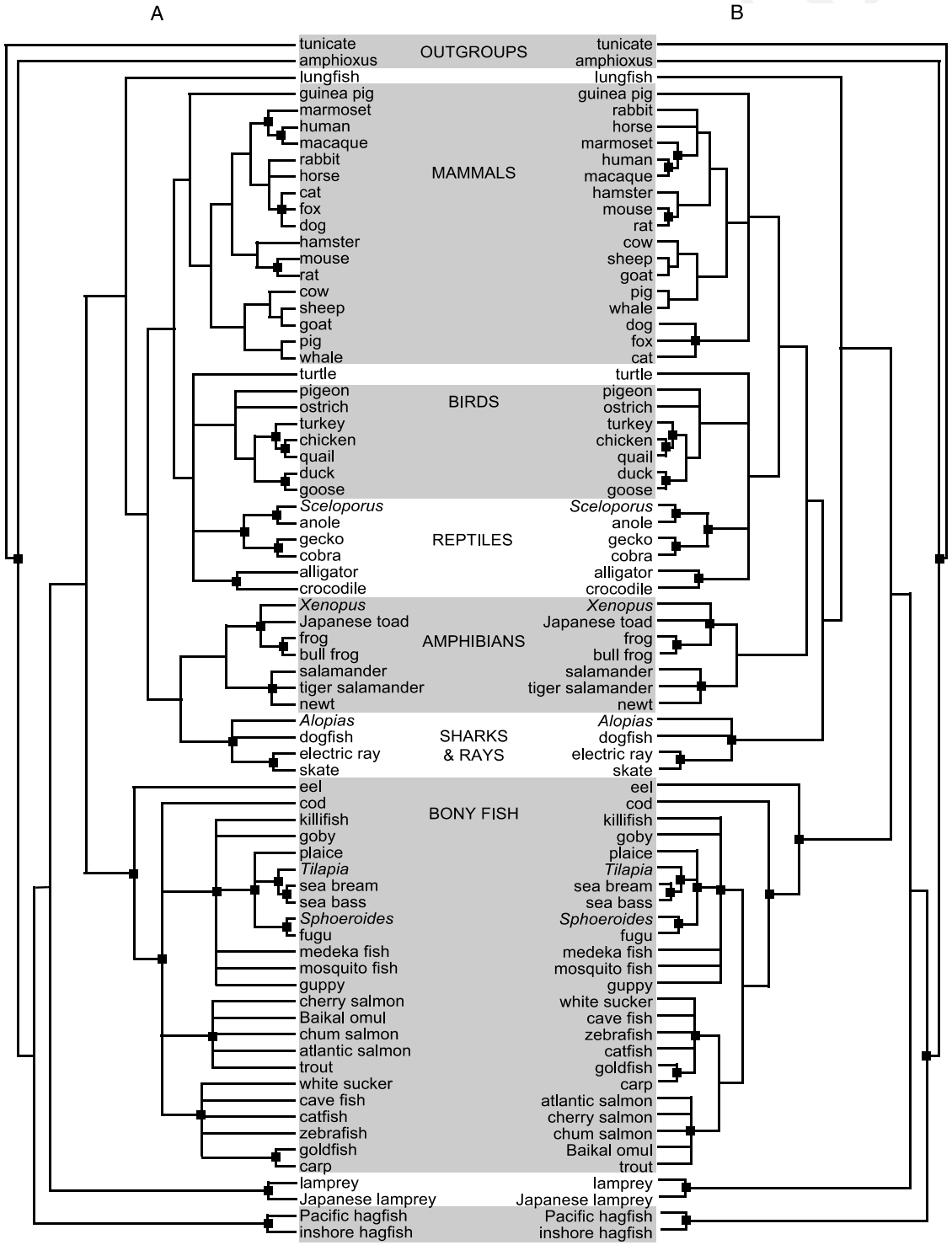

Fig. 3. Results of a gene tree parsimony search finding the species tree minimising the number of: (A) duplications and losses and (B) gene duplications across the most parsimonious trees from the gene families of Page (2000). Nodes marked with a square were constrained during the search.

