MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 7, No. 2, APRIL, pp. 231–240, 1997
ARTICLE NO. FY960390

# From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem

Roderic D. M. Page and Michael A. Charleston

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

Received August 7, 1996

**The processes of gene duplication, loss, and lineage sorting can result in incongruence between the phylogenies of genes and those of species. This incongruence complicates the task of inferring the latter from the former. We describe the use of reconciled trees to reconstruct the history of a gene tree with respect to a species tree. Reconciled trees allow the history of the gene tree to be visualized and also quantify the relationship between the two trees. The cost of a reconciled tree is the total number of duplications and gene losses required to reconcile a gene tree with its species tree. We describe the use of heuristic searches to find the species tree which yields the reconciled tree with the lowest cost. This method can be used to infer species trees from one or more gene trees.**

## INTRODUCTION

The oldest use for phylogenies of genes is inferring organismal phylogeny (Fitch, 1996). The implicit assumption made by most of this work is that gene trees are isomorphic with species trees—the former can be converted into the latter merely by substituting the name of the sequence with the name of the organism from which the sequence was obtained. As sequence data has accumulated it has come to be appreciated that not only does this kind of data present new and challenging problems of analysis, but that the relationship between gene trees and species trees may be more complex than a simple one-to-one correspondence (Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991; Doyle, 1992).

This has several implications, not least for the debate about whether nucleotide sequence data should be analyzed independently of, or jointly with, morphological data. One popular view advocated by, for example, Kluge (1989), is that sequence and morphological data should be combined in a single analysis ("total evidence"). This position is appealing, but runs the risk of confounding characters of genes with characters of organisms. Gene trees need not faithfully reflect species trees for a variety of reasons, including gene duplication (resulting in paralogous genes), lineage sorting, and horizontal transfer. These causes of incongruence between gene and species trees are distinct from the causes of homoplasy in sequence data (e.g., multiple substitutions).

For example, suppose the gene clade (*a*,*b*) in Fig. 1 is supported by 10 nucleotide sites, each with unique and unreversed substitutions. Optimizing these sequences on the organismal tree would require the changes at all these 10 sites to be interpreted as homoplasy as they would not fit perfectly on the organismal tree. However, this is entirely an artefact of treating nucleotide substitutions as characters of organisms rather than genes (Doyle, 1992). These 10 sites have not undergone multiple substitutions; rather the gene lineage they diagnose (*a*,*b*), has not tracked the species tree with absolute fidelity. We could postulate that, for example, gene clades (*a*,*b*) and (*c*,*d*) are paralogous, in which case the explanation of incongruence between gene and species tree requires an hypothesis of gene duplication and loss (see Fig. 2, below) rather than nucleotide substitution.

### *Gene Trees as Character Trees*

Recognition of the possibility of discordance between gene and species trees has led to proposals for treating genes (or whole genomes if they are linked, such as the mitochondrial genome) as single characters and the gene tree as a character state tree (Baum, 1992; Doyle, 1992). The "instinctive" (Rodrigo, 1993:635) response of some systematists to the hierarchical structure of gene trees has been to convert them into additive binary codes and analyze the resulting matrix using parsimony, following Brooks' (1981) approach to the analysis of host and parasite phylogenies (e.g., Baum, 1992; Doyle, 1992; Ragan, 1992). In this method nodes in the gene tree are represented as binary characters. For example, the gene tree shown in Fig. 1 can be represented as a set of binary codes (Table 1), where each code corresponds to an edge in the gene tree. While this begins to shift attention to the appropriate level, that

1055-7903/97 $25.00

is, the relationship between trees, rather than between characters and trees, the use of binary characters still reflects the powerful grip of the latter on systematists. Converting a tree into a suite of binary characters is essentially a device to render the problem in more familiar form, rather than a genuine solution to the problem itself. As Rodrigo (1993) points out, it is not at all obvious how to interpret "homoplasy" in the binary coded representation of a tree. In Table 1 binary code 5 will require an additional step to fit onto the species tree in Fig. 1. Just what biological process this extra "step" represents is not clear. Similar problems of biological interpretation beset application of this approach to host–parasite systems, where it was originally developed (see Page, 1994).

It would therefore be desirable to have a method for comparing gene and species trees that was biologically interpretable. The purpose of this paper is to outline the application of reconciled trees (Goodman *et al.*, 1979b; Page, 1994) to this problem, and to demonstrate how reconciled trees can be used to interpret the history of a gene by tracing its lineage as it ramifies through the species tree, and to employ reconciled trees to infer species trees. The method described here has been implemented in the computer program GeneTree, written by RDMP and available on the Internet from http://taxonomy.zoology.gla.ac.uk/rod/genetree.html. An earlier implementation is available in the program COMPONENT (Page, 1993a).

## RECONCILED TREES

### *Background*

The concept of a reconciled tree was first introduced by Goodman *et al.* (1979b) to account for discordance between mammalian haemoglobin gene trees and previously accepted notions of mammal phylogeny. Much the same idea was suggested independently by Nelson and Platnick (1981) in the context of biogeography. That similar methods have been developed independently in the fields of molecular systematics, host–parasite cospeciation, and biogeography suggests that these fields are all essentially instances of the same problem, and that instances of contained and containing trees (Maddison, submitted) recur at various hierarchical levels within biology, from the relationship between an organism and its genes, through to the relationships between geological and organismal differentiation (Page, 1993b).

