

## Chapter 1

# TANGLED TALES FROM MULTIPLE MARKERS

### *Reconciling conflict between phylogenies to build molecular supertrees*

James A. Cotton

jamec2@nhm.ac.uk

Roderic D. M. Page

r.page@bio.gla.ac.uk

*Division of Environmental and Evolutionary Biology*

*Institute of Biomedical and Life Sciences*

*University of Glasgow, Glasgow G12 8QQ, UK*

**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 seek to 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:** reconciled trees, gene duplication, gene tree parsimony

## 1. Introduction

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

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 of these trees (Aho et al., 1981; Semple, 2003; Steel, 1992). 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 either remove conflict by pruning leaves (e.g. in maximum agreement subtrees), represent the conflict through soft polytomies, resolve the conflict, or some combination of these.

In fact, the only supertree method that has been at all widely used by biologists is Matrix Representation with Parsimony (MRP, see chapter by Baum and Ragan in this volume), with an increasing number of supertrees constructed using this method appearing in the literature (e.g. Pisani et al., 2002; Kennedy and Page, 2002). MRP uses additive binary coding to represent the hierarchical structure 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 analysed using parsimony methods to construct a supertree or set of supertrees. While MRP supertrees have played an important part in stimulating the field of supertree research and may be reasonably successful in reconstructing relationships (Bininda-Emonds and Sanderson, 2001), there has been an increasing literature on the biases of MRP methods, and a similar number of proposed modifications to the original method (e.g. Bininda-Emonds and Bryant, 1998; Purvis, 1995; Ronquist, 1996; Thorley, 2000). There are similar problems with other supertree algorithms too, such as the min-cut supertree method (Semple and Steel, 2000), which has a number of 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).

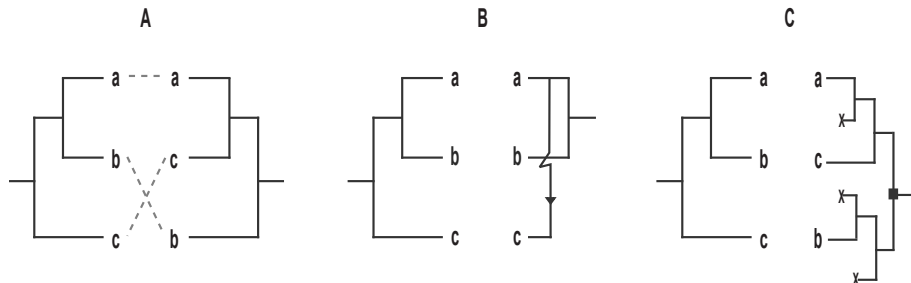
In an effort to classify the growing number of supertree methods available to systematists, at least two authors have characterised the

supertree problem in a distance framework (Thorley and Wilkinson, 2003; Chen et al., 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 minimising 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 in mind this framework, we should note that any problem of identifying an optimal tree is likely to be NP-complete (Wareham, 1993), including the maximum-parsimony problem used by MRP methods (Graham and Foulds, 1982), so heuristics 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 may have produced the differences observed between source trees. These events can be inferred using a co-phylogenetic method called reconciled trees. In this chapter, we introduce reconciled trees and their use to infer a species tree, or supertree, from a number of molecular source trees, which has become known as gene tree parsimony (GTP). We include a brief empirical example of a gene tree parsimony supertree. We then make a preliminary attempt to characterise the GTP method by suggesting some properties of the method, as has been attempted for other supertree methods. Lastly, we go beyond gene tree parsimony itself to argue that understanding the causes of conflict between source trees should help us resolve that conflict appropriately, and suggest that a model-based framework may enable systematic biologists to both understand the causes of conflict between trees and to construct accurate supertrees in the face of such conflict.

## **2. Tangled trees, or cophylogeny**

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 (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 re-



*Figure 1.1.* 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).

construct the host phylogeny, or to understand the pattern of evolution by optimizing the characters onto the second phylogeny (Brooks, 1981). Similarly 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 are non-independent.

