Chapter 5

TANGLED TALES FROM MULTIPLE MARKERS

Reconciling conflict between phylogenies to build molecular supertrees

James A. Cotton and Roderic D. M. Page

Abstract: Supertree methods combine information from multiple phylogenies into a larger, composite phylogeny. When there is no disagreement between the source phylogenies, constructing the supertree is straightforward. But in the (nearly universal) presence of disagreement between source trees, supertree methods seek to either represent or resolve this conflict. Existing supertree methods that resolve conflict between source trees do so in an *ad hoc* way. Gene tree parsimony is a supertree method that can combine molecular phylogenies for overlapping taxon sets and interprets conflict between these phylogenies in a biologically meaningful way. We review the method and discuss the relationship between gene tree parsimony and other supertree methods. Finally, we suggest that a better understanding of the causes of conflict between source trees should lead to appropriate ways of resolving this conflict when constructing supertrees.

Keywords: gene duplication, gene tree parsimony, reconciled trees

1. Introduction

Combining information from different sources of phylogenetic evidence can be important for two different reasons: 1) to increase the scope of the phylogenetic results by including a greater range of terminal taxa, or 2) to improve the accuracy of the results by incorporating more data for these taxa. Supertree methods have been used to achieve both these aims by incorporating source trees constructed from a wide range of relevant data.

Bininda-Emonds, O. R. P. (ed.) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, pp. 107–125. Computational Biology, volume 3 (Dress, A., series ed.).

Where source trees are rooted and compatible, supertree construction is relatively trivial: efficient algorithms exist to decide whether or not a set of trees are compatible and to construct the parent trees that contain all these trees (Aho *et al*., 1981; Steel, 1992; Semple, 2003). However, most practical applications of supertree methods involve source trees that are incompatible, and supertree workers have been less successful in designing algorithms to combine information from conflicting trees. Such algorithms can remove conflict by pruning leaves (e.g., in maximum agreement subtrees), represent the conflict through soft polytomies, resolve the conflict, or use some combination of these.

In fact, the only supertree method that has been at all used widely by biologists is matrix representation with parsimony (MRP; see Baum and Ragan, 2004), with an increasing number of supertrees constructed using this method appearing in the literature (e.g., Kennedy and Page, 2002; Pisani *et al*., 2002; see Baum and Ragan, 2004). MRP uses additive binary coding to represent the hierarchical structure of a set of trees as a series of matrix elements — each node on the trees is represented by a column of the matrix, with missing data for those taxa not present on a particular source tree. This matrix is then analyzed using parsimony methods to construct a supertree or set of supertrees. Although MRP supertrees have played an important part in stimulating the field of supertree research and might be reasonably successful in reconstructing relationships (Bininda-Emonds and Sanderson, 2001), there has been an increasing literature on the biases of MRP methods, and several proposed modifications to the original method (e.g., Purvis, 1995; Ronquist, 1996; Bininda-Emonds and Bryant, 1998; Thorley, 2000). There are similar problems with other supertree algorithms, such as the MINCUTSUPERTREE method (Semple and Steel, 2000), which has several undesirable properties (Page, 2002). These problems have prompted a widening interest in other methods of supertree construction, such as shown in this volume and elsewhere (Page, 2002).

In an effort to classify the growing number of supertree methods available to systematists, at least two authors have characterized the supertree problem in a distance framework (Chen *et al*., 2003; Thorley and Wilkinson, 2003). These authors suggest that the supertree problem can be seen as the problem of finding a tree (or set of trees) that is closest to a set of input trees under some measure of distance between trees. For example, as both sets of authors point out, MRP seeks to find the tree minimizing the number of steps required on the MRP matrix. Other distance measures are certainly possible, such as distances based on nearest-neighbour interchanges (NNIs; Waterman and Smith, 1978). Bearing this framework in mind, we note that all problems of identifying an optimal tree are likely to be NP-complete (Wareham, 1993), including the maximum-parsimony problem used by MRP methods (Graham and Foulds, 1982). Thus, heuristic strategies are likely to be needed.

In this framework, we suggest a new distance measure for supertree inference, one based on the number of actual biological events that might have produced the differences observed between source trees. These events can be inferred using the co-phylogenetic method of reconciled trees. In this chapter, we introduce reconciled trees and their use to infer a species tree, or supertree, from several molecular source trees, a procedure which has become known as gene tree parsimony (GTP; Slowinski and Page, 1999). We include a brief empirical example of a GTP supertree. We then make a preliminary attempt to characterize the GTP method by describing some properties of the method, as has been attempted for other supertree methods. Lastly, we go beyond GTP itself to argue that understanding the causes of conflict between source trees should help us resolve that conflict appropriately, and to suggest that a model-based framework might enable systematic biologists both to understand the causes of conflict between trees and to construct accurate supertrees in the face of such conflict.

2. Tangled trees, or co-phylogeny

Evolutionary biologists have long been interested in the relationship between ecologically associated entities, particularly hosts and their parasites. One important question in host-parasite biology is the extent to which these organisms co-evolve, and, more specifically, the extent to which they co-diverge (i.e., the extent to which speciation events in one lineage are mirrored by speciation events in the other). This led to interest in comparing the phylogenetic trees of associated organisms, along with a parallel interest in relating the phylogenies of organisms to their biogeography (Page and Charleston, 1998). The initial solution to this problem was to use a binary coding of the dependant tree, similar to those used in MRP supertree methods. This matrix was then used either to reconstruct the host phylogeny, or to understand the pattern of evolution by optimizing the characters onto the second phylogeny (Brooks, 1981). Similar to the problems with the binary coding used in MRP, various fixes failed to alleviate the fundamental problem that the characters produced by this coding for a given tree are non-independent.

