Molecular Phylogenetics and Evolution
Vol. 14, No. 1, January, pp. 89–106, 2000
Article ID mpev.1999.0676, available online at http://www.idealibrary.com on IDEAL®

Extracting Species Trees From Complex Gene Trees: Reconciled Trees And Vertebrate Phylogeny

Roderic D. M. Page

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

Received December 14, 1998; revised April 22, 1999

Paralogy is a pervasive problem in trying to use nuclear gene sequences to infer species phylogenies. One strategy for dealing with this problem is to infer species phylogenies from gene trees using reconciled trees, rather than directly from the sequences themselves. In this approach, the optimal species tree is the tree that requires the fewest gene duplications to be invoked. Because reconciled trees can identify orthologous from paralogous sequences, there is no need to do this prior to the analysis. Multiple gene trees can be analyzed simultaneously; however, the problem of nonuniform gene sampling raises practical problems which are discussed. In this paper the technique is applied to phylogenies for nine vertebrate genes (aldolase, α-fetoprotein, lactate dehydrogenase, prolactin, rhodopsin, trypsinogen, tyrosinase, vassopressin, and Wnt-7). The resulting species tree shows much similarity with currently accepted vertebrate relationships.
© 2000 Academic Press

INTRODUCTION

> "Regarding the analysis of nuclear genes, it is worth noting that, although the findings, in general, do not support rodent polyphyly, different genes have provided dissimilar answers to the question of rodent phylogeny. Such discrepancies are often observed when using nuclear genes, perhaps because some genes evolve under different evolutionary constraints in various tree branches, or because the genes analysed might be paralogous. In contrast, mitochondrial genomes contain only orthologous single-copy genes, and can thus provide more reliable phylogenies." (D'Erchia *et al.,* 1996, p. 597)

Mitochondrial and nuclear genes have different strengths and weaknesses when used to infer vertebrate phylogeny. There is a wealth of mitochondrial sequence data available for vertebrates, although in many taxa only one or two genes have been sequenced. Inferences made from individual mtDNA genes may differ (Cao *et al.,* 1994); hence, phylogenies from mtDNA are vulnerable to sampling error. This can be obviated by sequencing the complete genome (about 16,000 bp in vertebrates), which has been done for a small but growing number of vertebrate taxa. However, even analyses of complete mitochondrial genomes have failed to resolve key questions in vertebrate phylogeny, such as the relationships of lungfishes, tetrapods, and the coelacanth (Zardoya and Meyer, 1997). Furthermore, at the level of deep vertebrate and chordate phylogeny, analysis of mtDNA sequences fails to recover the generally accepted relationships among these taxa (Naylor and Brown, 1998).

Mitochondrial DNA has the virtue of comprising single-copy genes in a genome sufficiently large to yield numerous characters yet sufficiently small for its complete sequence to be readily determined. At the same time, this virtue can be seen as a limitation. In the absence of recombination, mtDNA is inherited as a single unit; hence, phylogenies derived from different mtDNA genes are not independent estimates of organismal phylogeny. Furthermore, its size pales in comparison with that of the nuclear genome.

The size of the nuclear gene database is potentially enormous, especially as genome sequencing projects gain momentum (Brown, 1996). However, nuclear genes are often present in multiple, paralogous copies, making it difficult to be certain that phylogenies are based on orthologous sequences (Fitch, 1970). Paralogy is pervasive. The extent of the problem is illustrated in Fig. 1, which shows the relationship between numbers of species and numbers of sequences for a range of vertebrate mitochondrial and nuclear genes. For mitochondrial genes the relationship is 1:1, each species has a single mitochondrial genome and hence a single sequence for each mitochondrial gene. However, in nuclear genes there are almost always many more sequences than species; so, each species may have several representatives of a particular gene family.

One approach to the analysis of nuclear genes is to concatenate the sequences of numerous putatively orthologous genes for the same species into one long sequence (e.g., Graur *et al.,* 1996). Apart from the possibility that such amalgamated data sets may obscure weaknesses in the data (Halanych, 1998), this does not directly address the problem of distinguishing orthology from paralogy—it merely hopes to over-

1055-7903/00 $35.00
Copyright © 2000 by Academic Press
All rights of reproduction in any form reserved.

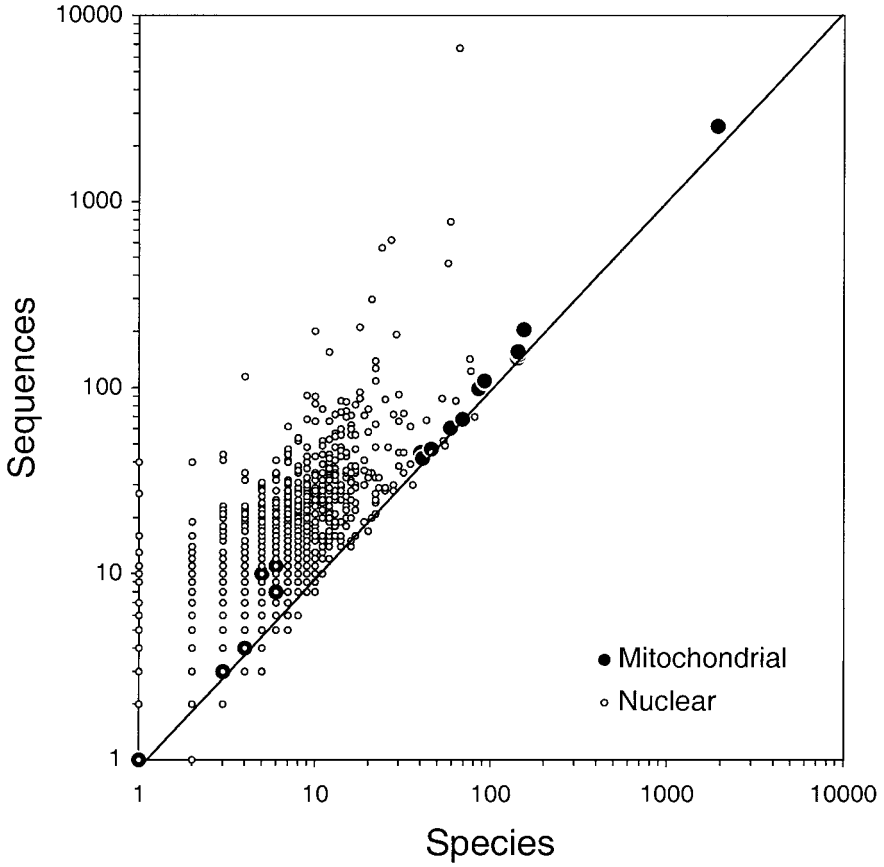

FIG. 1. Number of sequences plotted against number of species for vertebrate gene families in release 29 (March 17, 1998) of the HOVERGEN (Duret *et al.,* 1994) data base. Note that usually each species has a single mitochondrial sequence for a given gene (hence, the mitochondrial genes fall along the 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.

whelm it by sheer weight of data. Phylogenies of multigene families potentially contain a wealth of information about organismal phylogeny. However, the multiplicity of sequences from the same taxa combined with uneven taxonomic sampling greatly complicates the relationship between gene and species trees. If the nuclear database is to be fully exploited in phylogenetic studies we need appropriate analytical tools. Reconciled trees are one such tool.

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.*'s (1979) study of hemoglobin gene phylogeny, in which there were significant discrepancies between gene and organismal phylogenies. Suppose we have a phylogeny for four species and four genes sampled from those species, and suppose that the gene and species trees—which we believe to be correct—disagree (Fig. 2a), then the question is: how can the trees be both true and yet discordant? One approach to a resolution is to embed the gene tree in the species tree (Fig. 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 (Fig. 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 the three gene losses. An alternative explanation for these "losses" is that the other copy of the gene is present in these species but is 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 could be viewed as predictions about the existence of undiscovered genes. In the example shown, further sequencing may uncover copy