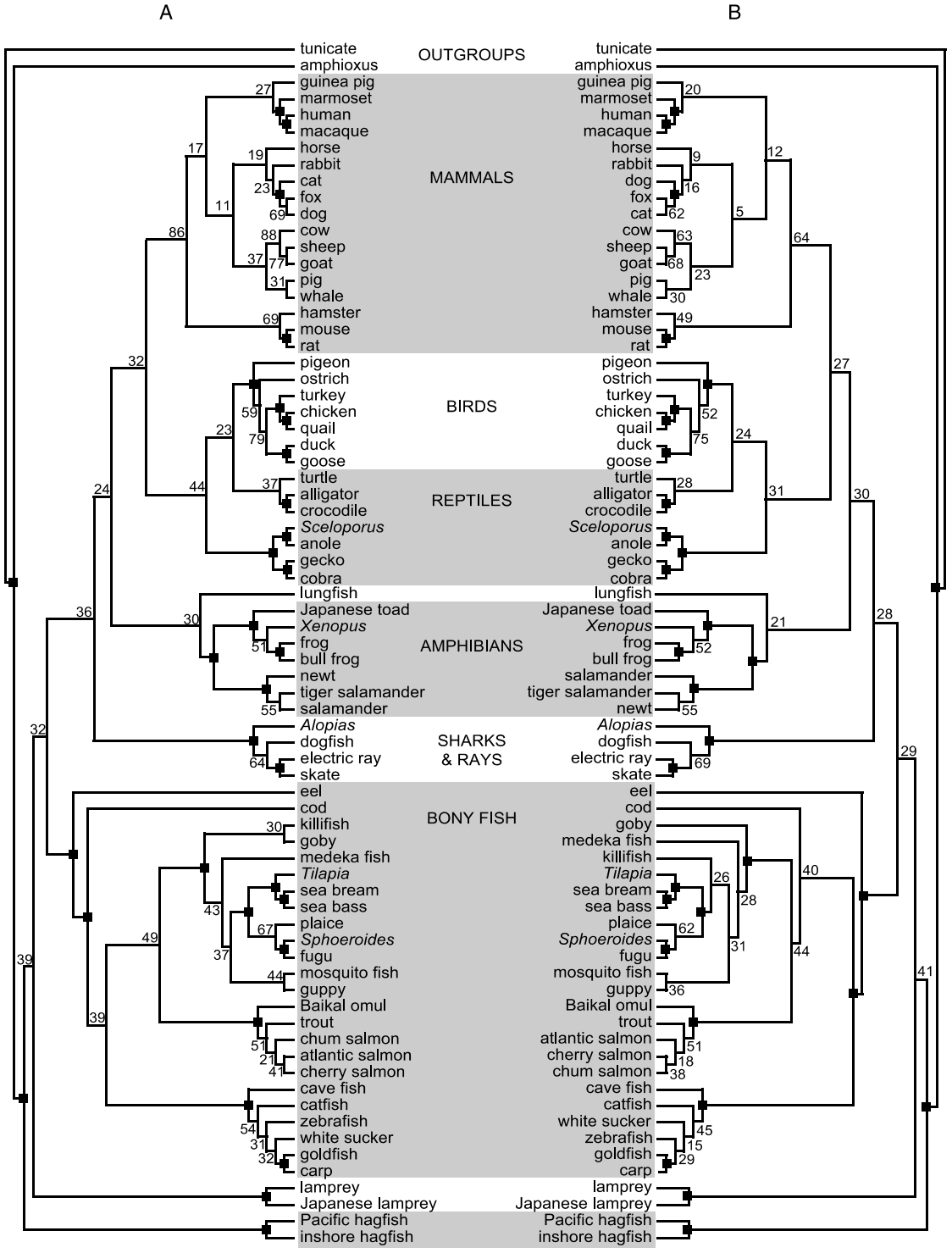

Fig. 4. Results of a gene tree parsimony bootstrap analysis. Show are majority-rule consensus trees (including compatible groups present in less than 50% of the trees) of 100 species tree obtained by minimising the number of: (A) gene duplications and losses and (B) duplications only, for each of 100 bootstrap trees for the gene families of Page (2000). Figures at nodes represent the number of times this node appeared in the 100 resulting species trees. Nodes marked with a square were constrained during the search.