In studies of co-phylogeny, the solution has been to map the dependant phylogeny explicitly into the host phylogeny, postulating directly events that lead to the differences between the two phylogenies (see Figure 1). This insight led to Page's (1994) formalization of the earlier concept of reconciled trees (introduced by Goodman *et al.*, 1979). Constructing a reconciled tree

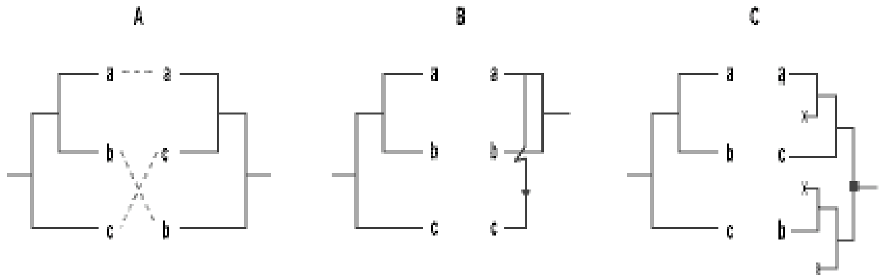

Figure 1. The incongruence between the species tree (A, left) and gene tree (A, right) in this example can be explained by postulating either a single lateral gene transfer from taxon a to taxon c (B) or a single gene duplication followed by three gene losses (C).

involves reconciling the differences between two trees by postulating certain co-phylogenetic events that introduced these differences. As shown in Figure 2, these events can be extinction of a lineage, independent speciation of a lineage, and horizontal transfer. Although co-phylogeny methods were developed in the context of biogeography and host-parasite evolution, similar events occur in the evolution of a gene lineage within a species (e.g., lateral gene transfer, gene duplications, and gene loss), so the same co-phylogeny mapping can also be used to study this system. Other evolutionary processes are also included under these co-phylogenetic events, with, for example, hybridization and some forms of recombination being indistinguishable from lateral gene transfer in this context.

The interest in supertree methods underlines the growing availability of phylogenies, and this increasing amount of data reflects both an increase in the taxonomic coverage of phylogenetic information ("width") and in the amount of data available for particular organisms ("depth"). This increasing depth is particularly a result of the rise of genome-level sequencing efforts for an increasing number of organisms, and an important corollary of this work is the increasing realization that phylogenies for different genetic loci for the same species frequently disagree. This has in turn prompted the realization that a range of evolutionary events can cause the correct phylogeny for a gene to be different from the correct phylogeny for the species it is sampled from, a problem known as the gene tree-species tree problem (Doyle, 1992; Maddison, 1997). Reconciled trees are a natural solution to this problem (Page and Charleston, 1997a) — we can use the reconciled-tree algorithm to score a species tree for a particular gene tree in terms of the number of gene duplications, gene losses and other evolutionary events that have introduced differences between the two trees. The numbers of these events is a distance between the trees that has a natural, biological interpretation (Mirkin *et al.*, 1996).

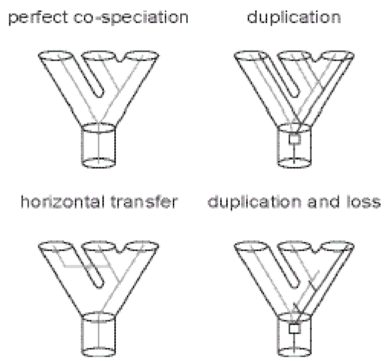

Figure 2. Some co-phylogenetic events, introducing differences between two associated phylogenies.

In principle, several different events can be scored in this way (Figure 2), including the number of deep coalescence events (Maddison, 1997). It should be noted that dealing with horizontal gene transfer correctly is complex and existing implementations of reconciled trees in this context exclude this possibility (Page, 1998). In particular, including horizontal transfer events makes tree reconciliation far more intensive computationally, and requires additional assumptions about the relative rates of gene duplication and loss and lateral gene transfer. Fortunately, solutions for co-phylogeny mapping incorporating horizontal transfer are available, and could be used in the context of GTP (Charleston, 1998; Ronquist and Nylin, 1990; Ronquist, 2003). Methods are also available for estimating optimal event costs for particular problems that suggest that reconciliation methods are robust to alternative weighting of different events (Ronquist, 2003). Even if just duplications and losses are included in the event set, different weightings of these two events are possible and will affect the result obtained (Ronquist, 2003). Fortunately, it seems that the duplication-and-loss optimal trees are a subset of the duplication-only optimal trees for a particular set of source trees (Page and Charleston, 1997b), so the consensus results with different weightings will differ only in degree of resolution. It is also often preferable to use the count of duplications alone (ignoring gene losses) as a distance function because gene losses are confounded with failure-to-sample in some kinds of study (e.g., due simply to the lack of a sequence in the sequence databases), and so do not represent a true biological cost (Cotton and Page, 2003). For the remainder of this chapter, we restrict ourselves to GTP using only duplication events or the sum of duplication and loss events, for which a software implementation is available (Page, 1998).

3. From reconciled trees to supertrees

When we have multiple gene trees, we can combine information from all these trees into a single tree by finding the species tree (or set of species trees) that minimizes the number of co-phylogenetic events required to reconcile the species tree with each source tree or minimizes some weighted sum of these events (assigning a cost to each event category). The resultant species tree can be on a larger taxon set than any of the source trees and is constructed using information from the topology of each source tree only. As such, it fits the definition of a conventional supertree. The set of GTP supertrees is thus the set of all supertrees that require a minimum number of the evolutionary events considered to explain the difference between the supertree and the set of source trees.

