Robustness of the Parsimonious Reconciliation Method in Cophylogeny

Laura Urbini, Catherine Matias, Marie-France Sagot, Blerina Sinaimeri

Robustness of the Parsimonious Reconciliation Method in Cophylogeny

Laura Urbini*[1], Blerina Sinaimeri†[1], Catherine Matias[2], and Marie-France Sagot[1]

[1] Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France and INRIA Grenoble Rhône - Alpes, France
[2] Sorbonne Universités, Université Pierre et Marie Curie, Université Paris Diderot, Centre National de la Recherche Scientifique, Laboratoire de Probabilités et Modèles Aléatoires, 4 place Jussieu, Paris, France

July 28, 2016

Abstract

The aim of this paper is to explore the robustness of the parsimonious host-symbiont tree reconciliation method under editing or small perturbations of the input. The editing involves making different choices of unique symbiont mapping to a host in the case where multiple associations exist. This is made necessary by the fact that no tree reconciliation method is currently able to handle such associations. The analysis performed could however also address the problem of errors. The perturbations are re-rootings of the symbiont tree to deal with a possibly wrong placement of the root specially in the case of fast-evolving species. In order to do this robustness analysis, we introduce a simulation scheme specifically designed for the host-symbiont cophylogeny context, as well as a measure to compare sets of tree reconciliations, both of which are of interest by themselves.

Keywords: cophylogeny, parsimony, event-based methods, robustness, measure for tree reconciliation comparison

1 Introduction

Almost every organism in the biosphere is involved in a so-called *symbiotic* interaction with other biological species, that is, in an interaction which is close

*laura.urbini@inria.fr
†blerina.sinaimeri@inria.fr

and often long term. Such interactions (one speaks also of *symbiosis*) can involve two or more species and be of different types, ranging from mutualism (when both species benefit) to parasitism (when one benefits to the detriment of the other). Some interactions may even become obligatory in the sense that neither species is able anymore to live without the other. This may in particular be the case when one of the species lives inside the cells of the other. We speak then of *endosymbiosis* (notice however that not all endosymbioses are obligatory). Understanding symbiosis in general is therefore important in many different areas of biology.

As symbiotic interactions may continue over very long periods of time, the species involved can affect each other's evolution. This is known as *coevolution*. Studying the joint evolutionary history of species engaged in a symbiotic interaction enables in particular to better understand the long-term dynamics of such interactions. This is the subject of *cophylogeny*.

The currently most used method in cophylogenetic studies is the so-called *phylogenetic tree reconciliation* [3, 4, 12, 16]. In this model, we are given the phylogenetic tree of the hosts H, the one of the symbionts S, and a mapping ϕ from the leaves of S to the leaves of H indicating the known symbiotic relationships among present-day organisms. In general, the common evolutionary history of the hosts and of their symbionts is explained through four main macro-evolutionary events that are assumed to be recovered by the tree reconciliation: (a) cospeciation, when host and symbiont speciate together; (b) duplication, when the symbiont speciates but not the host; (c) host switch, when after speciation of the symbiont, one of the new species of symbionts switches to a new host that is not related to the previous one; and (d) loss, which can describe three different and undistinguishable situations: (i) speciation of the host species independently of the symbiont, which then follows just one of the new host species due to factors such as, for instance, geographical isolation; (ii) cospeciation of host and symbiont, followed by extinction of one of the new symbiont species and; (iii) same as (ii) with failure to detect the symbiont in one of the two new host species. A reconciliation is a function λ which is an extension of the mapping ϕ between leaves to a mapping that includes all internal nodes and that can be constructed using the four types of events above. An optimal reconciliation is usually defined in a parsimonious way: a cost is associated to each event and a solution of minimum total cost is searched for. If timing information (*i.e.* the order in which the speciation events occurred in the host phylogeny) is not known, as is usually the case, the problem is NP-hard [15, 24]. A way to deal with this is to allow for solutions that may be biologically unfeasible, that is for solutions where some of the switches induce a contradictory time ordering for the internal nodes of the host tree. In this case, the problem can be solved in polynomial time [1, 6, 7, 13, 21]. In most situations, as shown in [6], among the many optimal solutions, some are time-feasible.

