

Phylogenetics

A congruence index for testing topological similarity between trees

Damien M. de Vienne^{1,*}, Tatiana Giraud¹ and Olivier C. Martin^{2,3}

¹Univ Paris-Sud, Laboratoire Ecologie Systématique et Evolution, UMR8079, Orsay, F-91405, CNRS, Orsay, F-91405, AgroParisTech, Paris, F-75231, ²Univ Paris-Sud, UMR8626, LPTMS, Orsay, F-91405, CNRS, Orsay, F-91405 and ³Univ Paris-Sud, UMR8120, Laboratoire de Génétique Végétale du Moulon, Gif-sur-Yvette, F-91190, France

Received on July 23, 2007; revised on September 18, 2007; accepted on September 30, 2007

Advance Access publication October 12, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Phylogenetic trees are omnipresent in evolutionary biology and the comparison of trees plays a central role there. Tree congruence statistics are based on the null hypothesis that two given trees are not more congruent (topologically similar) than expected by chance. Usually, one searches for the most parsimonious evolutionary scenario relating two trees and then one tests the null hypothesis by generating a high number of random trees and comparing these to the one between the observed trees. However, this approach requires a lot of computational work (human and machine) and the results depend on the evolutionary assumptions made.

Results: We propose an index, I_{cong} , for testing the topological congruence between trees with any number of leaves, based on maximum agreement subtrees (MAST). This index is straightforward, simple to use, does not rely on parametrizing the likelihood of evolutionary events, and provides an associated confidence level.

Availability: A web site has been created that allows rapid and easy online computation of this index and of the associated P -value at <http://www.ese.u-psud.fr/bases/upresa/pages/devienne/index.html>

Contact: damien.de-vienne@u-psud.fr

1 INTRODUCTION

Phylogenetic trees have taken a great importance in evolutionary biology and tree comparisons are used for multiple purposes, from unveiling the history of species to deciphering evolutionary associations among organisms and geographical areas. Tree comparisons are performed by testing the null hypothesis that the trees are not more congruent (topologically similar) than expected by chance.

In studies of host-parasite associations, testing for congruence helps one unravel the evolutionary processes underlying the emergence of new parasite species, for instance by determining whether host shift or cospeciation (host tracking by parasite lineages) was the prevailing mode of speciation in the parasites. The generally accepted idea is that, if new parasite species have emerged by multiple cospeciation events with their hosts, the phylogenies of the host and of the parasites will be highly congruent. In contrast, if the emergence of new parasite

species is mainly due to host-switches to distant hosts, the phylogenies will be incongruent. In recent years, such cophylogenetic analyses have taken a new direction with the development of powerful methods to reconstruct, from the phylogenies of interacting species, the history of their association. These methods include Brooks' parsimony analysis (Brooks, 1988; Brooks and McLennan, 1991), component analysis (Component, Page, 1993), tree reconciliation (Treemap, Page, 1994), event-based-methods (Jungles, Charleston, 1998; Treefitter, Ronquist, 1995, 1997), maximum-likelihood (Huelsenbeck *et al.*, 1997) and a method based on Bayesian inference (Huelsenbeck *et al.*, 2000). These methods seek the optimal evolutionary scenario for the association between a set of hosts and their parasites, by assigning to each type of evolutionary event (cospeciation, host-switch, but also duplication and extinction of the parasites lineages) a given probability of occurrence or a cost. The conclusions reached by such methods therefore depend on these evolutionary parameters for which one usually has no estimation. Furthermore, they generally assume that congruence between phylogenies implies that the interacting species have cospeciated. However, recent studies have shown that congruence between host and parasite trees can result from other evolutionary processes such as preferential switches to related hosts (Charleston and Robertson, 2002; de Vienne *et al.*, 2007; Hirose *et al.*, 2005).