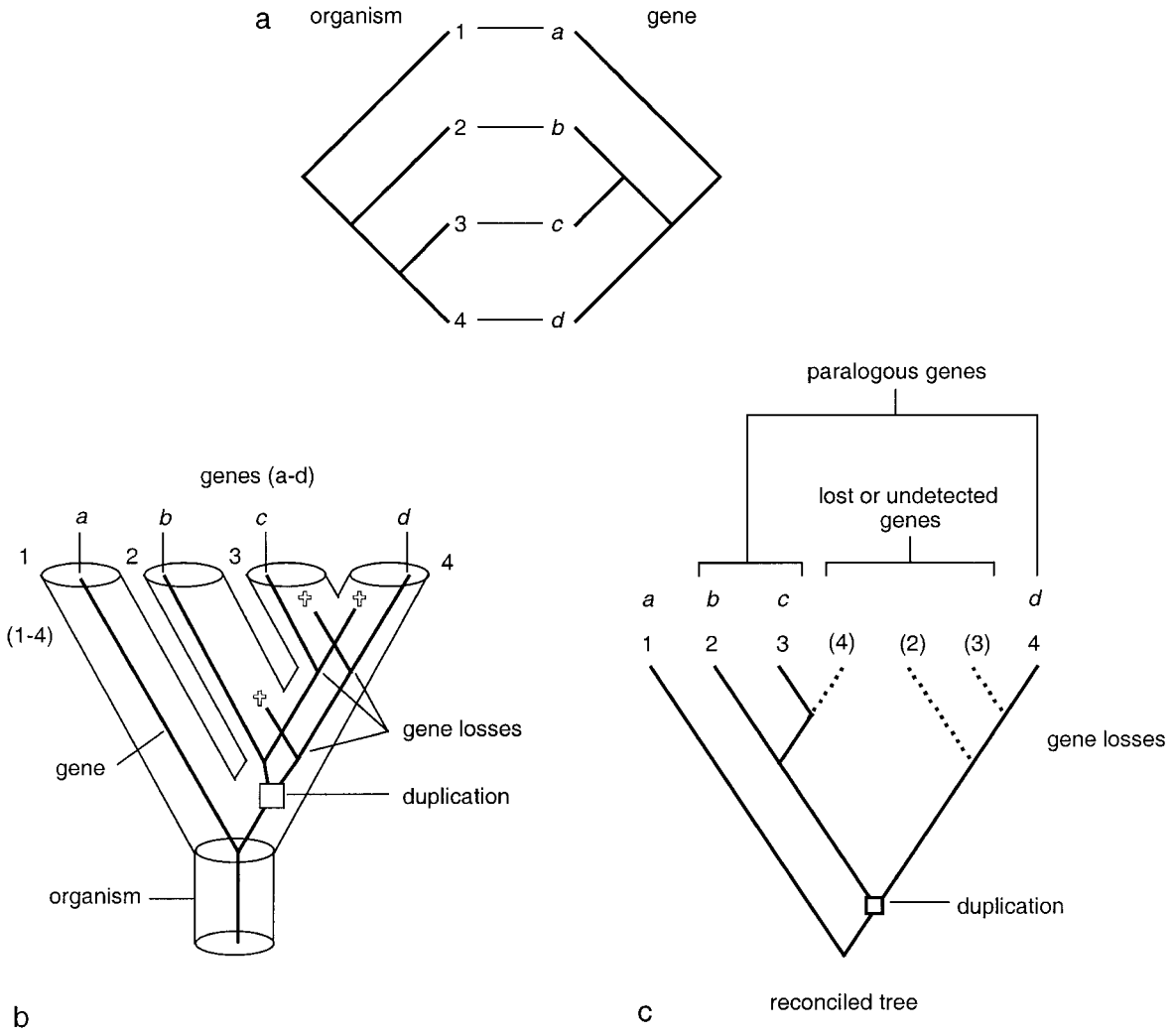

FIG. 2. (a) Incongruent gene and species trees. This incongruence can be explained by hypothesizing a gene duplication (□) at the base of the gene tree (b). The presence of only a single gene (*a–d*) extant in each of the present-day species (1–4) requires postulating three gene losses (✞). (c) The corresponding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

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 that 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., hemoglobin α and β in the same species). However, additional paralogous genes may be identified using reconciled trees.

Most applications of reconciled trees in molecular systematics have been to single-gene families for the purposes of illustrating the technique (e.g., Page, 1994; Page and Charleston, 1997a). To date, there has been only one large-scale attempt to use reconciled trees to analyze the evolutionary history of multiple genes. Guigó *et al.* (1996) took release 19 of the SWISS-PROT data bank (Bairoch and Apweiler, 1997) and constructed trees for 53 eukaryote genes. They then computed a species tree based on those gene trees and counted the number of gene duplications and losses required to reconcile these trees with the best organismal tree (although this study did not actually construct reconciled trees, their measure of fit between gene and species trees is formally identical to the reconciled tree procedure; see Eulenstein *et al.,* 1997). They used the resulting species tree to locate episodes of gene duplication and suggested that the observed duplications in the 53 genes could be accounted for by five episodes of whole-genome duplication. Guigó *et al.*'s pioneering study shows the potential of reconciled trees in studies of gene and genome evolution but has several serious flaws.

(1) The methods used to construct the individual gene trees were clustering techniques rather than explicitly phylogenetic methods, such as parsimony,

likelihood, or neighbor-joining, and the resulting trees were midpoint rooted, which assumes a molecular clock. This assumption was not tested.

(2) When more than one gene was present in the same taxon, Guigó *et al.* used the average distance between those genes and sequences from other species to construct a distance matrix. This use of "composite" sequences for some taxa almost certainly resulted in spurious trees for some genes.

(3) Their search strategy for finding the optimal species tree was ineffective—Page and Charleston (1997b) found substantially more parsimonious species trees for the same data using different search strategies. These species trees differed markedly from the tree found by Guigó *et al.*

My goal here is to illustrate how reconciled trees might be applied to a real systematic problem, in this case vertebrate phylogeny. I have not set out to "solve" this problem but rather to explore the usefulness of reconciled trees when applied to a real problem. There are significant practical differences between using reconciled trees to simply depict the history of a gene family (Page and Charleston, 1997a) and using them to investigate organismal phylogeny. These problems emerged during this study and are considered below.

Searching for Optimal Species Trees

The fit between a gene tree and a species tree can be used as an optimality criterion for choosing among competing species trees—the species tree that accommodates the gene tree with the least cost is the preferred species tree (Slowinksi and Page, 1999). This approach can be generalized to more than one gene, so that we can use evidence from multiple genes. The cost of reconciling a given gene and species tree can be computed efficiently (Eulenstein, 1997); however, the problem of finding which species tree has the optimal value of this cost is NP complete (L. Zhang, personal communication). Hence, we must rely on heuristic searches. Page and Charleston (1997a,b) showed that a combination of nearest neighbor interchanges (nni) and subtree pruning and regrafting (spr) (Swofford *et al.,* 1996) is effective in finding most-parsimonious species trees.

Missing Sequences

The extreme taxonomic bias of the sequence data bases toward a few model organisms (93% of vertebrate nucleotide sequences in GenBank come from humans, rats, or mice) means that 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 in which species will be grouped on the absence of genes, rather than any actual evidence of their relationship. This problem can be minimized by using only the number of duplications as the optimality criterion for selecting species trees (Page and Charleston, 1997a).

Another problem caused by missing sequences is the rapid increase in the number of species trees that are equally parsimonious explanations of the gene trees. As an example, consider the two gene trees shown in Fig. 3. Both trees have something to say about the relationships among amphibians, birds, and mammals; however, genes 1 and 2 have no mammal species in common. Hence, there are seven optimal species trees for these two gene trees, which correspond to the alternative placements of the horse on the subtree

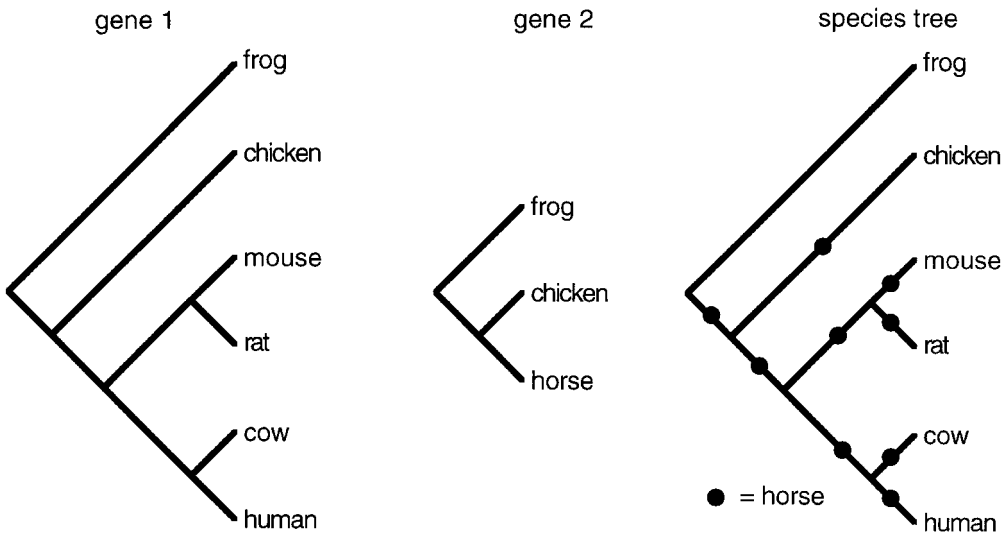

FIG. 3. Two gene trees containing sequences from amphibians, birds, and mammals. Whereas the amphibian and bird are the same in the two genes, no mammals are shared between the two genes. Hence, any species tree that is consistent with these two gene trees will be a parsimonious explanation of the data. In this example, there are nine such trees, which correspond to the alternative placements of the horse on the subtree (((mouse, rat), (cow, human)), chicken) indicated by ●.

((mouse, rat), (cow, human)). The practical dilemma here is whether to include gene 2. Its inclusion brings information about the relations among the higher-taxa amphibians, birds, and mammals. However, because there is no sequence for gene 1 from the horse, the relationships of this taxon are unconstrained with respect to the other mammals. There is a trade-off between adding additional or corroborating information about higher-taxon relationships and minimizing the ambiguity of multiple equally parsimonious trees due to taxa that are represented by few sequences "floating" with respect to other taxa.

Constrained Searches

Constraint trees (Constantinescu and Sankoff, 1986) can be used to address a major problem in inferring species trees from the current sequence data bases, namely their limited taxonomic coverage. A rather extreme hypothetical example is presented in Fig. 4. Suppose that we have two genes (1 and 2), and for both genes we have a single sequence from a teleost fish, a bird, and a mammal. However, in each instance the fish, bird, and mammal species are different for the two genes. Given these two gene trees, there are 105 six-taxon species trees that minimize the number of gene duplications and losses for these gene trees. The strict consensus of these trees is a star tree. This result is due to the lack of any shared information in the two trees that link, for example, the rat and the mouse, the salmon and the trout, or the chicken and the duck.