However, an important issue in this model is that it makes strong assumptions on the input data which may not be verified in practice. We examine two cases where this situation happens.

The first is related to a limitation in the currently available methods for tree

reconciliation where the association ϕ of the leaves is for now, to the best of our knowledge, required to be a function. A leaf s of the symbiont tree can therefore be mapped to at most one leaf of the host tree. This is clearly not realistic as a single symbiont species can infect more than one host. We henceforth use the term *multiple association* to refer to this phenomenon. For each present-day symbiont involved in a multiple association, one is currently forced to choose a single one. Clearly, this may have an influence on the solutions obtained.

The second case addresses a different type of problem related to the phylogenetic trees of hosts and symbionts. These indeed are assumed to be correct, which may not be the case already for the hosts even though these are in general eukaryotes for which relatively accurate trees can be inferred, and can become really problematic for the symbionts which most often are prokaryotes and can recombine among them [14, 20, 23]. We do not address the problem of recombination in this paper, but another one that may also have an influence in the tree reconciliation. This is the problem of correctly rooting a phylogenetic tree. Many phylogenetic tree reconstruction algorithms in fact produce unrooted trees [14, 19, 23]. The outgroup method is the most widely used in phylogenetic studies but a correct indication of the root position strongly depends on the availability of a proper outgroup [9, 18, 20]. A wrong rooting of the trees given as input may lead to an incorrect output.

The aim of this paper is, in the two cases, to explore the robustness of the parsimonious tree reconciliation method under "editing" (multiple associations) or "small perturbations" of the input (rooting problem). Notice that the first case is in general due to the fact that we are not able for now to handle multiple associations, although there could also be errors present in the association of the leaves that is given as input. The editing or perturbations we will be considering involve, respectively: (a) making different choices of single symbiont-host leaf mapping in the presence of multiple associations, and (b) re-rooting of the symbiont tree. In both studies, we explore the influence of six cost vectors that are commonly used in the literature (for a more detailed discussion, see for e.g. [2, 4]). The final objective is to arrive at a better understanding of the relationship between the input and output of a parsimonious tree reconciliation method, and therefore at an evaluation of the confidence we can have in the output.

We wish here to call attention to the fact that we will consider the robustness of the parsimonious method in the case where the solutions provided may be time-unfeasible. Our choice is driven by two reasons. The first is that, as already mentioned, finding time-feasible optimal tree reconciliations is an NP-hard problem, and therefore testing a significant number of large datasets is computationally impossible in practice. The second is that, as also indicated, it has been empirically observed that time-unfeasible methods when they are exhaustive, that is when they correctly output all optimal solutions, can be a good heuristic for finding optimal time-feasible solutions [6]. Many tree reconciliation algorithms exist, but only a few enumerate all solutions. The most commonly used are Notung [22], Jane 4 [5], and CoRe-Pa [13]. However, the first was designed for a gene/species context and imposes some restrictions on the costs

that may be given to some of the events, while the last two provide for most instances only a proper subset of all the optimal solutions [6]. Currently, only the method that we developed, called EUCALYPT [6], is exhaustive, and we therefore decided to use it exclusively in order to explore the robustness of the parsimonious tree reconcilation method.

Another important point is that we tested the parsimonious reconciliation method both on real and simulated datasets. There are not many methods available to simulate datasets that coevolved as these were mostly developed in a gene/species context [1, 7]. These are not suitable here for two reasons, the first being that they do not consider cospeciation as an event with its own parameter value (a gene *automatically* speciates within its species, *i.e.* when speciation occurs we consider that two different genes are automatically created, whether their sequences/functions already differ or not). The second reason is that these methods most often rely on a dating scheme of the host tree which might be difficult to tune so as to mimic real datasets. These limitations were already noticed in [10] where the authors attempted to provide their own simulation setup (to our knowledge, the only other one available in the cophylogeny context) by generating simultaneously a host and a symbiont tree relying on parameter values for the events. In this paper, we use a simulation method which we previously introduced in COALA [2] whose interest lies in that it uses parameter values (for the event probabilities) that are estimated on real datasets. Hence, this simulation scheme is more realistic and is designed for the cophylogeny context.