To avoid these problems, a cophylogenetic analysis should first test for topological congruence between two trees before trying to infer coevolutionary scenarios. For instance, a significant congruence between the host and parasite trees can be the result of at least two different alternative histories: the two groups of organisms may have mostly undergone cospeciations, or there may have been many host-switches to closely related hosts. Additional evidence, such as ages of the nodes, or assumptions on the probability of the different events, will be required for deciding between these two hypotheses. It would therefore be useful to be able to determine first whether the two trees are significantly congruent, before making any assumption on the likelihood of evolutionary scenarios. After such a first test, using the methods cited above will be informative to complete the picture given by the topological congruence measurement.

*To whom correspondence should be addressed.

Tree comparisons are not only useful in cophylogenetic analyses. They can also teach us about the processes that shape biodiversity across geographic areas. This is the aim of comparative phylogeography analyses, where phylogeographic trees of a wide range of co-distributed species are compared. A good congruence between these trees is thought to reflect common geological and historical processes shaping broad-scale patterns of biodiversity (Lapointe and Rissler, 2005).

Clearly a simple test for topological congruence between trees will be useful in both host-parasite association and phylogeographic studies. To test the null hypothesis that two given trees are not more congruent than expected by chance, one has to: (1) compute their topological congruence; (2) generate a large number of random pair of trees and compute their pairwise topological congruence and (3) compare the values of the given pair and the random pairs. If less than 5% of the random pairs of trees are topologically more congruent than the two given trees, one will conclude that these trees are more congruent than expected by chance. Unfortunately, generating a high number of random trees and analysing them is time consuming and has to be done *de novo* for each new comparison, as the number of leaves in the given trees changes.

Here we propose a simple index based on topological congruence, which provides a *P*-value for the null hypothesis that two given trees are not more similar than expected by chance. It can be used for arbitrary binary tree pairs, i.e. fully resolved, with more than seven leaves, without having to generate any random sample of trees and should therefore be very useful for cophylogenetic and comparative phylogeographic analyses. A web site has been created that allows rapid and easy online computation of this index and of the associated *P*-value.

To establish this index, a large number of random trees were generated, with a wide range of number of leaves. The topological congruence between the trees was then assessed using MAST (Maximum agreement subtrees). This method determines the minimum number of leaves that have to be removed in each phylogeny to render the trees identical; the size of the MAST is thus a simple yet powerful measure of topological congruence. The index we designed is based on a limit law for the distribution of the size of the MAST when trees are chosen at random.

To illustrate the usefulness of our index, we used it to compare real host-parasite phylogenies, the plant-fungal associations previously analysed by Jackson (2004), and we compared our results to those he obtained on the same trees using 'Jungles' implemented in TREEMAP 2.0 (Charleston, 1998). In addition, we used the program TREEFITTER (Ronquist, 1995) to compare the same tree pairs and understand the discrepancies between the results of the different analyses.

2 METHODS

2.1 Generating random trees

A total of 10 000 pairs of random trees were generated for each number of leaves analysed (7–50). As there was no a priori expectation on the

imbalance of the trees that will be compared, we used the model where all labelled binary rooted trees were equiprobable. The trees were generated using the software COMPONENT, version 2.0 (Page, 1993) that uses the algorithm described by Furnas (1984) to generate random trees. We did not analyse trees with more than 50 leaves because such phylogenies are absent in the cophylogeny literature.

2.2 Topological congruence

The topological congruence of two trees was assessed through their maximum agreement subtree (MAST), computed using COMPONENT v. 2.0 (Page, 1993). A MAST is the largest possible tree compatible with two given trees (Finden and Gordon, 1985) and is obtained by removing the minimum number of leaves in both trees for which perfect congruence occurs. The number of leaves in the MAST for a pair of trees quantifies their congruence: the larger this number, the more congruent the trees are.

2.3 Real phylogenies from the literature

To assess the reliability of our index, we compared the results on topological congruence obtained with our index to those obtained by Jackson (2004) on eight plant-fungal associations. Only associations where each plant species was associated with a single fungal species were retained in order to eliminate possible discrepancies due to the choice of the leaves that had to be removed or added.