**TABLE 1**

**Binary Coding of the Gene Tree in Fig. 1**

| | Code | | | | | | |
|---|---|---|---|---|---|---|---|
| Gene | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| *a* | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| *b* | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| *c* | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| *d* | 0 | 0 | 0 | 1 | 0 | 1 | 1 |

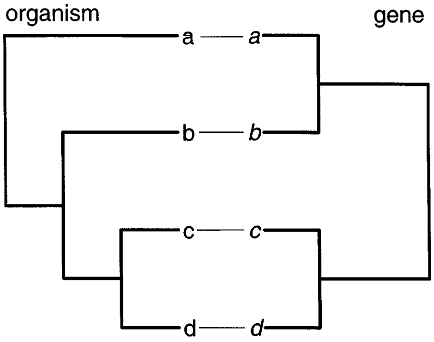

**FIG. 1.** Incongruent organism and gene phylogenies.

Page (1994) formalized reconciled trees and described an algorithm for their construction based on the concept of maps between trees. Mirkin *et al.* (1995) developed an alternative formalism for embedding gene trees in species trees that yields the same result but without explicitly constructing a reconciled tree. They offered a proof that tree mapping and their method identified the same duplications, and furthermore that the "cost" of each individual duplication in terms of the number of gene losses it requires (see below) is minimal. Mirkin *et al.* (p. 504) criticized mapping trees for missing their "information gap concept" and for failing to compute the number of losses a duplication incurs. However this criticism does not apply to the reconciled tree algorithm (Page, 1994), which computes the same cost as Mirkin *et al.*'s method (Mirkin, personal communication) and clearly identifies where the postulated losses occurred. Arguably reconciled trees also provide a clearer visualization of the relationship between gene and species trees than Mirkin *et al.*'s annotations of the species tree (e.g., their Fig. 3). However, for large gene trees the latter may produce more manageable diagrams than the reconciled tree algorithm.

### *Examples*

If we have a species tree and a gene tree which are mutually incongruent (Fig. 1), and we are confident that both are correct for the species and genes, respectively, then we might ask under what circumstances could both be true. If we regard genes as "tracking" species, then we can embed the gene tree in the species tree. In Maddison's (submitted) terminology, the species tree "contains" the gene tree. The incongruence between these two trees can be explained by postulating a gene duplication that gave rise to two sets of paralogous genes, of which only four have survived to the present day (Fig. 2). Genes *a* and *b* are orthologous, as are *c* and *d*. Given the duplication δ at the base of the gene tree we would have expected to find two copies of

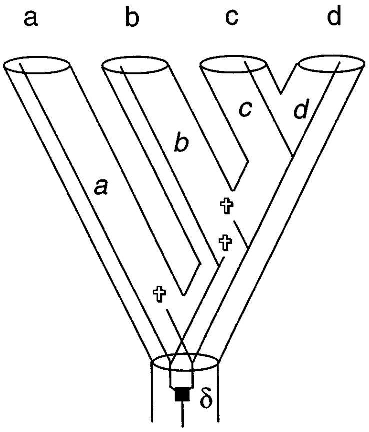

**FIG. 2.** The incongruence between the trees shown in Fig. 1 can be explained by hypothesizing a gene duplication (δ) at the base of the gene tree, with genes *a* and *b* being paralogous with genes *c* and *d*. The presence of only a single gene extant in each present day species requires postulating three gene losses (✞).

this gene in taxa a–d. The presence of only a single copy in each requires at least three independent gene losses.

Figure 3 shows the reconciled tree computed for the trees shown in Fig. 1. This tree can be thought of as the tree obtained by "unfolding" the gene tree embedded in the species tree in Fig. 2 and laying it flat on the page. This tree reconciles the incongruent gene and species trees by postulating that the observed gene tree is a relict of the larger gene tree that results from the gene duplication δ. This larger tree is the reconciled tree, and is the tree we would obtain if no gene loss or extinction occurred. Given that the gene tree is contained within the species tree, each of the two gene copies traces the species phylogeny exactly, hence the reconciled tree comprises two identical subtrees (a,(b,(c,d))). As before, given the duplication δ we would have expected two copies of the gene in each species. That we do not see these copies requires us to postulate a total of three losses, one from each of species a and b, and one from the ancestor of species c and d. Note that these genes may be present but as yet undetected (see later). The total number of events the reconciled tree postulates (one duplication plus three losses) is the "cost" of the tree, and can be written $c(G,T)$, where $G$ is the gene tree and $T$ is the species tree.

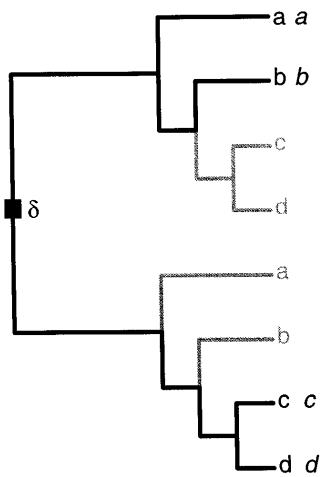

**FIG. 3.** Reconciled tree for the gene and species trees shown in Fig. 1. The tree has one gene duplication (δ) and three losses (represented by branches changing from black to gray).

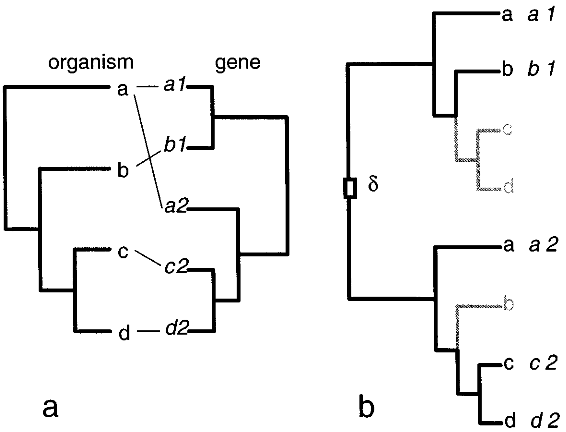

**FIG. 4.** (a) Organism and gene trees with one species containing two copies of a gene and (b) the corresponding reconciled tree. The reconciled tree has a cost of 3 (1 duplication and 2 losses).