In cophylogeny, the solution has been to explicitly map the dependant phylogeny into the host phylogeny, directly postulating events that lead to the differences between the two phylogenies (see figure 1.1). This insight led to Page's 1994 formalisation of the earlier concept of reconciled trees (introduced by Goodman et al., 1979). Constructing a reconciled tree involves reconciling the differences between two trees by postulating certain co-phylogenetic events that introduced these differences. As shown in figure 1.2 these events can be extinction of a lineage, independent speciation of a lineage and horizontal transfer. While cophylogeny 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 – lateral gene transfer, gene duplications and gene loss, so the same cophylogeny mapping can also be used to study this system. Other evolutionary processes are also included under these co-phylogenetic events, with, for example, hybridisation 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").

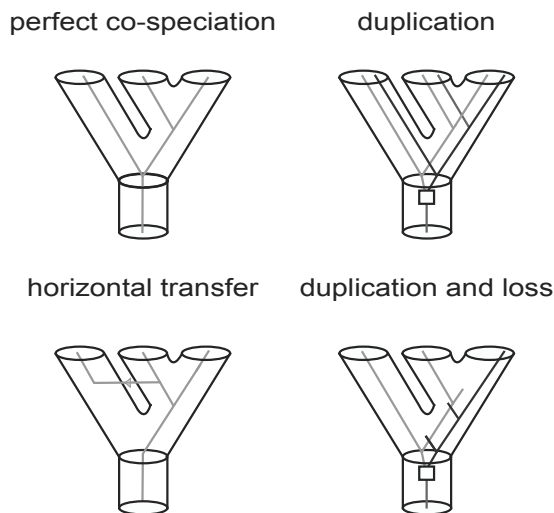


Figure 1.2. Some co-phylogenetic events, introducing differences between two associated phylogenies

This increasing depth is particularly due to the rise of genome-level sequencing efforts for an increasing number of organisms, and an important corollary of this work is the increasing realisation that phylogenies for different genetic loci for the same species frequently disagree. This has in turn prompted the realisation 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 (Maddison, 1997; Doyle, 1992). 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).

In principle, a number of different events can be scored in this way (figure 1.2), 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 computationally intensive and requires us to make more assumptions about the relative rates of gene duplication and loss and lateral gene transfer. For-

Fortunately, solutions for co-phylogeny mapping incorporating horizontal transfer are available, and could be used in the context of gene tree parsimony (Charleston, 1998; Ronquist and Nylin, 1990; Ronquist, 2003) and methods are available for estimating optimal event costs for particular problems which suggest that reconciliation methods are quite 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 effect the result obtained (Ronquist, 2003). Fortunately, it seems that the duplication-and-loss optimal trees will be 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 in some kinds of study, gene losses are confounded with failure-to-sample (simply 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) minimizing the number of co-phylogenetic events required to reconcile the species tree with each source tree, or minimizing 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 and so fits the definition of a conventional supertree. This method of combining data using reconciled trees has become known as "gene tree parsimony" (GTP, Slowinski and Page, 1999). 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 tree and species tree (Zhang, 1997; Eulenstein, 1997; Zmasek and Eddy, 2001), and while 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 are due to the evolutionary events included, then the gene tree parsimony supertree would be expected to accurately reconstruct the correct supertree (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 due to simple estimation error, due to the finite amount of data available from any single gene, while the inadequacy of existing models will also lead to some error, and so introduce some conflict between phylogenies. It may thus be that little of the error between phylogenetic estimates from different molecular markers is due to the kinds of evolutionary events dealt with by gene tree parsimony, and it is unclear how GTP will perform at resolving conflict from other 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. chapter in this volume). 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 under-perform compared to other methods when phylogenetic conflict is due to estimation error or model inadequacy. It seems that 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 is to incorporate some measure of uncertainty into the input source trees, for example from a bootstrap profile of trees from non-parametric bootstrapping (Salamin et al., 2002; Bininda-Emonds and Sanderson, 2001; Ronquist, 1996). 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' events rather than due 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. A number of 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***