Jackson (2004) considered that a significant number of cospeciation events and a significant congruence was equivalent and that, in contrast, a non-significant number of codivergence events was the same as incongruence (see Jackson, 2004 for more details). We also used the software TREEFITTER version 1.0 (Ronquist, 1995) to examine the same eight associations. The program TREEFITTER assigns to each type of evolutionary event a cost and evaluates the global cost for reconciling two trees. The smaller this cost, the more congruent the phylogenies are. The global cost is calculated for each reconciliation of the host tree with a number of random parasite trees obtained either by permutations of the leaves of the initial parasite tree or by permutations of the topology of the parasite tree (generation of random parasite trees). Here, we chose three different sets of costs for the four evolutionary events. In the first case, the costs are the same as those used by Jackson (2004) in TREEMAP (cospeciation: 0; duplication, loss and switch: 1); in the second case, cospeciations are more costly, and sorting and switching events are less costly (cospeciation: 0.2; duplication: 1; loss: 0 and switch: 0.5) and in the third case switches are very costly (cospeciation: 0.2; duplication: 1; loss: 0 and switch: 1.5). We performed 1000 permutations for each association and for each type of permutation that can be performed in TREEFITTER: permutation of the leaves of the parasite tree, and permutation of the topology of the parasite tree. If more than 50 trees of the 1000 obtained by permutations yielded a smaller global cost for reconciling the trees than the one found using the unpermuted parasite tree, the host and parasite trees were considered as incongruent; otherwise the trees were considered as significantly congruent.

3 RESULTS

3.1 Statistical properties of the number of common leaves (n_c)

For each number of leaves in the trees (N) and for each pair of random trees (10 000 for each choice of N), the number of leaves common to the two trees (n_c), which is just the size of the associated MAST, was calculated. Since n_c is a random variable, we study its distribution.

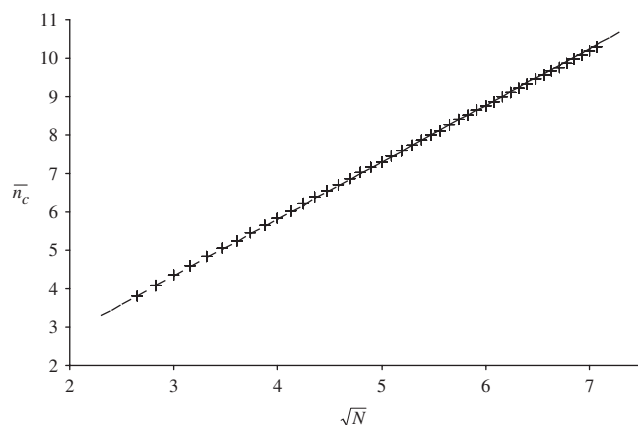


Fig. 1. Mean number of common leaves (\bar{n}_c), or MAST size, for random trees versus the squared root of the number of leaves (\sqrt{N}) in the trees compared. The dashed line represents the fit with a linear function growth in \sqrt{N} [Equation (1)].

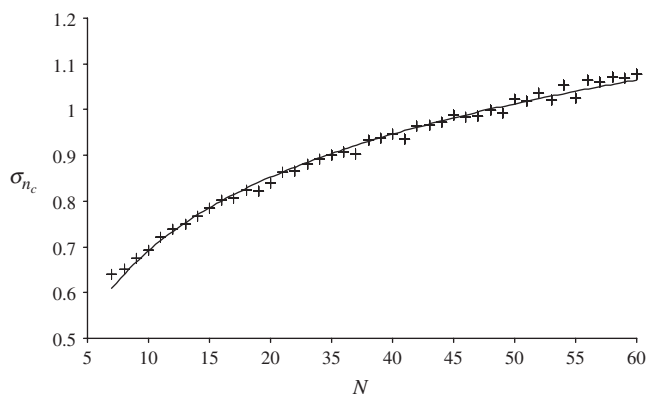


Fig. 2. Standard deviation (σ_{Nc}) of n_c versus N . The dashed curve represents the fit to a logarithmic law [Equation (2)].