Reconciled trees can be applied to any gene tree/species tree comparison where purely vertical transmission has occurred. Figure 4 shows a hypothetical example similar to that in Fig. 1 but where species a has retained both descendants of the gene duplication. As a consequence in this instance we have observational evidence for the duplication at the base of the gene tree, whereas the duplication in Fig. 3 is inferred solely on the basis of incongruence between gene and species trees. These two classes of duplication (hypothetical and observed) will be distinguished in this paper by shaded and open squares, respectively.

Note that although the duplications discussed so far are interpreted as actual gene duplications, similar reconciled trees may be obtained from orthologous sequences. For example, the gene tree shown in Fig. 4 may be for five alleles at the same locus. In this case the "duplication" represents not an actual duplication but a coalescence event between alleles which occurred within the common ancestor of the four species a–d, that is, independently of any cladogenesis of these species. This is the essential feature which gene duplications and this class of coalescence events share—their independence from species cladogenesis. They represent events where gene lineages arise within a species lineage such that the species contains more than one gene lineage (be they paralogous loci or orthologous alleles). The reconciled tree identifies these "duplications" and depicts their fate as they track their containing species. If the genes in Fig. 2 are alleles and not paralogs, then the three loss events depicted may be interpreted as either extinction of alleles or instances of lineage sorting.

### *Inferring Species Phylogeny from a Complex Gene Tree*

If we are only interested in the history of the gene relative to the organisms—for example, documenting how many gene duplications took place—then we need go no further, for as shown above, the reconciled tree provides this information. However, if we are interested in what the gene tree can tell us about species relationships then we have an additional problem to solve. For any gene tree and any species tree we can obtain a reconciled tree that will have a particular cost associated with it. We can use this cost as an optimality criterion for choosing a species tree or trees. Put more precisely, given a gene tree $G$, we can estimate the actual species phylogeny $T^*$ by finding that tree $\hat{T}$ which when reconciled with G has the lowest cost. Finding $\hat{T}$ requires searching the set of all possible trees.

To illustrate, suppose we had the gene tree shown in Fig. 4 but were ignorant of the species tree. By computing the cost of each of the 15 possible trees for four species we can identify the species tree (or trees) in which the gene tree can be embedded with the least cost. Figure 5 shows the distribution of $c(G,T)$ for the 15 trees, of which three have the minimal cost. One of these is shown in Fig. 4; the remaining two are the other possible resolutions of (a,b,(c,d)).

As is well known (Felsenstein, 1978) the number of evolutionary trees increases at an alarming rate with increasing numbers of species, making the exhaustive search just undertaken prohibitive in all but the smallest cases. Hence, for gene trees of the size typically reported in the literature we will have to rely on heuristics. The approach used here (and implemented in the program GENETREE) is to search tree space using the well-known tree perturbations of nearest neighbor interchanges and/or cut and paste (also known as "subtree pruning and regrafting") (e.g., Swofford and Olsen, 1990). An initial starting species tree is chosen, typically at random. Its cost is computed by reconciling it with the gene tree. The start tree is then perturbed in search of a better tree. If one is found, the search continues from the better tree, repeating until no perturbation produces an improvement. Random starting trees tend to be poor estimates of the optimal host tree, but using multiple random trees gives information on the landscape for the problem (Charleston, 1995). In particular, convergence on the same cost value from multiple starting points suggest the hypothesis that the cost value may be optimal. Multiple starting points also increase the chance of detecting multiple, equally good (or near equally good) solutions.

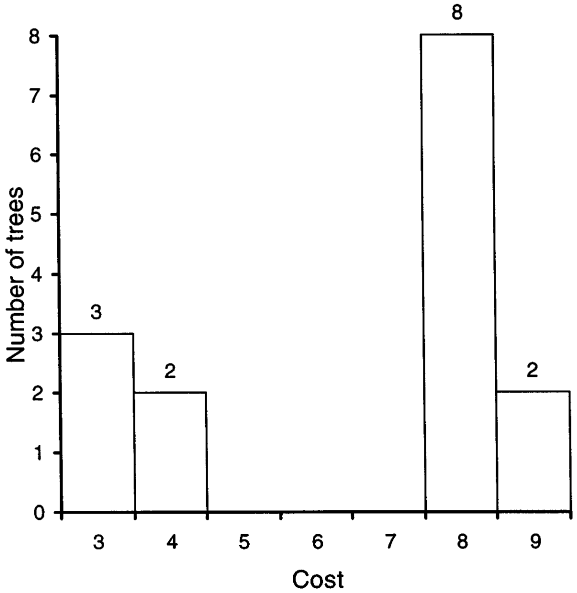

**FIG. 5.** Distribution of the cost of reconciling the gene tree shown in Fig. 4 with all 15 possible rooted trees for four species. Three trees have minimal cost.

### *Inference from Multiple Gene Trees*

The method described in the previous section can readily be generalized to more than one gene. Given $n$ gene trees $G_1, G_2, \ldots, G_n$, and species tree $T$, the cost of reconciling all $n$ gene trees is simply the sum of the costs of reconciling the individual gene trees, $c(G_1, T) + c(G_2,T) + \ldots + c(G_n, T)$. Note that this is analogous to parsimony analysis of individual characters (e.g., nucleotide sites) where the total length of the tree is the sum of the minimum number of changes required for each individual character to evolve on the tree. It is also analogous to cladistic biogeographers' use of multiple taxon cladograms to infer area relationships (Page, in press) or to parasitologists' inference of host phylogeny from multiple parasite phylogenies (Brooks, 1981).

### *Optimality Criteria*

The optimality criterion used in this paper is the total number of evolutionary events ("duplications" and "losses") required to reconcile a gene tree with its species tree. Duplications are postulated whenever we have multiple copies of a gene in the same taxon or when the gene and species trees are incongruent. Losses are a consequence of postulating duplications and are postulated when a taxon lacks a gene lineage the reconciled tree predicts it should have. For example, given the reconciled tree in Fig. 4b, taxa c and d should each have two copies of the gene. That they have only one copy ($c$ 2 and $d$ 2, respectively) requires two hypotheses of gene loss.

