Zoologica Scripta

Comparing tree shapes: beyond symmetry

Pablo A. Goloboff, Joan S. Arias & Claudia A. Szumik

Submitted: 23 August 2016
Accepted: 4 December 2016
doi:10.1111/zsc.12231

Goloboff, P.A., Arias, J.S. & Szumik, C.A. (2017). Comparing tree shapes: beyond symmetry. — *Zoologica Scripta*, *00*, 000–000.

This paper describes two types of problems related to tree shapes, as well as algorithms that can be used to solve these problems. The first problem is that of comparing the similarity of the unlabelled shapes instead of merely their degree of balance, in a manner analogous to that routinely used to compare topologies for labelled trees. There are possible practical applications for this comparison, such as determining, based on tree shape similarity alone, whether the taxa in two phylogenies are likely to have a correspondence (e.g. hosts and parasites with high specificity). It is shown that tree balance is insufficient for this task and that standard measures of topological difference (Robinson–Foulds distances, SPR distances or retention indices of the matrices representing the trees, MRPs) can be easily adapted to the problem. The second type of problem is to determine whether taxa of uncertain matching unique to two different phylogenies could correspond to each other (e.g. the same species in larvae and adults of metamorphic animals, fossils known from different body parts). This second problem can be solved by either relabelling taxa in such a way that the number of consensus nodes is maximized, or relabelling taxa in such a way that the sum of the number of steps in the MRP of each tree mapped onto the other is minimum.

Corresponding author: *Pablo A. Goloboff, Unidad Ejecutora Lillo, Miguel Lillo 251, 4000 S.M. Tucumán, Argentina. E-mail: pablogolo@csnat.unt.edu.ar*

Pablo A. Goloboff, Unidad Ejecutora Lillo, Fundación Miguel Lillo, CONICET, Miguel Lillo 251, 4000 San Miguel de Tucumán, Argentina. E-mail: pablogolo@yahoo.com.ar

Joan S. Arias, Unidad Ejecutora Lillo, Fundación Miguel Lillo, CONICET, Miguel Lillo 251, 4000 San Miguel de Tucumán, Argentina and Facultad de Ciencias Naturales, Universidad Nacional de Tucumán, Miguel Lillo 205, 4000 San Miguel de Tucumán, Argentina. E-mail: jsalarias@gmail.com

Claudia A. Szumik, Unidad Ejecutora Lillo, Fundación Miguel Lillo, CONICET, Miguel Lillo 251, 4000 San Miguel de Tucumán, Argentina. E-mail: szu.claudia@gmail.com

Introduction

This paper calls attention to possible implications, for practising taxonomists and systematists, of aspects related to tree shape. In phylogenetics, most of the relevant comparisons between trees are made for fully labelled trees. Unlabelled trees are rarely referred to, especially by taxonomists and systematists. Tree shapes have long been considered as a means to test speciation/extinction models, an active field of research which took momentum after Mooers & Heard (1997). Almost all of the literature on this subject is based on testing whether the degree of symmetry (or 'balance') in trees obtained from real data matches that expected under the model. Using the degree of symmetry in this way is so common that, in many papers, 'shape' and 'symmetry' are used interchangeably. The degree of symmetry is most commonly measured with either the Sackin (Sackin 1972) or Colless (Colless 1982) indices; other indices were proposed by Shao & Sokal (1990), Mooers & Heard (1997), McKenzie & Steel (2000) and Mir *et al.* (2013). Recent examples of practical applications of this approach can be found in Poon *et al.* (2013), and Frost & Volz (2013). Related to this is the work on diversification rates (e.g. Alfaro *et al.* 2009; Shah *et al.* 2012; Rabosky 2014), which uses indirect measures of balance (i.e. differences in number of branching events in sister clades). Several R-packages implement these indices, such as Ape (Paradis *et al.* 2004), apTreeshape (Bortolussi *et al.* 2006) and PhyloTempo (Norström *et al.* 2012). Stadler (2013) provided a recent general review of the subject.

Although measures of tree balance seem adequate for testing specific evolutionary models, the more general goal of comparing the shapes – as opposed to merely the degree of balance – has received little attention from biologists and phylogeneticists, and it is presently unclear how such comparisons should be made. As recently noted by Lewitus & Morlon (2016: 495–496), '[m]etrics like the Robinson–Foulds distance (Robinson & Foulds 1981) and nearest neighbor interchange (Moore *et al.* 1973) ... are used to compare different trees representing the same set of organisms ... They are not, however, built (or adapted) to function as comparative metrics between species trees representing different sets of organisms'. The present paper proposes such an adaptation of the Robinson–Foulds distance (and other measures of tree distance) and illustrates contexts in which this comparison may be useful.

A few papers in the field of computer science have dealt with the problem of comparing unlabelled tree shapes. These papers have studied the cost of editing the strings of parentheses representing the two trees, so that they become equivalent (e.g. Germain & Pallo 1996; Wu & Huang 2010), or the number of nearest neighbour interchange (NNI) moves needed to convert the trees themselves (e.g. Pallo 1990; in this field, the NNI rearrangement operation is often called a 'tree rotation'). Although these computer science papers deal with the problem of comparing tree shapes beyond the mere quantification of balance, they have had little influence on biologists.

In one of the most interesting papers on the subject, Matsen (2006) has pointed out that tree comparisons based only on the degree of balance can be misleading and that some more general evaluation of shape differences may be desirable, and makes some points similar to those in the present paper. As the present paper is less mathematically oriented than Matsen's (2006), presents specific instances of the relevance of tree shapes for systematics (with biological examples) and provides a computer implementation to compare tree shapes, we hope that it will be more accessible to biologists and more effective in calling attention to the relevance of tree shape in systematics.

Shape is not the same thing as balance

Measuring the balance of trees may be appropriate to test specific models of evolution, but there is little doubt that balance alone is simply one of the aspects of the 'shape' of a tree. As the total degree of balance is a summary over all the tree, trees which are closer in balance are not necessarily closer in shape. Matsen (2006) made the same point using a hypothetical diagram; Fig. 1 provides a specific example, with four trees (A–D) arranged so that asymmetry increases to the right. The values of the Sackin index (not normalized, not counting the root itself) are shown below each tree.

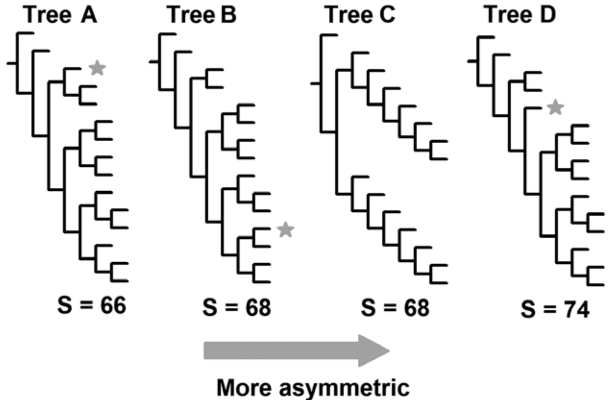

Fig. 1 A case of four tree shapes (A–D) with different degrees of balance, as measured by Sackin's index (S). Shape B is exactly as balanced as shape C, but it is very different, and more similar to the other shapes (A and D, more and less balanced, respectively). Tree shapes A, B and D can be interconverted by moving just the terminal branch marked with a star.