A number of empirical examples of using reconciled tree methods to infer phylogenies are available in the literature (Page, 2000; Cotton and Page, 2002; Slowinski et al., 1997; Martin and Burg, 2002), but here we present a small empirical example, of a small-scale supertree of *Drosophila* and some related genera based on five nuclear genes (figure 1.3). The trees were re-labelled with the species names and then a standard MRP matrix was built using the program supertree<sup>1</sup>. The standard MRP matrix was analysed using PAUP\* 4b10, using standard parsimony. The GTP analysis was performed using the program GENETREE (Page, 1998). For both GTP and MRP 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, suggesting that the five searches had each successfully sampled from across the large island of trees, finding trees of cost 97 parsimony steps in MRP and 63 duplications and losses under GTP. The set of gene tree parsimony supertrees thus includes all the trees found to be reconciled with the 5 source trees using just 63 duplications and losses (in fact, all the supertrees found require either 17 duplications and 46 losses, or 18 duplications and 45 losses). While 18 duplications sounds like a lot, 7 duplications are required by multiple gene copies being present on the *Alcohol dehydrogenase* gene tree (see figure 1.3), so only at most 11 duplications and due to incongruence between the source trees and supertrees.

Given that they come from the same data, it is reassuring that both the MRP and GTP analyses are rather similar (figures 1.4 and 1.5). Both analyses agree in supporting the monophyly of the subgenus *Sophophora*, and indeed show exactly the same relationships within *Sophophora*. It appears that GTP is rather more conservative than MRP, in that it is largely compatible with the MRP results but somewhat less well-resolved, although this is not the case for every clade. Both GTP and MRP find the other subgenera of *Drosophila* included to be para- or