However, an alternative interpretation is that the missing genes are actually present but as yet undetected. Given uneven sampling of taxa (one has only to think of the preponderance in sequence data bases of human and rodent genes among vertebrates) it is entirely likely that some, if not most, "losses" are only apparent, not real. This has implications for choosing optimal species trees, because minimizing both duplications and gene loss may group together species on the basis of absence of genes due to inadequate sampling, rather than actual evolutionary relationship. This problem can be illustrated by returning to the example shown in Fig. 4. If we consider each subset of orthologous genes separately, gene 1 implies the species tree (a,b) (and is hence uninformative), and gene 2 specifies the species tree (a,(c,d)). There are five trees for species a–d that are compatible with both of these two subtrees, each corresponding to a different placement of species b on the subtree (a,(c,d)) (Fig. 6). However, if we count both duplications and losses together only three of these trees are optimal (see Fig. 5). The two remaining trees each posit an additional loss and hence require a total of four events (Fig. 7).

The grouping of species c and d in trees 1, 4, and 5 (Fig. 6) is based on the shared absence of gene 1 from both species, which can be most parsimoniously accounted for if c and d are posited as sister taxa. However, if the absence was due to sampling rather than genuine absence we would be grouping those two species on the basis of negative evidence. If sampling is suspected to be poor or uneven it may be more defensible to use duplications alone as the optimality criterion. Because, as noted above, duplications can only be parsimoniously inferred from the presence of multiple copies of a gene, or incongruence between gene and species trees, they can not be inferred from negative evidence (although if sampling is poor their number may be underestimated). All five trees shown in Fig. 6 have a single duplication and hence if we ignore losses all would be considered as equally good candidates for the actual species tree.

## RECONSTRUCTING THE HISTORY OF A GENE FAMILY

A straightforward use of reconciled trees is to visualize the history of a gene embedded within an organismal phylogeny. For example, consider the eukaryote P type ATPase Na$^+$-K$^+$ ion pump genes, one of 25 gene families analysed by Iwabe *et al.* (1996) in a study of tissue evolution. In this example the presence of multiple copies in the same gene is *a priori* evidence for gene duplications, as it was in the example shown in Fig. 4. The reconciled tree (Fig. 8) for Na$^+$-K$^+$ indicates the three duplications required to explain the history of this gene and locates them on the gene tree.

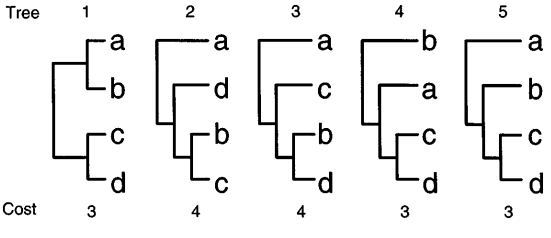

**FIG. 6.** Five of the 15 possible species trees for the taxa in Fig. 4a. When reconciled with the gene tree in Fig. 4a, each species tree requires a single duplication. However, trees 1, 4, and 5 require two losses, whereas trees 2 and 3 require three losses.

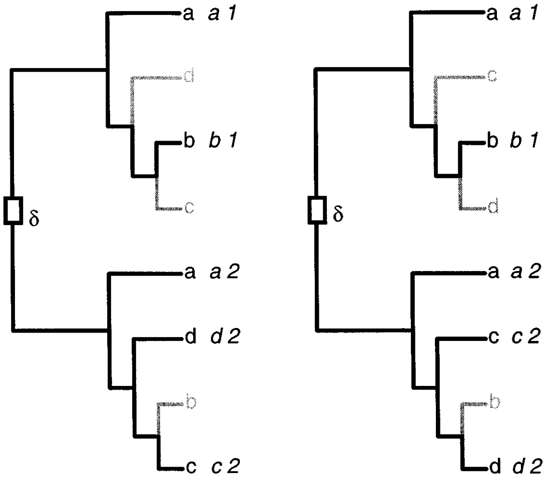

**FIG. 7.** Reconciled trees for trees 2 and 3 in Fig. 6 and the gene tree shown in Fig. 4. Both trees require three loss events and hence are less parsimonious than trees 1, 4, and 5 in Fig. 6 (compare these trees with the reconciled tree for tree 5 shown in Fig. 4b).

Iwabe *et al.* were interested in the temporal distribution of gene duplications, and noted that duplication $\delta_1$ occurred within invertebrates, duplication $\delta_2$ prior to the divergence of the vertebrates, and that duplication $\delta_3$ was of uncertain age. The reconciled tree locates $\delta_3$ after the split between amphibia and the amniotes, but prior to the bird–mammal split. This placement is the most parsimonious interpretation as it requires only single absences of the Na$^+$-K$^+$ pump gene from *Xenopus* and *Catastomus.* However, given that this likely reflects lack of sampling, we consider the placement to be based on absence of evidence, rather than evidence of absence.

In one sense this example is straightforward because each duplication is supported by the physical evidence of having multiple copies of the same gene present in a single species. Given that humans and chickens each have three Na$^+$-K$^+$ genes, *a priori* we require at least two duplications ($\delta_2$ and $\delta_3$). Further, one of Na$^+$-K$^+$ genes from *Artemia* is more closely related to that from *Drosophila* than it is to the other *Artemia* Na$^+$-K$^+$ gene, which implies that we require one more duplication ($\delta_1$). Constructing the reconciled tree serves to confirm that three duplications are all that are required.