We start by introducing the datasets that will be used, both real and simulated ones as well as in the latter case, the method to generate them. We also present a measure to compare sets of tree reconciliations which may be of independent interest. We then describe the methods used to explore small perturbations in the two cases considered here, and discuss the results obtained.

The implemented methods are included in the tree reconciliation method we previously developed, called EUCALYPT, and will be made freely available at `http://eucalypt.gforge.inria.fr/`. This webpage also contains the online Supplementary Material with exhaustive results on the datasets.

2 Materials and Methods

In what follows, a dataset is a pair of host and symbiont trees (H, S), together with the association ϕ of their leaves. The indexes c, d, s, l relate to the 4 different events: cospeciation, duplication, switch and loss, respectively.

To analyse the influence of a perturbation, we adopted a set of cost events that correspond to those most commonly used in the literature on cophylogeny. We thus considered the following cost vectors $c = \langle c_c, c_d, c_s, c_l \rangle \in \mathcal{C}$ where $\mathcal{C} = \{\langle -1, 1, 1, 1\rangle, \langle 0, 1, 1, 1\rangle, \langle 0, 1, 2, 1\rangle, \langle 0, 2, 3, 1\rangle, \langle 1, 1, 1, 1\rangle, \langle 1, 1, 3, 1\rangle\}$.

2.1 Materials

2.1.1 Biological Datasets.

To test the robustness of the method, we selected 15 biological datasets from the literature: AW - Arthropods (12 leaves) & *Wolbachia* (12 leaves), CT - *Cichlidogyrus* (19 leaves) & *Tropheini* (28 leaves), EC - *Encyrtidae* (7 leaves) & *Coccidae* (10 leaves), FD - Fishes (20 leaves) & *Dactylogyrus* (50 leaves), GL - Gophers (8 leaves) & Lices (10 leaves), IFL - Insects (17 leaves) & Flavobacterial endosymbionts (17 leaves), MP - *Myrmica* (8 leaves) & *Phengaris* (8 leaves), PML - Pelicans (18 leaves) & Lices (18 leaves) where both trees are generated through a maximum likelihood approach, PMP - Pelicans (18 leaves) & Lices (18 leaves) where both trees are generated through a maximum parsimony approach, PP - Primates (36 leaves) & Pinworms (40 leaves), RH - Rodents (34 leaves) & Hantaviruses (42 leaves), RP- Rodents (13 leaves) & Pinworms (13 leaves), SBL - Seabirds (15 leaves) & Lices (8 leaves), SC - Seabirds (11 leaves) & Chewing Lices (14 leaves) and SCF - Smut Fungi (15 leaves) & Caryophillaceus plants (16 leaves). The choice was dictated by: (1) the availability of the data in public databases, and (2) the desire to cover for situations as widely different as possible in terms of the topology of the trees and the presence of multiple associations. For a more detailed description of these biological datasets, see the online Supplementary Material. We call attention here to the fact that only 3 of these datasets present multiple associations (namely MP, SBL, SFC) and are the ones used for studying the robustness of the method in the case of multiple associations.

2.1.2 Simulated Datasets.

We generated simulated datasets using a method that we previously developed, called Coala [2], and the 15 biological datasets as follows.

For any such dataset, Coala first estimates the corresponding probability of each coevolutionary event (cospeciation, duplication, switch and loss) based on an approximate Bayesian computation approach. As we needed the datasets to be as realistic as possible, each time we ran Coala to obtain 50 vectors of probabilities $\gamma = \langle \gamma_c, \gamma_d, \gamma_s, \gamma_l \rangle$ that are in some sense a likely explanation of the observed data.

In a second step, we used these vectors and the symbiont tree generation algorithm in Coala (see Baudet *et al.* [2] for more details) to obtain, for each vector γ, a simulated symbiont tree S' whose evolution follows that of the host tree H. Each dataset (H, S, ϕ) and probability vector γ thus led to a simulated dataset (H, S', ϕ'). In total, we created $15 \times 50 = 750$ such datasets. For each of the 15 real datasets, we call the whole set of 50 simulated datasets (generated using the parameter estimates on the real dataset) by the name of the real dataset followed by *sim*, for instance AW-sim.

