VERTEBRATE PHYLOGENOMICS: RECONCILED TREES AND GENE DUPLICATIONS

R.D.M. PAGE, J.A. COTTON
Division of Environmental and Evolutionary Biology, IBLS,
Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK
E-mail: r.page@bio.gla.ac.uk

Ancient gene duplication events have left many traces in vertebrate genomes. Reconciled trees represent the differences between gene family trees and the species phylogeny those genes are sampled from, allowing us to both infer gene duplication events and estimate a species phylogeny from a sample of gene families. We show that analysis of 118 gene families yields a phylogeny of vertebrates largely in agreement with other data. We formulate the problem of locating episodes of gene duplication as a set cover problem: given a species tree in which each node has a set of gene duplications associated with it, the smallest set of species nodes whose union includes all gene duplications specifies the locations of gene duplication episodes. By generating a unique mapping from this cover set we can determine the minimal number of such episodes at each location. When applied to our data, this method reveals a complex history of gene duplications in vertebrate evolution that does not conform to the "2R" hypothesis.

1 Introduction

Most genes belong to large gene families, so the analysis of the gene family evolution represents a considerable challenge for the study of genome evolution. Within vertebrates, paralogy (the relationship between genes within a family) is pervasive, and gene duplication has clearly been particularly common [1], but a broadly similar pattern is found in prokaryotes. The timing and frequency of gene duplications is of particular interest, given that gene (and genome) duplication has been posited as a major factor in the evolution of complexity in vertebrates [2]. A popular – and controversial [3,4] – hypothesis of vertebrate genome evolution postulates two successive genome duplications early in vertebrate evolution (the "2R" hypothesis). Understanding the evolution of vertebrate genomes requires a well supported phylogenetic framework for vertebrates, and methods for locating episodes of gene duplication. In this paper we explore the use of reconciled trees [5,6] to address the latter question.

1.1 Reconciled trees

Conventional phylogenetic methods use molecular sequences as characters of organisms, which conflates organismal and gene phylogenies. However, gene phylogenies are not species phylogenies - processes such as gene duplication,

gene loss, and lineage sorting can introduce important differences between the correct phylogenetic tree for a set of genes and the correct tree for the corresponding species. An alternative is to investigate the relationship between gene trees and species trees using reconciled trees. A reconciled tree [5,6] is a map between a gene tree and a given species tree, with gene duplications and losses being postulated to explain any incongruence between the two trees. If the species tree is unknown then the most parsimonious estimate of the species tree is that minimizing the number of gene duplications required on a gene tree [7,8]. We can extend the method to many genes, so the most parsimonious species tree is that which implies the minimum number of gene duplication (or duplication and loss) events over the set of gene families ("gene tree parsimony" [7]). The map between a gene tree and a species can be computed in linear time [9], making reconciled trees practicable for very large analyses, and potentially even for genome-wide comparisons.

1.2 Vertebrate phylogeny

To test the performance of gene tree parsimony on a real dataset, we constructed a data set of 118 vertebrate gene families [a] based on data from the HOVERGEN database [10]. The higher-level phlyogeny and ancient evolution of the vertebrate in many ways represents an ideal test-case for these methods, because there has been considerable recent interest in both their phylogeny and in evolution by gene duplication in the group. A fairly robust consensus on the main relationships within the group had emerged, based on morphological evidence from both fossil and extant taxa [11], but analyses of whole mitochondrial genomes have produced unorthodox and controversial phylogenies, provoking new debate [12].

The species tree we obtained using gene tree parsimony (Fig. 1) differs little from a conventional view of vertebrate phylogeny [11], in marked contrast to the unorthodox trees obtained from mitochondrial genomes [12]. This result confirms preliminary findings [13] that reconciled tree methods can reconstruct phylogeny accurately in the face of gene duplication and loss.

1.3 Genome duplications

The timing and location of gene duplications is a key problem in understanding the evolution of gene families and genomes. Existing techniques for mapping gene trees onto species trees can identify gene duplications, but do not necessarily locate them precisely on the species tree. Furthermore, gene duplication

[a] The GeneTree file and individual alignments and gene trees are available from http://kimura.zoology.gla.ac.uk/vertebrate_data.