We can see that all four analyses shown in Fig. 5 support different relationships among the higher vertebrate taxa, suggesting (as our bootstrap values reflect) that these genes do not give very strong support for any picture of vertebrate relationships. As Fig. 5 shows, none of the analyses correctly reproduces the traditional picture of vertebrate phylogeny, a view supported by a great weight of morphological work (e.g., Bishop and

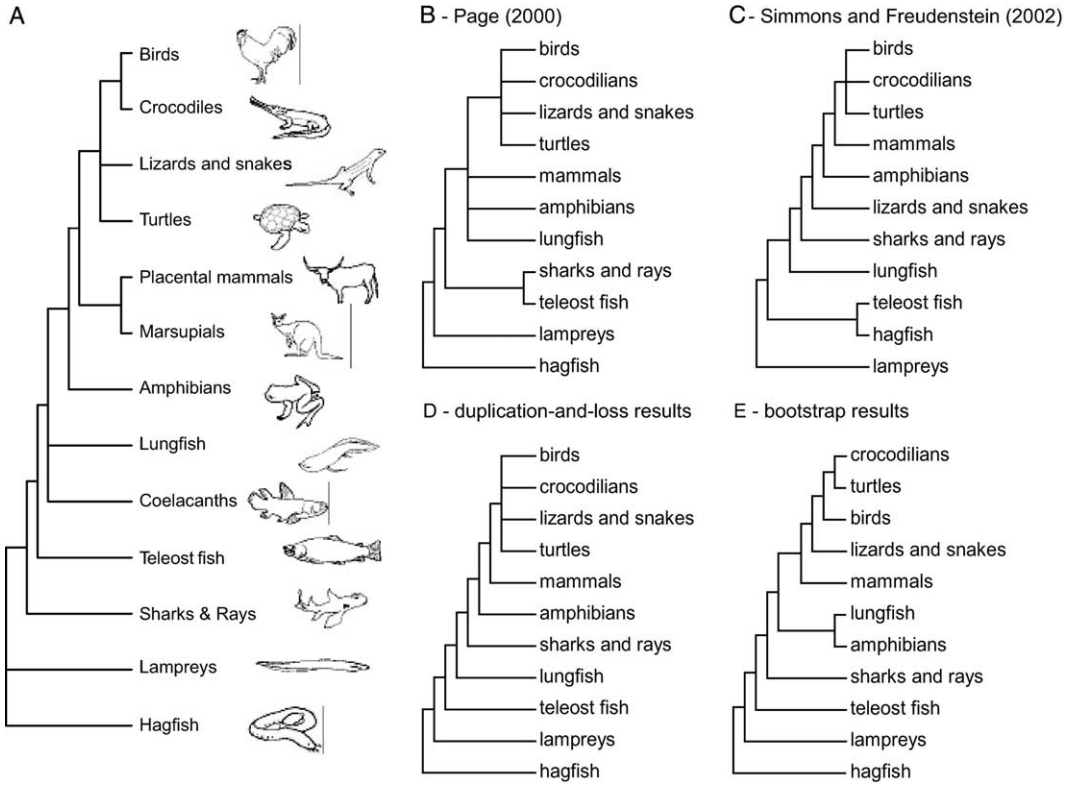

Fig. 5. Summary of the results of (B) Page (2000), (C) Simmons and Freudenstein (2002) and this study: (D) shows the strict consensus of duplication-only optimal trees, (E) the majority-rule consensus of the bootstrap replicates. Part (A) shows a traditional picture of vertebrate phylogeny based on morphological and paleontological evidence (Bishop and Friday, 1988) (part A is from Cotton and Page, 2002).

Friday, 1988; Løvtrup, 1977) and by gene tree parsimony analysis of a much larger data set (Cotton and Page, 2002). Furthermore, none of the results are wholly congruent with phylogenies based on whole mitochondrial genome data (Rasmussen and Arnason, 1999; Zardoya and Meyer, 2001).

All of the four results share some weaknesses—all misplace the sharks and rays, placing them in too derived a position in the vertebrate tree. The trees also all fail to resolve relationships within the reptiles, or present a somewhat unusual phylogeny within this group. While most workers would agree that the turtles are the most basal of the extant reptiles, with lizards and snakes (the lepidosaurs) forming a sister-group to an archosaur clade of crocodiles and birds, relationships within the group have become somewhat uncertain in the light of molecular evidence, which tends to place turtles as relatives of the archosaurs (Hedges and Poling, 1999; Rieppel, 2000), as suggested by Simmons and Freudenstein's result—the placement of turtles within the archosauria as shown in Fig. 5D is not supported by other evidence.