The simulated datasets will be used only for testing the rooting of the trees. Indeed, using simulated datasets in the multiple associations context would require a model that allows for such multiple associations by considering

additional events. To the best of our knowledge, such a model does not exist yet. We therefore did not use such datasets to test the robustness of the associations.

2.2 Methods

2.2.1 Generating All the Optimal Solutions.

We used EUCALYPT [6], which for a given dataset (H, S, ϕ) and vector $c = \langle c_c, c_d, c_s, c_l \rangle$ specifying the costs of the events, generates all the optimal reconciliations in polynomial-delay, meaning that the computation time between two outputs is polynomial in the input size.

2.2.2 Comparing Two Sets of Reconciliations.

To estimate the similarity of the outputs of two different runs of the tree reconciliation algorithm, we needed a measure to compare two sets of tree reconciliations. Most studies summarise a reconciliation as a *pattern* of integers $\pi = \langle n_c, n_d, n_s, n_l \rangle$, representing the number of each event that it contains. The set of optimal solutions for a given dataset (H, S, ϕ) and cost vector c can thus be viewed as a multiset $\Lambda_{H,S,\phi,c}$ of patterns in $\mathbb{N}^4$. Notice that we needed to consider multisets as different reconciliations may induce the same pattern of events.

There is a wide literature on distances for sets of points. One of the best-known metrics between subsets, the Hausdorff metric, does not take into account the overall structure of the point sets. Other distances used for mining multisets, such as the Jaccard or Minkowski distance (see for example Chapter 6 in [11]), have the drawback of taking into account not the distance between the elements in the sets but only the number of different elements and their multiplicity.

Hence, for our purpose, we decided to introduce the following measure. Given a tree reconciliation Λ (i.e. a multiset of patterns), we define its representative $v_\Lambda = \sum_{\pi \in \Lambda} \pi$. Notice that such sum takes into account the multiplicities of a pattern. Given two tree reconciliations Λ_1 and Λ_2, we define a *dissimilarity measure* $d(\Lambda_1, \Lambda_2)$ as follows:

$$d(\Lambda_1, \Lambda_2) = \frac{||v_{\Lambda_1} - v_{\Lambda_2}||}{(|\Lambda_1| + |\Lambda_2|) \max_{\pi \in \Lambda_1 \cup \Lambda_2} ||\pi||} \tag{1}$$

where $||\cdot||$ is the L_1 norm and $|\Lambda|$ is the cardinality of the multiset Λ. Observe that $d(\Lambda_1, \Lambda_2) = 0$ whenever $\Lambda_1 = \Lambda_2$ while the converse is not necessarily true. Note also that we normalised this dissimilarity measure so that it takes values in $[0, 1]$. This dissimilarity measure, while not being a distance, enables us to summarize the comparison between two multisets of reconciliations. In particular, it takes into account both the multiplicity of the patterns and their actual values (patterns are vectors in $\mathbb{N}^4$ that might be close to each other).

2.2.3 Choosing Among Multiple Associations.

Three of the real datasets we selected present multiple associations. For each of them, we considered all the datasets that may be obtained by resolving the multiple associations in all the possible ways. More precisely, for each symbiont associated with more than one host, we chose one and only one of the possible associations, and we did this in all the possible ways. For instance, in the SBL dataset, 5 out the 8 leaves of the symbiont tree have multiple associations, each connected to 2, 2, 4, 5, and 7 leaves of the host tree respectively (see Figure 1 in the online Supplementary Material). By choosing in all possible ways among the multiple associations, we thus obtain 560 datasets.