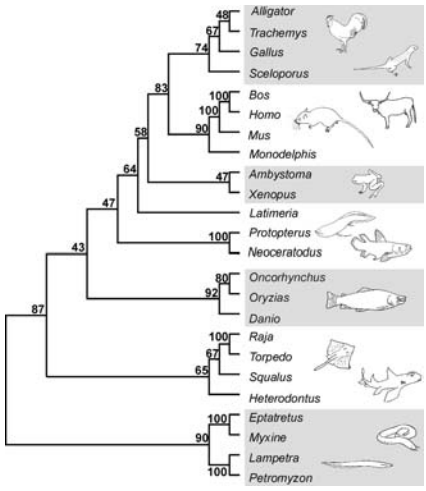

Figure 1: Phylogeny of vertebrates reconstructed using gene tree parsimony in GeneTree [15] on a set of 118 nuclear genes. Bands of shading identify higher taxonomic groups of vertebrates. This is the majority-rule consensus of 100 species trees, generate from 100 bootstrap trees for each gene tree. Figures on nodes are bootstrap percentages.

events can occur on any scale, from small pieces of DNA carrying fragments of genes right up to polyploidisation events due to hybridisation or incorrect division, so duplications on individual gene trees could be correlated, occurring as a result of the same molecular events. Identifying these events is complicated by the fact that most gene families are known from only some species, so there can be considerable uncertainty in where particular duplications occurred on the species tree. We need techniques that can identify these "duplication episodes" by clustering individual gene duplications [16,17]. We now present a method for achieving this and apply the technique to our vertebrate data set.

2 Locating gene duplications

2.1 Terminology

We will restrict ourselves to rooted trees. The immediate ancestor of a node in a tree is its *parent*, and the immediate descendants of a node are its *children*. A node with no children is a *leaf*. Let G be a rooted tree for m genes obtained from $n \leq m$ species (a *gene tree*), and S be a rooted tree for the species (a *species tree*). For each node in S the set of nodes that are its descendants form that nodes *cluster*. The cluster of the root is $\{1, \ldots, n\}$, the clusters of the

leaves are $\{1\}, \{2\}, \ldots, \{n\}$. Following Margush and McMorris [14], we use the shorthand of treating the node and its cluster as synonymous. Hence, for any pair of nodes x and y in S, if $x \subset y$ then x is a descendant of y. For any node $g \in G$, let $\eta(g)$ be the set of species in which occur the extant genes descendant from g (if g is a leaf then $\eta(g)$ is the species from which gene g was obtained). For any $g \in G$, let $M(g)$ be the node in S with the smallest cluster satisfying $\eta(g) \subseteq M(g)$. A map from G into S associates each node $g \in G$ with a node $M(g) \in S$, and can be visualized using a reconciled tree [6]. Let l and r be the left and right children of a node $g \in G$. If either l or r (or both) map onto $M(g)$ (i.e., $M(l) = M(g)$ and/or $M(r) = M(g)$) then we infer that g is a gene duplication [5].

2.2 The problem

The problem of locating gene duplications using reconciled trees was first addressed by Guigó *et al.* [16], who noted that the map between gene tree and species tree puts bounds on the location of a given duplication, rather than necessarily locating the duplication precisely. Whereas the map between gene and species tree associates each node g in the gene tree with a single node $M(g) = s$ in the species tree, the actual gene duplication may have occurred anywhere along the path between $M(g)$ and $M(parent(g))$ [b]. Given this ambiguity, our task is to find the optimal placement of the duplications required to reconcile a set of gene trees $G_1, G_2, \ldots, G_k$ with a species tree S. It is important to clearly distinguish between episodes of gene duplication and *genome* duplication. Guigó *et al.* refer to *any* clustering of gene duplications as a "genome duplication," regardless of whether the whole genome or only a part of it duplicated. Here we use the term "episode" as the generic term for two or more duplications in different gene families that can be explained by a single event.