There are at least two strategies that we could use to circumvent this problem. The simplest, but least attractive, solution is to rename each sequence with the name of the corresponding higher taxon (i.e., the rat and the mouse both become "mammals," the chicken and duck "birds," etc.). This would allow us to recover a single species tree (fish, (birds, mammals)) but has two draw-

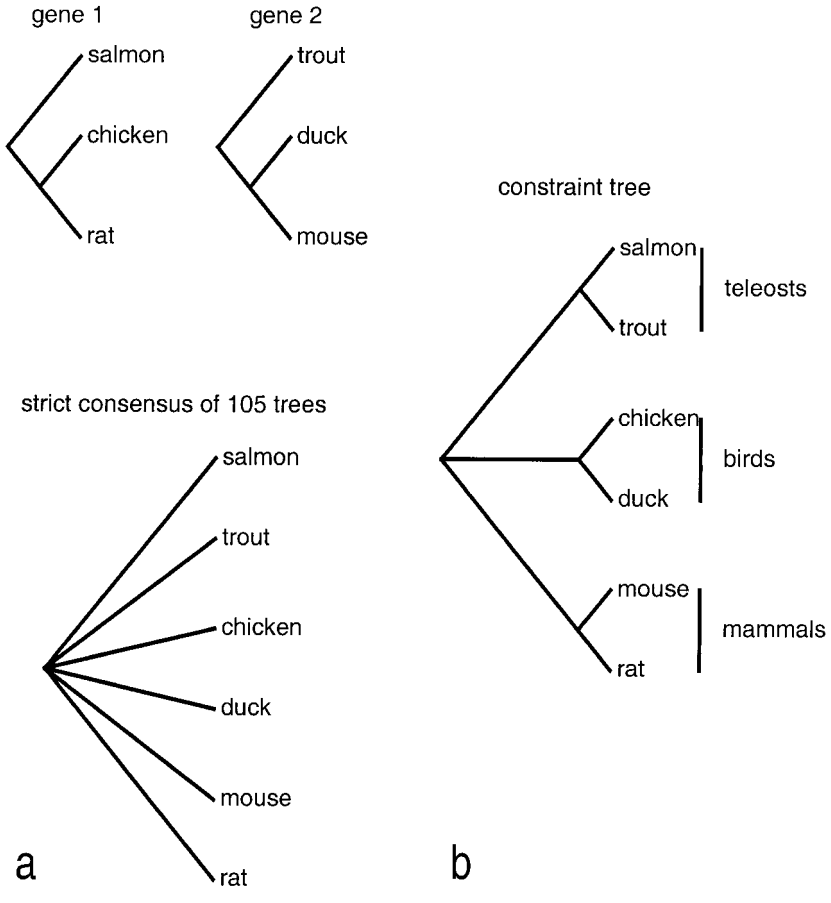

FIG. 4. (a) Trees for two different genes (1 and 2) for which sequences have been obtained from a teleost fish, a bird, and a mammal. The sequences have been obtained from different representatives of these higher taxa. For these two gene trees, there are 105 equally parsimonious species trees which yield an unresolved strict consensus tree. This is because there is no information in the gene trees that the salmon and the trout are, for example, both teleost fishes (and that there are two birds and two mammals). One solution is to enforce constraints on the set of possible optimal species trees by accepting only those trees that preserve the higher-taxon relationships among the sequences (b).

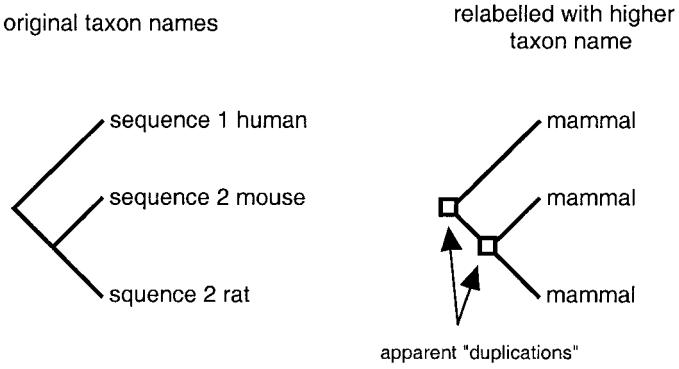

FIG. 5. An example in which relabeling a set of sequences with the corresponding higher taxon (in this instance "mammals") implies false gene duplications.

backs. The first is that it prevents us from investigating relationships among taxa within any of these higher clades; if all mammalian sequences are simply labeled "mammals," regardless of whether they are from elephants, whales, or mice, then any information about mammalian relationships from a gene with sequences from different mammals is lost. The second problem is that if we rename sequences we may infer spurious gene "duplications." For example, a clade of three mammalian sequences would require the occurrence of two duplications to explain the three sequences in the same taxon (Fig. 5). However, if the sequences are from different mammals, this would be simply an artefact of relabeling the sequences. A better solution is to specify interrelationships among species using a constraint tree (Fig. 4b). This enables uncontested groupings, such as birds and mammals, to be specified without loosing potential information on relationships within these groups.

Kinds of Duplications

Gene duplications identified in reconciled trees can be classified into two categories (Goodman *et al.,* 1979): those that are based on direct physical evidence, such as sequencing more than one copy in the same species, and those that are inferred from incongruence between gene and species trees. Estimates of gene trees are subject to uncertainty but current algorithms for computing reconciled trees do not take this into account—any incongruence, no matter how weakly supported, will result in duplications being inferred. It would be desirable to incorporate some measure of the robustness of the gene tree when constructing reconciled trees (see Discussion) but no method for doing this is currently available. The approach used in this paper to assess the degree of support for an inferred duplication is based on the close relationship between the number of duplications required to fit a gene tree into a species tree and the nni distance between the two trees (Ma *et al.,* 1998). Given two trees, the nni distance between them is the number of nni's required to transform one tree into the other (Waterman and Smith, 1978) (Fig. 6a). Consider the gene and species trees shown in Fig. 6b. Gene tree 1 is incongruent with the species trees; hence, a duplication is inferred. However, if the gene tree were really (((a,b),c),d) instead of (((a,c),b),d), there would be no incongruence and hence no duplication. The trees (((a,b,),c),d) and (((a,c),b),d) are similar to each other, differing in a single nearest neighbor interchange. Hence, a fairly trivial modification to the gene tree would undermine the evidence for the inferred duplication. In contrast, gene tree 2 is more dissimilar to the species tree, differing by 2 nni's and requiring two duplications to fit the species tree. Given the greater mismatch between these two trees, we might have more confidence in the existence of those duplications—it would take a more drastic rearrangement of the gene tree to make it concordant with the species tree. However, using nni's alone as a measure of incongruence ignores the degree of support for the nodes in the tree. If the bootstrap values are as shown in Fig. 6, then, although only a single nni will suffice to eliminate the duplication required by gene tree 1, the nni is across a very well-supported node (bootstrap value of 95%). By comparison, the two nni's required to eliminate the duplications implied by gene tree 2 involve relatively weakly supported nodes. Hence, we might have rather less confidence in these duplications. A crude measure of confidence in duplication would be the average bootstrap value of the edges in the tree that are involved in the nni's required to "undo" the duplication. For the example in Fig. 6, the duplication required by gene tree 1 has a score of 0.95 (a single nni with a bootstrap value of 95), whereas gene tree 2 has a score of 0.30 (2 nni's with bootstrap values of 40 and 20). This measure has limitations (among them the nonindependence of bootstrap values) but allows a quick assessment of the degree of support for inferred duplications.

Supertrees

Using reconciled trees to infer species trees from multiple gene trees is superficially similar to methods for assembling "supertrees" (Sanderson *et al.,* 1998) from suites of smaller trees. However, apart from the differences discussed earlier (Page, 1994), reconciled trees use an explicitly biological criterion (numbers of gene duplications and losses) to choose the optimal species tree (Mirkin *et al.,* 1995), in contrast to consensus supertree methods, which do not. Furthermore, a supertree approach to constructing a species tree from complex gene trees will encounter the problem of multiple occurrences of genes from the same species in the gene tree.

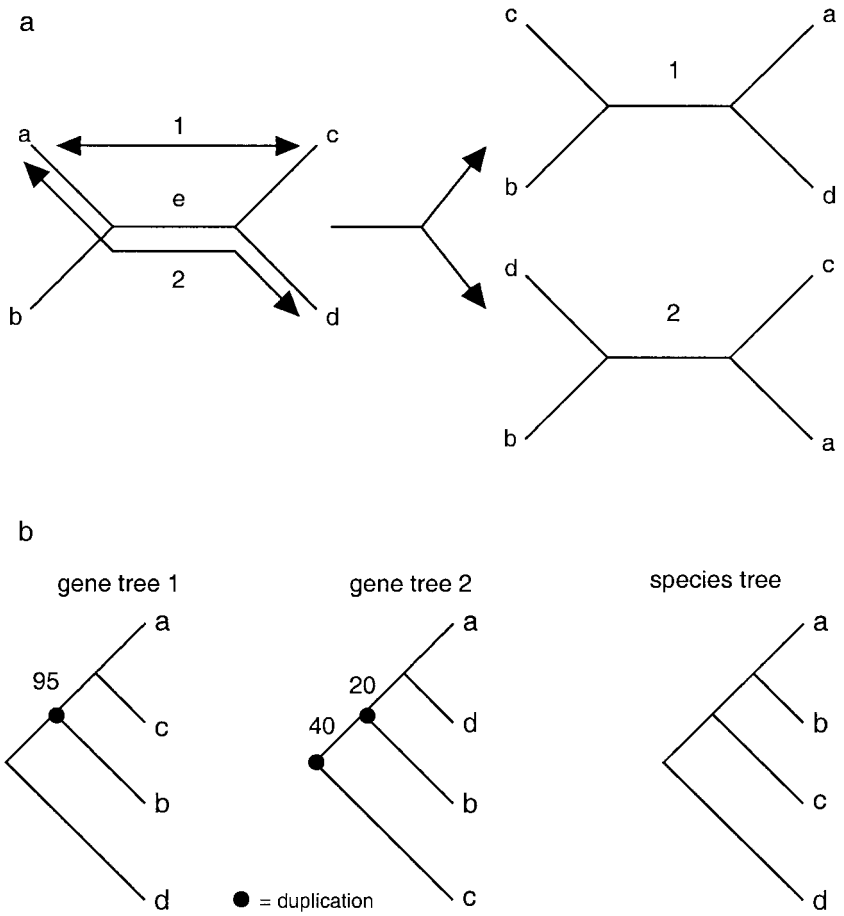

FIG. 6. (a) An unrooted tree for four objects. Interchanging any pair of nodes either side of the internal edge *e* results in one of two trees; hence, these trees are one nearest neighbor interchange (nni) away from the original tree. (b) Two gene trees which are both incongruent with the species tree and hence require gene duplications to be postulated. For each gene the number of nni's required to make the gene tree match the species are shown. Gene tree 1 requires only a single nni, whereas gene tree 2 requires two nni's. However, the nni in gene tree 1 is across a well-supported node (bootstrap value of 95), whereas the two nni's in gene tree 2 are across poorly supported nodes (bootstrap values of 20 and 40). Hence the duplications inferred from the mismatch between gene tree 2 and the species tree are more likely to be artifacts of an incorrect gene tree than the duplication inferred from gene tree 1.