2.2.4 Re-Rooting of the Symbiont Tree.

Most phylogenetic reconstruction algorithms produce unrooted trees, or rooted ones that have an unreliable root [9]. Rooting a phylogenetic tree is especially challenging for fast-evolving organisms. We therefore studied the influence on the optimal tree reconciliation of an erroneous rooting of the symbiont tree. More precisely, given a host tree H and a symbiont tree S, the association of their leaves ϕ, and a cost vector c, we compute all the optimal reconciliations for the pair H, S' where S' is obtained by positioning the root of S in an edge of S. Intuitively, one would expect that the correct positioning of the root would correspond to the reconciliation(s) having the minimum cost among all the ones that could be obtained by other rootings. This is indeed motivated by the same parsimony principle as for the tree reconciliation itself. Although slightly less immediate to grasp, one could expect also that positioning the root "near" to what would be the real one would lead to optimal reconciliation costs that are near the minimum.

Both cases were in fact observed by Gorecki *et al.* [8] who showed the existence of a certain property in models such as the Duplication-Loss for the gene/species tree reconciliation. Such property, which the authors called the *plateau property*, states that if we assign to each edge of the parasite tree a value indicating the cost of an optimal reconciliation when considering the parasite tree rooted in that edge, the edges with minimum value form a connected subtree in the parasite tree, hence the name of plateau. Furthermore, the edge values in any path from a plateau towards a leaf are monotonically increasing. In the presence of host switches, it was however not known whether such plateau property was satisfied.

Here, for both biological and simulated datasets, we count the number of plateaux (i.e. subtrees where rootings lead to minimal optimum cost), and we further keep track whether the original root belongs to a plateau. To study the robustness, we define a "small perturbation" of the rooting as follows. Given a dataset (H, S, ϕ), let $k = \max(5\%|V(S)|, 3)$. We compute all the optimal reconciliations for the pair H, S' where S' is obtained from S by positioning the root of S in an edge $(x, y) \in E(S)$ at a distance exactly k from the root, the latter being defined as the minimum distance between the node and the edge endpoints.

The variable k captures the "closeness" of the new root to the original one. We compare the sets of reconciliations obtained with the true positioning of the root and with the positioning at distance k using our dissimilarity measure (1). We then analyse the variations of these dissimilarities with respect to the variation of the distance k.

3 Results and Discussions

For both the editing of host-symbiont associations and perturbations of the symbiont tree root, we present only part of the results obtained in our analysis (in terms of datasets and/or of cost vectors) for reasons of space. In every case, the choice of which results to show was dictated either by the most interesting case observed among all those explored for the purposes of a discussion of the effect of edits and small perturbations on a parsimonious tree reconciliation, or, in the case of the cost vectors, by the one(s) that are more commonly used in the literature. An exhaustive presentation appears in Supplementary Material. Here, time-unfeasible reconciliations have been filtered-out. For each result appearing in Supplementary Material, we specify whether this is the case or not.

3.1 Perturbation of the Present-Day Host-Symbiont Associations

We present here the results for the SBL dataset analysed with cost vector $\langle 0, 1, 1, 1\rangle$. The TreeMap analysis of this dataset performed in [17] tried to maximise the number of cospeciations between hosts and symbionts but found out that sometimes host switches must be postulated to maximise cospeciation. Thus in some sense the choice of this cost vector is in accordance with the TreeMap philosophy. Our results for this dataset with the other cost vectors together with the two other datasets (MP and SFC) are presented in Section 2.1 from the online Supplementary Material.

Figure 1 (left) shows the optimal reconciliation costs obtained for the 560 datasets that were simulated from the SBL one by resolving the multiple associations in all the possible ways. We observe that when we change the associations, most often the optimum cost remains the same, namely 70% of the datasets have the same cost (of 7). However, in many cases (30%), changing association of the leaves results in a change of the optimum cost value (from 7 to a value in {6,8,9}).

To go further and analyse whether two datasets with same optimum cost have the same evolutionary history, we compared their sets of reconciliation patterns as described in Section 2.2.2. Figure 1 (right) shows the pairwise dissimilarities (see Eq. 1) between the reconciliation sets of the 392 datasets with same optimum cost of 7. Even if often the dissimilarity between two reconciliation sets is 0 (and we checked that the multisets of reconciliations are in fact exactly the same in those cases), in 65.5% of the cases this is not so, and the value instead ranges

inside [0.05,0.6], the largest dissimilarity (value of 0.6) being observed in 8.5% of the cases.