Finding an optimal species tree under either the duplication-only or duplication-and-loss score has been the focus of some attention by mathematicians and computational biologists. Linear-time algorithms exist for computing these scores for a particular pair of gene and species trees (Eulenstein, 1997; Zhang, 1997; Zmasek and Eddy, 2001), and although it is known (as expected) that finding the minimum-cost species tree is NP-complete (Ma *et al*., 1998), there is a polynomial-time (fixed-parameter tractable) algorithm to find this tree where the maximum number of gene lineages extant at any point on the tree has an upper bound (Hallett and Lagergren, 2000).

If we restrict the source trees to be molecular trees, the duplication count (or duplication cost) is a biologically interpretable measure of the evolutionary difference between the source tree (or gene tree) and supertree (or species tree). If all the differences between source trees were a result of the evolutionary events included, then the GTP supertree would be expected to reconstruct the correct supertree accurately (at least as far as the methodological assumptions of parsimony hold). Unfortunately, little is understood about the causes of disagreement between molecular phylogenies. Clearly, some error will be a result of simple estimation error owing to the finite amount of data available from any single gene. The inadequacy of existing models will also lead to some error and so introduce conflict between phylogenies. Thus, it could be that little of the error between phylogenetic estimates from different molecular markers is because of the kinds of evolutionary events dealt with by GTP, and it is unclear how GTP will perform at resolving conflict from other (non-molecular) sources. It is, however, similarly unclear exactly how well other supertree methods perform in practice, although a start has been made on using simulation studies to address this for some methods (Bininda-Emonds and Sanderson, 2001, Burleigh *et al*., 2004; Lapointe *et al*., 2004). It is clearly an empirical question how well any supertree method performs in practice, and there seems no reason to suspect that GTP will necessarily underperform compared with other methods when phylogenetic conflict is a result of estimation error or model inadequacy. More work is needed in comparing supertree methods in a range of situations before the strengths and weaknesses of different supertree methods will be understood.

One modification to standard supertree methods that has been shown to be highly effective in improving the accuracy of results (Ronquist, 1996; Bininda-Emonds and Sanderson, 2001; Salamin *et al*., 2002) involves incorporating some measure of uncertainty into the input source trees (e.g., from a bootstrap profile of trees from non-parametric bootstrapping). An idea akin to this "weighted MRP" has also been mentioned in the reconciled-tree literature, where it seems particularly apposite. If reconciled-tree methods rely on identifying evolutionary events that lead to incongruence between trees, it is clearly crucial to incorporate some idea of the uncertainty in tree estimates if these events are to be "real" rather than owing to this uncertainty (Page, 2000; Page and Cotton, 2000; Ronquist, 2003). Using a bootstrap profile of trees for each gene has been shown to improve the species tree estimate in at least one empirical study (Cotton and Page, 2002), and also provides analogous bootstrap support values for the species tree or supertree itself. Several other methods for incorporating uncertainty in source tree estimates into reconciled tree analyses have also been proposed (Page, 2000; Page and Cotton, 2000).

4. An empirical example: a small supertree of *Drosophila*

Several empirical examples of using reconciled-tree methods to infer phylogenies exist in the literature (Slowinski *et al*., 1997; Page, 2000; Cotton and Page, 2002; Martin and Burg, 2002), but we present here a novel empirical example of a small-scale supertree of *Drosophila* and some related genera based on five nuclear genes (Figure 3). The source trees were relabeled with the species names and a standard MRP matrix was built using the program Supertree (available at http://darwin.zoology.gla.ac.uk/~rpage/supertree/). The MRP matrix was analyzed using PAUP* v4b10 (Swofford, 2002) using standard parsimony. The GTP analysis was performed using GeneTree (Page, 1998). For both analyses, a large number of equally optimal trees were found, so five separate searches were performed, with each one swapping on a maximum of 50 000 (for MRP) or 15 000 (for GTP) trees. Consensus trees for each of the five searches were very similar,