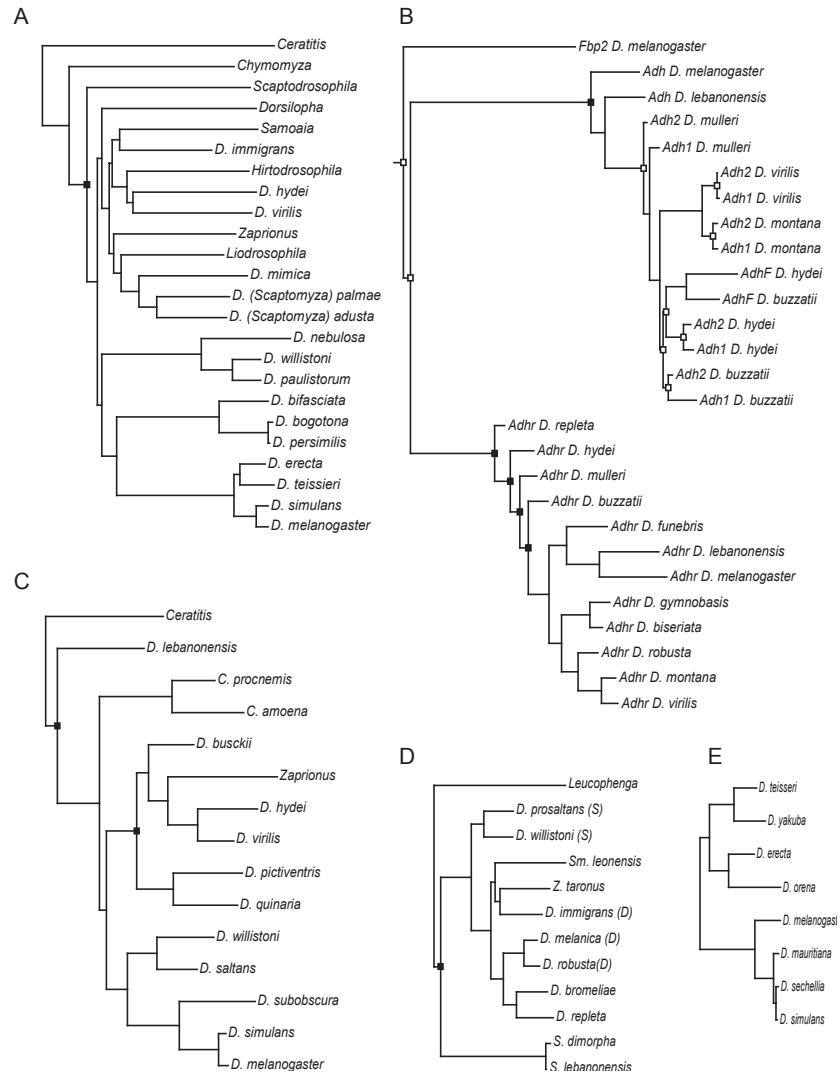


Figure 1.3. The five gene trees used in building the *Drosophila* supertree presented here. (A) is for *Dopa decarboxylase* (Tatarenkov et al., 1999). (B) is for *Alcohol dehydrogenase* and the *Alcohol dehydrogenase-related* gene (Betrán and Ashburner, 2000). (C) is for Cu-Zn superoxide dismutase (Kwiatowski et al., 1994). (D) is for 28S rRNA (Russo et al., 1995). (E) is for 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, while closed boxes are those duplications inferred from conflict between the gene tree and the supertree. All the duplications are implied by every supertree except the duplication on the 28S rRNA.