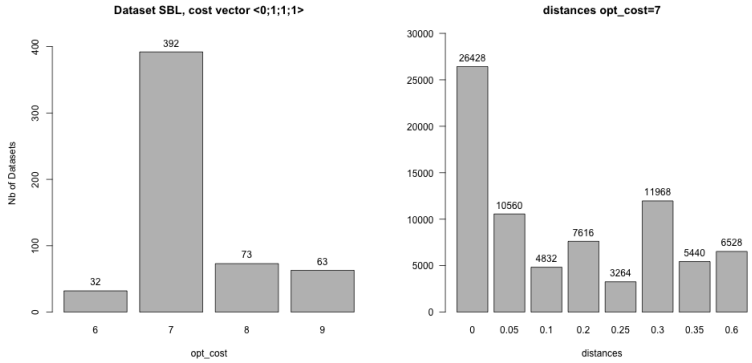

Figure 1: Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with optimum cost 7 (right) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$.

3.2 Re-Rooting of the Symbiont Tree

3.2.1 Testing the Plateau Property.

Table 1 in the online Supplementary Material presents the results for the 15 biological datasets evaluated with the 6 cost vectors in $\mathcal{C}$. Most of the datasets present only 1 plateau and only 2 datasets (CT and EC) present 2 plateaux. Moreover for 5 out of the 6 cost vectors tested, there is always a biological dataset for which 2 plateaux are observed.

The plateau property therefore does not hold in the presence of host switches for real datasets analysed with biologically plausible setups. It is interesting to observe that among the 15 biological datasets, there were never more than 2 plateaux. This may be due to the relatively small size of the trees.

We also note that in 37% of the cases, the original root is not in a plateau. Moreover, the difference between the optimal cost obtained for the original rooting and the cost obtained by placing the root inside the plateau is quite large (difference between columns D and B in Table 1 in online Supplementary Material). Among these 37%, in addition, for the datasets AW, FD, RH, and SFC, the original root of the symbiont tree is never in a plateau. This may indicate that either the original root is not at its correct position, or that there is not enough evolutionary dependence between the two organisms to allow for a correct inference of the symbiont tree root.

The simulated datasets present similar results as the biological ones (Table 2 in the online Supplementary Material). The number of datasets with more than

one plateau however increases, as does in some cases the number of plateaux observed. Indeed, some simulated datasets from the sets AW-sim, MP-sim, and SFC-sim exhibit up to 5 plateaux. In 17% of the simulations, the original root does not belong to a plateau (data not shown).

3.2.2 Rerooting at Distance k.

We show in Figure 2 the results obtained with the biological dataset MP. Similar figures are presented with other biological datasets in Section 2.3 from the online Supplementary Material. Here the dissimilarity of the reconciliation globally increases as k also increases. The farther is the new root from the original one, the more dispersed the patterns tend to be (*i.e.* the values of d have larger variance). These conclusions extend for 8 of the remaining biological datasets (EC, FD, GL, PML, PP, RP, SBL, SC). However, no such global trend is obtained for the other biological datasets for which we only observe variability (neither increasing nor decreasing) in the dissimilarities.

As concerns the simulated datasets, we observe a bigger dispersion between patterns with larger values taken by the dissimilarities (see Section 2.4 from the online Supplementary Material). This might be due to the fact that there are much more datasets (50 simulated datasets corresponding to one biological dataset). The trend of a global increase of the values and the variance of the dissimilarity when k increases is observed again.

4 Conclusions and Open Problems

In this paper, we explored the robustness of the parsimonious tree reconciliation method to some editing of the input required in order to associate a symbiont to a unique host in the case where multiple associations exist, as well as to small perturbations linked to a re-rooting of the symbiont tree.

In the first case, we observed that the choice of leaf associations may have a strong impact on the variability of the reconciliation output. Although such impact appears not so important on the cost of the optimum solution, probably due to the relatively small size of the input trees, the difference becomes more consequent when we refine the analysis by comparing, not the overall cost, but instead the patterns observed in the optimal solutions. Notice that this highlights the great interest in finding measures for the dissimilarity of sets of reconciliations such as the new one we proposed in this paper.