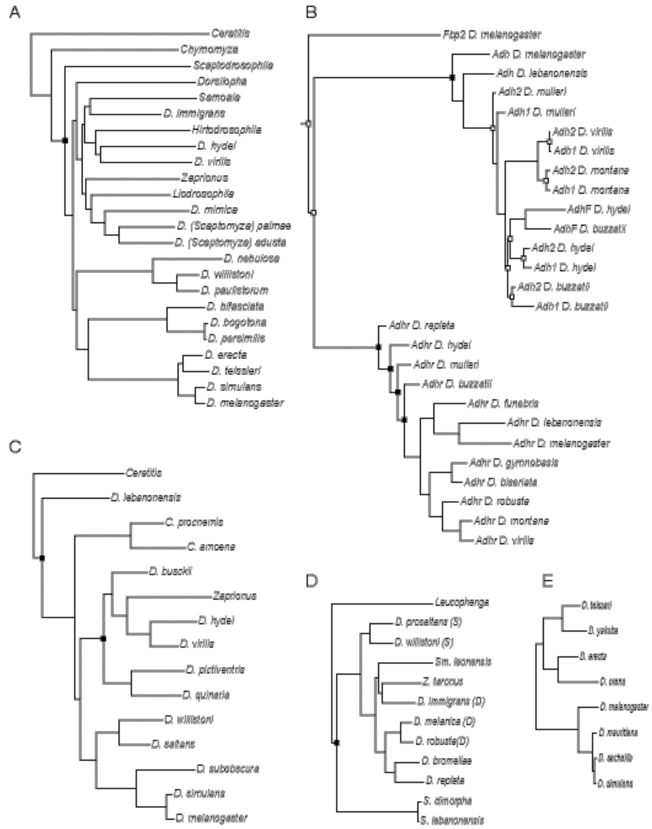

Figure 3. The five gene trees used in building the *Drosophila* supertree presented here: A) *Dopa decarboxylase* (Tatarenkov *et al*., 1999), B) *Alcohol dehydrogenase* and the *Alcohol dehydrogenase-related* gene (Betrán and Ashburner, 2000), C) *Cu-Zn superoxide dismutase* (Kwiatowski *et al*., 1994), D) 28S rRNA (Russo *et al*., 1995), and E) the regulatory gene *roughex* (Avedisov *et al*., 2001). Boxes show positions of gene duplications implied by the supertrees. Open boxes are duplications necessitated by the multiple copies of *Alcohol dehydrogenase* genes, whereas closed boxes are those duplications inferred from conflict between the gene tree and the supertree. All the duplications, except that for 28S rRNA, are implied by every supertree.

suggesting that the five searches had each sampled successfully from across the large island of trees. Trees of cost 97 parsimony steps were found under in MRP, and 63 duplications and losses under GTP. The set of GTP supertrees thus included all the trees reconciled with the five source trees using all combinations of 63 duplications and losses (in fact, all the supertrees found required either 17 duplications and 46 losses, or 18 duplications and 45 losses). Although 18 duplications sounds like a lot, nine duplications are required by multiple gene copies being present on the

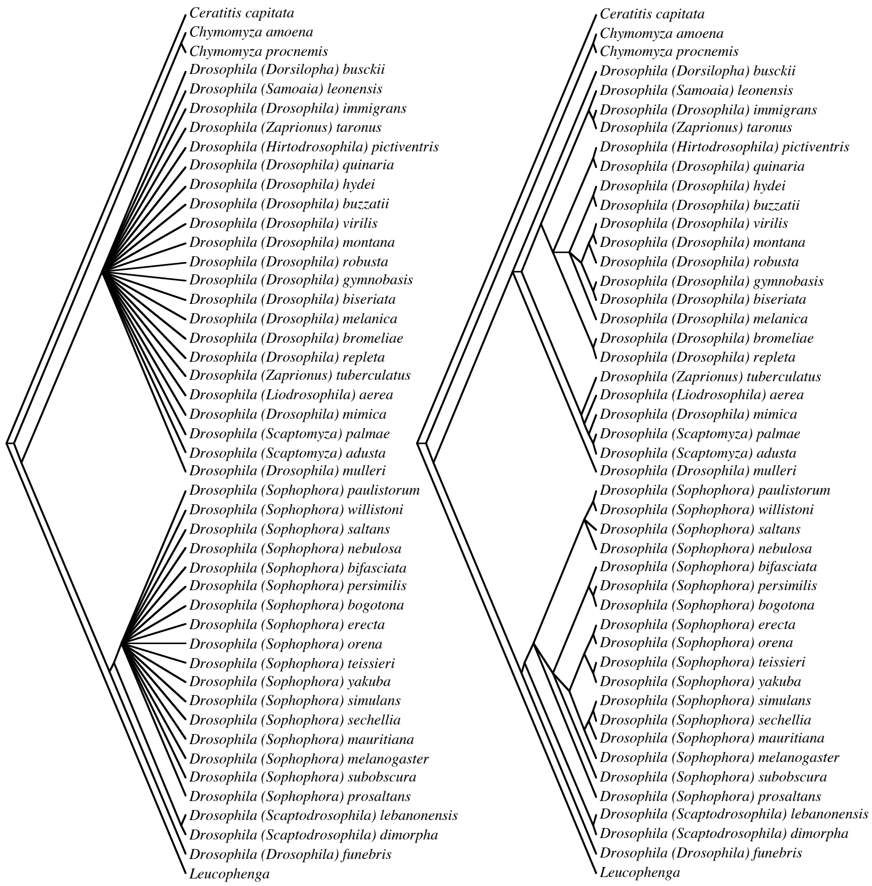

Figure 4. The strict component (left) and Adams (right) consensus of the GTP supertrees from the *Drosophila* gene trees under the duplication-and-loss criterion.

Alcohol dehydrogenase gene tree (Figure 3). Thus, only nine duplications, at most, are a result of incongruence between the source trees and supertrees.

Given that they are derived from the same data, it is reassuring that both the GTP and MRP analyses are similar (Figures 4 and 5, respectively). Both analyses support the monophyly of the subgenus *Sophophora*, and indeed show exactly the same relationships within *Sophophora*. It appears that GTP is more conservative than MRP, in that the results it produces are largely compatible with those from MRP, but somewhat less resolved (although this is not the case for every clade). Both GTP and MRP find the other subgenera of *Drosophila* included to be paraphyletic or polyphyletic, but there are some differences between the two sets of trees. One instructive difference is in the way *D. melanica*, *D. robusta*, *D. biseriata* and *D. gymnobasis* cluster