If differences in Sackin's index were to be interpreted as a measure of differences in shape, the two trees that differ the most would be trees A and D. It is clear, however, that trees A and D are the two trees with most similar shapes (at least for most possible measures of similarity; see below). For example, it is possible to interconvert between trees A and D by an NNI move of a single branch (marked with an asterisk in both trees). Tree C would be, under that interpretation of Sackin's index, identical to tree B, but it is clear that tree C is the most different tree: all the other trees can be interconverted between each other with a single SPR move.

Thus, using Sackin's index to measure the similarity of shapes – instead of the degree of balance – is obviously inappropriate. Note that similar examples could be constructed using any other measure of tree balance. Colless' statistic, for example, would arrange the four trees in the same sequence, also indicating that trees A and D are the most different ones and that D is most similar to C (the only difference is that trees B and C would not be identical in that case; other sets of four trees, however, reproduce the exact same situation of Fig. 1 for Colless' statistic).

For comparing labelled trees, taxonomists routinely use a number of measures of distance and similarity. Felsenstein (2004, pp. 528–535) gives an overview; a more recent review of the performance of some measures is provided by Kuhner & Yamato (2015). These measures of distance between labelled trees are well known by biologists, and their meaning and definition are for the most part intuitive. Two widely used measures are the Robinson–Foulds distance (RF; Robinson & Foulds 1979, 1981; the 'rooted' version is used in this paper) and the SPR distance (number of subtree–pruning–regrafting moves needed to interconvert the trees; Hein 1990). A third measure is the retention index R of Farris (1989, a modification of the 'distortion coefficient' of Farris 1973), based on calculating the number of steps for the matrix representing one of the trees mapped onto the other, or MRP ('matrix representation with parsimony' of Baum & Ragan 1993; what Farris 1973 had called 'group membership variables'). Although this measure, R, is not commonly used, it has interesting properties (including the ease of calculation and ability to detect subtle difference in the trees); Wheeler (1999) proposed extensions of Farris' (1973) original measure; Goloboff *et al.* (2008a) used R to compare trees in their analysis of the influence of weighting. In the present paper, the totals of G (maximum possible), M (minimum possible) and S (observed) are used (as in Farris 1989) instead of the average (as in Farris's 1973 original proposal); this is a measure of similarity, with 1 indicating identity. Farris' (1989) original measure is asymmetrical when used to measure tree similarity: the steps of the MRP for tree A mapped onto tree B may not the same as the reciprocal (Goloboff 2005), and thus, the measure cannot be a metric. To eliminate that problem of asymmetry, the sum of the maxima, minima and steps of the reciprocally mapped matrices is used (as implemented in the *tcomp* command of TNT; Goloboff *et al.* 2008b), so that the measure can become a metric.

Of course, as these (and other) measures of similarity between labelled trees may arrange trees in slightly different sequences, it is clear that they capture slightly different aspects of 'the shape' (see, e.g., Bansal *et al.* 2010: 10), but the interpretation and limitations of these measures are generally well understood. For example, it is well known that RF may be increased more by a single terminal moving to a far away position than by several terminals moving to nearby locations, even if the second case can be considered as changing the tree more. In what is in a sense the opposite situation, the SPR distance between the original tree and each of the trees resulting from moving a terminal to alternative locations (farther or closer away) will always be the same, even if some of those moves represent a more radical transformation of the tree (e.g. Goloboff 2008). The retention index R will be intermediate in any of these situations. In general, biologists are well aware of the contexts in which a given measure may be problematic.

Coming back now to the shape of unlabelled trees, how could these be compared, beyond mere balance? For practising phylogeneticists to easily grasp the meaning of a given tree comparison, the ideal situation would be that some of the measures already established for labelled trees are adapted to the unlabelled case (cf. Lewitus & Morlon 2016). If the tree shapes are identical, then it will be possible to label both trees in such a way that the measure of distance d indicates that no difference exists between the trees. Likewise, when the shapes are not exactly identical, labelling the trees in such a way that d produces the maximum possible similarity between the shapes will give an indication of the similarity between shapes. More properly, it gives an indication which is as good as the measure used; for example, just one taxon switching between two distant positions will strongly increase the RF, will count as just one SPR move and will mildly decrease the value of R (more strongly so for more distant moves). In this paper, to indicate that these measures of distance or similarity between trees are being used to compare shapes, through optimal relabelling of the two trees, the subindex s ('shape') will be used: RF_s, R_s and SPR_s distances.

When applied to the example of Fig. 1, this relabelling produces a sensible comparison between the tree shapes. Figure 2 shows a possible optimal labelling. Note that the optimal labelling will normally be different for different pairwise comparisons between trees (in this particular example, the four trees can be optimally labelled at the same time). Figure 2 shows the consensus of possible pairs of trees, together with the values of R_s and SPR_s distances. With the labelling shown, tree C is identified as the most different tree (with the lowest number of groups shared, the lowest R_s and the largest number of SPR_s moves, when compared to any of the other trees), and trees A–D are identified as the most similar. Tree B is more similar to trees A and D than it is to tree C (even if it is exactly as balanced as tree C, according to Sackin's index).

When is tree shape relevant for systematists?

In general, comparing tree shapes may be of interest when the correspondence between the taxa in one tree and the taxa in the other is not known with certainty, but might be established as consequence of topological correspondence. These will generally be special circumstances, of cases where lack of information forces us to rely on the similarity of phylogenies to establish a correspondence. It is hard to know how often the situation may arise in the practical work of taxonomists, especially because publication of phylogenies fulfilling this requirement may have been impeded by the general lack of discussion on the problem, and the lack of relevant phylogenetic tools. The case is certainly possible, and even if uncommon, interesting from the methodological point of view.

Consider a hypothetical case where phylogenies are known for a genus of fish, and for a group of parasites suspected (but not known) to parasitize that genus of fish with high specificity. Of course nothing is better than having actual data on the association, but assume for the sake of the argument that only the phylogenies are available, for the same numbers of terminal taxa in each case. The shape of the parasite phylogeny is as in tree D of Figs 1–2; the