As indicated, we were able to do the analysis on the choice of leaf associations only for the real biological datasets because we are currently not capable of simulating the coevolution of symbionts and hosts following the phylogenetic tree of the latter and allowing for an association of the symbionts to multiple hosts. This is an interesting and we believe important open problem in the literature on reconciliations which we are currently trying to address.

As concerns the problem of the rooting, we were able to show that allowing for host switches invalidates the plateau property that had been previously

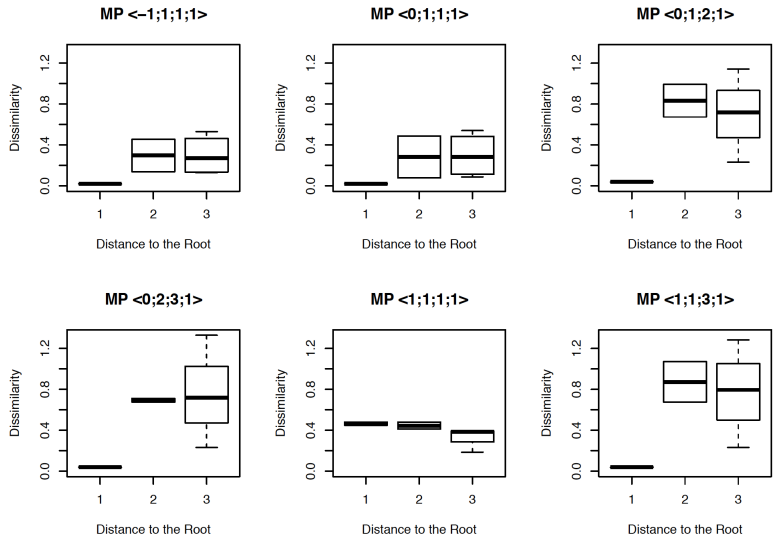

Figure 2: Boxplots of the dissimilarities between reconciliations obtained for the original dataset MP and all datasets simulated from MP by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between new and original root. The y-axis shows the value d of the dissimilarity of the reconciliation patterns.

observed (and actually also mathematically proved) in the cases where such events were not considered. Again here, the number of plateaux observed is small for the real datasets (this number is indeed 2). Moreover, such increase from 1 to 2 does not concern all pairs of datasets and of cost vectors, even though for all, except one of the cost vectors tested, there is always a biological dataset for which 2 plateaux are observed. We might be tempted to say that this is once more due to the small sizes of the input trees. However, the sizes are of the same order for the simulated datasets, but there the differences are greater: we may indeed reach up to 5 plateaux in some cases. We are currently not able to explain this difference between the two types of datasets (this might be just chance related to the fact that we have 50 times more simulated than biological datasets). For both of them, we also observe that the original root may not be inside a plateau, and that the proportion for which this is observed is approximately the same (3 cases out of 15 as compared to 17% respectively) for real or simulated datasets. We hypothesised that for the real datasets, this might indicate that the original root is not at its correct position. It would be interesting in future to try to validate this hypothesis. If it were proved to be true, an interesting, but hard open problem would be to be able to use as input for a cophylogeny study unrooted trees instead of rooted one, or even directly

the sequences that were originally used to infer the host and symbiont trees. In this case, we would then have to, at a same time, infer the trees and their optimal reconciliation.

Re-rooting the symbiont tree at distance k leads in many cases to an increase in both the values and variance of the dissimilarity measure in the patterns (9 out of 15 biological datasets and all sets of simulations). The dispersion and the values of dissimilarity are also greater in the simulated datasets than in the biological ones (here again, this could be an artefact due to the large number of simulated datasets).

Clearly, the effect in terms of number of plateaux depends on the presence of host switches since this number was proved to be always one when switches are not allowed [8]. Perhaps the most interesting open problem now is whether there is a relation between the number of plateaux observed as well as the level of dissimilarity among the patterns obtained on one hand, and the number of host switches in the optimal solutions on the other hand. Actually the relation may be more subtle, and be related not to the number of switches but to the distance involved in a switch, where by distance of a switch we mean the evolutionary distance between the two hosts involved in it. This could be measured in terms of the number of branches (as is the case in our method EUCALYPT) or in terms of the sum of the branch lengths, that is of estimated evolutionary time.