2.3 Guigó et al.'s algorithm for placing duplications

Guigó *et al.* partition gene duplications into three categories:

free: if g is the root of G.
locked: if g is not the root of G.
absolutely locked: if g is locked and $M(parent(g)) = parent(M(g))$.

[b] Note that moving a duplication down the species tree towards the root will require additional losses to be postulated. However, given that many apparent "losses" in reconciled trees may be due to lack of knowledge (such as poor taxonomic or genomic sampling), rather than actual gene loss, invoking additional losses does not seem unreasonable.

Examples of these three categories can be seen in Figure 2b. Guigó *et al.* sketched an algorithm to cluster gene duplications into the minimum number of locations on the species tree. First we identify the set of allowed locations A_g in the species tree for a duplication g. If g is the root of the gene tree then $A_g = \{s \in S : M(g) \subseteq s\}$ (the set of all nodes in the species tree from $M(g)$ down to the root). If g is not the root of the gene tree then $A_g = \{s \in S : M(g) \subseteq s \subseteq M(parent(g))\}$ (the set of nodes in the species tree from $M(g)$ down to, but not including, the node into which the parent of g is mapped). Duplications are placed as follows:

> Step 1: Place on the species tree S all absolutely locked duplications (for which $A_g = M(g)$). The set of locations of absolutely locked duplications is $D_{absolute}$.
>
> Step 2: For all locked duplications g_l for which $A_{g_l} \cap D_{absolute} \neq \emptyset$ find the absolutely locked duplication(s) ($g_a : A_{g_l} \cap A_{g_a} \neq \emptyset$). If $|A_{g_l} \cap A_{g_a}| > 1$ place g_L at the $s \in A_{g_l} \cap A_{g_a}$ that is furthest from the root of S. The set of locations of locked duplications is D_{locked}.
>
> Step 3: For all locked duplications g_l for which $A_{g_l} \cap D_{absolute} = \emptyset$, if $A_{g_l} \cap D_{locked} = \emptyset$ then g_l is placed at the node $M(g)$, otherwise the duplication is placed such that the total number of locations of gene duplications is minimal.
>
> Step 4: Free duplications g_f for which $A_{g_f} \cap D_{locked} \neq \emptyset$ are placed at the node $s \in A_{g_f} \cap D_{locked}$ that is furthest from the root of S, otherwise they are placed at the root of S.

The result of applying these steps is a clustering of gene duplications into episodes, and a final mapping of duplications onto the species tree. Note that although Guigó *et al.* gave hints about how to minimize the number of gene duplications (Step 3) they did not present a formal algorithm for doing this.

2.4 An alternative formulation

Fellows *et al.*[17] define the MULTIPLE GENE DUPLICATION problem as being the mapping of a set of gene trees $G_1, G_2, \ldots, G_k$ into a species tree S such that the number of multiple gene duplication events is minimal. They go on to show that this problem is *NP*-hard. Their formulation of the problem is somewhat different from Guigó *et al.*'s – those authors aim to minimize the number of locations in S where gene duplications have occurred, but do not postulate any additional duplications over and above those required to reconcile each gene tree G_i with S. Fellows *et al.*, however, will invoke additional duplications if it

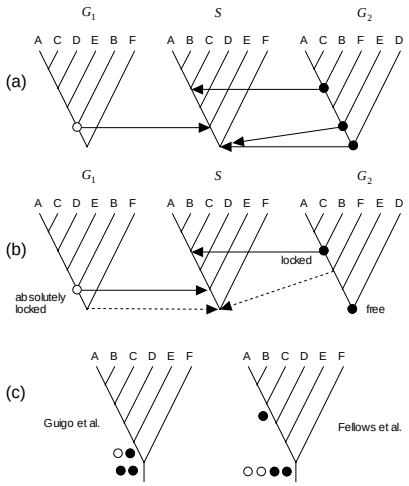

Figure 2: (a)Two gene trees and their species tree with nodes mapped onto S. (b) Node ABCDE in G_1 is absolutely locked, whereas node ABC in G_2 is locked. (c) Comparison of how Guigó *et al.*[16] and Fellows *et al.*[17] would place the duplications on S to minimise the number of multiple gene duplications.