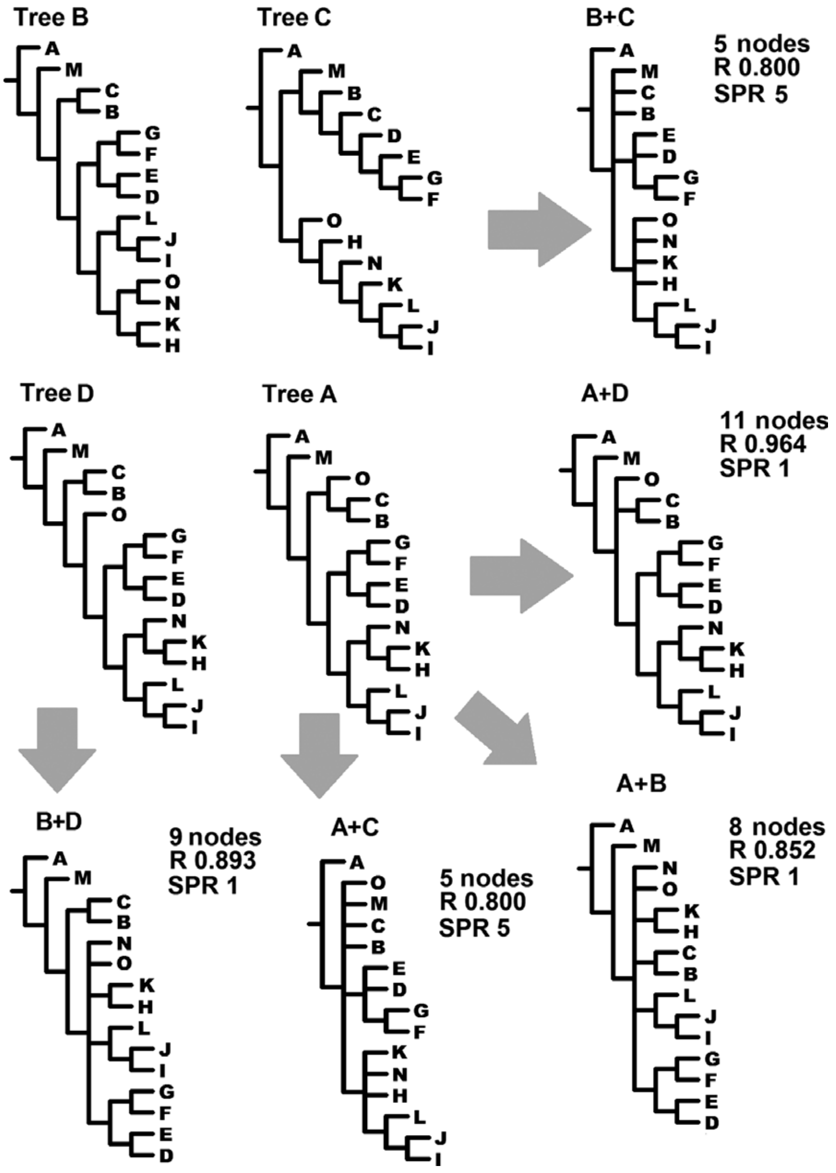
Tree B
A M C B G F E D L J I O N K H
Tree C
A M B C D E G F O H N K L J I
B+C
A M C B E D G F O N K H L J I
5 nodes
R 0.800
SPR 5
Tree D
A M C B O G F E D N K H L J I
Tree A
A M O C B G F E D N K H L J I
A+D
A M O C B G F E D N K H L J I
11 nodes
R 0.964
SPR 1
B+D
A M C B N O K H L J I G F E D
9 nodes
R 0.893
SPR 1
A+C
A O M C B E D G F K N H L J I
5 nodes
R 0.800
SPR 5
A+B
A M N O K H C B L J I G F E D
8 nodes
R 0.852
SPR 1

Fig. 2 The four shapes shown in Fig. 1, labelled so that the trees become as similar as possible. The consensus of different tree pairs is shown (as well as the number of nodes in the strict consensus). With this labelling, tree C is the most distinct tree, for the Robinson–Foulds distances, the retention index of the MRPs (R) and the number of SPR moves. Note that for two binary trees of t taxa, the Robinson–Foulds distance is directly proportional to the number of nodes n in their strict consensus (with $RF = 2 \times (t - n - 3)$), so that fewer nodes in the consensus indicate a larger RF distance.

shape of the fish phylogeny is as in tree C. If the parasites are highly specific, then the shapes of the two phylogenies should be similar. But the shapes of trees D and C are very different; this situation is not what could be expected if the parasites are highly specific for that genus of fish. Trees D and C differ in shape more (at the 5% significance level) than expected if the topologies of the fish and parasite are totally independent of one another, when measured with RF_s. The proper test to address this problem is the generation of multiple pairs of random trees (i.e. all trees equiprobable), relabelling the trees so that RF is minimal, and counting the proportion of cases in which RF_s is larger than the observed. In TNT, this is easily achieved with the following commands:

```
shpcomp =+ 0 1 ;
var: OBS count ;
set OBS rrfdist 0/1 ;
set count 0;
loop 1 1000
  keep 0; rseed*; randtree 2;
  shpcomp =+ 0 1;
  if ( rrfdist 0/1 >='OBS')
    set count ++ ; end
    stop
quote'count' cases;
```

The pairs of random trees had differences in shape equal to or greater than the observed one (0.53846) in only 3.4% per cent of the cases; thus, by generating pairs of random trees, it is unlikely that we will obtain trees that are as *different* in shape as trees C and D. If the phylogeny of another fish genus is available, with shape as in tree A of Figs 1–2, then that other fish genus is a much better candidate for being the host of these parasites; even if the shapes of the trees are not exactly identical, only 0.1% of the pairs of random trees had a RF_s smaller than the observed (0.07692; we used 'smaller than or equal' in this comparison, instead of 'larger than or equal'). Thus, the shape of the phylogeny of the second genus of fish resembles the shape of the phylogeny of the parasite more than expected by chance alone.

Similar examples can be made regarding the association between larvae and adults in animals with metamorphosis, or males and females that have not been unequivocally associated in groups with strong sexual dimorphism. In palaeontology, it is common for fossil taxa to be known from disarticulated remains and the association between different body parts is often uncertain. Applying a reasoning like the one illustrated for hosts and parasites may serve to determine whether cranial and postcranial remains (for example) are likely to belong to the same group of taxa.

Calculating optimal relabellings

Although the approach above is conceptually simple, the relabelling that maximizes a given measure of tree similarity is not easy to find. This section describes the algorithm used in recent versions of TNT (Goloboff *et al.* 2003; Goloboff & Catalano 2016), which does a sort of 'heuristic search' for better labellings.

The algorithms used in TNT take advantage of the fact that the MRP can be updated quickly, using incremental reoptimization (as in Goloboff 1996) of the variables representing groups to measure changes in fit (i.e. differences in sums of observed steps, *S*, in the formula for the retention index, *R*; note that the relabelling can affect neither *G* nor *M*).

Unlike the case for searching trees, the shape of both trees must remain the same, and this requires that all the 'moves' are instead switches between taxa in one of the trees. The exchange between two taxa A and B in one tree proceeds in the following steps: first, the states of A are made identical to those of B for each character representing a group in the other tree and then each character is reoptimized incrementally; second, the states for B are made identical to those that A had, and each character is reoptimized incrementally; third, the actual tree structure is changed, switching taxa A and B. The other tree does not change, but the matrix representing the changed tree must be changed (by switching the states assigned to taxa A and B), then incrementally reoptimizing the other tree below A and B. Differences in length for each of the characters are calculated as the incremental reoptimization proceeds.

