GENETREE: A TOOL FOR EXPLORING GENE FAMILY EVOLUTION

Roderic D. M. Page

James A. Cotton

Molecular biologists interested in the evolution of gene families and molecular systematists interested in the evolution of whole organisms are both concerned with the relationship between gene phylogenies and organism phylogenies. We present reconciled trees as a tool for exploring this relationship. In discussing recent developments, we focus on techniques which enable researchers to take account of uncertainty in the underlying gene phylogenies and to locate gene duplications and episodes of gene duplication on the species tree. Implementation of these methods should allow rapid, automated analysis of large sets of gene families and even of whole genomes, producing well supported organism phylogenies and allowing us to quantitatively investigate patterns of gene family evolution.

1 Introduction

Evolutionary trees for gene sequences are studied from two complementary, but distinct, perspectives. Molecular biologists seek to understand the evolution of the structure and function of a particular gene, and discover relationships among families of genes. Molecular systematists use gene trees to recover organismal phylogeny. Central to both perspectives is the relationship between gene and organismal phylogeny.

The key assumption that motivates molecular systematics is that evolutionary trees for genes also contain information about the evolutionary relationships of organisms. Indeed, it is often assumed that gene trees are the same as species trees – hence one can obtain a species tree simply by sequencing the same gene in a range of species, and replacing the names of the genes with the names of the corresponding species. However, two observations contradict this assumption: (1) species may contain more than one copy of the same gene, and (2) different gene trees may imply different species trees. If two or more copies of a gene are sequenced (for example, haemoglobin α and β from *Homo sapiens*) then replacing the genes by the corresponding species will result in the same species occurring more than once in the tree. In this case there is no longer a one-to-one correspondence between the gene and species trees, raising the problem of how to extract

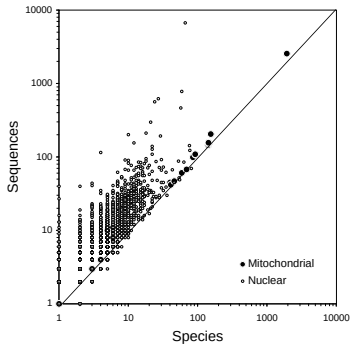

Figure 1: Number of sequences plotted against number of species for vertebrate gene families in release 29 (March 17, 1998) of the HOVERGEN (Duret et al., 1994) data base. Note that usually each species has a single mitochondrial sequence for a given gene (hence, the mitochondrial genes fall along the 1:1 line), whereas most nuclear genes are present in multiple copies. Due to redundancy in species names (for example, "human" and "*Homo sapiens*" being used to describe the source of different genes in the same family), some gene families appear to have fewer sequences than species. From Slowinksi and Page (1999, fig. 1).

the latter from the former. If different gene trees support different species trees (i.e. the gene trees are incongruent) then this raises the question of how to choose among these alternative species trees.

For molecular biologists, the relationship between gene and organismal phylogeny can be crucial in identifying orthologous genes. If only single copies of a gene have been sequenced in a range of taxa, it may not be obvious from the gene tree alone whether the genes are orthologous or paralogous. Comparison of gene and species trees can identify unrecognised instances of paralogy among genes. Once the history of gene duplication and loss events is determined for a set of genes, broader evolutionary questions can be asked, such as rates of gene duplication and loss, and the relative timing of duplications in different gene families.

The analysis of gene family phylogenies represents a considerable challenge for the study of genome evolution, especially when one considers how common gene duplication has clearly been in some taxa. Within vertebrates, paralogy is pervasive (Figure 1) and a similar picture is found in the Eubacteria and Archaea when data from HOBACGEN (Perrière et al., 2000) are examined.

Our goal here is to explore some issues in the analysis of gene family evolution using reconciled trees as implemented in GENETREE (Page, 1998). This software package is freely available for Windows 95/NT and MacOS operating systems from `http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html`. To illustrate specific points we use the L-lactate dehydrogenase (L-LDH) gene family (`http://www.expasy.ch/cgi-bin/nicezyme.pl?1.1.1.27`), which has often served as a model data set for developing ideas about reconciled trees