The next example concerns the mammalian interleukin-1 (IL-1) gene (Hughes, 1994). Mammalian phylogeny is somewhat uncertain (see recent review by Allard *et al.*, 1996), but here we follow Hughes and use the tree supported by the best sampled interleukin locus (IL-1α). The reconciled tree (Fig. 9) identifies four duplications, three of which are supported by the presence of more than one gene in the same organism. However, unlike the $Na^+$-$K^+$ pump gene presented earlier, one duplication ($\delta_3$) is hypothetical and is inferred solely from the incongruence between the mammalian tree and the phylogeny of the IL-β genes. Specifically, mouse IL-1β is more closely related to human IL-β than are bovine and sheep IL-β. This indicates that if the mammalian phylogeny used to construct Fig. 9 is correct then the IL-β genes are not orthologous (Hughes, 1994; p. 10).

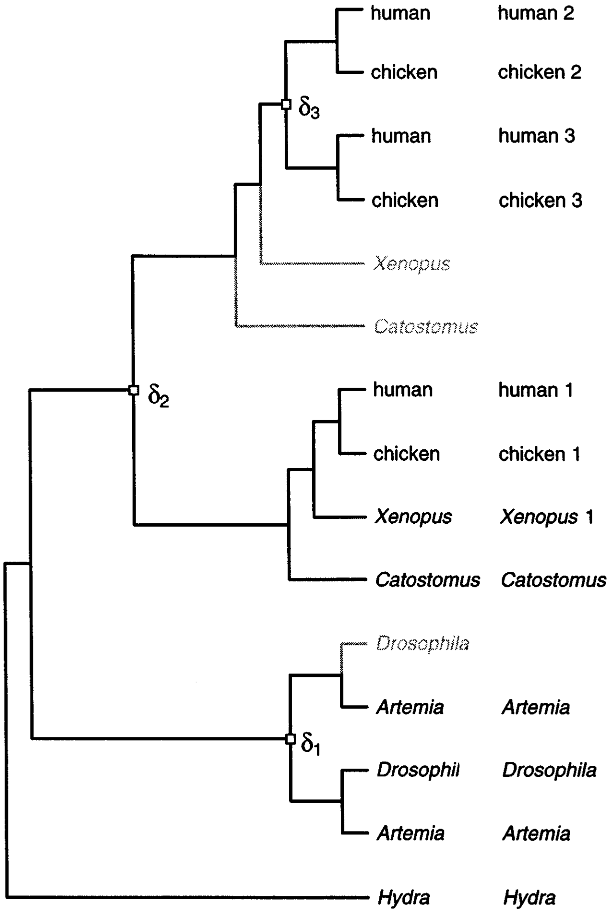

**FIG. 8.** Reconciled tree for $Na^+$-$K^+$ ion pump genes showing three gene duplications ($\delta_1$–$\delta_3$). (Gene tree taken from Iwabe *et al.*, 1996: Fig. 1a.)

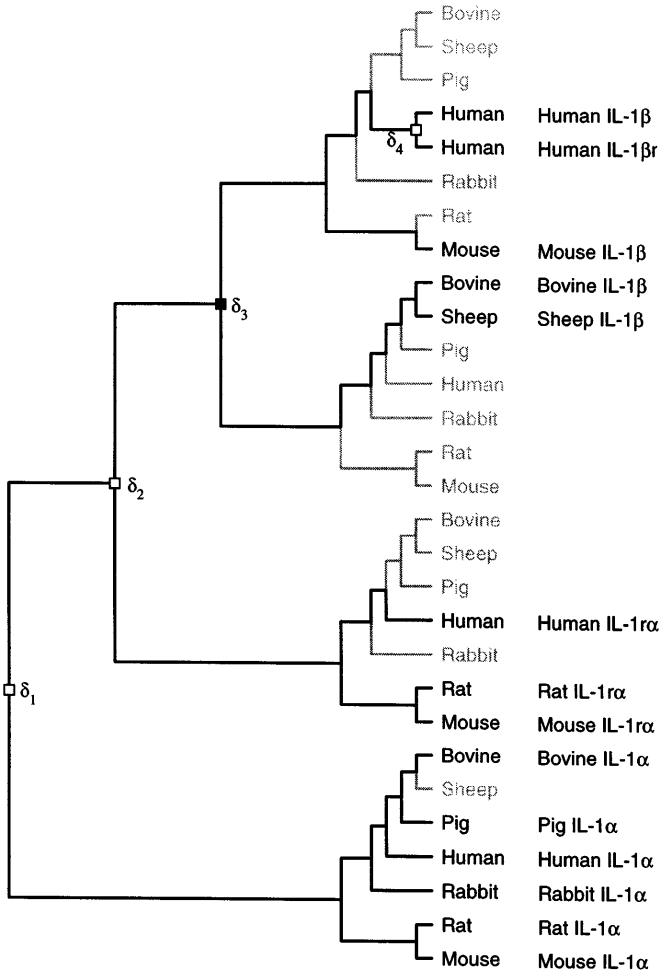

**FIG. 9.** Reconciled tree for mammalian interleukin-1 (IL-1) genes. The tree has a cost of 14 (4 duplications $\delta_1$–$\delta_4$ and 10 losses) and was computed by reconciling the C-terminal region tree for IL-1 (Hughes, 1994; Fig. 2c) with the mammal tree ((((bovine,sheep),pig),human), rabbit,(mouse,rat)). Of the four duplications, three are supported by the presence of multiple copies of IL in the same species, and one ($\delta_3$) is required to explain the incongruence between IL-1β and mammalian phylogeny.

## INFERRING ORGANISMAL PHYLOGENY FROM COMPLEX GENE TREES

The example of interleukins discussed above (Fig. 9) highlights the problem that interpretation of the history of gene family requires some knowledge the species tree that contains the gene tree. Other mammalian phylogenies may alter the interpretation of which sequences were orthologous or paralogous. As a final example of the application of reconciled trees we shall use lactate dehydrogenase (LDH), which was briefly discussed by Page (1994) based on data from Quattro *et al.* (1993). Although Tsuji *et al.* (1994) have since published a larger LDH tree we shall persist in using Quattro *et al.*'s tree as its smaller size makes it more manageable.