Rather than attempting switches at random, the switches are attempted on the basis of the mapping on one tree of each character representing a group of the other tree. By switching together independent derivations of the same state, or switching down to plesiomorphies what are mapped as reversals, the number of steps needed to fit the MRP is more likely to be decreased. Such guided switches consists of three phases (a)–(c):

(a) For every character of the MRP representing one of the trees (tree A), mapped onto the other (tree B), list all the branches of tree B with 1→0 changes (including potential ones, resulting from ambiguous optimization). From all pairs of branches in the list (b_i, b_j, where $i > j$), for each of the sisters (in tree B) of b_i that do not have state 0, attempt a label switch between b_j and the sister of b_i. If the switch produces a better MRP score, accept it, remove b_j from the list and update ancestral assignments for both MRPs. This is illustrated in Fig. 3A. After trying all the sisters of b_i, if none produced a better MRP score, then for each of the sisters of b_j that do not have state 0, attempt a label switch in the same way.

(b) For every character of the MRP, list all the branches with 0→1 changes. Operate similarly: for all b_i, b_j in the list, attempt label switches for the sisters that do not have state 1 (instead of 0, as before), first for the sisters of b_i, then if no exchange produced any improvement, for the sisters of b_j that do not have state 1.

(c) For each branch b_i of tree B that remained in the first list (i.e. the list of branches with 1→0 changes), travel down the tree (i.e. from b_i towards root) until finding the first node n which there is a change 0→1. Let b_j be the descendant of that node n in the path to b_i. Then, for each of the sisters of b_j, attempt to switch labels between b_i and sister of b_j. This is illustrated in Fig. 3B. If the switch produces a better score, remove b_i from the list and update ancestral assignments on each tree B for the MRP representing the other tree.

A cycle of guided switches consists of performing routines (a–c), in both directions (i.e. exchanging trees A and B once). In the illustration, only terminal taxa are being exchanged. When b_i or b_j are groups (internal nodes) instead of terminal taxa, the exchange is performed by switching, one at a time, each of the terminals from the two groups (updating the MRPs for each exchange of terminal taxa); these individual pairwise exchanges are tried in an arbitrary sequence. Note that when the groups are of different size, some terminal taxa in the larger group will not be exchanged.

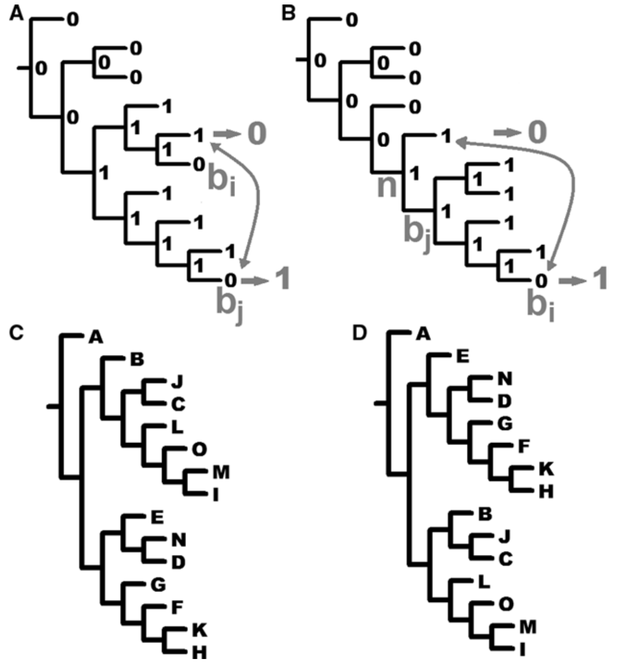

Fig. 3 —A. Example to show how taxon switches in a tree can be guided by parallel 1→0 changes in a character representing a group in the other tree; exchanging b_j and the sister of b_i reduces the number of steps in the character from 3 to 2, thus making the groups in the trees more similar. A similar reasoning can be applied for parallel 0→1 changes. —B. Example to show how taxon switches can be guided by reversals; moving bi down the tree, to be the first sister of the rest of the group (n) with a change 0→1, causes the number of steps for the character to go from 2 to 1. —C,D. Two identical shapes, labelled so that the taxon-switching algorithm cannot produce any improvement in the similarity of the tree (i.e. a local optimum for the relabelling algorithm).

Every individual switch can be carried out (or undone) rather quickly, using incremental reoptimization; in a standard computer, TNT can evaluate about 87 800 switches per second for 50 taxa and about 36 500 for 100 taxa. These guided switches quickly improve the mutual MRP fit, but can get trapped in 'local optima'. An example is in the two trees shown in Fig. 3C,D. Note that the two tree shapes are identical, and the first split below the out-group comprises two groups of the exact same size. It is possible to obtain two identical labelled topologies by exchanging the taxa in those two groups, but this would have to be carried out with a specific sequence. That is, B must be exchanged with E, C with D, F with O, etc., at the same time. Recall that the exchange of internal nodes is performed by exchanging the terminals descended from the nodes in an arbitrary sequence; there is no way to know which specific sequence of exchanges will produce the desired identity. Thus, if the guided switches start from a random labelling, the two trees in Fig. 3C,D become identical in about half of the cases, the rest of the times getting trapped in a 'local optimum' like the one shown. It must be noted that there are simple algorithms (e.g. based on a renumbering of terminal taxa within groups, with groups examined in a postorder traversal, so that the numbering depends on the size of the group) that can quickly detect whether two tree shapes are identical; these algorithms have been purposefully avoided in the present implementation, so as to allow testing the ability of the heuristic based on guided switches to detect identity.

The guided switches thus work reasonably well in some cases, although they can be easily trapped in local optima. The solution adopted in TNT is switching a low number of terminal taxa at random, and applying the guided switches again, a number of times. Every certain number of tries of this milder perturbation, a larger number of random switches (so as to start again completely from scratch) is effected. The labelling that produces the best (lowest) MRP score of all these cycles is stored and reported at the end.

With this randomization (using by default four starting points, randomly exchanging 25% of the taxa every time, and analysing each of those with 30 rounds of random exchanges of 10% of the taxa, and five cycles of guided switches), the identity of trees with 50–60 taxa is detected in the vast majority of cases within a second. Beyond those numbers of taxa, the times needed to approximate optimal labellings increase rapidly, and the heuristic begins to fail more frequently.

To us, one of the unexpected results of applying this algorithm to a number of examples is that all trees can be made to agree substantially by appropriate taxon relabellings, even on random trees. The typical result of producing random labelled trees is an unresolved consensus, but when the trees are relabelled, they can normally be made to share more groups without altering their shapes. Although unanticipated, this result actually follows from the fact that (under the uniform model) the number of cherries for random trees of *n* taxa tends to *n*/4 (McKenzie & Steel 2000: 88), and all cherries can be made equivalent. As a consequence, for random trees of 50 or more taxa, the value of R_s is usually above 0.90 (indicating a high degree of similarity), and the RF$_s$ distance is often below 0.50. In a similar vein, for maximum agreement subtrees, it has also been observed that random trees can be expected to share identical subtrees of a substantial size (e.g. Bryant *et al.* 2003). We conjecture that maximum differences will be obtained between completely balanced and completely pectinate trees; for such trees (with 64 taxa plus an outgroup), the R_s is 0.88309, but RF$_s$ is (more appropriately) 0.90476. The problem is thus more acute for R_s; perhaps this statistic could be rescaled when used to compare tree shapes, so that the values that indicate maximum difference are more easily interpretable.