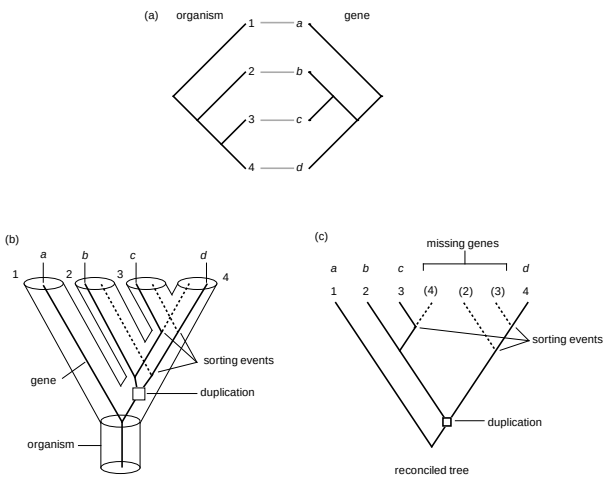

Figure 2: (a) Incongruent gene and species trees. This incongruence can be explained by hypothesizing a gene duplication (h) at the base of the gene tree (b). The presence of only a single gene (a-d) extant in each of the present-day species (1-4) requires postulating three gene losses. (c) The corresponding reconciled tree. After Page (2000).

(Page, 1994; Page and Charleston, 1997a; Martin, 1999a) and about gene family evolution more generally (Holmes, 1972; Li et al., 1983).

2 Reconciled trees

A reconciled tree is the simplest embedding of a gene tree within a species tree. The technique has its origins in Goodman et al. (1979), a study of haemoglobin gene evolution where there were significant discrepancies between gene and organismal phylogenies. Suppose we have a phylogeny for four species and a phylogeny for four genes sampled from those species, and that the gene and species trees – which we believe to be correct – disagree (Figure 2a).

The question is, how can the trees both be true, and yet be discordant? One approach is to embed the gene tree in the species tree (Figure 2b), which requires us to postulate a number of gene duplications and subsequent gene losses (in this instance one duplication and three losses). This embedding can also be represented using a reconciled tree (Figure 2c), which simply takes the embedded gene tree

and "unfolds" it so that it lies flat on the page. The reconciled tree depicts the complete history of the gene if there had been no gene losses. In this example, given the gene duplication we would expect species 2, 3, and 4 to each have two copies of the gene. It is the presence of only one copy of the gene in each of these species that leads us to infer three gene losses. An alternative explanation for these "losses" is that the other copy of the gene is present in these species, but as yet undetected. Given the unevenness of the sampling of different organisms (indicated by the preponderance of a few model organisms in the sequence data banks), this may often be the case. Indeed, the "losses" indicated by the reconciled tree can be viewed as predictions about the existence of undiscovered genes. In the example shown, further sequencing may uncover copy 1 in species 4, and copy 2 in species 2 and 3. The reconciled tree also shows that genes b and c are paralogous to gene d, which is not apparent from the gene phylogeny alone. This highlights the role organismal phylogeny can play in identifying homology relationships among genes. Direct evidence for paralogy is the presence of multiple genes in the same species (e.g., haemoglobin α and β in the same species), but many additional paralogous genes may be identified using reconciled trees.

3 Inferring species trees

One basic goal of analysing gene families is to shed light on the evolutionary relationships of the organisms from which those genes were obtained. Given one or more gene trees we can ask what species tree would accommodate those gene trees with the fewest number of duplications and losses (Page and Charleston, 1997a). The problem of finding the optimal species tree is NP-complete (Ma et al., 1998), so we must rely on heuristics for all but the smallest problems.

GeneTree implements a simple "hill-climbing" heuristic, where an initial species tree (either a random tree or one supplied by the user) is rearranged in search of a species tree with a better cost. Random trees provide a useful tool for exploring the tree landscape (Charleston, 1995), but searches that start from a random tree tend to be time consuming. Often it is substantially quicker to start from a species tree based on some other evidence, such as the currently accepted taxonomic classification. However, this may bias the results, especially if a poor rearrangement strategy is used. The importance of effective search strategies is emphasised by Page and Charleston (1997b), who used GeneTree to find substantially more parsimonious species trees than those found by Guigo et al. (1996) using the same set of eukaryote gene trees.

The extreme taxonomic bias of the sequence data bases towards a few model organisms (93% of vertebrate nucleotide sequences in GenBank come from humans, rats or mice) means it is almost certainly the case that not all genes will have been discovered (or, indeed, looked for) in all the taxa of interest. This can lead to cases where species will be grouped on the absence of genes, rather than on actual evidence of their relationship. This problem is avoided by using the number of duplications alone as the optimality criteria for selecting species

trees (Page and Charleston, 1997a), but this could lead to incorrect assumptions of orthology if actual gene loss events are common. Missing sequences also lead to a rapid increase in the number of species trees that are equally parsimonious explanations of the gene trees (Page, 2000). Where some taxa are sampled for only one or few gene families, this poor taxonomic overlap will result in some of these many parsimonious species trees being biologically absurd. One solution to this problem is to use constraint trees (Constantinescu and Sankoff, 1986) to enforce some species groupings that are considered incontrovertible (such as "mammals"), but clearly this requires us to accept some species relationships *a priori*.