First, we determined the *mean number* of common leaves (\bar{n}_c); this quantity is plotted as a function of \sqrt{N} in Figure 1. The correlation was highly significant ($R^2=0.9996$, $P<10^{-10}$), the equation of the regression being:

$$\bar{n}_c = -0.11 + 1.48\sqrt{N} \quad (1)$$

Note that the squared root law of this particular regression had already been showed by Bryant *et al.* (2003).

Similarly, we determined the SD of n_c as a function of N . This function is plotted in Figure 2. The relationship is compatible with a logarithmic law, and the fit to the following equation is indeed highly significant ($R^2=0.9924$, $P<10^{-10}$):

$$\sigma_{n_c} = 0.232 \ln(N) + 0.159 \quad (2)$$

Finally, the distributions of n_c for the 44 different values of N tested (from 7 to 50) were centred and rescaled as shown in

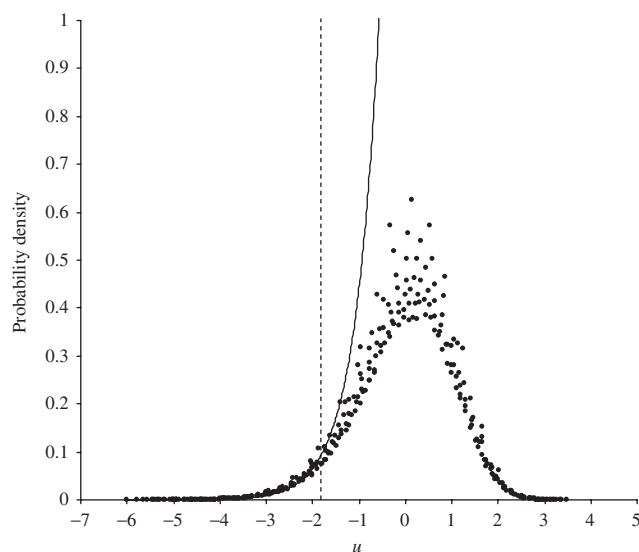


Fig. 3. Centred and rescaled distributions of n_c [c.f. Equation (3)] for the 44 different tree sizes tested (from 7 to 50 leaves). The plain curve represents the exponential function fitting the curve for values of u smaller than -1.5 [Equation (5)]. The vertical dashed line represents the 5% significance threshold of the index (see text).

Figure 3. More specifically, we introduce the shifted and rescaled random variable

$$u = \frac{\bar{n}_c - n_c}{\sigma_{n_c}} \quad (3)$$

and display its distribution $\rho_N(u)$ for the different N . By construction, the mean of u is 0 and its variance 1. What is remarkable is that $\rho_N(u)$ is very insensitive to N , seemingly becoming independent of N when N grows larger than 7. This kind of limit law arises in many systems and can be exploited here for our index. We consider $\rho_N(u)$ to be independent of N for large enough N and call it $\rho(u)$. Note that $\rho(u)$ is far from normal (Shapiro–Wilk normality test: $W=0.9618$, $P\text{-value}=1.2 \times 10^{-7}$) and therefore should not be approximated by a Gaussian.

3.2 A congruence index

The index we propose is calculated as the size of the MAST of two observed phylogenies (n_c^{obs}), normalized by \bar{n}_c , that is the mean expected from random trees [c.f. Equation (1)]. We call this index I_{cong} :

$$I_{\text{cong}} = \frac{n_c^{\text{obs}}}{\bar{n}_c} \quad (4)$$

As this index is for testing a higher congruence between trees than expected by chance, it is the left part of $\rho(u)$ that is relevant for computing a P -value. From $u=-\infty$ to $u=-1.5$, we find that $\rho(u)$ can be approximated ($R^2=0.9791$, $P<10^{-10}$) by an exponential function (solid curve in Fig. 3) of equation:

$$\rho(u) \approx 2.96e^{1.9u} \quad (5)$$

Table 1. Results of congruence tests for eight plant-fungal associations using three different methods: I_{cong} , ‘Jungles’ in TREEMAP 2.0 and TREEFITTER

