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 8QO, 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 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
species. If the genes in Fig. 2 are alleles and not logs, then the three loss events depicted may be interpreted as either extinction of alleles or instances of age sorting.

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

we are only interested in the history of the gene tree of the organisms—for example, documenting exactly how many gene duplications took place—then we need no further, for as shown above, the reconciled tree encodes this information. However, if we are interested in what the gene tree can tell us about species stances among when 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 $T$ which when reconciled with $G$ has the lowest cost. Inducing $T$ 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.

**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, $c(G_1, T) + c(G_2, T) + \cdots + c(G_n, T)$. 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 subsets, 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 organism. 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.

Iwabe et al. were interested in the temporal distribution of gene duplications, and noted that duplication $δ_1$ occurred within invertebrates, duplication $δ_2$ prior to the divergence of the vertebrates, and that duplication $δ_3$ was of uncertain age. The reconciled tree locates $δ_1$ 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 *Catalosonius*. 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 ($δ_2$ and $δ_3$). 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 ($δ_4$). 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 phylog-
eny 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 (δ₂) is hypothetical and is inferred solely from the incongruence between the mammalian tree and the phylogeny of the IL-1β genes. Specifically, mouse IL-1β is more closely related to human IL-1β than are bovine and sheep IL-1β. This indicates that if the mammalian phylogeny used to construct Fig. 9 is correct then the IL-1β genes are not orthologous (Hughes, 1994; p. 10).

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

**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 2873 ± 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 (δ₁ 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).

**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.
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 scaphocephalan 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 incongruence of 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 Sowinski 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


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