MATERIALS AND METHODS

Genes were selected for this study by surveying recent publications on molecular phylogeny of vertebrates and gene families (e.g., Caspers *et al.,* 1996; Mannen *et al.,* 1997; Stock *et al.,* 1997) and by browsing the HOVERGEN (Duret *et al.,* 1994) and SWISS-PROT data banks, supplemented by BLAST (Altschul *et al.,* 1990) searches. The primary selection criteria were breadth of taxonomic coverage (which eliminated the vast majority of gene families) followed by robustness of the inferred gene trees. Protein sequences were aligned and neighbor-joining trees were inferred using CLUSTALX (Thompson *et al.,* 1997). The alignments are available from the author's Web site (*http://taxonomy.zoology.gla.ac.uk/rod/data/vertebrates/*). Robustness of the gene trees was assessed using bootstrapping (Felsenstein, 1985). Throughout this paper, a sequence is referred to by a data base identifier (such as an accession number or locus name) and the name of the species from which the sequence was obtained.

The reconciled tree algorithm requires rooted gene trees; so, the trees were rooted using one of two criteria. First, where possible, the gene tree was rooted using one or more obviously paralogous sequences. For example, the tree for prolactin sequences was rooted using somatotropin sequences. Alternatively, a sequence from a closely related nonvertebrate taxon, such as a tunicate or amphioxus, was used as the outgroup (see lactate dehydrogenase below).

A total of nine gene families were selected for analysis. Table 1 lists the species for which sequences were available. Most taxa are represented by only a few genes. In five cases ("whale," "turtle," "frog," "*Sclerophus,*" and "goose") a higher taxon was represented by different species in different genes but none of these

TABLE 1

Taxonomic Distribution of Sequences for the Nine Genes Included in This Study

Taxon	Scientific name	ALD	AFP	LDH	PRL	OPS	TRYP	TYR	VASS	WNT-7
Human	*Homo sapiens*	●	●	●	●	●	●	●	●	●
Macaque	*Macaca mulatta*				●					
Marmoset	*Callithrix jacchus*					●				
Mouse	*Mus musculus*	●	●	●	●	●	●	●	●	●
Rat	*Rattus norvegicus*	●	●	●	●	●	●		●	
Hamster	*Mesocricetus auratus*				●	●				
Guinea pig	*Cavia porcellus*								●	
Rabbit	*Oryctolagus cuniculus*	●	●	●	●	●				
Cow	*Bos taurus*		●	●	●	●	●		●	
Sheep	*Ovis aries*	●	●		●	●			●	
Pig	*Sus scrofa*			●	●		●		●	
Goat	*Capra hircus*				●					
Fox	*Vulpes vulpes*			●						
Dog	*Canis familiaris*		●			●	●	●		
Cat	*Felis silvestris catus*				●					
Horse	*Equus caballus*		●		●				●	
Whale	*Balaenoptera borealis*				●					
	Balaenoptera physalus								●	
Chicken	*Gallus gallus*	●	●	●	●	●	●	●	●	
Duck	*Anas platyrhynchos*			●						
Goose	*Anser anser*								●	
	Anser caerulescens									●
Pigeon	*Columba livia*			●		●				
Quail	*Coturnix coturnix*							●		
Turkey	*Meleagris gallopavo*				●				●	
Ostrich	*Struthio camelus*								●	
Scleropus	*Sceloporus undulatus*			●						
	Sceloporus occidentalis									●
Anole	*Anolis carolinensis*					●				
Gecko	*Gecko gecko*					●				
Alligator	*Alligator mississippiensis*			●	●	●				
Crocodile	*Crocodylus novaeguineae*				●					
Turtle	*Trachemys scripta*			●						
	Trionyx sinensis							●		
	Chelonia mydas caranigra				●					
Cobra	*Naja naja*		●							
Xenopus	*Xenopus laevis*	●	●	●		●				●
Bull frog	*Rana catesbeiana*		●			●				
Japanese toad	*Bufo japonicus*				●				●	
Frog	*Rana pipiens*					●				
	Rana esculenta								●	
	Rana nigromaculata							●		
Tiger salamander	*Ambystoma tigrinum*					●				
Salamander	*Plethodon jordani*									●
Newt	*Pleurodeles waltl*									●
Lungfish	*Protopterus aethiopicus*				●					
	Neoceratodus forsteri								●	
Killifish	*Fundulus heteroclitus*			●						
Goldfish	*Carassius auratus*	●				●				
Sea bream	*Sparus aurata*	●								
Sphoeroides	*Sphoeroides nephelus*	●								
Fugu	*Takifugu rubripes*						●		●	
Cod	*Gadus morhua*						●			
Plaice	*Pleuronectes platessa*						●			
Atlantic salmon	*Salmo salar*		●		●		●			
Cherry salmon	*Oncorhynchus masou*								●	
White sucker	*Catostomus commersoni*								●	
Chum salmon	*Oncorhynchus keta*				●				●	
Medeka fish	*Oryzias latipes*					●		●		●
Cave fish	*Astyanax fasciatus*					●				
Eel	*Anguilla anguilla*				●	●				
Bailkal omul	*Coregonus autumnalis*				●	●				
Goby	*Pomatoschistus minutus*									

TABLE 1—*Continued*

Taxon	Scientific name	ALD	AFP	LDH	PRL	OPS	TRYP	TYR	VASS	WNT-7
Mosquito fish	*Gambusia affinis*					●				
Guppy	*Poecilia reticulata*					●				
Zebrafish	*Brachydanio rerio*					●				
Carp	*Cyprinus carpio*				●					
Catfish	*Ictalurus punctatus*				●					
Trout	*Oncorhynchus mykiss*				●					
Tilipia	*Oreochromis mossambicus*				●					
Sea bass	*Dicentrarchus labrax*				●					
Dogfish	*Squalus acanthias*			●			●			
Alopias	*Alopias vulpinus*									●
Electric ray	*Torpedo marmorata*								●	
Skate	*Raja erinacea*					●				
Japanese lamprey	*Lampetra japonica*	●				●			●	
Lamprey	*Petromyzon marinus*		●	●		●	●			
Pacfic hagfish	*Eptatretus stouti*								●	●
Inshore hagfish	*Eptatretus burgeri*								●	

Note. Gene abbreviations: ALD, aldolase; AFP, α-fetoprotein; LDH, lactate dehydrogenase; PRL, prolactin; OPS, rhodopsin; TRYP, trysinogen; TYR, tyrosinase; VASS, vassopressin.

species were sequenced for the same gene. In these instances I combined the taxa.

Optimal species trees were inferred using GeneTree (Page, 1998). The optimality criterion was number of duplications. The constraint tree used (Fig. 7) enforced the monophyly of the Vertebrata, Amphibia (frogs and salamanders), Squamata (lizards and snakes), Aves (birds), and Actinopterygii (bony fish). Because of the very patchy coverage of the Actinopterygii within this clade, the constraint tree is moderately resolved. Other constraints involved terminal taxa which are obviously closely related (e.g., ducks and geese) but for which there were no sequences from the same gene available. Although the constraint tree imposes some structure on the result, note that it does not specify any relationships among the vertebrate classes.

An initial search was undertaken to find a good starting point for more extensive branch swapping. This preliminary search was performed using 20 different random-start trees. The heuristic search employed alternating nearest neighbor interchanges and subtree pruning and regrafting (Swofford *et al.,* 1996) rearrangements; with the constraint tree shown in Fig. 7 enforced, equally good trees were not retained. This last option was chosen to avoid spending time rearranging what would eventually prove to be suboptimal trees. As discussed above (see Fig. 3), even moderate numbers of missing sequences can result in huge numbers of equally good species trees. Rather than spend search time trapped in local "branch swapping eddies" (Novacek, 1992) due to missing data, I preferred to survey the search landscape (Charleston, 1995) more thoroughly from multiple starting points. The best tree from this initial search was rearranged with all equally good trees retained (up to GeneTree's limit of 1000 trees). Consensus trees were computed using COMPONENT 2.0 (Page, 1993).