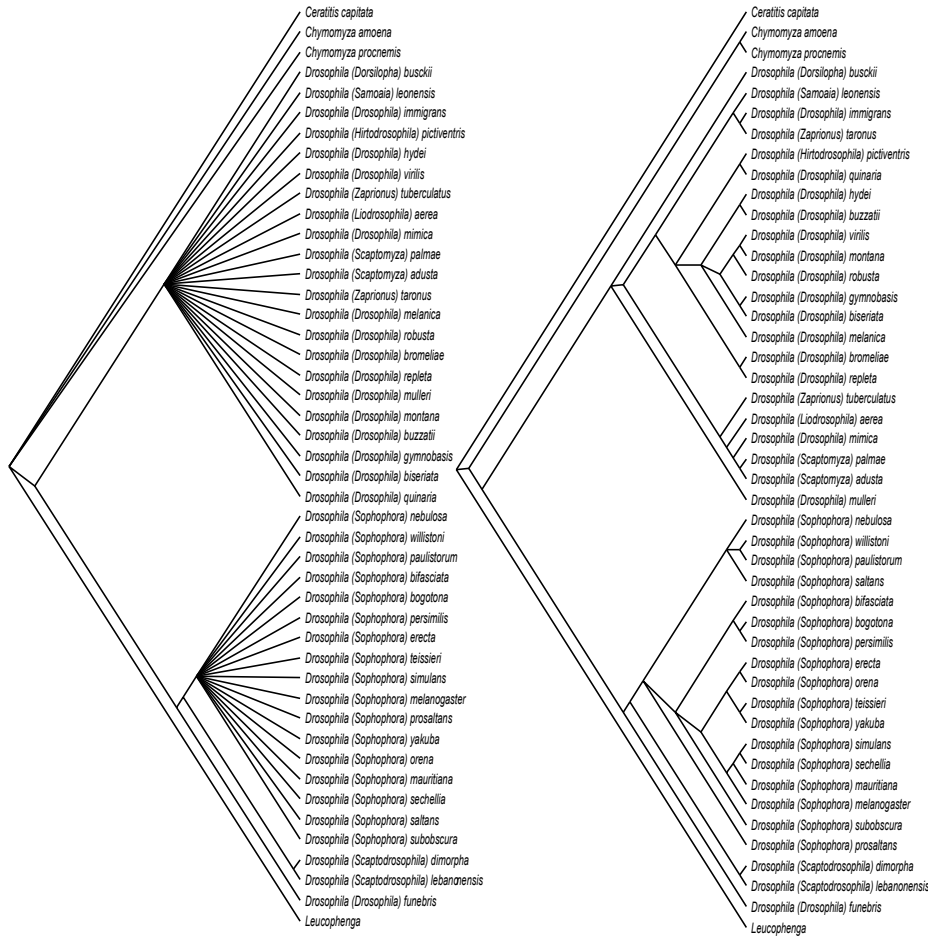


Figure 1.4. The strict component (left) and Adams (right) consensus of the Gene Tree Parsimony supertrees from the *Drosophila* gene trees, under the duplication-and-loss criterion.

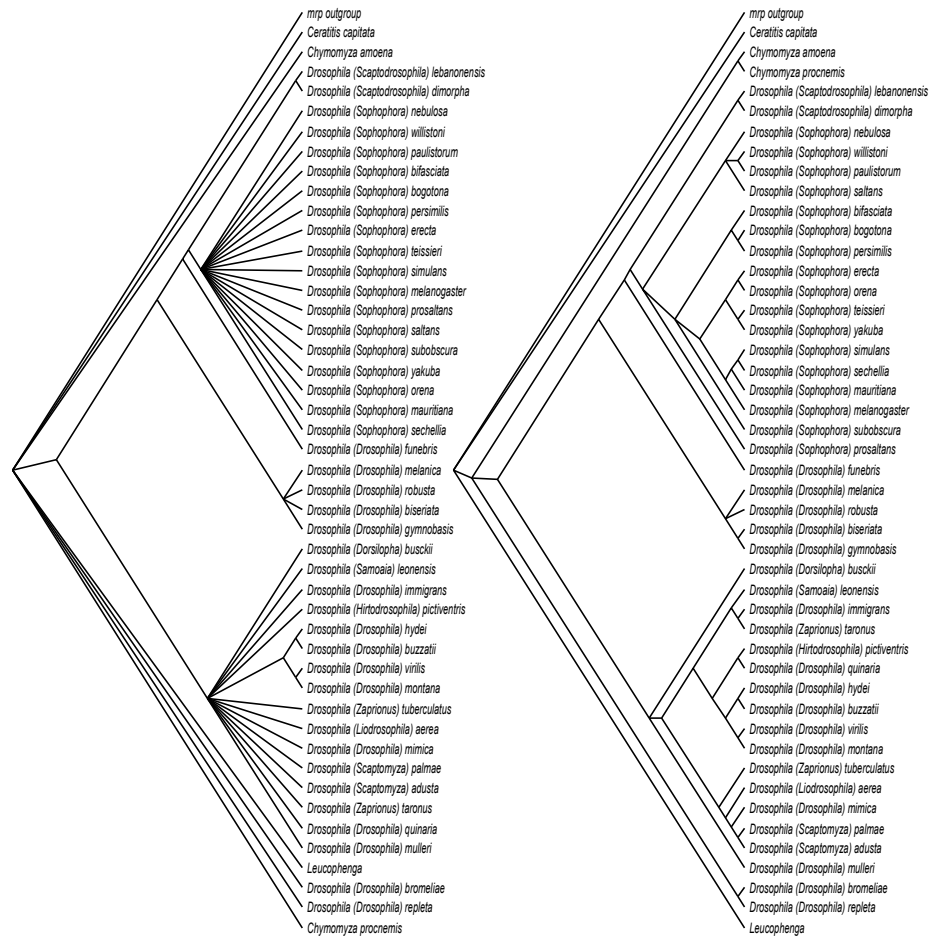


Figure 1.5. The strict component (left) and Adams (right) consensus of the standard MRP supertrees.