Simmons and Freudenstein's result shows some problems not present in any of the gene tree parsimony results. Their results fail to correctly unite the lizards and snakes with the other archosaurs, and fail to place the hagfish as a basal vertebrate lineage. There is no doubt that lizards and snakes form part of a monophyletic radiation of diapsid reptiles, although there has been some debate about the exact relationships between the different extant lineages within this radiation, as discussed above. Similarly, there has been debate about the exact relationships between hagfish, lampreys and gnathostomes (Delarbre et al., 2002; Janvier, 1996), but the only hypotheses supported by recent work are that lampreys and hagfish form a monophyletic cyclostome group, or that hagfish are the most basal vertebrates, with lampreys a sister-group to the gnathostomes. In conclusion, the results of this study are a better estimate of correct vertebrate phylogeny than those of Simmons and Freudenstein. It is striking that Simmons and Freudenstein find high bootstrap support for some clearly erroneous relationships, such as 87% support for a monophyletic clade of amphibians and tetrapods, but excluding the lizards and snakes, and 90% support uniting the hagfish and teleost fish.

8. Conclusion

Differences between uninode coding and gene tree parsimony are largely ones of perspective—uninode

coding is a combined analysis method, modified to allow the use of multiple genes for each taxon. The relative effectiveness of gene tree parsimony methods and uninode coding will partly depend on the extent of hidden paralogy—the extent to which the signal from different clades coded in the uninode matrix conflict—and to what extent noise makes the individual gene trees inaccurate. This is an empirical issue, and not one decided by Simmons and Freudenstein's criticisms of gene tree parsimony methods. For the data analysed here, gene tree parsimony gives a more reasonably vertebrate phylogeny, suggesting that for these data it is important to correctly identify hidden paralogy. Finally, gene tree parsimony methods can identify gene duplications despite widespread gene loss, and so are valuable tools in the study of the pattern and process of gene duplication itself (Page and Cotton, 2002).

References

Barrett, M., Donoghue, M.J., Sober, E., 1991. Against consensus. Syst. Zool. 40, 486–493.

Bishop, M.J., Friday, A.E., 1988. Estimating the interrelationships of tetrapod groups on the basis of molecular sequence data. In: Benton, M.J. (Ed.), The Phylogeny and Classification of the Tetrapods, vol. 1, 2 vols. Clarendon Press, Oxford.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L., Waddell, P.J., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42, 384–497.

Cognato, A.I., Vogler, A.P., 2001. Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). Syst. Biol. 50 (6), 758–781.

Cotton, J.A., Page, R.D.M., 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R Soc. Lond. B 269 (1500), 1555–1561.

de Queiroz, A., Donoghue, M.J., Kim, J., 1995. Seperate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26, 657–681.

Delarbre, C., Gallut, C., Barriel, V., Janvier, P., Gachelin, G., 2002. Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. Mol. Phylogenet. Evol. 22 (2), 184–192.

Eulenstein, O., 1997. "A linear time algorithm for tree mapping." Arbeitspapiere der GMD No. 1046.

Fellows, M., Hallett, M., Stege, U., 1998. On the multiple gene duplication problem. In: Kyung-Yongn, C., Ibarram, O.H. (Eds.), Proceedings of the Ninth International Symposium on Algorithms and Computation.

Fitch, W.M., 1979. Cautionary remarks on using gene expression events in parsimony procedures. Syst. Zool. 28, 375–379.

Gatesy, J., O'Grady, P., Baker, R.H., 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 15, 271–313.

Gonnet, G.H., Hallett, M.T., Korostensky, C., Bernardin, L., 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. Bioinformatics 16, 101–103.

Guigo, R., Muchnik, I., Smith, T.F., 1996. Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 6 (2), 189–213.

Hallett, M.T., Lagergren, J., 2000. New algorithms for the duplication-loss model. RECOMB '00, the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan.

Hedges, S.B., Poling, L.L., 1999. A molecular phylogeny of reptiles. Science 283 (5404), 998–1001.

Holmes, S., 2003. Statistics for phylogenetic trees. Theoretical Population Biology 63, 17–32.

Huelsenbeck, J.P., Bull, J.J., 1996. A likelihood ratio test for detection of conflicting phylogenetic signal. Syst. Biol. 45, 92–98.