RESULTS

Gene Trees

Trees for each of the nine genes are shown in Fig. 8. Aldolase sequences were taken from Nikoh *et al.* (1997), supplemented by additional sequences from the data banks. The vertebrate genes were rooted using a sequence from the amphioxus *Branchistoma belcheri.* α-Fetoprotein sequences were obtained from HOVERGEN and rooted using rat vitamin D-binding protein precursor. Lactate dehydrogenase sequences were obtained by combining those from Stock *et al.* (1997) and Mannen *et al.* (1997). The tree was rooted using the tunicate *Styela plicata.* Prolactin sequences from SWISS-PROT were rooted with the related gene somatotropin. The rhodopsin data comes from Yokohama (1994), supplemented with sequences from the data banks. The tree was rooted using invertebrate sequences. Roach *et al.* (1997) was the source for the trypsinogen sequences, which were rooted using the tunicate *Boltenia villosa.* Tyrosinase is among the genes used by Caspers *et al.* (1996) in their study of turtle relationships. The tree was rooted using "tyrosinase-related proteins" 1 and 2. The vassopressin gene family was extracted from the HOVERGEN data base, with further sequences obtained from SWISS-PROT. The tree was rooted using the vasopressin-related peptide Lys-conopressin from the pond snail *Lymnaea stagnalis* (van Kesteren *et al.,* 1995). Deeper branches within the tree show poor bootstrap values but the gene tree was included in this analysis because of the breadth of taxa included (including lungfish and hagfish). Wnt-7 is part of a larger gene family (Sidow, 1992). This gene was chosen for its relatively broad taxonomic coverage relative to other Wnt genes. The tree was rooted with echinoderm sequences.

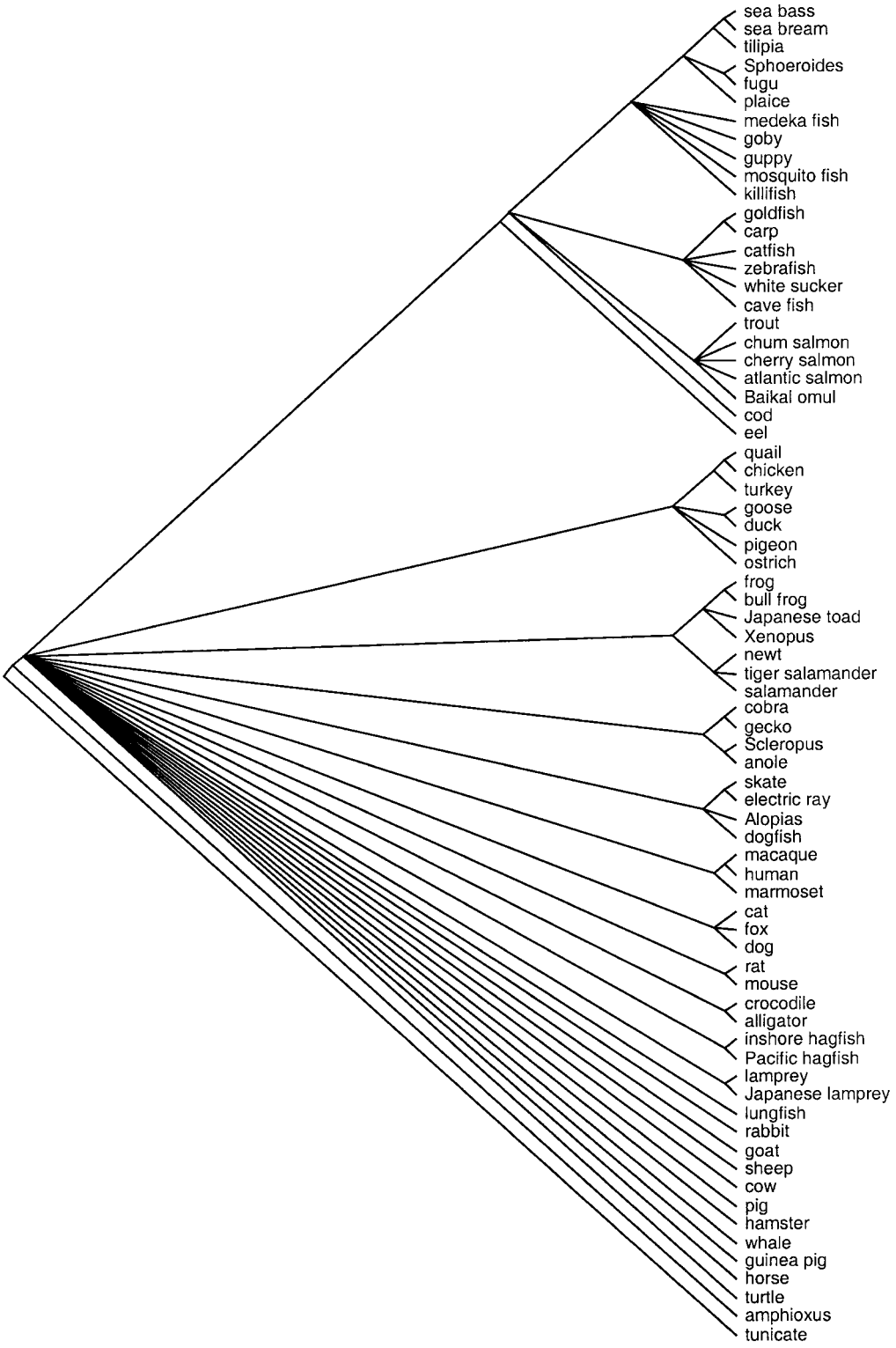

FIG. 7. Constraint tree used in this analysis. Only species trees compatible with this constraint were accepted. The tree specifies the monophyly of some vertebrate classes but does not specify any relationship among those classes.

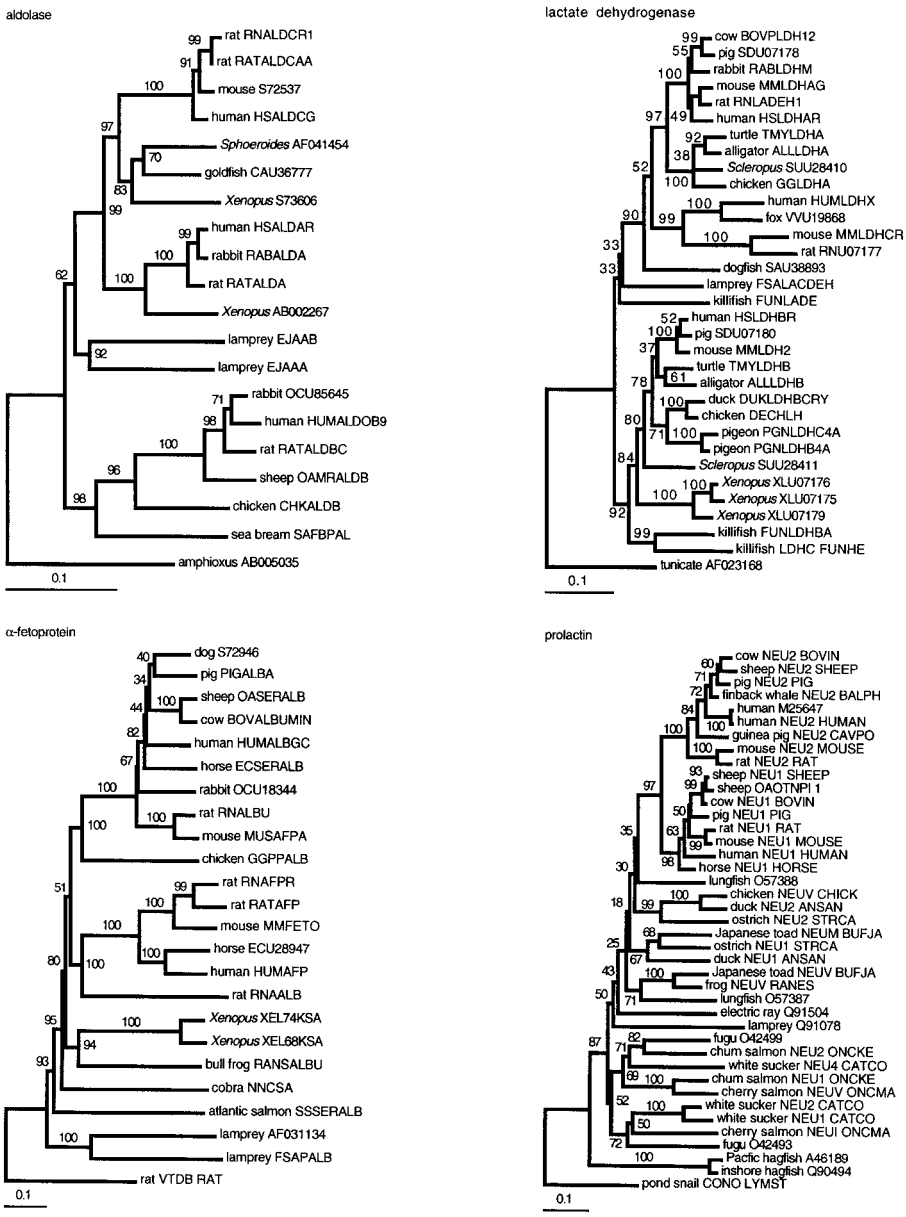

FIG. 8. Neighbor-joining trees for the vertebrate genes included in this study. Numbers above branches are bootstrap values. Scale bar represents 0.1 amino acid replacements per amino acid site.

Species Trees

The initial search found 2 trees with 87 duplications. Rearranging these trees produced 1000 trees with the same cost. The strict consensus of these trees is shown in Fig. 9. It should be emphasized that the topology of this species tree depends entirely on the topology of the 9 gene trees (and the constraint tree); no reference is made to the underlying sequence data. Much of the lack of resolution in the consensus tree is due to lack of shared information among the gene trees (see Fig. 3), rather than conflicts among the gene trees.