New algorithms for finding optimal species trees are appearing. Stege (1999) presents a fixed-parameter tractable algorithm (Downey and Fellows, 1998) for finding the species tree that minimises the number of duplications for a set of gene trees, parameterised by the number of duplications needed. Hallett and Lagergren (2000) have developed an algorithm minimising both duplications and losses where the parameter is the "width" – the maximum number of gene lineages that coexist in a species at any one time. These algorithms can find the globally optimal species trees in cases where their parameter values are small – generally in fairly simple cases – and the latter has been used to show that the species trees found by Page and Charleston (1997b) were indeed the most parsimonious.

4 Uncertain gene trees

Gene trees inferred from sequence data are estimates of the true gene tree. So far we have assumed that the gene tree is obtained without error, but this will rarely be the case. Figure 3 shows a phylogeny for vertebrate L-LDH sequences. Some of the species relationships implied by this tree (figure 4b) seem anomalous: the two amphibians are not grouped together, the shark is basal to tetrapods and the relationships between mammalian orders are unconventional. This suggests that the gene tree may not be entirely accurate.

It may be that an alternative gene tree - less parsimonious or less likely than the optimal tree - is the actual gene tree, and the fit between gene and species tree could be used as an additional criterion for selecting among competing gene trees. Goodman et al. (1979) suggested such a strategy in their pioneering work on reconciled trees, in which they preferred less parsimonious haemoglobin gene trees which had better fit to accepted species trees than most parsimonious trees that required more duplications and losses. Their approach assigned each gene tree a total score based on the length of the tree in terms of number of nucleotide substitutions plus the number of gene duplications and losses, where each type of event had the same cost. This drew immediate criticism from Fitch (1979), who argued that there was no obvious way of determining the relative cost of a nucleotide substitution versus a gene duplication. Another approach would be to consider a set of gene trees for each gene, such as those comprising a "confidence interval" around the optimal gene tree (Sanderson, 1989; Page, 1996). The best estimate of the gene tree would be that tree within the confidence interval that had