Huelsenbeck, J.P., Rannala, B., Masly, J.P., 2000. Accomodating phylogenetic uncertainty in evolutionary studies. Science 288, 2349–2350.

Janvier, P., 1996. The dawn of the vertebrates: characters versus common ascent in the rise of current vertebrate phylogenies. Palaeontology 39 (Pt 2), 259–287.

Jow, H., Hudelot, C., Rattray, M., Higgs, P.G., 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. 19 (9), 1591–1601.

Kluge, A., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst. Zool. 37, 315–328.

Levausser, C., Lapointe, F.-J., 2001. War and peace in phylogenetics: a rejoinder on total evidence and consensus. Syst. Biol. 50 (6), 881–892.

Løvtrup, 1977. The Phylogeny of the Vertebrata. Wiley, New York.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290 (5494), 1151–1155.

Maddison, W.P., 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5 (365–377).

Martin, A.P., Burg, T.M., 2002. Perils of paralogy: ssing HSP70 genes for inferring organismal phylogenies. Syst. Biol. 51 (4), 570–587.

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

Nei, M., Rogozin, I.B., Piontkivska, H., 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc. Natl. Acad. Sci. USA 97 (20), 10866–10871.

Nixon, K., Carpenter, J., 1996. On simultaneous analysis. Cladistics 12, 221–241.

Page, R.D.M., 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. Syst. Biol. 43 (1), 58–77.

Page, R.D.M., 1996. On consensus, confidence, and "total evidence". Cladistics 12 (1), 83–92.

Page, R.D.M., 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14 (9), 819–820.

Page, R.D.M., 2000. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol. Phylogenet. Evol. 14 (1), 89–106.

Page, R.D.M., Charleston, M.A., 1997. Reconciled trees and incongruent gene and species trees. In: Mirkin, B., McMorris, F.R., Roberts, F.S., Rzhetsky, A. (Eds.), Mathematical Hierarchies in Biology, vol. 37. American Mathematical Society, Providence, RI, pp. 57–71.

Page, R.D.M., Cotton, J.A., 2000. GeneTree: a tool for exploring gene family evolution. In: Sankoff, D., Nadeau, J.H. (Eds.), Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families. Kluwer Academic Publishers, Dordrecht, pp. 525–536.

Page, R.D.M., Cotton, J.A., 2002. Vertebrate phylogenomics: reconciled trees and gene duplications. Pac. Symp. Biocomput., 536–547.

Rasmussen, A.S., Arnason, U., 1999. Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fishes within the tree of bony fishes. J. Mol. Evol. 48 (1), 118–123.

Rieppel, O., 2000. Turtles as diapsid reptiles. Zool. Scr. 29 (3), 199–212.

Simmons, M.P., Bailey, C.D., Nixon, K.C., 2000. Phylogeny reconstruction using duplicate genes. Mol. Biol. Evol. 17 (4), 469–473.

Simmons, M.P., Freudenstein, J.V., 2002. Uninode coding vs gene tree parsimony for phylogenetic reconstruction using duplicate genes. Mol. Phylogenet. Evol. 23 (481–498).

Slowinski, J.B., Knight, A., Rooney, A.P., 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 (3), 349–362.

Slowinski, J.B., Page, R.D.M., 1999. How should phylogenies be inferred from sequence data? Syst. Biol. 48 (4), 814–825.

Swofford, D.L., 1991. When are phylogenetic estimates from molecular and morphological data incongruent? In: Miyamato, M.M., Cracraft, J. (Eds.), Phylogenetic Analysis of DNA sequences. Oxford University Press, Oxford, pp. 295–333.

Swofford, D.L., 1998. PAUP*—Phylogenetic Analysis Using Parsimony (* and Other Methods). Sinauer Associates, Sunderland, MA.

Zardoya, R., Meyer, A., 2001. Vertebrate phylogeny: limits of inference of mitochondrial genome and nuclear rDNA sequence data due to an adverse phylogenetic signal/noise ratio. In: Ahlberg, P.E. (Ed.), Major Events in Early Vertebrate Evolution. Taylor and Francis, London, pp. 135–156.

Zhang, L., 2000. Inferring a species tree from gene trees under the deep coalescence cost. Poster, RECOMB2000, Tokyo, Japan.