The species tree shows a basal split between hagfish and the rest of the vertebrates, with lampreys as the sister taxon to the Gnathostomes (jawed vertebrates). Within the Gnathostomes the basal split is between chondrichthyans (sharks and rays) and actinopterygians (ray-finned fish) on one side and lungfish and tetrapods on the other. Because of the poor taxonomic overlap in sequences from actinopterygians, the constraint tree imposed considerable structure on this clade, with the relationships among only the Gadiformes, Salmoniformes, Ostariophysi, and Acanthopterygii left unspecified. The consensus tree groups the Acanthopterygii and the Salmoniformes together. Rela-

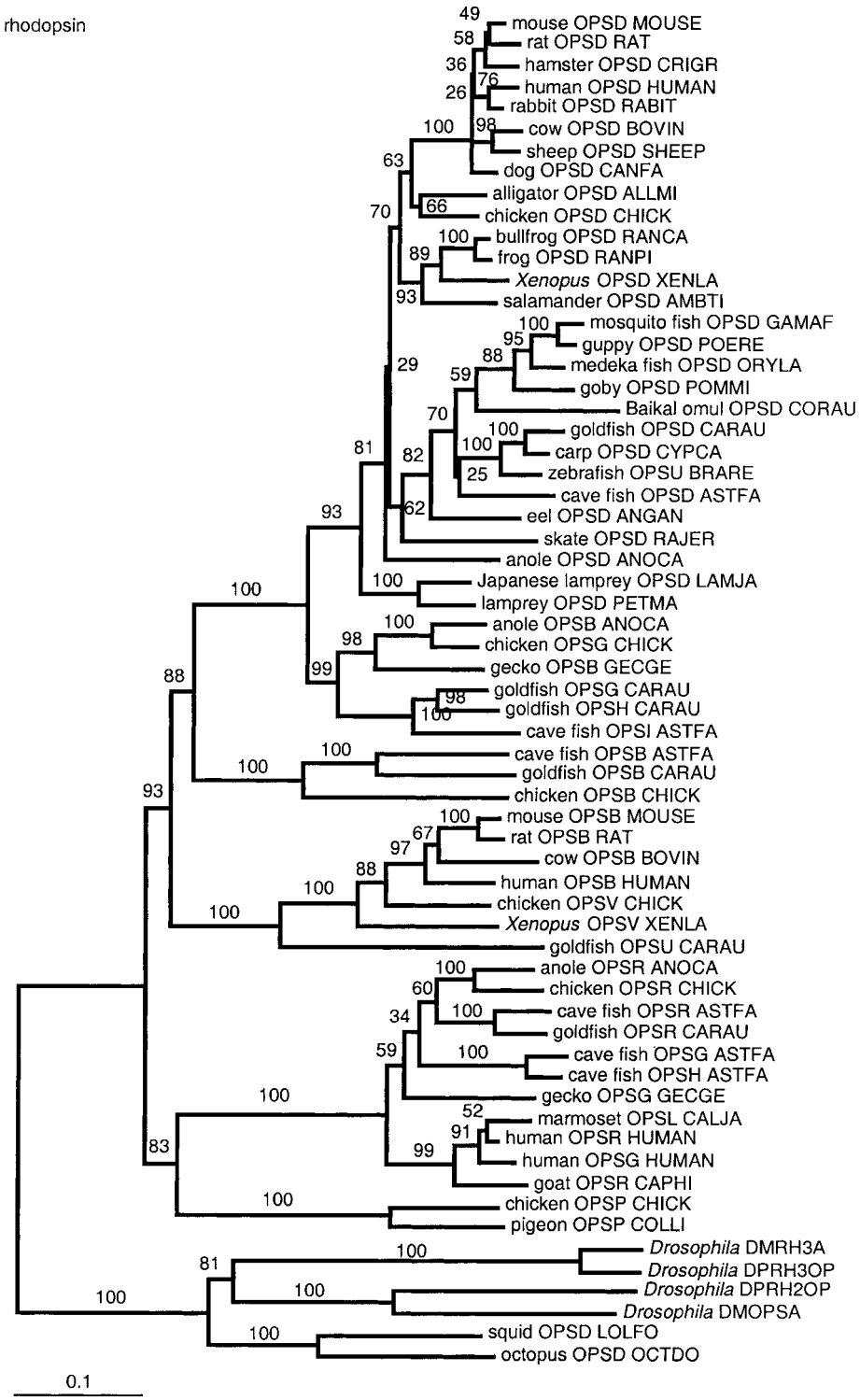

FIG. 8.—*Continued*

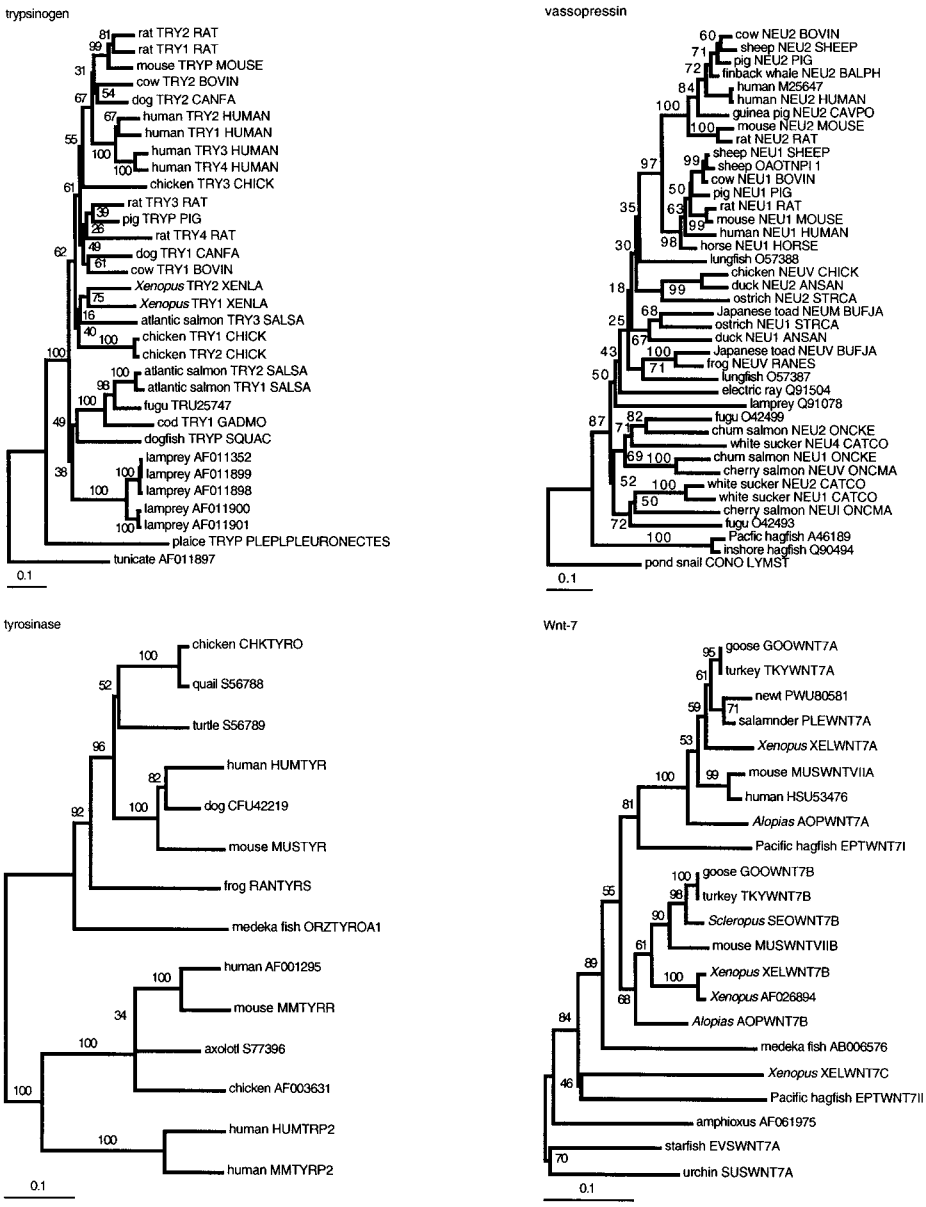

FIG. 8—*Continued*

tionships between lungfish, amphibians, and amniotes are unresolved. The amniotes are divided into mammals and the diapsids (turtles, crocodilians, snakes + lizards, and birds). The four diapsid groups form an unresolved clade. Mammalian relationships are reasonably resolved, with myomorph rodents (rats, mice, and their relatives) being basal to the remaining sampled placental mammals.

Duplications

The numbers of duplications for each gene are listed in Table 2. The duplications are divided into those that are based on direct physical evidence and those that are inferred from incongruence between gene and species trees. For the latter, the total number of nni's required to "undo" the duplications (Fig. 6) and the mean cost of those nni's in terms of bootstrap values are listed. The degree of bootstrap support for individual inferred duplications spans the range from very weak to values of 100% (Fig. 10). This suggests that some inferred duplications are likely to be artifacts of erroneous gene trees. For example, the apparent duplications among mammalian prolactin genes have either low or very weak bootstrap support, suggesting that the sequence data support alternative gene trees that do not require duplications. In contrast, the duplications inferred for aldolase have strong (>80%) bootstrap support.