reduces the number of multiple gene duplication events. For example, given the two gene trees in Figure 2a, using the rules of Guigó *et al.* the duplication at node ABCDE in G_1 is absolutely locked and hence cannot be moved. However, Fellows *et al.* move this duplication to the root of the species tree (at the cost of an additional duplication). Similarly, Fellows *et al.* state that "it is not beneficial" to move node ABC in G_2. However, in Guigó *et al.*'s terminology, this duplication is not absolutely locked and could be placed anywhere along the path from ABC to ABCDEF in S. Moving it to node ABCDE in S reduces the number of multiple gene duplications from 4 to 3, the same score as for the Fellows *et al.* reconstruction, but without invoking an extra duplication.

3 Placing duplications using set cover

We can reformulate Guigó *et al.*'s algorithm as a set cover problem. Let D be the set of all nodes $g \in G_i, i = 1, \ldots, k$ that are gene duplications. Each $s \in S$ has associated with it a set of duplications $D_s = \{d : d \in D, s \in A_d\}$. Finding the smallest number of locations at which gene duplication has taken place corresponds to finding the smallest number of sets such that their union is D. The set cover problem is NP-complete, but heuristics are available[19].

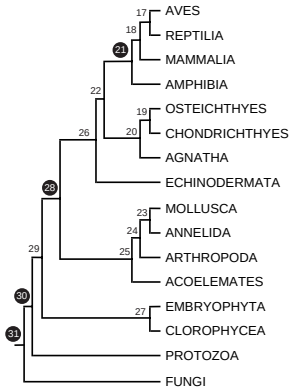

Figure 3: A species tree for 16 eukaryotes from Guigó *et al.* [16]. Internal nodes are labelled $17-31$ in postorder. The locations of the "genome" duplications inferred by Guigó *et al.*[16] are highlighted.

We illustrate this approach using Guigó *et al.*'s data set. This has played an important role in developing methods of tree reconciliation. Previous work has shown that they miscount the number of gene losses [18] and that their species tree is not optimal for the 53 gene trees [18,20].

The species tree shown in Figure 3 requires 46 gene duplications, which are distributed over 7 nodes in the species tree:

$$
\begin{aligned}
D_{21} &= \{2, 22, 36, 37, 44, 46\} \\
D_{22} &= \{8, 9, 13, 32, 33, 35-38, 44\} \\
D_{26} &= \{8-9, 13, 32, 33, 35-38, 44\} \\
D_{28} &= \{1, 4, 6, 8, 9, 13-17, 19, 20, 25, 26, 29, 32, 33, 35-38, 41\} \\
D_{29} &= \{1, 6, 8, 9, 13-17, 19, 20, 24-26, 30, 32-38, 41\} \\
D_{30} &= \{1, 7-9, 13-17, 19, 20, 24-26, 30, 32-38, 40, 42, 43\} \\
D_{31} &= \{1, 3, 57-18, 21, 23-28, 31-39, 45\}
\end{aligned}
$$

The duplications are arbitrarily numbered $1-46$. The minimal set cover for D is $\{D_{21}, D_{28}, D_{30}, D_{31}\}$. These are the same four locations of the "genome" duplications identified by Guigó *et al.* (Figure 3).