The heuristic implemented in TNT is a first, proof-of-concept approximation to the problem of finding optimal relabellings, useful only for relatively small problems. No doubt, it can be improved using smarter algorithms. An obvious possibility for improvement is in trying to divide the trees to compare in sectors, identifying subtrees of identical shape (e.g. the upper subtree in both Fig. 3C,D; recall that shape identity can be established easily), so that the subtree can be replaced by a single label, piecemealing the problem.

Note that the actual minimization performed is of the sum of the number of steps implied by each tree on the MRP of the other. This will normally also increase the number of shared nodes between the trees, thus decreasing RF and other measures of distance. When the RF is to be minimized, then the exchanges are attempted based on the criteria (a–c) above, but the actual RF distance is used instead of the raw number of steps to decide whether a better labelling has been found (the RF distance is the number of characters with more than a single step, easily obtained from the routines described above).

Other uses for shape comparisons: taxon correspondences

The examples given above consider the case of fully unlabelled trees. There is another case where considerations of tree shape may be relevant, and this is when a plausible correspondence or synonymy needs to be identified. In that case, the two trees may share some or most of the taxa, and the goal will be to match those taxa that are found in only one of the trees. In the absence of additional information (e.g. unique morphological characters, or specimens collected together, in the case of matching sexes), the tree topologies for two separate phylogenies may suggest alternative pairings.

Consider the example of Fig. 4A, with two trees for five taxa each. Taxa E and C are present in only one of the trees, and taxa F and C are present only in the other; there is the suspicion that E and C might be synonyms with F and G, but the precise correspondence is not known with certainty. Assuming that the phylogenies for both sets of taxon names are indeed equivalent, taxon E can be seen to be in the same position – relative to the general shape of the tree – as G, and taxon C in the same position as F. Thus, the expected result is E = G, and C = F.

Aside from visual inspection of both trees, the only existing tool to vaguely approximate this conclusion is a supertree (Fig. 4D), which allows combining trees with different taxon sets. For the present example, species E and G on the one side, and C and F on the other, will appear as part of two trichotomies, thereby suggesting that they occupy analogous positions in the trees and that the proper matching would be E with G, and C with F.

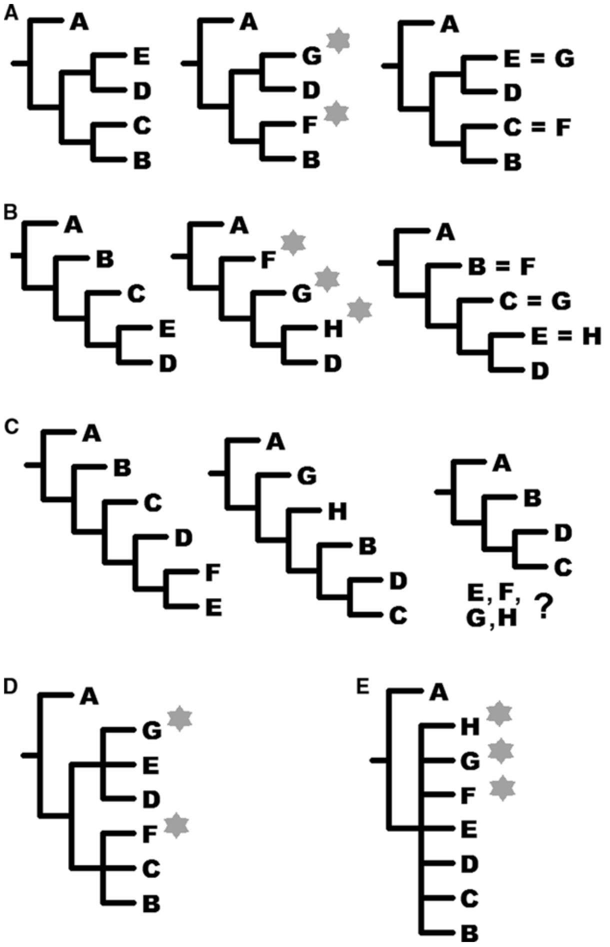

Fig. 4 Different cases in which taxon correspondence might be determined. —A–C. Input trees and expected correspondence (taxa of uncertain correspondence marked with a star); (D) semi-strict supertree (Goloboff & Pol 2002) for the input trees in (A); (E) semi-strict supertree for the input trees in (B).

The supertree, however, only allows establishing proper conclusions in some cases, not in general. Figure 4B shows another example, where B, C and E are suspected synonyms of F, G, H, but the specific pairings are not known. The expected result, considering the shapes of the trees, is B = F, C = G and E = H. The supertree for the eight taxa is a complete bush (Fig. 4E), thus providing no information as to possible correspondences.

Finding matchings that maximize the number of consensus nodes

Two possible criteria to deal with this situation have been explored and implemented. The first criterion for deciding the best matchings is in counting the number of nodes in the consensus of the two trees, after names of matched taxa have been made equivalent in both trees, and unmatched taxa are removed from the trees. Those matchings that maximize the number of nodes in the consensus are chosen. There may be, of course, several alternative matchings that maximize the number of consensus nodes. In the implementation of TNT, the user can optionally choose to maximize the number of nodes in either the combinable component (Bremer 1990) or the strict consensus tree (obviously, when input trees are fully resolved, both options produce identical results).

Counting the number of consensus nodes will properly solve cases like the one shown in Fig. 4C. Taxa E and F might match taxa G and H, but these are placed in distant positions in the trees. Any matching will produce fewer than the two nodes (not counting root node) obtained by not matching at all. Thus, the best conclusion seems in principle that taxa E, F and taxa G, H do not match. However, the criterion of simply counting the number of consensus nodes has the drawback that the position of the taxa to be matched must be identical in both trees, or differ at the most by a single node, for the matching to be preferred over a non-matching. Consider the case of Fig. 5, with a single taxon unmatched in each of two otherwise identical trees (F in the first tree, Fig. 5A, G in the second, Fig. 5B–D). Comparing trees A and B, in which taxa F and G occupy exactly the same position, the synonymy F = G is always preferred. When G is placed one node apart (as in the tree of Fig. 5C), the synonymy produces a trichotomy for DEF, while the non-matching produces a group DE (with F and G removed from the consensus); thus, both matching and non-matching are optimal. When G is placed one more node apart in the second tree (as in Fig. 5D), the number of nodes on the consensus when G is synonymized with F (with a tetrachotomy for CDEF) is lower than the number of nodes when G and F are considered distinct (with F and G pruned from the input trees, the two input trees become identical).