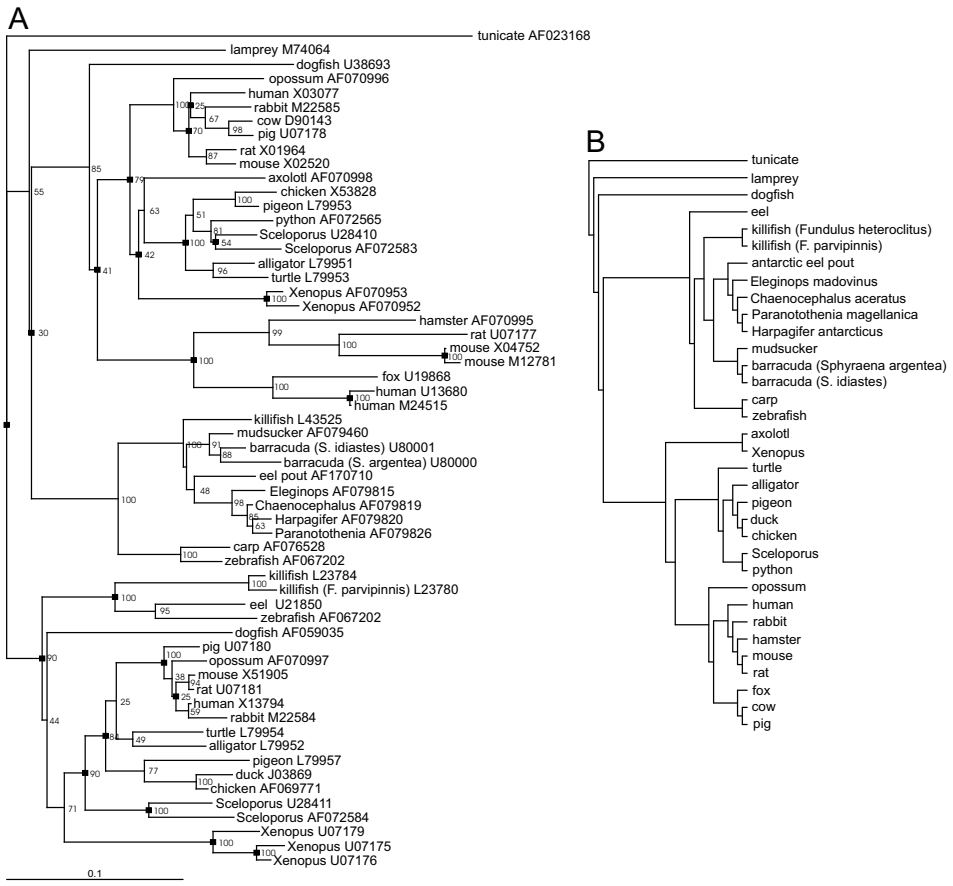

Figure 3: (a) Neighbour joining tree for vertebrate L-LDH sequences, rooted with a tunicate ("sea squirt") as the outgroup, with GenBank accession numbers. The numbers on the internal nodes of the tree are bootstrap values, the scale bar represents 0.1 amino acid replacements per site. Gene duplications required by reconciling this tree with currently accepted relationships amongst the species (b) are shown as filled boxes.

the best fit to the species tree. Martin (1999a) chose the L-LDH gene phylogeny with lowest duplication and loss cost that was not significantly worse than the most parsimonious gene tree, effectively giving a greater weight to duplications and losses than to substitution events.

Alternative approaches to the problem of uncertainty in gene trees deserve to be explored. One method would be to rearrange the optimal gene tree to improve its fit to the species tree. This idea has been formalised by Chen et al. (2000), who describe a simple greedy rearrangement algorithm that takes the initial estimate of the gene tree and performs nearest neighbour rearrangements (Waterman and Smith, 1978) around nodes with bootstrap support less than some specified value. This inverts the problem from one of finding the optimal species tree given a gene tree to one of finding the optimal gene tree, within certain constraints, given a species tree. A maximum likelihood framework has been suggested in the context of coalescence models by Maddison (1997). However, while reasonable statistical models of nucleotide substitution exist, there are none yet for gene duplication, and any such model would need to incorporate the extreme sampling bias that exists in the sequence databases (and hence that many gene "losses" are sampling artifacts).

Uncertainty in gene trees also has implications for inferring species trees. Presently available implementations of reconciled trees do not give any measure of the degree of support for any nodes in the species tree. This makes it difficult to evaluate competing hypotheses, such as the relationships among hagfish and lampreys. Reconciled tree analysis of nine vertebrate gene families supported grouping the lamprey with the rest of the vertebrates, to the exclusion of the hagfish (Page, 2000), whereas analyses of ribosomal genes suggest hagfish and lampreys are sister taxa (Mallatt and Sullivan, 1998). One brute force approach to coping with uncertainty in gene trees would be to construct species trees for each tree in the set of bootstrap trees for a gene family and use the majority rule consensus (Margush and McMorris, 1981) of those resulting trees as the best estimate of species relationships. Applying this to the L-LDH sequences, we get the species tree shown in figure 4a, revealing which relationships are only weakly supported by the L-LDH data.

If one has a set of gene families one could apply resampling methods to those familes. This is analgous to Felsenstein's use of the bootstrap on sequence data (Felsenstein, 1985), however, we would resample the gene families rather than the nucleotide or amino acid sites for each gene family. This amounts to treating each gene family as a single character.

5 Locating gene duplications

> Take four, or maybe eight, decks of 52 playing cards. Shuffle them all together and then throw some cards away. Pick 20 cards at random and drop the rest on the floor. Give the 20 cards to some evolutionary biologists and ask them to figure out what you've done.(Skrabanek and Wolfe, 1998, p. 698)