3.1 Final mapping

The minimal set cover might not yield an unambiguous mapping between the gene trees and the species tree; for example, duplication 36 is an element

of all four sets in the minimal cover. This node occurs at the root of the gene tree for β-Nerve growth factor precursor (NGF) which has the topology (REPTILIA,(MAMMALIA,(AMPHIBIA,AVES))), and hence in Guigó *et al.*'s terminology is "free." Its set of allowable locations comprises vertex S_{21} and all its ancestors in the species tree (Figure 3). Following Guigó *et al.*, any duplication g which occurs in more than one set in the minimal set cover is mapped onto the node closest to $M(g)$. This can be easily done as follows:

Step 1: Let F be a set of duplications. Initially $F \leftarrow \emptyset$.
Step 2: Process each node in S in postorder. For each node s for which $D_s \neq \emptyset$ go to Step 3.
Step 3: If $F = \emptyset$ then $F \leftarrow D_s$, otherwise $D_s \leftarrow D_s \setminus F$ and $F \leftarrow F \cup D_s$

The result of this procedure is a unique mapping from the gene trees into the species tree, consistent with the minimal set cover. Applying this to Guigó *et al.*'s data we obtain the following mapping, where duplications are labeled by the abbreviated gene family name from Guigó *et al.*'s table 2.

D_{21} = {ACHG, GLUC, NGF, NGF, PAHO, TBB2, TPMA}.
D_{28} = {ACH2, ACT2, ACT3, ACTB, ANFC, COLI, CYLA, CYLA, CYLB, CYLB, G3P, G3P2, H2B, H2B, H4, HBA1, HBA2, PRVA, TBA1}.
D_{30} = {ACT3, H2A3, H4, HMDH, TBA1, TBA1, TBB}.
D_{31} = {ACT, ACT2, ATPB, CATA, CISY, CYLH, G6PI, H2A2, H2B1, H31, H4, RLA2, TOP2}.

This mapping differs from that shown by Guigó *et al.* (their fig. 4), in that those authors assign one duplication in gene NGF to D_{28}, and one duplication of the genes CYLA, CYLB, and TBA1 to D_{30}. However, these placements violate Guigó *et al.*'s own rule that "free duplications are placed at the closest location preceding the node in which the duplication is mapped where a duplication – absolutely locked or locked –, if any, has already been placed" (Step 4 in section **??** above).

3.2 Counting the number of episodes of gene duplication

If more than one duplication in a gene tree G is associated with the same node s in the species tree S (i.e., $|G \cap D_s| > 1$) then we may have to postulate multiple episodes of gene duplication occurring at s. For example, given two nodes g_1 and g_2 where g_1 is ancestral to g_2, if both nodes are in D_s then two duplication episodes are needed. However, if neither g_1 nor g_2 is ancestral to the other then both could be explained by the same event. Let the duplication

height, $h(g)$, of a node $g \in G$ be the number of nodes along the path between g and the root of G for which are in D_s. Any duplication $g \in D_s$ with the same height can be explained by the same duplication event. Hence, the minimum number of distinct episodes of duplication at node s in gene family G is then $E_{(G,s)} = \mathrm{MAX}(h(g) : g \in G, g \in D_s) + 1$. The minimum number of episodes of duplication at node s across all k gene families is then $\mathrm{MAX}(E_{(G,s)} : G_1, \ldots, G_k)$.

For the Guigó *et al.* example, we require two episodes of gene duplication at D_{21}, D_{28}, and D_{30}, and one at D_{31}. This differs from their finding single duplications at all locations except D_{30}, where they postulate that a double duplication occurred. This difference stems from their misplacing the duplications for genes NCF, CYLA, CYLB, and TBA1 (see Sec. 3.1).

3.3 Duplication patterns in vertebrates

The locations of the 1380 inferred gene duplications in our 118 gene family data set (Sec. 1.2) were found using the above algorithm (Sec. 3), showing that they can be strongly clustered on the species tree (Fig. 4). Many apparent duplications occur near the tips of the tree in the mouse and human lineages, but the bulk of these "duplications" actually represent multiple alleles at polymorphic loci, rather than gene duplications. Figure 4 shows that substantial numbers of duplication events have occurred throughout vertebrate evolution, often affecting many gene families simultaneously. The largest single such event (duplicating 58 out of 118 families) occurred after the divergence of sharks and rays and prior to the divergence of teleosts and lobe fin fish. Gene duplication is clearly an important feature of vertebrate evolution, but the pattern shown in figure 3 is more complex than that expected from the "2R hypothesis". Some gene families have undergone as many as 11 successive episodes of duplication, and at no point in vertebrate phylogeny can we explain all gene duplications that occurred at that time by a single genome-wide event.