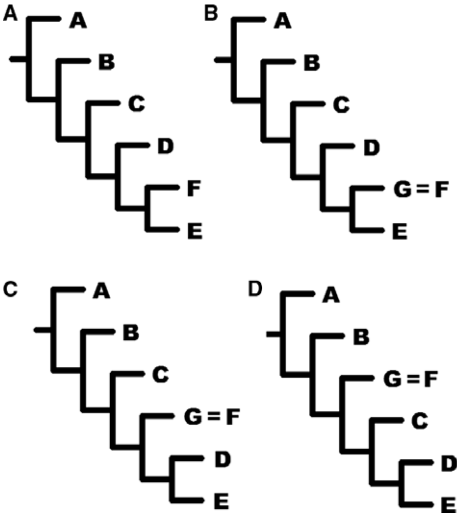

Fig. 5 Two possibly synonymous taxa (F and G), with taxon G placed at different location of an otherwise similar tree. For the algorithms described, the cost of synonymizing taxa F (in tree a) and G increases as G is further away (trees b–d) from the position of F in tree a. See text for discussion.

To the extent that one has more confidence that all the species in the two trees are to be matched (i.e. that no species is truly unrepresented in one of the trees), the method of just counting consensus nodes may produce fewer matches than desired. This can be solved by considering a penalty, P, for every case of an unmatched taxon. Then, instead of simply counting the number of consensus nodes C, the matchings chosen are those that maximize $C - (P \times n)$, where n is the number of unmatched taxa. When $P = 1$, the synonymy F = G produces a better score as the non-matching (instead of the same score), for the comparison between trees A and C, and the same score as the non-matching (instead of an inferior one) for the comparison between trees A and D. When $P \geq 2$ (or more), the best conclusion is always F = G, for each of the tree comparisons. This allows establishing conclusions when the two trees cannot be made identical by matching taxa.

The implementation of this criterion in TNT works by brute force, enumerating all possible matchings (and non-matchings). Thus, the solution it produces is guaranteed to optimize the number of consensus nodes, but is slow for large numbers of unmatched taxa, quickly becoming impractical beyond 15 unmatched taxa. It serves only as a proof-of-concept implementation; it is possible that smarter algorithms can be devised to produce faster exact solutions.

Finding matchings that minimize steps in the MRPs

The second criterion is a heuristic, based on minimizing the steps of the MRPs, implemented in TNT with a modification of the taxon-switching algorithm described in the previous section. The exchanges between terminal taxa (guided, as before, by the most parsimonious optimization of the variables representing the groups) are applied only to taxa of uncertain correspondence, skipping the taxa present in both trees. For the taxa present in only one tree, an initial set of arbitrary correspondences is set (using a randomized list); as these taxa exchange positions, the list of correspondences is updated. This algorithm is order dependent, and it takes into account ambiguity in the correspondences through repetition with different random seeds. For the cases of Fig. 4A,B, where there is no ambiguity, the expected result is produced.

A problem with such simple modification of the taxon-switching algorithm is that the approach, given N unmatched terminals in each of the trees, will necessarily match each of those – the algorithm described in the previous section does not consider the possibility that the best conclusion may be that some of the taxa of uncertain correspondence do not match at all. The algorithm is based on minimizing number of steps (and thus maximizing retention index, *R*). The retention index is rescaled between 0 and 1 (see Farris 1989), but the values are not comparable for different data sets; a retention index of 1 might be achieved by excluding (i.e. non-matching) almost all taxa, and this – despite the congruence – is not a very useful result. The raw number of steps of the MRP matrix is not useful, either, because a decrease in number of steps might correspond to a modification of the matrix that eliminates informative characters.

Although the number of steps in the MRP does not suffice for deciding whether some taxa are better left unmatched, adding a step penalty for every pair of taxa left unmatched will take into account the decrease in steps produced by the elimination. That decrease cannot exceed one step per each character representing a group to which the taxon belongs in one tree (and not in the other). This provides a natural criterion to decide whether taxa that are placed far apart in the tree should be matched, illustrated in Fig. 5 with a single taxon unmatched in each of two otherwise identical trees (F in the first tree, Fig. 5A, G in the second, Fig. 5B–D). When the unmatched taxon G is in exactly the same position in the second tree (Fig. 5B), the two MRPs require no extra steps in each of the trees when F and G are considered synonyms; the two taxa are thus always matched, regardless of penalty. When G is one node apart in the second tree (Fig. 5C), synonymizing it with F will incur in two steps for the MRPs, one on tree C for the character representing group EF from tree A and the other on tree A for the character representing group DE in tree C. As G is more nodes apart in the second tree, synonymizing it requires two additional steps per node; thus, when G is in the position shown in tree D, two more steps (in addition to those two just discussed) are required for the MRPs, one on tree D for the group DEF from tree A and the other on tree A for the group CDE. When the user sets a penalty to a value *P*, for every taxon that is not matched to some taxon in the other tree, a value *P* is added to the number of steps of the MRPs. For the pair of trees A and C, the synonymy F = G is preferred over a non-match when $P \geq 1$; for the pair A and D, the synonymy is preferred when $P \geq 2$. This properly takes into account the case where the number of unmatched taxa differs in the

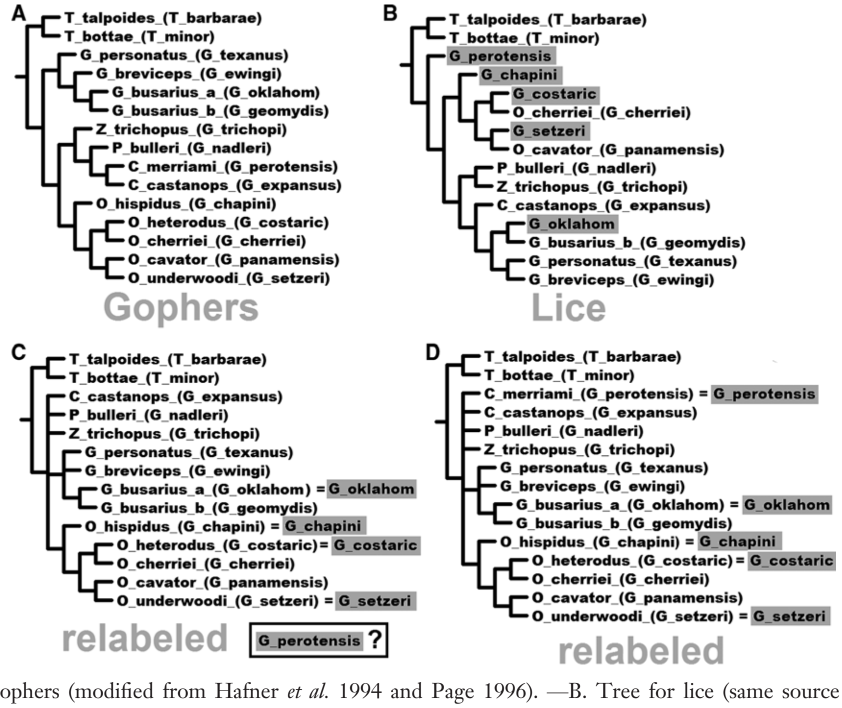

Fig. 6 —A. Tree for gophers (modified from Hafner *et al.* 1994 and Page 1996). —B. Tree for lice (same source as A), with five species (highlighted in grey) for which the host is assumed to be unknown. —C,D. The taxon-matching algorithm based on maximizing the number of consensus nodes either matches correctly four pairs of taxa and leaves the fifth with unknown correspondence (C), or correctly matches all species with their host (D).

two trees (so that not all of the unmatched taxa can be simultaneously matched).