poly-phyletic, but there are some differences between the two sets of trees. One instructive difference is in the way the species *D. melanica*, *D. robusta*, *D. biseriata* and *D. gymnobasis* cluster with respect to one another – in the MRP results, these taxa are placed strikingly away from the other members of subgenus *Drosophila*, as a sister clade to the subgenera *Sophophora* and *Scaptodrosophila* and *D. funebris*, in sharp contrast to the GTP tree, where these taxa are embedded within the paraphyletic assemblage around subgenus *Drosophila*, to which they all belong. The MRP results seems surprising on examination of the source trees – the four taxa in question appear only on trees 1.3B and 1.3D, and in both of these trees they are grouped with other members of the subgenus *Drosophila*. This odd placement is probably partly due to the different treatment of *D. bromeliae* and *D. repleta*, the other principal difference between the two sets of trees. These two taxa are the sister taxa to *D. melanica* and *D. robusta* on the 28S tree (1.3D). The unresolved position of *D. bromeliae* and *D. repleta* on the MRP tree is understandable, given that these two taxa are placed (with three other members of the clade containing subgenus *Drosophila* and others), somewhere between subgenera *Sophophora* and *Scaptodrosophila* on the 28S tree. However, given the support for the three other taxa as related to members of the subgenus *Drosophila* on two other trees (1.3A and 1.3C), 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., as it contains the vast majority of duplications and losses) and help us to understand the general pattern of genetic evolution for this group. Furthermore, the hypothesised duplications and losses may be testable using other evidence: for example, do the suggested paralogs have different functions, occur in different parts of the genome or have different genetic architecture? Another approach might be to use the GTP supertree to inform a search for additional gene copies – 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 gene tree parsimony as a supertree method

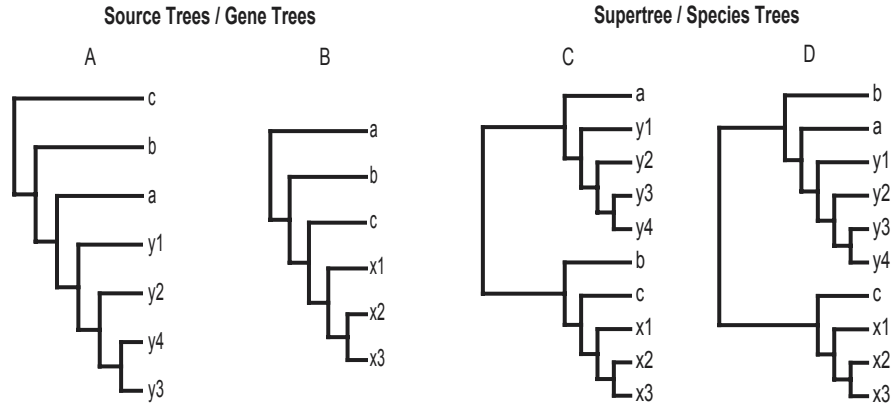
Recently, some progress has been made in thinking about desirable properties of supertree methods (see Wilkinson et al. chapter, this volume). These properties are characteristics of particular methods that would seem to be desirable in any supertree method, and which seem likely to correlate with the accuracy of a method's results. Fairly little has been done to formally characterising supertree methods in terms of these properties or more formal axioms. In particular, it may be of interest to see how gene tree parsimony resolves conflict between source trees when compared to those variants of MRP that are already characterised in terms of some of these properties. The properties named in italics below are used in the sense of Wilkinson et al., in this volume. Aside from the three properties discussed below, GTP methods are *assessable*, *weightable*, *plenary*, and show *order invariance*, and seems to be *pareto*, on components. They do not show *generality*, or *uniqueness*, and are not particularly *speedy*, compared to polynomial-time methods. Their behaviour in terms of being *co-pareto* and *independent of irrelevant alternatives* is unclear.