Acknowledgments

The authors would like to express their gratitude to Christian Gautier for fruitful preliminary discussions on this work.

References

[1] M. S. Bansal, E. J. Alm, and M. Kellis. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinf.*, 28(12):i283–i291, 2012.

[2] C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C. Gautier, C. Matias, and M.-F. Sagot. Cophylogeny Reconstruction via an Approximate Bayesian Computation. *Syst. Biol.*, 64(3):416–431, 2015.

[3] M. A. Charleston. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.*, 149(2):191–223, 1998.

[4] M. A. Charleston. Recent results in cophylogeny mapping. *Adv. Parasitol.*, 54:303–330, 2003.

[5] C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. Jane: a new tool for the cophylogeny reconstruction problem. *Algo. Mol. Biol.*, 5(16):1–10, 2010.

[6] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, and M.-F. Sagot. EUCALYPT: efficient tree reconciliation enumerator. *Algo. Mol. Biol.*, 10(1):3, 2014.

[7] J.-P. Doyon, C. Scornavacca, K. Y. Gorbunov, G. J. Szöllősi, V. Ranwez, and V. Berry. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In E. Tannier, editor, *Proceedings of the 8th annual RECOMB Satellite Workshop on Comparative Genomics (RECOMB-CG 2010)*, volume 6398 of *LNB*, pages 93–108. Spring-Verlag Berlin Heidelberg, 2011.

[8] P. Górecki, O. Eulenstein, and J. Tiuryn. Unrooted tree reconciliation: A unified approach. *IEEE/ACM Trans. Comput. Biology Bioinf.*, 10(2):522–536, 2013.

[9] B. Holland, D. Penny, and M. Hendy. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock: A simulation study. *Syst. Biol.*, 52(2):229–238, 2003.

[10] S. Keller-Schmidt, N. Wieseke, K. Klemm, and M. Middendorf. Evaluation of host parasite reconciliation methods using a new approach for cophylogeny generation. Technical report, Univ. of Leipzig, 2011.

[11] W. A. Kosters and J. F. J. Laros. Metrics for mining multisets. In M. Bramer, F. Coenen, and M. Petridis, editors, *Research and Development in Intelligent Systems XXIV*, pages 293–303. Springer London, 2008.

[12] D. Merkle and M. Middendorf. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory Biosci.*, 123(4):277–299, 2005.

[13] D. Merkle, M. Middendorf, and N. Wieseke. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinf.*, 11(Suppl 1):S60, 2010.

[14] M. Nei and S. Kumar. *Molecular evolution and phylogenetics.* Oxford Univ. Press, 2000.

[15] Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas. The cophylogeny reconstruction problem is NP-complete. *J. Comput. Biol.*, 18(1):59–65, 2011.

[16] R. D. M. Page. Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, 10(2):155–173, 1994.

[17] A. M. Paterson, R. D. Gray, D. H. Clayton, and J. Moore. Host-parasite co-speciation, host switching, and missing the boat. In D. H. Clayton and J. Moore, editors, *Host-parasite evolution: general principles and avian models*, pages 236–250. Oxford University Press, Oxford, 1997.

[18] Y.-L. Qiu, J. Lee, B. A. Whitlock, F. Bernasconi-Quadroni, and O. Dombrovska. Was the anita rooting of the angiosperm phylogeny affected by long-branch attraction? *Mol. Biol. Evol.*, 18(9):1745–1753, 2001.

[19] M. J. Sanderson and H. B. Shaffer. Troubleshooting molecular phylogenetic analyses. *Annu. Rev. Ecol. Syst.*, pages 49–72, 2002.

[20] J. Stavrinides and D. S. Guttman. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J. Virol.*, 78(1):76–82, 2004.

[21] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinf.*, 28(18):i409–i415, 2012.

[22] M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinf.*, 28(18):i409–i415, 2012.

[23] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular systematics*, pages 407–514. Sinauer Associates, Inc., 1996.

[24] A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB)*, 8(2):517–535, 2011.