Empirical examples

The degree to which two trees allow matching taxa depends on their similarity. In applied studies, it would be unrealistic to expect that the two trees have exactly the same shape (due to either error in the estimation of one or both phylogenies, or the parasites tracking the host phylogeny less than perfectly), so that a compromise solution will usually be necessary.

Platnick & Shadab (1978), in their revision of the genus *Anapis* (Araneae, Anapidae), presented two trees obtained independently (see their figs. 9–10), one in which species with unknown males are excluded, and the other where species with unknown females are excluded. Anapid spiders are sexually dimorphic and matching males with females when they have not been collected together may be problematic. Through a careful consideration of morphology, Platnick & Shadab (1978) established possible correspondences, with two species known only from males (*A. chiriboga* and *A. castilla*), and nine known only from females (*A. atuncela*, *A. circinata*, *A. choroni*, *A. digua*, *A. discoidalis*, *A. felidia*, *A. guasca*, *A. hetchski* and *A. meta*). If the matching method described in this paper is used to establish possible correspondences between the two male-only species with the female-only species (either with the exact procedure maximizing number of strict or semi-strict consensus nodes, or with the heuristic based on minimizing MRP steps), the most likely correspondence for the males is either with no species at all (i.e. Platnick and Shadab's hypothesis), or with *A. choroni*, *A. circinata* or *A. meta*. These matchings generally correspond to Platnick & Shadab's (1978) discussion of morphological characters supporting the different groups in each of the two separate phylogenetic trees, and are congruent with geographical distribution (e.g. they do not include *A. hetchski*, known from Southern Atlantic instead of Amazonian forest). As no additional taxonomic work on these spiders has been published in the ca. 40 years elapsed since Platnick & Shadab's (1978) work, there is no way to test whether the male–female matchings implied by the present method are reasonable.

As a more definitive test of the approach, it seems desirable to use a case where the actual matchings are known with certainty, comparing to the results that would be obtained if some of the matchings were unknown. An interesting example is provided by the trees for gophers and lice used by Page (1996) to illustrate cospeciation. The lice are highly specific, and the trees contain about as many species of lice as gophers. In all cases, the associations between host and parasite are known from actual observations. For the comparisons below, we have simplified the example so that the gopher *Thomomys bottae*, parasitized in fact by *Geomydoecus actuosi* and *Thomomydoecus minor*, are parasitized only by the latter; *G. actuosi* is removed from the lice tree. Similarly, *Thomomys talpoides* is parasitized by two species of lice, *Geomydoecus thomoyus* and *Thomomydoecus barbarae*; the former louse is removed from the tree, and only *T. barbarae* is considered to occur on *T. talpoides*.

What if some of the associations between lice and gopher were not known, and the host–parasite correspondence was established on the basis of tree shapes, using the criterion of maximizing number of consensus nodes described in the previous section? A possible case is shown in Fig. 6. The names of the gophers include (in parentheses) the name of the corresponding louse. In the case of the lice tree, the name of the gopher in which the species is known to occur is used instead of the louse name. By doing this, the degree of congruence when all host and parasite associations are known can be visualized with a consensus tree. In Fig. 6B, a tree for lice is shown, but there are five species (*Geomydoecus perotensis*, *Geomydoecus*

Table 1 Results of applying the taxon-matching algorithm based on maximizing consensus nodes to the trees for gopher and lice, for different numbers of unmatched taxa in each of the trees ('n' column). The 'wrong' column indicates the proportion of cases where a taxon could be paired only with the wrong taxon (or taxa); this does not include the cases where a taxon could be either paired with the wrong taxon (or taxa) or left unmatched (which are simply ambiguous). The 'perfect' column indicates the proportion of cases where the pairing of a gopher was done only with the correct louse; columns '1/2', '1/3' and '1/>3' indicate the cases where the correct pairing was as optimal as other (1, 2 or more) possible pairings (including non-matching at all), thus indicating ambiguity but with the correct results among other possibilities. The number of possible *n* unknown associations is fixed; the results reported are exact, based on exhaustive enumeration of all combinations; the numbers of possible combinations of unknown associations are 105 (for *n* = 2), 455 (*n* = 3), 1365 (*n* = 4), 3003 (*n* = 5), 5005 (*n* = 6) and 6435 (*n* = 7).

Penalty	*n*	Wrong	Perfect	1/2	1/3	1/>3
0	2	0.004762	0.552381	0.342857	0.085714	0
0	3	0.013919	0.506227	0.306227	0.134799	0.015385
0	4	0.026557	0.461905	0.285897	0.156410	0.040842
0	5	0.042158	0.419647	0.277323	0.159307	0.071595
0	6	0.060440	0.379620	0.276257	0.150916	0.103796
0	7	0.081185	0.342369	0.278788	0.137329	0.134221
2	2	0	0.89524	0.10476	0	0
2	3	0	0.79927	0.19194	0.00879	0
2	4	0.00073	0.71117	0.26172	0.02381	0.00256
2	5	0.00293	0.63017	0.31442	0.04249	0.00999
2	6	0.00733	0.55558	0.35058	0.06234	0.02418
2	7	0.01465	0.48678	0.37110	0.08099	0.04649

chapini, *Geomydoecus costarricensis*, *Geomydoecus setzeri* and *Geomydoecus oklahomensis*) for which the host is not known; this also leaves five species of gophers (*Cratogeomys merriami*, *Orthogeomys hispidus*, *Orthogeomys heterodus*, *Geomys busarius* and *Orthogeomys underwoodi*) with no known parasite. Applying the taxon-matching algorithm described in the previous section, two labellings are found to be optimal (Fig. 6C–D), one which correctly matches all five species of lice and gopher, and another which correctly matches four pairs of species but provides no answer for the remaining one (the lice *G. perotensis* and the gopher *C. merriami*).

The number of possible scenarios of unknown associations is finite; with 15 taxa in each tree, there are $\binom{15}{n}$ possible combinations of *n* unknown associations. All possible combinations of unknown associations from $n = 2$ to $n = 7$ were enumerated, finding the matchings that maximize the number of consensus nodes. The results are shown in Table 1; very few combinations of unknown associations result in an erroneous matching, especially when a penalty for non-matching is used. Even in the case of seven unknown associations, the matchings with a penalty where the correct association is either the only optimal one, to one of three possible matchings, sum up to 0.939. A very low proportion of cases indicates incorrect associations as optimal. The method is often ambiguous regarding the associations, but rarely misleading.