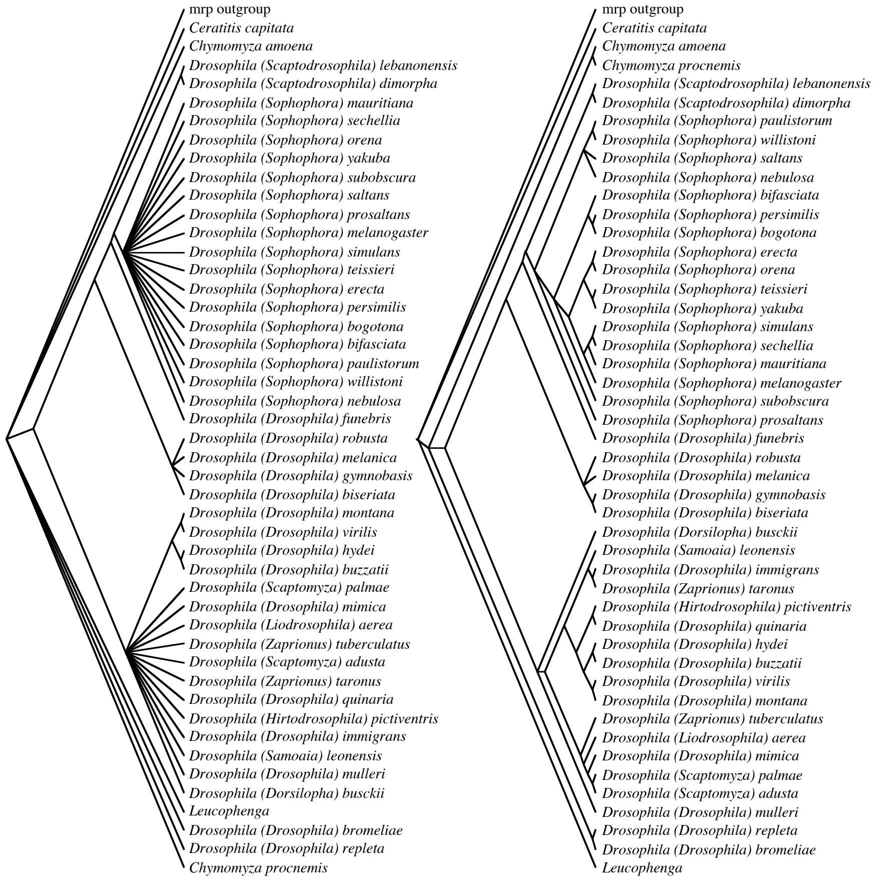

Figure 5. The strict component (left) and Adams (right) consensus of the standard MRP supertrees.

with respect to one another. In the MRP results, these species are placed strikingly away from the other members of subgenus *Drosophila* as a sister clade to *D. funebris* and the subgenera *Sophophora* and *Scaptodrosophila*. This is in contrast to the GTP tree, where these species are embedded within the paraphyletic assemblage around subgenus *Drosophila* to which they all belong. The MRP results seem surprising given the source trees: the four species in question appear only on trees in Figures 3B and D, where they group with other members of the subgenus *Drosophila*. This odd placement is probably partly a result of the different treatment of *D. bromeliae* and *D. repleta*, the other principal difference between the two sets of trees. These two species are the sister taxa to *D. melanica* and *D. robusta* on the 28S tree (Figure 3D). The unresolved position of *D. bromeliae* and *D. repleta* on the MRP tree is understandable given that they are placed (with three other members of the clade containing subgenus *Drosophila* and others), between subgenera *Sophophora* and *Scaptodrosophila* on the 28S tree. However, given the support for the three other taxa as being related to members of the subgenus *Drosophila* on two other source trees (Figures 3A, C), the more-resolved position on the GTP trees seems at least as reasonable.

Investigators can examine incongruence in the GTP supertree in terms of duplications and losses in specific genes. This can both help assess whether incongruence is restricted to a single gene (i.e., because it contains the vast majority of duplications and losses) and help to understand the general pattern of genetic evolution for this group. Furthermore, the hypothesized duplications and losses might be testable using other evidence: for example, do the suggested paralogues have different functions, occur in different parts of the genome, or have different genetic architectures? Another approach might be to use the GTP supertree to inform a search for additional gene copies. For example, the proposed duplication in *Dopa decarboxylase* could be confirmed by finding an additional copy of the gene in *Scaptodrosophila*, although it would be wise to examine the strength of support for a particular duplication before expending much laboratory effort on such a search!

5. Properties of GTP as a supertree method

Progress has been made recently in thinking about desirable properties of supertree methods (see Wilkinson *et al*., 2004). These properties are characteristics that would seem to be desirable in all supertree methods, and which seem likely to correlate with the accuracy of the results of a method. Comparatively little has been done to characterize supertree methods formally in terms of these properties or more formal axioms. In particular, it might be of interest to see how GTP resolves conflict between source trees when compared with those variants of MRP that are already characterized in terms of some of these properties. The properties named in italics below are used in the sense of Wilkinson *et al*. (2004). Aside from the three properties discussed below, GTP methods are *assessable*, *weightable*, *plenary*, show *order invariance*, and seem to be *Pareto* on components. They do not show *generality* or *uniqueness*, and are not particularly *speedy* compared with polynomial-time methods. Their behaviour in terms of being *co-Pareto* and *independent of irrelevant alternatives* is unclear.