Figure 10 shows a phylogeny for 22 LDH sequences from organisms as diverse as humans and bacteria, but with a decided bias toward vertebrates. The presence of multiple copies of LDH already tells us that duplications must be postulated, and the letters A, B, and C indicate hypotheses of orthology for the different se-

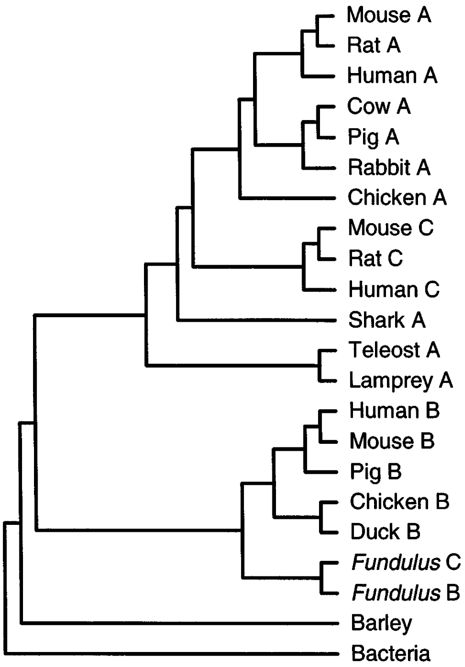

**FIG. 10.** Phylogeny for lactate dehydrogenase (LDH) sequences (from Quattro *et al.*, 1993; Fig. 2).

quences. However, a complete reconstruction of the history of this gene will require embedding it in the appropriate species phylogeny. Given the reasonably broad taxonomic spread of these sequences (at least among vertebrates) we could attempt to find the species tree (or trees) that most parsimoniously contains this gene tree.

### *Searching for Optimal Species Trees*

Charleston (1995) developed a simple way of visualizing the landscape of the tree search problem. The landscape of a given instance of this problem can be characterized by the distribution of "maximal steepest climb" (MSC) length values. Given randomly chosen starting positions we perform a simple hill-climbing search (Reeves, 1995) from each, proceeding to the best possible adjacent tree at each step until there can be no further improvement. The number of steps required is the length of the climb, and the frequency distribution of these lengths can tell us whether the landscape comprises few optima, close together, or many widely spread optima. The terrain in the latter case makes it less tractable to search, as hill climbing methods may frequently become trapped in local optima far from the globally optimal solution. The former landscape with a pronounced peak is much more desirable. The topography of the landscape itself is a function of the data, the optimality criterion, and the tree perturbation(s) used. The effect of the latter can be investigated by using different tree perturbations for the hill-climbing.

We performed three sets of 100 tree searches using nearest neighbor interchanges (NNI), cut and paste (CP) (also known as subtree pruning and regrafting, Swofford and Olsen, 1990), and alternate NNI and CP (ALT). Each search began with a different randomly chosen tree (the same set of 100 random trees was used for all three sets of searches). The distribution of maximal steepest climb lengths (Fig. 11) shows that the NNI's typically have much shorter climbs than do CP and ALT.

The distribution of MSC lengths for NNI's suggests that the landscape corresponding to this perturbation is rugged and comprises many local optima which are not globally optimal. This is borne out by the observation that no search using NNI obtained the minimal value of 12 which was consistently found by the CP and ALT searches. While both CP and ALT always found

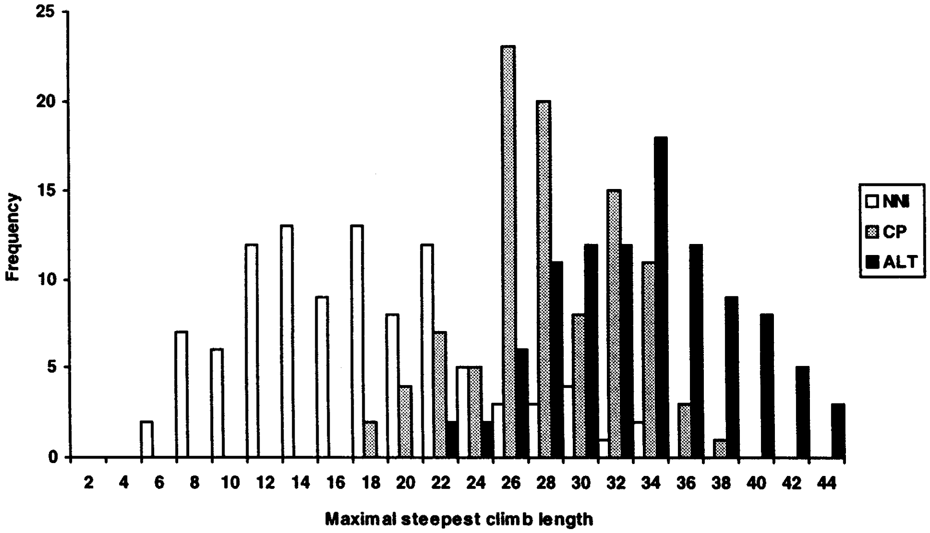

**FIG. 11.** Distribution of maximal steepest climb lengths for three tree perturbations used in the heuristic search for the optimal species tree for the LDH sequences shown in Fig. 10.

the same minimal cost, ALT perturbations were more efficient requiring on average 3355.5 ± 82.1 rearrangements per search compared to 5676.3 ± 60.6 for CP. The former also had a higher average MSC (Fig. 11), indicating a better structured search landscape. Hence, for this data set NNI's are a poor choice, despite being much quicker than CP and ALT (average number of rearrangements 287.3 ± 10.2). Alternating NNI's and CP's gives the best landscape and also is more efficient than CP alone.