### 5.1 GTP correctly displays unique subtrees

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. Gene tree parsimony appears to correctly include unique subtrees in the supertree or species tree, a property shared by MRP methods, but not by the original formulation of mincut supertrees (Page, 2002). Using Page's example (figure 1.6) we can see that gene tree parsimony correctly reconstructs these groupings under both duplication-only and duplication-and-loss criteria. Both gene tree parsimony and MRP perform better than the modified mincut method, in correctly placing taxon a 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 any supertree method. 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 suggested for MRP matrices produces supertrees biased towards including those relationships shown



*Figure 1.6.* Trees C and D are the two supertrees for source trees A and B under both the duplication-only and duplication-and-loss costs. The same two trees (C and D) are also the standard and Purvis coding MRP supertrees for trees A and B. (source trees taken from Page, 2002).

on larger source trees (Purvis, 1995) because of redundant information in the matrix. Purvis shows that some matrix entries are redundant in the sense of not being needed to reconstruct the original source trees, but this information may not be redundant in a different sense (see Ronquist, 1996). We use Purvis's example to show that gene tree parsimony 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 figure 1.7 A and B support just a single species tree under the duplication-and-loss criterion, that of fig.1.7C. This tree places taxon d in the position supported by tree A, the larger of the two source trees, despite the very different position of this taxon in tree B, and so effectively ignores the conflicting signal from this smaller tree. Under the duplication-only criterion an additional species tree (fig 1.7D) 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 more accurately. This source of bias disappears under the duplication-only criterion.

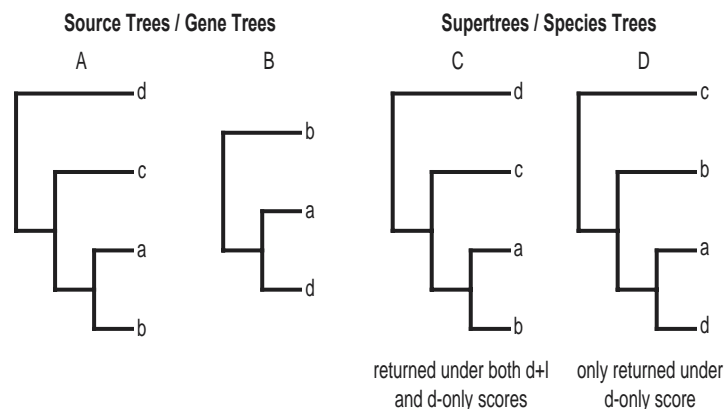


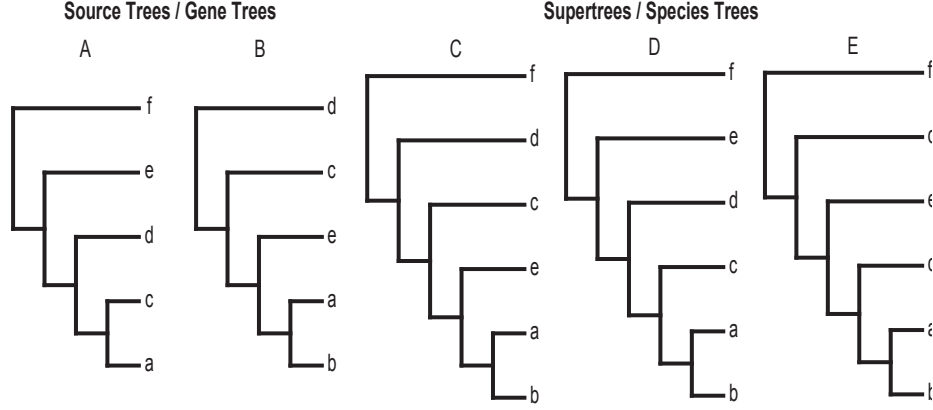
Figure 1.7. Trees C and D are the two supertrees for source trees A and B under the duplication-only cost. Tree C is the unique supertree under the duplication-and-loss cost. Tree C is the unique supertree under standard MRP, while both C and D are purvis coding MRP supertrees

### 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 (Purvis, 1995) suggested modification to the original MRP encoding, as is shown by the example in figure 1.8 (from Thorley, 2000). The figure 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 (fig 1.8C). This places taxon e in the more crownward position, as suggested by source tree B, overruling the conflicting position suggested by source tree A. It seems that GTP thus shows a bias towards placing taxa in the more crownward position.