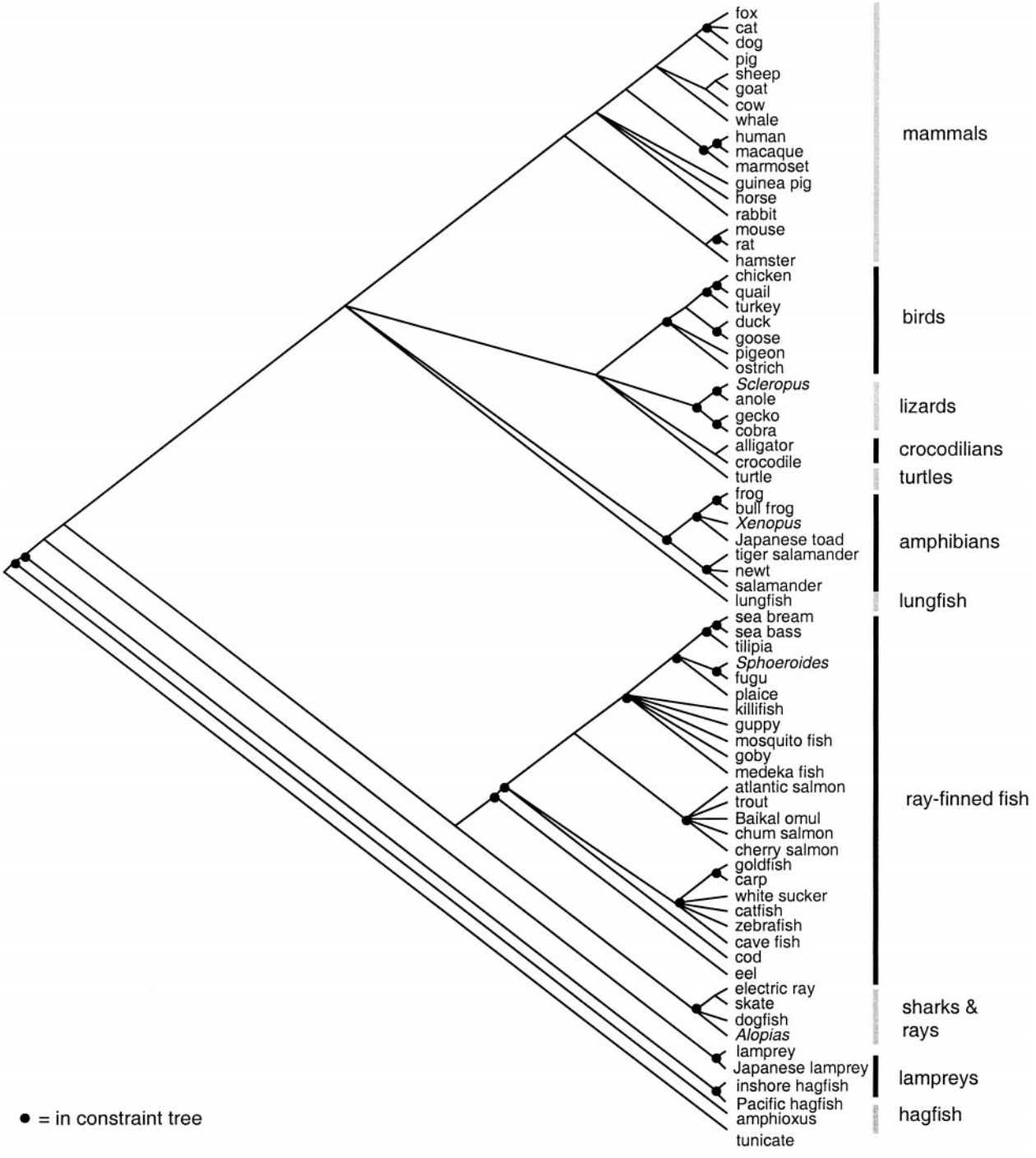

FIG. 9. Strict consensus tree of 1000 equally parsimonious species trees for the gene trees shown in Fig. 8. Nodes that were defined in the constraint tree (Fig. 7b) are indicated by ●.

DISCUSSION

Vertebrate Phylogeny

Although the broad outlines of vertebrate relationships are generally agreed, some aspects are subjects of considerable debate (Janvier, 1998; Patterson *et al.,* 1993). Relationships among major vertebrate clades inferred from the nine gene trees (Fig. 11) agree in many respects with the currently accepted view of vertebrate phylogeny. The paraphyly of the agnatha (hagfishes and lampreys) agrees with Forey and Janvier's (1993) morphological study and with Rasmussen *et al.*'s (1998) analysis of complete mitochondrial genomes, although not with trees inferred from nuclear

TABLE 2

Number of Gene Duplications for Each Gene

Gene	Duplications			nni	Mean bootstrap
	Observed	Inferred	Total		
Aldolase	4	2	6	3	0.84
α-Fetoprotein	5	2	7	4	0.78
Lactate dehydrogenase	6	5	11	6	0.54
Prolactin	4	4	8	5	0.43
Rhodopsin	9	7	16	9	0.54
Trypsinogen	16	3	19	3	0.24
Tyrosinase	0	0	0	0	—
Vassopressin	8	6	14	7	0.51
Wnt-7	3	3	6	4	0.61

Note. Duplications are divided into those for which there is the direct evidence of multiple sequences from the same taxon ("observed") and duplications inferred from mismatches between gene and species tree ("inferred"). The total number of nni's required to make the gene tree congruent with the species tree and the mean bootstrap value for the edges involved in these nni's (see Fig. 6) are listed.

ribosomal genes (Mallatt and Sullivan, 1998; Stock and Whitt, 1992). The grouping of sharks and teleost fish is unconventional; the current consensus is that teleosts are more closely related to lungfish and tetrapods than to chondrichthyians. None of the genes (lactate dehydrogenase, rhodopsin, trypsinogen, Wnt-7) that have both chondrichthyian and teleost sequences support this conventional relationship. Interestingly, trees inferred from complete mitochondrial genomes place chondrichthyians within the teleosts (Rasmussen and Arnason, 1999).

The close relationship between lungfish and tetrapods is supported by reconciled tree analysis, although the monophyly of the tetrapods is neither supported

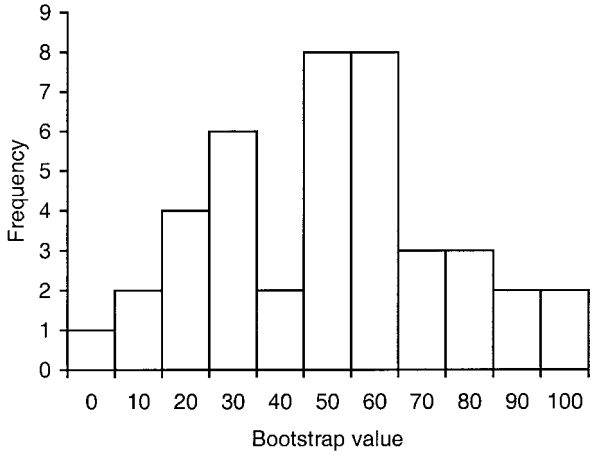

FIG. 10. Distribution of bootstrap values for each edge across which a nni must be made to make the gene trees shown in Fig. 8 congruent with the species tree shown in Fig. 9. The mean bootstrap value for each gene is shown in Table 2.

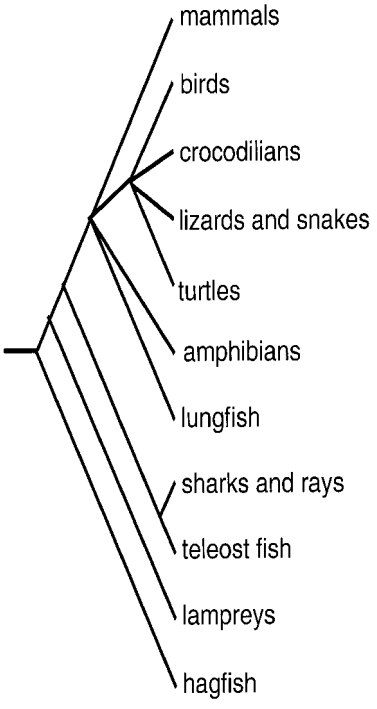

FIG. 11. Summary of the relationships among major vertebrate groups suggested by the reconciled trees for nine gene families.

nor contradicted. The uncertainty in lungfish relationships reflects the contradictory evidence offered by the two genes in this study for which lungfish have been sequenced. Lungfish prolactin is sister to a monophyletic clade of tetrapod prolactins, whereas the two lungfish vassopressin sequences group with mammals and amphibians. The grouping of turtles with other diapsids agrees with recent morphological (Rieppel and deBraga, 1996) and molecular (Caspers *et al.,* 1996) studies. This is not surprising as two of the genes included here (prolactin and tyrosinase) were also studied by Caspers *et al.* (1996). Although the widely accepted sister grouping of birds and crocodilians is supported by prolactin and rhodopsin, this relationship is contradicted by lactate dehydrogenase; consequently, the strict consensus tree is unresolved for these taxa. Mammalian phylogeny is somewhat uncertain (de Jong, 1998). The species relationships recovered here are not unreasonable for most taxa, although the grouping of carnivores and the pig is unorthodox.

Limitations and Future Directions

Perhaps the two greatest limitations of the method used here are its reliance on fully resolved gene trees and its inability to distinguish among nodes within a tree based on their degree of support—weakly supported nodes may have as much influence on the result as those that are much more robust. This problem could be addressed in a number of ways, which are currently being investigated. One solution would be to incorpo-

rate some measure of node support, such as bootstrap values (Felsenstein, 1985). In this study I have attempted to measure the relative support for inferred duplications using bootstrap values and nni's (Table 2 and Fig. 10), given a species tree inferred under the assumption that all nodes in the gene trees have equal weight. While this approach can be used to identify weakly supported (and hence possibly erroneous) gene duplications at the end of the analysis, it could be improved by taking the bootstrap values into account during the search for the optimal species tree, such that duplications and associated losses inferred from weakly supported nodes would be down-weighted. In this way poorly supported gene tree nodes would have less influence on the inferred species tree.