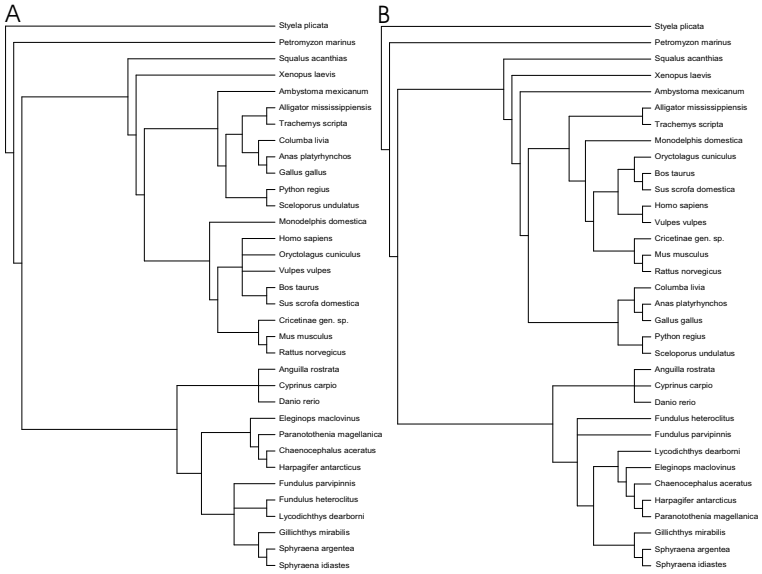

Figure 4: (a) Majority rule consensus tree for selected vertebrate species based on 100 bootstrap gene trees for L-LDH. (b) Strict consensus of 9 optimal species tree for the L-LDH data, requiring 12 duplications and 32 losses.

Although the mapping between a gene and species tree is unique (Page and Charleston, 1997b) – and hence each node in the gene tree is mapped onto a single node in the species tree – if the species tree is poorly sampled then there will still be ambiguity in the actual location of a duplication on the species tree. This ambiguity means that many gene duplications may cluster together, indicating DNA duplication events affecting large stretches of sequence, or even whole genomes. Genome duplication has been posited as a major factor in the evolution of complexity in vertebrates, although there is considerable debate as to the number and location of these putative duplications (Figure 5). Recent analyses (Martin, 1999b) using an earlier implementation of reconciled trees (Page, 1993) suggest that gene duplications within vertebrates have been largely independent.

Guigo et al. (1996) encountered this ambiguity in their study of eukaryote gene families. They reconciled 53 gene family trees with a species tree comprising 16 taxa. Because many of their gene trees were small (comprising 4-5 genes) there was some ambiguity in the placement of some of these duplications. Using a heuristic algorithm to cluster together the duplications, they found that the 46 duplications could be accounted for by five genome duplications at four different points on the species tree.

Currently implemented algorithms for reconciled trees assume that duplications in different gene families are independent, that is, the algorithms seek to minimise the number of gene duplications. Minimising the number of episodes of gene

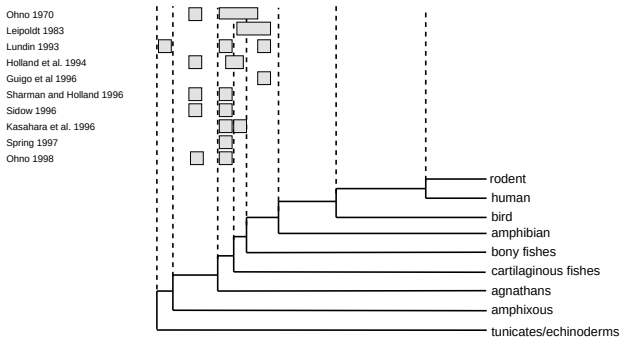

Figure 5: Alternative hypotheses of genome duplication in vertebrates. The phylogeny is drawn with branch lengths proportional to time. From Martin (1999b, fig. 1).

duplication is a significantly harder problem (Fellows et al., 1998).

6 Future

As more and more gene trees are assembled, the metaphor of a simple tree of life becomes increasingly strained, leading us to view organism phylogeny as a "cloud" or statistical distribution of gene histories, largely congruent with one another but showing significant variance (Maddison, 1997). Gene duplication and loss may not be the only cause of this variance. Horizontal transfer of genes makes reconstructing the history of a gene much more difficult, but can be addressed with reconciled trees using techniques developed for an analagous situation in the context of host-parasite coevolution (Charleston, 1998). Horizontal transfer seems unlikely to be of any great importance in vertebrate gene families, but would certainly have to be addressed in other cases, e.g. in bacteria (Martin, 1999c).

There is also the inevitable lag between theoretical developments and their implementation in software. The current release of GeneTree has some of these developments, such as a linear time algorithm for tree mapping (Eulenstein, 1997), but has yet to include more recent results.

Another pragmatic issue is how well the software can cope with the ever growing flood of sequence data. GeneTree was originally conceived as a test bed for algorithms for displaying reconciled trees. There is now a need to enable it to handle numerous, large gene families. For example, it would be very useful to be able to extract gene trees from data bases like Hovergen (Duret et al., 1994) and input these directly into GeneTree. It would then be possible to obtain the best estimates of species phylogeny based on simultaneous analysis of thousands of gene families, and to locate episodes of gene duplication in these families. Work on this is currently in progress.