## 6. A probabilistic view of the supertree problem

We can usefully view the supertree problem in a probabilistic setting, a view which makes a number of the themes of this paper particularly clear. This is a fairly natural extension of the distance-based view expressed earlier – instead of 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, which is proportional to the probability that the source trees come from the supertree. There are a number of different ways we could frame this likelihood function based on how similar the source trees are to the



*Figure 1.8.* 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-E are MRP supertrees under standard coding.

subtrees of the proposed supertree induced by their leaf sets. For example, if we assume every Nearest-Neighbour Interchange (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 source trees would then be the likelihood of this supertree under this simple NNI-binomial model. This model has only a single parameter,  $q$ , for the probability of an NNI difference between a source tree and the supertree, that must be estimated from the data. For a supertree,  $T_s$ , from a set of  $n$  subtrees,  $T_1 \dots T_n$ , 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  $G$ , the likelihood of the supertree is given by:

$$L(T_s | T_1, T_2, \dots, T_n) \propto \prod_{i=1}^n p(T_i | T_s) \quad (1.1)$$

where the probability of each source tree is simply:

$$p(T_i | T_s, q) = \binom{G}{d_{T_i, T_s}} q^{d_{T_i, T_s}} (1 - q)^{\Delta G - d_{T_i, T_s}} \quad (1.2)$$

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 v6wbu6ytri-

## Conclusion

We are clearly at an early stage in the development of supertree methods – many methods are being proposed, but little is known about the relative merits of the different approaches. While most supertree methods treat conflict between source trees in a fairly ad-hoc way, it is possible to treat such incompatibility in a biologically realistic way, at least for some causes of incompatibility. 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 likely that different methods will be preferable for different data, reflecting the different causes of conflict in these different data. For example, reconciled tree methods may be the most appropriate if all conflict between source trees is caused by gene duplication and gene loss (probably a rather unlikely scenario), whereas MRF (matrix representation with flipping, see chapter by Burleigh et al.), may 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, Frederik Ronquist, Gordon Burleigh and Mark Wilkinson for comments on the manuscript. JAC has been supported by a NERC studentship and by BBSRC grant 40/G18385.

## Notes

1. available from <http://darwin.zoology.gla.ac.uk/~rpage/supertree/>

## 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.

- 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 Shamir, R., Miyano, S., Istrail, S., Pevzner, P., and Waterman, M., editors, *RECOMB '00, Proceedings of the fourth annual international conference on computational molecular biology*. 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. *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.
- Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses. In Istrail, S., Pevzner, P. A., and Waterman, M. S., editors, *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pages 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 Guigó, R. and Gusfield, D., editors, *Proceedings of WABI 2002*, volume 2452 of *Lecture Notes in Computer Science*, pages 537–551. Springer-Verlag.
- 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 Mirkin, B., McMorris, F., Roberts, F., and Rzhetsky, A., editors, *Mathematical Hierarchies in Biology*, volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 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 Sanko, D. and Nadeau, J. H., editors, *Comparative Genomics*, pages 525–536. Kluwer Academic Publishers, Utrecht.
- 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 *Tangled Trees: phylogeny, cospeciation and coevolution*, pages 22–64. University of Chicago Press.
- 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. *Systematic Biology*, 51:136–150.
- Semple, C. (2003). Reconstructing minimal rooted trees. *Discrete Applied Mathematics*, In Press.
- 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.
- 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*. PhD thesis, University of Bristol.
- Thorley, J. L. and Wilkinson, M. (2003). Reduced consensus and super-tree methods. In Janowitz, M., Lapointe, F.-J., McMorris, F., Mirkin, B., and Roberts, F., editors, *Bioconsensus, In Press*. DIMACS-AMS.
- 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.
- 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.