The heuristic searches described above yielded 5 equally parsimonious species trees with a cost of 12 events (3 duplications and 9 losses). These five trees were the best found whether duplications and losses, or duplications alone were counted, and differ only in the placement of the fish *Fundulus* with respect to the lamprey, teleost, and shark (Fig. 12). This ambiguity arises because the LDH tree (Fig. 10) contains two major subtrees, one including the two *Fundulus* sequences and the other including the lamprey, teleost, and shark sequences. These two subtrees are rooted at a gene duplication ($\delta_1$ in Fig. 13) and hence the sequences are paralogous. In the absence of an orthologous set of sequences from all four "fish" taxa, we lack sufficient information to resolve their interrelationships.

The reconciled tree shown in Fig. 13 shows that the letters "A," "B," and "C" do not correspond exactly to orthologous sequences, as noted by Quattro *et al.* (1993). Therefore the homology of these sequences must be reconsidered. Some of this information could already be gained by simply considering the gene tree alone (Fig. 10); for example, the nonmonophyly of LDH-C sequences. However, the gene tree alone is insufficient. None of the five species trees that most parsimoniously contain the gene tree accord with current hypotheses of vertebrate relationships. In particular, *Fundulus* is a teleost fish, and the lamprey is the sister taxon to all the other vertebrates from which LDH has been sequenced. As this species tree is not one of the five optimal species trees found above, then at least one additional gene duplication must be postulated, further reducing the degree of orthology implied by the names of the sequences (see also Tsuji *et al.*, 1994; 9396).

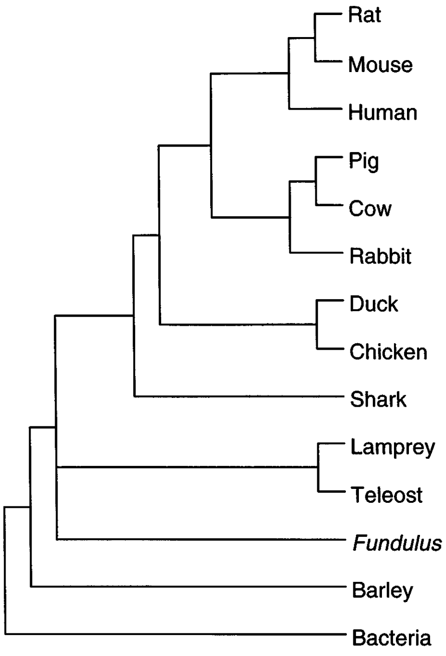

**FIG. 12.** Adams (1986) consensus tree for the five optimal species trees for the LDH gene tree shown in Fig. 10.

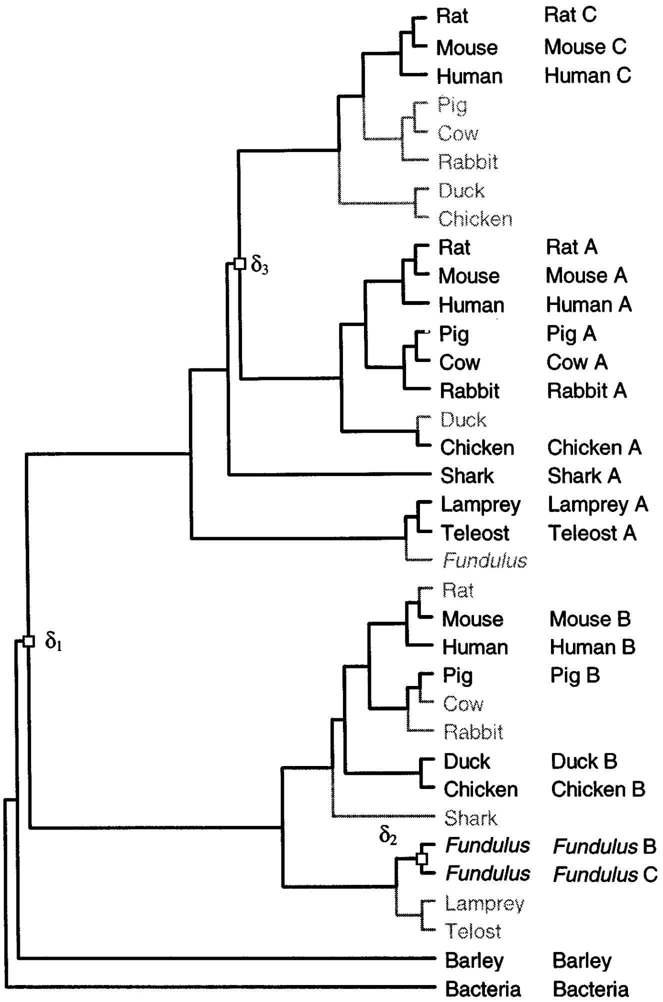

**FIG. 13.** Reconciled tree for the LDH gene tree in Fig. 10 and one of the five species trees whose consensus is shown in Fig. 12.

*Gene Trees from Species Trees*

That the optimal species trees obtained from the LDH gene tree do not accord exactly with the widely accepted species tree raises the possibility that the gene tree shown in Fig. 10 is incorrect. Quattro *et al.*

(1993; 244) expressed reservations about their tree, and Tsuji *et al.*'s tree based on more sequences (Tsuji *et al.*, 1994; Fig. 3) groups lamprey A and teleost A (from the scorpaenid *Sebastolobus alascanus*) with LDH-B. This arrangement means that lamprey and teleost LDH-A are not orthologous with the remaining LDH-A sequences and is consistent with the accepted species tree. However, Tsuji *et al.* found that grouping lamprey-A with shark-A added only one additional step to their most parsimonious tree, and they suggested that because this longer tree required fewer gene duplications to arise before the divergence of the vertebrates it is the "more correct" gene tree. This argument echoes one of the original motivations for Goodman *et al.*'s development of reconciled trees (Goodman *et al.*, 1979b). Faced with incongruent gene and species trees, they argued that less than optimal gene trees, which were a better fit to the species tree might be better estimates of the actual gene tree. With this goal in mind, Goodman *et al.* proposed that the parsimony criterion for choosing the optimal gene tree should include the number of gene duplications and losses each gene tree requires. While this proposal is not without difficulties, especially the problem of assigning weights to these events (Fitch, 1979; Goodman *et al.*, 1979a), it does suggest that while the bulk of molecular phylogenetic inference is from gene trees to species trees, the implications of the species tree for the gene tree should also be considered.