Assoc.*	N	Congruence index I_{cong}		‘Jungles’		TREEFITTER						
		I_{cong}	p	Congruence	p	Congruence	CO:0; DU: 1; SO: 1; SW: 1		CO:0.2; DU: 1; SO: 0; SW: 0.5		CO:0.2; DU: 1; SO: 0; SW: 1.5	
		I_{cong}	p	Congruence	p	Congruence	p	Congruence	p	Congruence	p	Congruence
A	13	1.148	0.420	NO	0.260	NO	0.01	YES	0.187	NO	0.054	NO
B	13	1.531	0.003	YES	0.310	NO	0.013	YES	0.158	NO	0.016	YES
C	13	1.339	0.034	YES	0.04	YES	0.001	YES	0.019	YES	0.006	YES
D	10	1.750	2.41E-04	YES	<0.01	YES	0	YES	0	YES	0	YES
E	16	2.065	1.24E-06	YES	<0.01	YES	0	YES	0	YES	0.001	YES
F	17	1.661	2.59E-04	YES	<0.01	YES	0.001	YES	0	YES	0.002	YES
G	14	0.921	8.504	NO	>0.16	NO	0.893	NO	0.363	NO	0.1	NO
H	13	1.148	0.420	NO	0.58	NO	0.82	NO	0.5	NO	0.363	NO

* Associations: A: Rosidae/*Uromyces*; B: Pooideae/*Tilletia*; C: Rosidae/*Monilinia*; D: Monotropeae/*Homobasidiomycetes*; E: Asteraceae/*Golovinomyces*; F: Asteraceae/*Exobasidiales*; G: Rosaceae/*Erysiphe*; H: Pooideae/*Epichloë*. N refers to the number of leaves in the compared phylogenies; I_{cong} represents the value of the congruence index; P refers to the P -value associated with the method used to test for congruence: I_{cong} , ‘Jungles’ in TREEMAP 2.0 or TREEFITTER. For the first case, the P -values are computed using Equation (6), for the last cases, the P -values are those obtained with permutation of the terminals of the parasite tree. NO means that there is no significant congruence between the trees and YES means that there is a significant congruence between the trees (we use the 5% significance threshold value). For the TREEFITTER test, CO, DU, SO and SW represent the costs assigned to cospeciation, duplication, sorting and switching events, respectively. Bold characters represent cases where the conclusion on congruence between the trees depends on the method used.

Then given any two trees to compare, the probability (P -value) that they are topologically unrelated (the null hypothesis) is

$$p = \int_{-\infty}^{u^{\text{obs}}} 2.96e^{1.9u} du \quad (6)$$

where

$$u^{\text{obs}} = \frac{\bar{n}_c - n_c^{\text{obs}}}{\sigma_{n_c}}$$

3.3 A web tool to compute I_{cong}

A web site has been made available for calculating the value of the index I_{cong} online, at <http://www.ese.u-psud.fr/bases/upresa/pages/devienne/index.html>. The MASTs in the web site are calculated following (Berry and Nicolas, 2006). The two trees to be compared must be in Newick format; they can be typed or pasted into boxes. A web page also explains how to format the trees. The calculation of I_{cong} and of its P -value is very easy and very rapid; the user does not wait for the answer.

3.4 Illustrative example

We calculated I_{cong} and the associated P -value for eight plant-fungal associations and compared our results of congruence to those of Jackson (2004) for the same trees using the ‘Jungles’ method implemented in TREEMAP 2.0 (Charleston, 1998). We also compared to the results obtained with the software TREEFITTER for different sets of costs (Table 1).