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 (Page, 1996; Sanderson, 1989). The cost of a given species tree would be computed for all gene trees within the confidence interval, and one value (such as the minimum cost for all gene trees) would be assigned to the species tree. Indeed, the fit between gene and species tree could be used as an additional criterion for selecting among competing gene trees that cannot be discriminated amongst on the basis of nucleotide or protein sequence data alone. Goodman *et al.* (1979) suggested such a strategy in their pioneering work on reconciled trees in which they preferred less-parsimonious hemoglobin gene trees which had better fit to accepted species trees than most-parsimonious trees that required more duplications and losses. Their approach assigned to a 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. A likelihood framework may provide one solution to this problem, as has been suggested in the context of coalescence models by Maddison (1997). However, while reasonable models of nucleotide substitution exist, there are none for gene duplication. Furthermore, any model would need to incorporate the extreme sampling bias that exists in the sequence data bases (and hence many gene "losses" are sampling artifacts).

The optimality criterion being minimized in this study is the number of gene duplications. A key assumption made is that gene duplications in different genes are mutually independent and hence can be minimized independently. Given that vertebrate genes have been duplicated in blocks of various sizes, up to entire genomes (Holland *et al.,* 1994; Pébusque *et al.,* 1998), it might be more appropriate to minimize episodes of multiple duplication, rather than the individual duplications themselves. However, this greatly increases the computational complexity of the problem. Minimizing individual gene duplications can be done in linear time (Eulenstein, 1997), whereas minimizing episodes of multiple duplication is NP-complete (Fellows *et al.,* 1998). Even if the frequency of block duplications has been overestimated (Hughes, 1998), there is a need to develop techniques that incorporate this process.

Given the reasonable success that reconciled trees had in recovering vertebrate phylogeny from a small number of gene trees of variable quality, I think the method merits further study and application. Other taxa that are good candidates for study are angiosperms and eukaryotes as a whole. Another obvious extension is to apply the method to much larger sets of gene trees. In this study I have restricted my attention to a few genes that have good coverage of the vertebrate classes. Given the large (and ever increasing) number of gene families available for analysis, there is considerable scope for automating the analysis. For example, it would be useful to be able to extract gene trees from data bases like HOVERGEN and input this directly into software such as GeneTree. Hence, it would be possible, in principle, to obtain the best estimates of species phylogeny based on simultaneous analysis of thousands of gene families.

ACKNOWLEDGMENTS

Access to the HOVERGEN data base was kindly provided by the UK Human Genome Mapping Project Resource Centre, Hinxton, Cambridge. I thank Michael Fellows and Louxin Zhang for sending preprints of their work.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Bairoch, A., and Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TREMBL. *Nucleic Acids Res.* **25:** 31–36.

Brown, J. R. (1996). Preparing for the flood: Evolutionary biology in the age of genomics. *Trends Ecol. Evol.* **11:** 510–513.

Cao, Y., Adachi, J., Janke, A., Pääbo, S., and Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. *J. Mol. Evol.* **39:** 519–527.

Caspers, G.-J., Reinders, G.-J., Leunissen, J. A. M., Wattel, J., and de Jong, W. W. (1996). Protein sequences indicate that turtles branched off from the amniote tree after mammals. *J. Mol. Evol.* **42:** 580–586.

Charleston, M. A. (1995). Towards a characterization of landscapes of combinatorial optimisation problems, with special reference to the phylogeny problem. *J. Comput. Biol.* **2:** 439–450.

Constantinescu, M., and Sankoff, D. (1986). Tree enumeration modulo a consensus. *J. Classif.* **3:** 349–356.

de Jong, W. W. (1998). Molecules remodel the mammalian tree. *Trends Ecol. Evol.* **13:** 270–275.

D'Erchia, A., Gissi, C., Pesole, G., Saccone, C., and Arnason, U. (1996). The guinea-pig is not a rodent. *Nature* **381:** 597–600.

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

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

Eulenstein, O., Mirkin, B., and Vingron, M. (1997). Comparison of annotating duplications, tree mapping, and copying as methods to compare gene trees with species trees. *In* "Mathematical Hierarchies in Biology" (B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, Eds.), Vol. 37, pp. 71–93. Am. Math. Soc. Providence.

Fellows, M., Hallett, M., and Stege, U. (1998). On the multiple gene duplication problem. *Ninth Int. Symp. Algorithms and Computation,* Taejon, Korea.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783–791.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19:** 99–113.

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

Forey, P., and Janvier, P. (1993). Agnathans and the origin of jawed vertebrates. *Nature* **361:** 129–134.

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. *Syst. Zool.* **28:** 132–168.

Graur, D., Duret, L., and Gouy, M. (1996). Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). *Nature* **379:** 333–335.

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

Halanych, K. M. (1998). Lagomorphs misplaced by more characters and fewer taxa. *Syst. Biol.* **47:** 138–146.

Holland, P. H., Garcia-Fernàndez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development* **1994 (Suppl.):** 125–133.

Hughes, A. L. (1998). Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* **15:** 854–870.

Janvier, P. (1998). A cold look at odd vertebrate phylogenies. *J. Mol. Evol.* **46:** 375–377.

Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in terms of duplications and losses. *Proc. 2nd Int. Conf. Computational Molecular Biology,* New York.

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

Mallatt, J., and Sullivan, J. (1998). 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.* **15:** 1706–1718.

Mannen, H., Tsoi, S. C.-M., Krushkal, J. S., Li, W.-H., and Li, S. S.-L. (1997). The cDNA cloning and molecular evolution of reptile and pigeon lactate dehydrogenase isozymes. *Mol. Biol. Evol.* **14:** 1081–1087.

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

Naylor, G. J. P., and Brown, W. M. (1998). Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47:** 61–76.

Nikoh, N., Iwabe, N., Kuma, K., Ohno, M., Sugiyama, T., Watanabe, Y., Yasui, K., Shi-cui, Z., Hori, K., Shimura, Y., and Miyata, T. (1997). An estimate of divergence time of parazoa and eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. *J. Mol. Evol.* **45:** 97–106.

Novacek, M. J. (1992). Fossils, topologies, missing data, and the higher level phylogeny of eutherian mammals. *Syst. Biol.* **41:** 58–73.

Page, R. D. M. (1993). "COMPONENT, Tree comparison software for Microsoft Windows," Version 2.0, The Natural History Museum, London.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.* **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., and Charleston, M. A. (1997a). From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* **7:** 231–240.

Page, R. D. M., and Charleston, M. A. (1997b). Reconciled trees and incongruent gene and species trees. *In* "Mathematical Hierarchies in Biology" (B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, Eds.), Vol. 37, pp. 57–70. Am. Math. Soc., Providence.

Patterson, C., Williams, D. M., and Humphries, C. J. (1993). Congruence between molecular and morphological phylogenies. *Annu. Rev. Ecol. Syst.* **24:** 153–188.

Pébusque, M.-J., Coulier, F., Birnbaum, D., and Pontarotti, P. (1998). Ancient large-scale genome duplications: Phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* **15:** 1145–1159.

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

Rasmussen, A.-S., Janke, A., and Arnason, U. (1998). The mitochondrial DNA molecule of the hagfish (*Myxine glutinosa*) and vertebrate phylogeny. *J. Mol. Evol.* **46:** 382–388.

Rieppel, O., and deBraga, M. (1996). Turtles as diapsid reptiles. *Nature* **384:** 453–456.

Roach, J. C., Wang, K., Gan, L., and Hood, L. (1997). The molecular evolution of the vertebrate trypsinogens. *J. Mol. Evol.* **45:** 640–652.

Sanderson, M. J. (1989). Confidence limits on phylogenies: The bootstrap revisited. *Cladistics* **5:** 113–129.

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

Sidow, A. (1992). Diversification of the *Wnt* gene family on the ancestral lineage of vertebrates. *Proc. Natl. Acad. Sci. USA* **89:** 5098–5102.

Slowinksi, J., and Page, R. D. M. (1999). How should species trees be inferred from sequence data? *Syst. Biol.,* in press.

Stock, D. W., Quattro, J. M., Whitt, G. S., and Powers, D. A. (1997). Lactate dehydrogenase (LDH) gene duplication during chordate evolution: The cDNA sequence of the LDH of the tunicate *Styela plicata. Mol. Biol. Evol.* **14:** 1273–1284.

Stock, D. W., and Whitt, G. S. (1992). Evidence from 18S ribosomal-RNA sequences that lampreys and hagfishes form a natural group. *Science* **257:** 787–789.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. *In* "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), 2nd ed., pp. 407–514. Sinauer, Sunderland, MA.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The ClustalX windows interface: Flexible

strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

van Kesteren, R. E., Smit, A. B., de Lange, R. P., Kits, K. S., van Golen, F. A., van der Schors, R. C., de With, N. D., Burke, J. F., and Geraerts, W. P. (1995). Structural and functional evolution of the vasopressin/oxytocin superfamily: Vasopressin-related conopressin is the only member present in *Lymnaea,* and is involved in the control of sexual behavior. *J. Neurosci.* **15:** 5989–5998.

Waterman, M. S., and Smith, T. F. (1978). On the similarity of dendrograms. *J. Theor. Biol.* **73:** 789–800.

Yokoyama, S. (1994). Gene duplications and evolution of the short wavelength-sensitive visual pigments in vertebrates. *Mol. Biol. Evol.* **11:** 32–39.

Zardoya, R., and Meyer, A. (1997). The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (*Latimeria chalumnae*). *Genetics* **146:** 995–1010.