## SUMMARY

Recognition that gene trees might not simply be isomorphic with species trees leads to the need for a method for describing and quantifying the relationship between the two kinds of trees. The method we present here meets these twin needs by computing the number of gene duplications and losses required to embed a gene tree in a species tree and providing a simple means of depicting the history of the genes with respect to the species (the reconciled tree). Given that we can compute a measure of fit between any pair of gene and species trees, this measure can be used as an optimality criterion for choosing the species tree or trees within which the gene tree can be embedded with the least cost. This method can be applied to one or more genes for the same species (e.g., Slowinski *et al.*, in preparation). Investigation of the landscape of this problem led to the design of more efficient heuristic search methods, which have been implemented in the program GeneTree by RDMP.

We should note two limitations of the reconciled tree algorithm employed here. First, it requires fully resolved (i.e., binary) trees, hence uncertainty in the relationships of either genes or species must be represented by multiple, binary trees rather than by a single tree with one or more polytomies. Second, the method assumes that gene transmission has been entirely vertical; horizontal transmission (such as horizontal gene transfer or introgression) is excluded *a priori*. Incorporating horizontal transmission in the method is not a trivial task, as preliminary efforts in the context of host–parasite systems demonstrate (Page, 1995; Ronquist, 1995). This is an area we are currently investigating.

## ACKNOWLEDGMENTS

We thank Joe Slowinski and Boris Mirkin whose empirical and theoretical work, respectively, prompted the writing of this paper. We thank both for commenting on the manuscript. This work was supported by NERC Grant GR3/10095 to R.D.M.P.

## REFERENCES

Adams, E. N. (1986). N-trees as nestings: Complexity, similarity, and consensus. *J. Classif.* **3:** 299–317.

Allard, M. W., McNiff, B. E., and Miyamoto, M. M. (1996). Support for interordinal eutherian relationships with an emphasis on primates and their archontan relatives. *Mol. Phylogenet. Evol.* **5:** 78–88.

Baum, B. R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41:** 3–10.

Brooks, D. R. (1981). Hennig's parasitological method: A proposed solution. *Syst. Zool.* **30:** 229–249.

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.

Doyle, J. J. (1992). Gene trees and species trees: Molecular systematics as one-character taxonomy. *Syst. Bot.* **17:** 144–163.

Felsenstein, J. (1978). The number of evolutionary trees. *Syst. Zool.* **27:** 27–33.

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

Fitch, W. M. (1996). Uses for evolutionary trees. *In* "New Uses for New Phylogenies" (P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee, Eds.), pp. 116–133, Oxford Univ. Press, Oxford.

Goodman, M., Czelusniak, J., and Moore, G. M. (1979a). Further remarks on the parameter of gene duplication and expression events in parsimony reconstructions. *Syst. Zool.* **28:** 379–385.

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

Hughes, A. L. (1994). Evolution of the interleukin-1 gene family in mammals. *J. Mol. Evol.* **39:** 6–12.

Iwabe, N., Kuma, K., and Miyata, T. (1996). Evolution of gene families and relationship with organismal evolution: Rapid divergence of tissue-specific genes in the early evolution of chordates. *Mol. Biol. Evol.* **13:** 483–493.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* **38:** 7–25.

Maddison, W. P. (Submitted). Gene trees in species trees. *Syst. Biol.*

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

Nelson, G., and Platnick, N. I. (1981). "Systematics and Biogeography: Cladistics and Vicariance," Columbia Univ. Press, New York.

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

Page, R. D. M. (1993b). Genes, organisms, and areas: the problem of multiple lineages. *Syst. Biol.* **42:** 77–84.

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. (1995). Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics* **10:** 155–173.

Page, R. D. M. On consensus, confidence, and "total" evidence. *Cladistics,* in press.

Pamilo, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5:** 568–583.

Quattro, J. M., Woods, H. A., and Powers, D. A. (1993). Sequence analysis of teleost retina-specific lactate dehydrogenase C: Evolutionary implications for the vertebrate lactate dehydrogenase gene family. *Proc. Natl. Acad. Sci USA* **90:** 242–246.

Ragan, M. A. (1992). Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1:** 53–58.

Reeves, C. R. (1995). "Modern Heuristic Techniques for Combinatorial Problems," McGraw Hill, New York.

Rodrigo, A. G. (1993). A comment on Baum's method for combining phylogenetic trees. *Taxon* **42:** 631–636.

Ronquist, F. (1995). Reconstructing the history of host-parasite associations using generalised parsimony. *Cladistics* **11:** 73–89.

Swofford, D. L., and Olsen, G. J. (1990). Phylogeny reconstruction. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, Eds.), pp. 411–501, Sinauer Associates, Sunderland.

Takahata, N. (1989). Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics* **122:** 957–966.

Tsuji, S., Qureshi, M. A., Hou, E. W., Fitch, W. M., and Li, S. S.-L. (1994). Evolutionary relationships of lactate dehydrogenases (LDHs) from mammals, birds, an amphibian, fish, barley, and bacteria: LDH cDNA sequences from *Xenopus,* pig, and rat. *Proc. Natl. Acad. Sci. USA* **91:** 9392–9396.

Wu, C.-I. (1991). Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127:** 429–435.