For six out of the eight associations tested, the conclusions on congruence using I_{cong} were concordant with both those of Jackson (2004) and those obtained using TREEFITTER, regardless of the costs set for the different evolutionary events

(Table 1). However, for two associations, (*Tilletia* species infecting Pooideae plants and *Uromyces* infecting Rosidae; bold characters in Table 1), the conclusions given by the three methods were not concordant. Nevertheless, regarding the *Uromyces*/Rosidae association, all three methods (I_{cong} , TREEMAP 2.0 and TREEFITTER) concluded that there was no congruence between the trees when the costs in TREEFITTER were set to the same values as those used by Jackson (2004) in TREEMAP 2.0. Regarding the *Tilletia*/Pooideae association, the pattern was more complex: I_{cong} and TREEMAP 2.0 gave opposite conclusions (lack of congruence according to ‘Jungles’ and significant congruence according to I_{cong}) and the results given by TREEFITTER depended on the costs assigned to each evolutionary event. If the costs were the same as those set by Jackson (2004) in TREEMAP 2.0, or if switches were very costly compared to other events (1.5 compared to 0.2, 1 and 0 for cospeciation, duplication and sorting, respectively), the conclusion was the same as with our index. But if switches were less costly (0.5), the conclusion was the same as Jackson’s (Table 1). Finally, note that the results given by TREEFITTER were the same whether the random trees were obtained by permutation of the leaves of the parasite tree or by permutation of the topology of the parasite tree.

4 DISCUSSION

4.1 Limitations of the index

4.1.1 Significance threshold The exponential function we use does not fit the distribution $\rho(u)$ for values of u greater than -1.5 . Fortunately, the position in u where the P -value of I_{cong} is equal to 0.05 (c.f. vertical dashed line in Fig. 3) lies close to -1.8 , so our approach based on Equation (6) can indeed be used as long as the threshold level is 5% or less.

By monotonicity, if u is significantly above -1.5 , then the P -value will be larger than 0.05 indicating that the two trees are not more congruent than expected by chance. This result holds even though Equation (6) is no longer accurate in that regime, sometimes even leading to values greater than 1.

4.1.2 Trees with less than seven leaves The index we propose here is not designed for trees containing less than seven leaves because then the function $\rho_N(u)$ deviates significantly from its large N limit. However, this is not much of a limitation as assessment of topological congruence between such small trees is rarely performed and in any case it is not very informative.

4.1.3 Widespread species The index we propose does not take into account cases where leaves on one tree are associated with multiple leaves on the other tree (one parasite associated with multiple host species in cophylogenetic analysis or one species associated to multiple geographic areas in phylogeographic analysis), despite the fact that this situation seems quite common. However, as discussed by Ronquist (2003), we considered the *one-host-per-parasite assumption* as a biologically relevant simplification that can be extended to the phylogeographic context. Thus if some leaves in one tree are associated with multiple leaves in the other, one can (1) remove these particular leaves and their associated ones from the trees, or (2) ‘divide’ the leaves that are associated with multiple ones into as many leaves as necessary (thereby creating new leaves). These methods are commonly used in parsimony analysis of coevolving species associations (Page, 2003).

4.2 Comparison between the different tests for congruence

Here we compare the conclusions given by using I_{cong} , ‘Jungles’ and TREEFITTER on eight plant-fungal associations. In all cases, the computations required only seconds of cpu time. For most of the host-parasite associations analysed, the result on the tree congruence obtained using our index was the same as the one obtained using TREEMAP 2.0 or TREEFITTER. There were however two exceptions, for which the three methods gave conflicting results. An explanation of the discrepancy between our conclusion and Jackson’s on the *Tilletia/Pooidae* association could come from the way random trees are generated. Indeed, to test the null hypothesis (the two trees are not much more congruent than expected by chance), the program TREEMAP 2.0 (in which ‘Jungles’ is implemented) compares their congruence with the congruence between the host tree and random trees obtained by permutations of the leaves of the parasite tree; in contrast, we performed our comparison using completely random pairs of trees. On the other hand, it is worth noting that TREEFITTER provided the same results when using permutations of the topologies and when using permutations of the leaves; thus the discrepancy there is not due to the choice of random trees but to the sensitivity of TREEFITTER to the cost parameters. As is already known, the assumptions made on the relative probabilities of cospeciation, duplication, extinction and host shift heavily influence the conclusions in this kind of approach;

unfortunately, these probabilities are very difficult to estimate. This clearly shows that our index fills a gap as it focuses on the topologies of the trees without parametrizing the likelihood of evolutionary events.