4 Future directions

Further work on this problem is needed. There are two limitations of our algorithm that we are aware of. Our algorithm for the final mapping (Sec. 3.1) minimizes the number of location in the species tree at which gene duplications occur, but it does not guarantee to minimize the total number of episodes of gene duplication. It is possible to construct examples where spreading gene duplications across more locations will reduce the overall number of episodes of duplication.

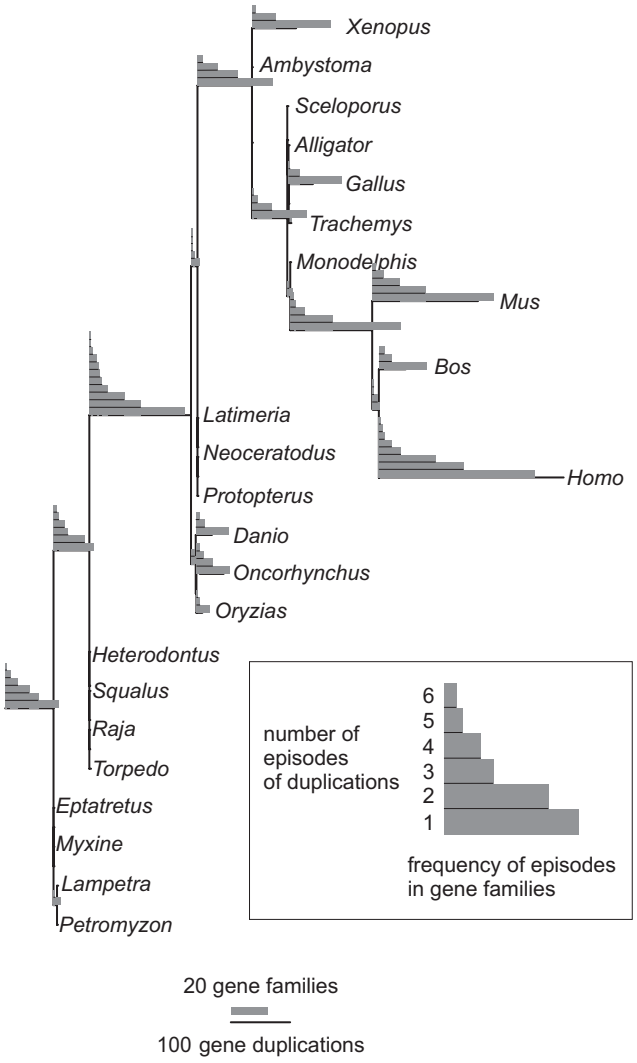

Figure 4: Distribution of gene duplications during vertebrate evolution. The species tree is one of three most parsimonious trees from a GeneTree [15] search. Branch lengths represent the number of separate gene duplications inferred to have occurred along each branch. Stacked bars represent the number of distinct episodes of gene duplication in each of the gene families that have duplicated along the branch. For clarity, bars have been omitted where only a single duplication episode is inferred for each gene family.

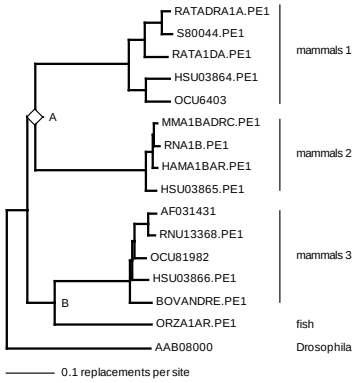

Figure 5: Phylogeny for vertebrate adrenergic receptor $\alpha 1$ sequences. The method for locating gene duplications described in this paper would place node A somewhere after the split of fish and mammals, but prior to the last common ancestor of mammals. Based on relative amount of sequence divergence with respect to node B (the split between fish and mammals), node A in fact pre dates the separation of fish from the ancestors of mammals. Data supplied by Xun Gu [21] . Sequence names are those used in the HOVERGEN database [10], in which ADRA1 is family FAM000048.