5.1 GTP displays unique subtrees correctly

Here, I define a unique subtree as one that appears in a single source tree, where no other source tree contains any of the taxa of the subtree. GTP

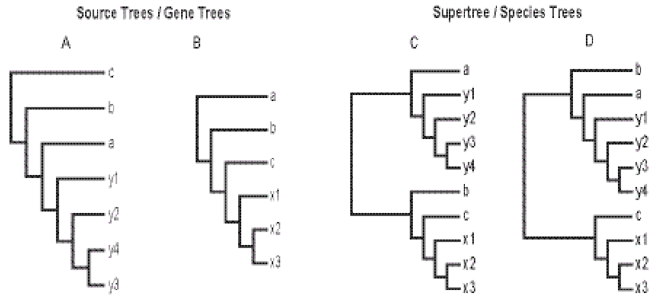

Figure 6. Trees C and D are the two supertrees for source trees A and B under both the duplication-only and duplication-and-loss costs. Trees C and D are also the standard and Purvis coding MRP supertrees for trees A and B (source trees taken from Page, 2002).

appears to include unique subtrees in the supertree or species tree correctly, a property shared by MRP methods, but not by the original formulation of MINCUTSUPERTREE (Page, 2002). Using Page's example (Figure 6), we see that GTP reconstructs these groupings correctly under both duplication-only and duplication-and-loss criteria. Both GTP and MRP perform better than the modified MinCut method in placing taxon a correctly as sister-group to the clade (x1, …, x3) and taxon c as sister-group to the clade (y1, …, y4), rather than collapsing these relationships to a polytomy (Page, 2002). Clearly, reconstructing clades that are unique to a single tree is a desirable property for all supertree methods. This property is a special case of property P7 of Steel *et al*. (2000), which they showed no rooted supertree method that produces a single output tree can possess.

5.2 GTP is not *sizeless*

It has been noted that the original coding for MRP matrices produces supertrees biased towards including those relationships on larger source trees because of redundant information in the matrix (Purvis, 1995). Purvis showed that some matrix entries are redundant in the sense of not being needed to reconstruct the original source trees, but this information might not be redundant in a different sense (see Ronquist, 1996). We use Purvis's example to show that GTP also suffers from this bias when the duplication-and-loss criterion is used, but not under the duplication-only criterion. The two gene trees shown in Figures 7A and B support just a single species tree under the duplication-and-loss criterion, that of Figure 7C. In this tree, taxon d occurs in the position supported by tree A, the larger of the two source trees, thereby ignoring effectively the conflicting signal from the very different position of this taxon in the smaller tree B. Under the duplication-

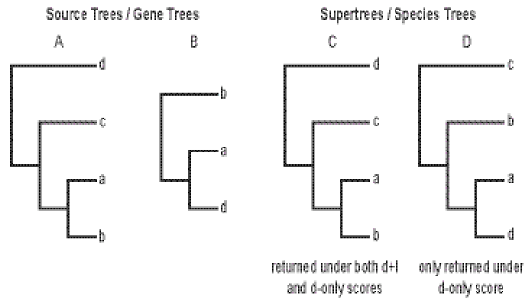

Figure 7. Trees C and D are the two supertrees for source trees A and B under the duplication-only cost. Tree C is the unique supertree under the duplication-and-loss cost. Tree C is the unique supertree under standard MRP, while both C and D are Purvis-coding MRP supertrees (source trees taken from Purvis, 1995).

only criterion, an additional species tree (Figure 7D) has an equal cost and shows taxon d in the position suggested by the smaller input tree.

The reason for this bias under the duplication-and-loss criterion is clear: duplications inferred on larger gene trees will tend to infer more gene losses than those on smaller trees. Under this criterion, the species tree will thus be selected to minimize gene duplications on larger gene trees more than on smaller ones, and so will tend to reflect relationships in larger gene trees. This source of bias disappears under the duplication-only criterion.

5.3 GTP is not *positionless*

Several suggested variants of MRP appear to suffer from a bias towards placing species in the most crownward position displayed by the input trees. This bias was first noticed by Ronquist (1996) as being a problem with Purvis's (1995) suggested modification to the original MRP encoding. Figure 8 shows two source trees, A and B. Under both the duplication-and-loss and duplication-only criteria, there is only a single optimal species tree (Figure 8C). This tree places taxon e in the more crownward position, as suggested by source tree B, overruling the conflicting position suggested by source tree A. Thus, it seems that GTP also shows a bias towards placing taxa in the more crownward position.

6. A probabilistic view of the supertree problem

We can view the supertree problem usefully in a probabilistic setting, a view that makes several themes of this paper particularly clear. This is a fairly natural extension of the distance-based view expressed earlier. Instead of

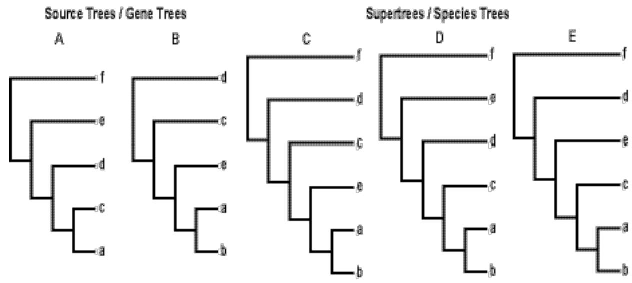

Figure 8. Tree C is the unique GTP supertree for source trees A and B under both duplication-only and duplication-and-loss costs. Tree C is also the unique MRP supertree under Purvis coding, while the all three trees C to E are MRP supertrees under standard coding. Adapted from Thorley (2000).