5 CONCLUSION

We created an index for testing topological congruence between trees to answer the question: ‘are two given trees more congruent than expected by chance?’, without referring to the likelihood of evolutionary events. The aim of this index is not to replace existing methods that are useful to investigate evolutionary scenarios, but to provide an easy and rapid answer to the question posed above. This index should be seen as a first step in studies where trees have to be compared, to assess the topological congruence between trees, before investigating evolutionary scenarios. The index we propose should therefore be useful in all fields that compare trees, such as host-parasite association studies and phylogeographic studies.

ACKNOWLEDGEMENTS

We thank A. P. Jackson who provided the plant-fungal association’s phylogenies and V. Berry who gave us a source code for the calculation of the MAST. We also thank J.-P. Briane for his help on the creation of the web site. This work was supported by EEC’s FP6 Programme under contract IST-001935 (EVERGROW) to O.C.M. and an ACI JC from the French Ministère de la Recherche to T.G.

Conflict of Interest: none declared.

REFERENCES

- Berry, V. and Nicolas, F. (2006) Improved parametrized complexity of maximum agreement subtree and maximum compatible tree PROBLEMS. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 289–302.
- Brooks, D.R. (1988) Macroevolutionary comparisons of host and parasite phylogenies. *Annu. Rev. Ecol. Syst.*, **19**, 235–259.
- Brooks, D.R. and McLennan, D.H. (1991) *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology*. [University of Chicago Press, Chicago].
- Bryant, D. et al. (2003) The size of a maximum agreement subtree for random binary trees. *Theor. Comput. Sci.*, **61**, 55–65.
- Charleston, M.A. (1998) Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.*, **149**, 191–223.
- Charleston, M.A. and Robertson, D.L. (2002) Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol.*, **51**, 528–535.
- de Vienne, D.M. et al. (2007) When can host shifts produce congruent host and parasite phylogenies? A simulation approach. *J. Evol. Biol.*, **20**, 1428–1438.
- Finden, C.R. and Gordon, A.D. (1985) Obtaining common pruned trees. *J. Classif.*, **2**, 225–276.
- Furnas, G.W. (1984) The generation of random, binary unordered trees. *J. Classif.*, **1**, 187–233.
- Hirose, S. et al. (2005) Molecular phylogeny and evolution of the maple powdery mildew (Sawadaea, Erysiphaceae) inferred from nuclear rDNA sequences. *Mycol. Res.*, **109**, 912–922.
- Huelsenbeck, J.P. et al. (1997) Statistical tests of host-parasite cospeciation. *Evolution*, **51**, 410–419.
- Huelsenbeck, J.P. et al. (2000) A bayesian framework for the analysis of cospeciation. *Evolution*, **54**, 352–364.

- Jackson, A.P. (2004) A reconciliation analysis of host switching in plant-fungal symbioses. *Evolution*, **58**, 1909–1923.
- Lapointe, F.-J. and Rissler, L.J. (2005) Congruence, consensus, and the comparative phylogeography of codistributed species in California. *Am. Nat.*, **166**, 290–299.
- Page, R.D.M. (1993) *User's Manual for COMPONENT, Version 2.0*. [The Natural History Museum, London].
- Page, R.D.M. (1994) Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, **10**, 155–173.
- Page, R.D.M. (2003) *Tangled Trees. Phylogeny, Cospeciation and Coevolution*. [The University of Chicago Press, Chicago].
- Ronquist, F. (1995) Reconstructing the history of host-parasite associations using generalised parsimony. *Cladistics*, **11**, 73–89.
- Ronquist, F. (1997) Phylogenetic approaches in coevolution and biogeography. *Zool. Scr.*, **26**, 313–322.
- Ronquist, F. (2003) Parsimony analysis of coevolving species associations. In Page, R.D.M., (ed.) *Tangled Trees*. [The University of Chicago Press, Chicago], pp. 22–64.