Our algorithm uses only the topology of the tree, and hence may make erroneous placements of duplications. For example, Figure 5 shows a gene tree for vertebrate adrenergic receptor $\alpha 1$ (ADRA1). The descendants of the duplication at node A are all mammalian sequences, hence a reconciled tree would place this duplication at the base of mammals. The set of allowed location for this duplication includes the common ancestor of mammals, and every node ancestral to that node that postdates the split between mammals and fish (equivalent to node B in Figure 5)[c]. However, if we consider the branch lengths in the tree, node A is deeper than node B in the tree and hence pre dates the oldest node in its allowed set of locations. One way to address this problem would be to refine the rules for determining sets of allowed location for gene duplications to take into account amounts of molecular sequence divergence (if they are sufficiently clock-like).

[c] This problem will be more prevalent in those gene families that have poorly sampled taxonomically, or have undergone substantial gene loss. Finding a single fish ADRA1 sequence related to either of the group 1 or group 2 mammal sequences would result in the method described here correctly inferring that node A pre dates the split between fish and mammals.

Acknowledgments

This work was supported the NERC, Wolfson Foundation, and the EMBO. Mike Charleston and two anonymous reviewers provided helpful comments.

References

1. R. D. M. Page, J. A. Cotton, in *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics*, (Kluwer Academic Publishers, 2000)
2. S. Ohno, Cell Devel. Biol. **10**, 517 (1999)
3. A. L. Hughes, J. Mol. Evol. **48**, 565 (1999).
4. L. Skrabanek, K. H. Wolfe, Curr. Opin. Genet. Dev. **8**, 694 (1998).
5. M. Goodman *et al*, Syst. Zool. **28**, 132 (1979).
6. R. D. M. Page, Syst. Biol. **48**, 53 (1994)
7. J. B. Slowinski, R. D. M. Page, Syst. Biol. **48**, 81 (1999).
8. R. D. M. Page, M. A. Charleston, Trends Ecol. Evol. **13**, 356 (1998).
9. L. Zhang, J. Comput. Biol. **4**, 177 (1997).
10. L. Duret, D. Mouchiroud, M. Gouy, Nucleic Acids Res. **22**, 2360 (1994).
11. M. J. Benton, *The Phylogeny and Classification of the Tetrapods* (Clarendon Press, Oxford, 1988).
12. R. Zardoya, A. Meyer, in *Major Events in Early Vertebrate Evolution*, ed. P. Ahlberg (Taylor and Francis, London, 2001).
13. R. D. M. Page, Mol. Phylogent. Evol **14**, 89 (2000)
14. T. Margush, F. R. McMorris, Bull. Math. Biol. **43**, 239 (1981).
15. R. D. M. Page, Bioinformatics **14**, 819 (1998).
16. R. Guigo, I. Muchnik, T. F. Smith, Mol. Phylogenet. Evol. **6**, 270 (1996).
17. M. Fellows, M. Hallett, U. Stege, in *Proceedings of the 9th International Symposium on Algorithms and Computation*, eds. C. Kyung-Yong, O. H. Ibarra (Springer, Heidelberg, 1998).
18. R. D. M. Page, M. A. Charleston, in *Mathematical Hierarchies in Biology*, eds. B. Mirkin, F. R. McMorris, F. S. Roberts, A. Rzhetsky (American Mathematical Society, Providence, RI, 1997).
19. T. H. Cormen, C. E. Leiserson, R. L. Rivest, *Introduction to algorithms* (MIT Press, Cambridge, MA, 1990).
20. M. T. Hallett, J. Lagergren, in *RECOMB '00, Proceedings of the fourth annual international conference on computational molecular biology*, (Association for Computing Machinery, 2000).
21. Y. Wang, X. Gu, J. Mol. Evol. **51**, 88 (2000).