seeking the closest tree to a set of source trees, we can look for the maximum likelihood or most probable supertree for this set. To do this, we need a likelihood function for the supertree that is proportional to the probability that the source trees come from the supertree. There are several different ways we could frame this likelihood function based on how similar the source trees are to the subtrees of the proposed supertree induced by their leaf sets. For example, if we assume every NNI needed to move from the induced subtree to the source tree is equally likely, it is relatively trivial to construct this function using a binomial distribution. To do this, we need only calculate the NNI distance between source tree and its induced subtree in the supertree, and the maximum possible distance between the trees under this operation. The product of these probabilities across all sources trees would then be the likelihood of this supertree under this simple NNI-binomial model. The model has only a single parameter that must be estimated from the data: q, the probability of an NNI difference between a source tree and the supertree. The likelihood of a supertree T_s from a set of n subtrees $T_1 \ldots T_n$, where the NNI distance between the source tree T_i and the subtree induced on T_s by the leaves of T_i is d_{T_i,T_s} and the maximum NNI distance between two trees of this size is ΔG, is given by

(1) $$L\left(T_s \middle| T_1, T_2, \ldots, T_n\right) \propto \prod_{i=1}^{n} p\left(T_i \middle| T_s\right),$$

where the probability of each source tree is simply

(2) $$p\left(T_i \middle| T_s, q\right) = \binom{\Delta G}{d_{T_i,T_s}} q^{d_{T_i,T_s}} \left(1-q\right)^{\Delta G - d_{T_i,T_s}}.$$

Constructing this likelihood function allows us to find a maximum likelihood supertree under this model using standard heuristic methods. It would also be easy to estimate the supertree in a Bayesian framework using Markov chain Monte Carlo. To do this, we need to propose a prior probability distribution on the supertree and place a prior on the NNI probability parameter of the model. A Bayesian method would let us construct a credible interval of trees within which the true supertree lies with high probability. Sampling from this posterior probability distribution of supertrees will also allow the use of correct probability distributions for trees, improving the accuracy of the various evolutionary studies in which supertrees have been used (Huelsenbeck *et al*., 2000b; also Ronquist *et al*., 2004). It should be noted that the above discussion assumes that source trees are known without error. If character data are available for all the subtrees, an obvious approach would be to calculate the probabilities of these trees using a model of sequence evolution, providing a natural way to incorporate uncertainty in the source tree estimates.

More importantly, formulating the supertree problem in this way shows that a wide range of likelihood functions relating a subtree to the supertree could be used to build supertrees. We emphasize that models such as the NNI-binomial model are likely to be gross simplifications and inadequate for most estimation purposes, so more complex models (such as that of Ronquist *et al*., 2004) will be needed. More interestingly, probabilistic models of gene duplication and gene loss have been developed recently (Arvestad *et al*., 2003) that could be extended to the supertree setting. Even horizontal transfer events can be incorporated (Huelsenbeck *et al*., 2000a), although this is more difficult to model mathematically (Charleston and Robertson, 2002). It seems probable that simplifying assumptions akin to those of single base substitutions in DNA sequence phylogeny models will be needed for supertree models. Perhaps the greatest advantage of both likelihood and Bayesian methods is that both provide a natural framework for comparing models, and so permit rational choice between different methods. As discussed earlier, relatively little is known about how different methods perform on real data, and it could be in this probabilistic framework that competing methods, and their different assumptions, can be compared the most rigorously. The simplistic model presented here might be a useful null model against which more realistic models can be tested.

7. Conclusion

We are clearly at an early stage in the development of supertree methods: many methods are being proposed, but little is known about their relative merits. While most supertree methods treat conflict between source trees in an *ad hoc* way, it is possible to treat at least some causes of incompatibility in a biologically realistic way. We hope that this chapter will encourage biologists to think more about how incongruence between trees can be investigated, and about the possible causes of this incongruence beyond simple estimation error. It is clearly an empirical question how different supertree methods will perform on real data, and it is probable that different methods will be preferable for different data, reflecting the different causes of conflict in them. For example, reconciled-tree methods might be the most appropriate if all conflict between source trees is caused by gene duplication and gene loss (probably a rather unlikely scenario), whereas matrix representation with flipping (Burleigh *et al*., 2004) might perform best where conflict results in randomly distributed errors on some binary matrix representation of the source trees. Much more work is clearly needed to understand both the causes and consequences of conflict between phylogenies from different data.

Acknowledgements

We thank Olaf Bininda-Emonds for inviting us to write this chapter and also Olaf Bininda-Emonds, Fredrik Ronquist, Gordon Burleigh, and Mark Wilkinson for comments on the manuscript. This work was performed while JAC was supported by a NERC studentship while at the Division of Environmental and Evolutionary Biology at University of Glasgow, and by BBSRC grant 40/G18385.

References

AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing* 10:405–421.

ARVESTAD, L., BERGLUND, A.-C., LAGERGREN, J., AND SENNBLAD, B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19:i7–i15.

AVEDISOV, S. N., ROGOZIN, I. B., KOONIN, E. V., AND THOMAS, B. J. 2001. Rapid evolution of a cyclin A inhibitor gene, *roughex*, in *Drosophila*. *Molecular Biology and Evolution* 18:2110–2118.

BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.

BETRÁN, E. AND ASHBURNER, M. 2000. Duplication, dicistronic transcription, and subsequent evolution of the Alcohol dehydrogenase and Alcohol dehydrogenase-related genes in *Drosophila*. *Molecular Biology and Evolution* 17:1344–1352.

BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497– 508.

BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.

BROOKS, D. R. 1981. Hennig's parasitological method: a proposed solution. *Systematic Zoology* 30:229–249.

BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.

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

CHARLESTON, M. A. AND ROBERTSON, D. L. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic Biology* 51:528–535.

CHEN, D., DIAO, L., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2003. Flipping: a supertree construction method. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 135–160. American Mathematical Society, Providence, Rhode Island.