7 Acknowledgments

This work has been supported the Wolfson Foundation and the Natural Environment Research Council.

References

Charleston, M. A. (1995). Towards a characterization of landscapes of combinatorial optimisation problems, with special reference to the phylogeny problem. *Journal of Computational Biology*, 2:439–50.

Charleston, M. A. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149:191–223.

Chen, K., Durand, D., and Farach-Colton, M. (2000). Notung: dating gene duplications using gene family trees. *RECOMB2000*.

Constantinescu, M. and Sankoff, D. (1986). Tree enumeration modulo a consensus. *Journal of Classification*, 3:349–56.

Downey, R. G. and Fellows, M. R. (1998). *Parameterized Complexity*.

Duret, L., Mouchiroud, D., and Gouy, M. (1994). Hovergen: a database of homologous vertebrate genes. *Nucleic Acids Research*, 22:2360–2365.

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

Fellows, M., Hallett, M., and Stege, U. (1998). On the multiple gene duplication problem. In *Proceedings of the 9th International Symposium on Algorithms and Computation (ISAAC'98), Taejon, Korea*, volume 1533 of *Lecture Notes in Computer Science*, pages 347–356.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39:783–91.

Fitch, W. M. (1979). Cautionary remarks on using gene expression events in parsimony procedures. *Systematic Zoology*, 28:375–9.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–168.

Guigo, R., Muchnik, I., and Smith, T. F. (1996). Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189–213.

Hallett, M. T. and Lagergren, J. (2000). New algorithms for the duplication-loss problem. *RECOMB2000*.

Holmes, R. S. (1972). Evolution of lactate dehydrogenase genes. *FEBS Letters*, 28:51–55.

Li, S. S.-L., Fitch, W. M., Pan, Y.-C. E., and Sharief, F. S. (1983). Evolutionary relationships of vertebrate lactate dehydrogenase isozymes A_4 (muscle), B_4 (heart), and C_4 (testis). *The Journal of Biological Chemistry*, 258:7029–7032.

Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses. *RECOMB98*.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46:523–536.

Mallatt, J. and Sullivan, J. (1998). 28s and 18s rdna sequences support the monophyly of lampreys and hagfishes. *Molecular Biology and Evolution*, 15:1706–1718.

Margush, T. and McMorris, F. R. (1981). Consensus n-trees. *Bulletin of Mathematical Biology*, 43:239–44.

Martin, A. (1999a). Choosing among alternative trees of multi-gene families. *Molecular Phylogenetics and Evolution*, (In Press).

Martin, A. (1999b). Increasing genomic complexity by gene duplication and the origin of the vertebrates. *American Naturalist*, 154:111–128.

Martin, W. (1999c). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *BioEssays*, 21:99–104.

Page, R. D. M. (1993). *COMPONENT, Tree comparison software for Microsoft Windows*. The Natural History Museum, London.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77.

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

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

Page, R. D. M. (2000). Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, 14:89–106.

Page, R. D. M. and Charleston, M. A. (1997a). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7:231–240.

Page, R. D. M. and Charleston, M. A. (1997b). Reconciled trees and incongruent gene and species trees. In Mirkin, B., McMorris, F., Roberts, F., and Rzhetsky, A., editors, *Mathematical Hierarchies in Biology*, volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 57–70. American Mathematical Society, Providence, Rhode Island.

Perrière, G., Duret, L., and Gouy, M. (2000). Hobacgen: Database system for comparative genomics in bacteria. *Genome Research*, 10:379–385.

Sanderson, M. J. (1989). Confidence limits on phylogenies: The bootstrap revisisted. *Cladistics*, 5:113–29.

Skrabanek, L. and Wolfe, K. H. (1998). Eukaryotic genome duplication - where's the evidence? *Current Opinion in Genetics and Development*, 8:694–700.

Slowinksi, J. and Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48:814–825.

Stege, U. (1999). Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable. Technical Report 319, Department of Computer Science, ETH Zurich.

Waterman, M. S. and Smith, T. F. (1978). On the similarity of dendrograms. *Journal of Theoretical Biology*, 73:789–800.

Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom.
E-mail: r.page@bio.gla.ac.uk

*E-mail:*j.cotton@udcf.gla.ac.uk