Conclusions

This paper shows that the shape of trees, so far considered only in the realm of testing evolutionary models, may also help solve some problems that are more specific to taxonomy and systematics. The degree of balance in the trees, often used to test different models of speciation and extinction, is insufficient for comparing tree shapes in wider contexts. Some algorithmic methods to deal with the problem of comparing shapes (crude and primitive, but better than nothing at all) are presented here and implemented in recent versions of the computer program TNT (Goloboff & Catalano 2016). Areas of further inquiry are, obviously, in the development of faster algorithms to find optimal labellings and optimal correspondences. Another standing problem is that relabelling tree shapes always produces trees with a high degree of similarity (as measured by the statistics normally used to compare phylogenetic trees) and a proper rescaling might help produce values that can be more easily intuited.

Comparing shapes may be used to gain some insight on whether two phylogenies are likely to correspond to taxa with a strong association, in the absence of observations other than the shape of the phylogeny. The example of gophers and lice analysed here, although slightly modified, suggests that (for discordances between taxa and differences in tree shape taken from a real example) the associations, when inferred from the comparison of tree shapes, may be quite reliable.

Acknowledgements

Support from CONICET (PIP 00687 to PAG, PIP 0260 to CAS), as well as NSF, NASA and FAPESP (via a grant to J. Cracraft and L. Lohman, entitled ‘Assembly and evolution of the Amazonian biota and its environment: an integrated approach’, with participation from PAG and CAS) was deeply appreciated. We also acknowledge useful comments from Eric Lewitus, Lutz Bachmann and two anonymous reviewers.

References

Alfaro, M., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D., Carnevale, G. & Harmon, L. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Procedings of the National Academy of Sciences of the United States of America*, *106*, 13410–13414.

Bansal, M., Gordon Burleigh, J., Eulenstein, O. & Fernández-Baca, D. (2010). Robinson-Foulds supertrees. *Algorithms in Molecular Biology*, *5*, 1–18.

Baum, B. & Ragan, M. (1993). A reply to A.G. Rodrigo’s A comment on Baum’s method for combining phylogenetic trees. *Taxon*, *42*, 637–640.

Bortolussi, N., Durand, E., Blum, M. & François, O. (2006). ApTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*, *22*, 363–364.

Bremer, K. (1990). Combinable component consensus. *Cladistics*, *6*, 369–372.

Bryant, D., McKenzie, A. & Steel, M. (2003). The size of the maximum agreement subtree for random binary trees. *BioConsensus (Dimacs Series in Discrete Mathematics and Theoretical Computer Science)*, Vol. *61* (pp. 55–65). Providence, RI: American Mathematical Society.

Colless, D. (1982). Phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, *31*, 100–104.

Farris, J. S. (1973). On comparing the shapes of taxonomic trees. *Systematic Zoology*, *22*, 50–54.

Farris, J. (1989). The retention index and the rescaled consistency index. *Cladistics*, *5*, 417–419.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, 664 pp.

Frost, S. & Volz, E. (2013). Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society*, *368*, 20120208.

Germain, C. & Pallo, J. (1996). Two shortest path metrics on well-formed parentheses strings. *Information Processing Letters*, *60*, 283–287.

Goloboff, P. (1996). Methods for faster parsimony analysis. *Cladistics*, *12*, 199–220.

Goloboff, P. (2005). Minority rule supertrees? MRP, compatibility, and minimum flip may display the least frequent groups. *Cladistics*, *21*, 282–294.

Goloboff, P. & Catalano, S. (2016). TNT version 1.5, including a full implementation of geometric morphometrics. *Cladistics*, *32*, 221–238.

Goloboff, P. & Pol, D. (2002). Semi-strict supertrees. *Cladistics*, *18*, 514–525.

Goloboff, P. (2008). Calculating SPR-distances between trees. *Cladistics*, *24*, 591–597.

Goloboff, P., Farris, J. & Nixon, K. 2003. TNT: Tree Analysis Using New Technology. Program and documentation. Available via http://www.lillo.org.ar/phylogeny.

Goloboff, P., Carpenter, J., Arias, J. S. & Miranda-Esquivel, D. (2008a). Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics*, *24*, 758–773.

Goloboff, P., Farris, J. & Nixon, K. (2008b). TNT, a free program for phylogenetic analysis. *Cladistics*, *24*, 774–786.

Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastes, J. W. & Nadler, S. A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science*, *265*, 1087–1090.

Hein, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, *98*, 185–200.

Kuhner, M. & Yamato, J. (2015). Practical performance of tree comparison metrics. *Systematic Biology*, *64*, 205–214.

Lewitus, E. & Morlon, H. (2016). Characterizing and comparing phylogenies from their Laplacian spectrum. *Systematic Biology*, *65*, 495–507.

Matsen, F. (2006). A geometric approach to tree shape statistics. *Systematic Biology*, *55*, 652–661.

McKenzie, A. & Steel, M. (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences*, *164*, 81–92.

Mir, A., Rosselló, F. & Rotger, L. (2013). A new balance index for phylogenetic trees. *Mathematical Biosciences*, *241*, 125–136.

Mooers, A. & Heard, S. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, *72*, 31–54.

Moore, G. W., Goodman, M. & Barnabas, J. (1973). An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *Journal of Theoretical Biology*, *38*, 423–457.

Norström, M., Prosperi, M., Gray, R., Karlsson, A. & Salemi, M. (2012). PhyloTempo: a set of R scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. *Evolutionary Bioinformatics*, *8*, 261–269.

Page, R. (1996). Temporal congruence revisited: comparison of mitochondrial DNA sequence divergence in cospeciating pocket gophers and their chewing lice. *Systematic Biology*, *45*, 151–167.

Pallo, J. (1990). A distance metric on binary trees using lattice-theoretic measures. *Information Processing Letters*, *34*, 113–116.

Paradis, E., Claude, J. & Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Platnick, N. & Shadab, M. (1978). A review of the spider genus *Anapis* (Araneae, Anapidae), with a dual cladistic analysis. *American Museum Novitates*, *2663*, 1–23.

Poon, A., Walker, L., Murray, H., McCloskey, R., Harrigan, P. & Liang, R. (2013). Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS One*, *8*, e78122.

Rabosky, D. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One*, *9*, e89543.

Robinson, D. & Foulds, L. (1979). Comparison of weighted labeled trees. *Lecture Notes in Mathematics*, *748*, 119–126.

Robinson, D. & Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131–147.

Sackin, M. J. (1972). “Good” and “bad” phenograms. *Systematic Zoology*, *27*, 159–188.

Shah, P., Fitzpatrick, B. & Fordyce, J. (2012). A parametric method for assessing diversification-rate variation in phylogenetic trees. *Evolution*, *67*, 368–377.

Shao, K. & Sokal, R. (1990). Tree balance. *Systematic Zoology*, *39*, 266–276.

Stadler, T. (2013). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, *26*, 1203–1219.

Wheeler, W. (1999). Measuring topological congruence by extending character techniques. *Cladistics*, *15*, 131–135.

Wu, C.-S. & Huang, G.-S. (2010). A metric for rooted trees with unlabeled vertices based on nested parentheses. *Theoretical Computer Science*, *411*, 3923–3931.