COTTON, J. A. AND PAGE, R. D. M. 2002. Going nuclear: vertebrate phylogeny and gene family evolution reconciled. *Proceedings of the Royal Society of London B* 269:1555–1561.

COTTON, J. A. AND PAGE, R. D. M. 2003. Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. *Molecular Phylogenetics and Evolution* 29:298–308.

DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163.

EULENSTEIN, O. 1997. A linear time algorithm for tree mapping. Arbeitspapiere der GMD, No. 1046.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA-sequences – a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.

GOODMAN, M., CZELUSNIAK, J., MOORE, G. W., ROMERO-HERRERA, A. E., AND MATSUDA, G. 1979. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology* 28:132–168.

GRAHAM, R. L. AND FOULDS, L. R. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computation time. *Mathematical Biosciences* 60:133–142.

HALLETT, M. T. AND LAGERGREN, J. 2000. New algorithms for the duplication-loss problem. In R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman (eds), *RECOMB '00, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, pp. 138–146. Association for Computing Machinery.

HUELSENBECK, J. P., RANNALA, B., AND LARGET, B. 2000a. A Bayesian framework for the analysis of cospeciation. *Evolution* 54:352–364.

HUELSENBECK, J. P., RANNALA, B., AND MASLY, J. P. 2000b. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288:2349–2350.

KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.

KWIATOWSKI, J., SKARECKY, D., BAILEY, K., AND AYALA, F. J. 1994. Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the cu,zn sod gene. *Journal of Molecular Evolution* 38:443–454.

LAPOINTE, F.-J. AND LEVASSEUR, C. 2004. Everything you always wanted to know about the average consensus, and more. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 87–105. Kluwer Academic, Dordrecht, the Netherlands.

MA, B., LI, M., AND ZHANG, L. 1998. On reconstructing species trees from gene trees in term of duplications and losses. In S. Istrail, P. A. Pevzner, and M. S. Waterman (eds), *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pp. 182–191. ACM, New York.

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

MARTIN, A. P. AND BURG, T. M. 2002. Perils of paralogy: using hsp70 genes for inferring organismal phylogenies. *Systematic Biology* 51:570–587.

MIRKIN, B., MUCHNIK, I., AND SMITH, T. F. 1996. A biologically consistent model for comparing molecular phylogenies. *Journal of Computational Biology* 2:493–507.

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

PAGE, R. D. M. 1998. GeneTree: comparing gene and species trees using reconciled trees. *Bioinformatics* 14:819–820.

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

PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.

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

PAGE, R. D. M. AND CHARLESTON, M. A. 1997b. Reconciled trees and incongruent gene and species trees. In B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky (eds), *Mathematical Hierarchies in Biology*, pp. 57–70. American Mathematical Society, Providence, Rhode Island.

PAGE, R. D. M. AND CHARLESTON, M. A. 1998. Trees within trees: phylogeny and historical associations. *Trends in Ecology and Evolution* 13:356–359.

PAGE, R. D. M. AND COTTON, J. A. 2000. GeneTree: a tool for exploring gene family evolution. In D. Sankoff and J. H. Nadeau (eds), *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, pp. 525–536. Kluwer Academic Publishers, Dordrecht, The Netherlands.

PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.

PURVIS, A. 1995. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.

RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. *Systematic Biology* 45:247–253.

RONQUIST, F. 2003. Parsimony analysis of coevolving species associations. In R. D. M. Page (ed.), *Tangled Trees: Phylogeny, Cospeciation and Coevolution*, pp. 22–64. University of Chicago Press, Chicago.

RONQUIST, F., HUELSENBECK, J. P., AND BRITTON, T. 2004. Bayesian supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 193–224. Kluwer Academic, Dordrecht, the Netherlands.

RONQUIST, F. AND NYLIN, S. 1990. Process and pattern in the evolution of species associations. *Systematic Zoology* 39:323–344.

RUSSO, C. A. M., TAKEZAKI, N., AND NEI, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution* 12:391–404.

SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.

SEMPLE, C. 2003. Reconstructing minimal rooted trees. *Discrete Applied Mathematics* 127:489–503.

SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.

SLOWINSKI, J. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.

SLOWINSKI, J. B., KNIGHT, A., AND ROONEY, A. P. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Molecular Phylogenetics and Evolution* 8:349–362.

STEEL, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.

STEEL, M., DRESS, A. W. M., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.

SWOFFORD, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer, Sunderland, Massachusetts.

TATARENKOV, A., KWIATOWSKI, J., SKARECKY, D., BARRIO, E., AND AYALA, F. J. 1999. On the evolution of *Dopa decarboxylase* (Ddc) and *Drosophila* systematics. *Journal of Molecular Evolution* 48:445–462.

THORLEY, J. L. 2000. *Cladistic Information, Leaf Stability and Supertree Construction*. Ph.D. dissertation, University of Bristol.

THORLEY, J. L. AND WILKINSON, M. 2003. A view of supertree methods. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 185–193. American Mathematical Society, Providence, Rhode Island.

WAREHAM, H. T. 1993. On the computational complexity of inferring evolutionary trees. Technical Report 9301, Department of Computer Science, Memorial University of Newfoundland.

WATERMAN, M. S. AND SMITH, T. F. 1978. On the similarity of dendrograms. *Journal of Theoretical Biology* 73:789–800.

WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.

ZHANG, L. 1997. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology* 4:177–187.

ZMASEK, C. M. AND EDDY, S. R. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17:821–828.