Exploring the Robustness of the Parsimonious Reconciliation Method in Host-Symbiont Cophylogeny

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

Exploring the Robustness of the Parsimonious Reconciliation Method in Host-Symbiont Cophylogeny

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

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 the tree reconciliation model is currently unable 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.

Index Terms—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 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* [1], [2], [3], [4]. 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 [5], [6]. 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 [7], [8], [9], [10], [11]. In most situations, as shown in [8], 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 avail-

- *L. Urbini, B. Sinaimeri and M. -F. Sagot are with 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*
  E-mail: {laura.urbini, blerina.sinaimeri}@inria.fr
- *C. Matias is with Sorbonne Université, Université Paris Diderot, Centre National de la Recherche Scientifique, Laboratoire de Probabilités, Statistique et Modélisation, 4 place Jussieu, Paris, France*

able 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. Note that some reconciliation tools (e.g. JANE 4 [12], CORE-ILP [13], CORE-SYM [14]) have been equipped with ad-hoc methods to deal with multiple associations but the underlying mathematical model does not include this possibility. 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 if they happen to be prokaryotes and can recombine among them [15], [16], [17]. 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 [15], [17], [18]. 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 [16], [19], [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 all possible choices of single symbiont-host leaf mapping in the presence of multiple associations (we call this *resolving* the multiple associations into simple ones), 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], [21]).

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 on the output.

Many tree reconciliation algorithms exist, but only a few enumerate all solutions. The most commonly used are NOTUNG [22], ecceTERA [23], JANE 4 [12], and CORE-PA [10]. However, the first two were 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 [8]. In the context of host/symbiont only the method that we developed, called EUCALYPT [8], is exhaustive, and we therefore decided to use it exclusively in order to explore the robustness of the parsimonious tree reconciliation method. We recall that our objective is not to test the robustness of each of the tools but of the underlying mathematical model they are all based on.

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 [7], [9]. 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 [24] where the authors provided a first basic simulation setup (to our knowledge, the only other one available in the host-symbiont 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 [21] 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.

In an earlier version of this work [25], we relied on 15 biological datasets extracted from the literature and considered simulated datasets (following the structure of the real ones) for studying the robustness in the re-rooting case. We also presented a measure to compare sets of tree reconciliations which may be of independent interest. In the present work, we extend our analysis with 13 new biological datasets in addition to the 15 previous ones, thus summing up to a total of 28 biological datasets explored. We also simulate datasets for studying the robustness in the multiple-associations case, relying on a method recently proposed in [26]. Our previous analysis introduced a dissimilarity measure to compare two sets of reconciliations and we used it to characterise how different the reconciliations were under editing of the multiple associations. However, we were lacking a way of assessing whether the observed values of this dissimilarity were big or small. In this paper, we describe a permutation method to obtain the empirical distribution of the dissimilarities between pairs of sets of reconciliations obtained through editing the datasets with multiple associations. The empirical distribution is obtained under the null hypothesis of random association between the leaves of the two trees. In our previous work, we also considered the robustness of the parsimonious method in the case where the solutions provided may be time-unfeasible. In the present one, only the time-feasible solutions are retained. In the end, we analyse and discuss the results obtained for 28 real biological datasets and their simulated counterparts from the viewpoint of exploring small perturbations on the datasets.

The organisation of the paper is as follows. We start by introducing the datasets that will be used, both real and simulated ones, as well as in the latter case the methods 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 are freely available at http://eucalypt.gforge.inria.fr/. This webpage also contains the online Supplementary Material with exhaustive results on the datasets.

2 Material and Methods

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

2.1.1 Biological Datasets

To test the robustness of the method, we selected 28 biological datasets from the literature:

AP - Acacia & Pseudomyrmex. This dataset was extracted from the work of Gómez-Acevedo *et al.* [27]. The host tree includes 9 leaves and the symbiont tree includes 7 leaves.

AS - Aves & Syringophilopsis. This dataset was extracted from the work of Hendricks *et al.* [28]. The host tree includes 19 leaves and the symbiont tree includes 16 leaves.

AW - Arthropod & Wolbachia. This dataset was extracted from the work of Simões *et al.* [29], [30] and is composed of a pair of host and symbiont trees which have each 12 leaves.

CA - Carex & Anthracoidea. This dataset was extracted from the work of Escudero [31]. The host tree includes 41 leaves and the symbiont tree includes 30 leaves.

CP - Cichlidae & Platyhelminthes. This dataset was extracted from the work of Mendlová *et al.* [32]. The host tree includes 6 leaves and the symbiont tree includes 29 leaves.

CT - Cichlidogyrus & Tropheini. This dataset was extracted from the work of Vanhove *et al.* [33]. The host tree includes 19 leaves and the symbiont tree includes 28 leaves.

EC - Encyrtidae & Coccidae. This dataset was extracted from the work of Deng *et al.* [34]. The host tree includes 7 leaves and the symbiont tree includes 10 leaves.

FA - Ficus & Agaonidae. This dataset was extracted from the work of McLeish and Van Noort [35]. The host tree includes 7 leaves and the symbiont tree includes 8 leaves.

FD - Fishes and Dactylogyrus. This dataset was extracted from the work of Juan *et al.* [36]. The host tree includes 20 leaves and the symbiont tree includes 50 leaves.

FE - Formicidae & Eucharitidae. This dataset was extracted from the work of Murray *et al.* [37]. The host tree includes 4 leaves and the symbiont tree includes 5 leaves.

GL - Gopher & Lice. This dataset was extracted from the work of Hafner and Nadler [38]. The host tree includes 8 leaves and the symbiont tree includes 10 leaves.

GM - Goodeinae & Margotrema. This dataset was extracted from the work of Martinez *et al.* [39] and is composed of a pair of host and symbiont trees which have each 14 leaves.

IFL - Insect & Flavobacterial endosymbionts. This dataset was extracted from the work of Rosenblueth *et al.* [40] and is composed of a pair of host and symbiont trees which have each 17 leaves.

MF - Mycocepurus smithii & Fungi. This dataset was extracted from the work of Kellner *et al.* [41]. The host tree includes 11 leaves and the symbiont tree includes 9 leaves.

MP - Myrmica & Phengaris. This dataset was extracted from the work of Jansen *et al.* [42] and is composed of a pair of host and symbiont trees which have each 8 leaves.

PML - Pelican & Lice ML. This dataset was extracted from the work of Hughes *et al.* [43] and is composed of a pair of host and symbiont trees which have each 18 leaves. The trees here were generated through a maximum likelihood approach.

PMP - Pelican & Lice MP. This dataset was extracted from the work of Hughes *et al.* [43] and is composed of a pair of host and symbiont trees which have each 18 leaves. The trees here were generated through a maximum parsimony approach.

PP - Primates & Pinworms. This dataset was extracted from the work of Hugot [44]. The host tree includes 36 species and the symbiont tree includes 40 leaves.

RH - Rodents & Hantaviruses. This dataset was extracted from the work of Ramsden *et al.* [45]. The host tree includes 34 leaves and the symbiont tree includes 42 leaves.

RM - Ramphastidae & Mallophaga. This dataset was extracted from the work of Weckstein [46]. The host tree includes 11 leaves and the symbiont tree includes 5 leaves.

RP - Rodents & Pinworms. This dataset was extracted from the work of Hugot [47] and is composed of a pair of host and symbiont trees which have each 13 leaves.

SBL - Seabirds & Lice, This dataset was extracted from the work of Paterson *et al.* [48]. The host tree includes 15 leaves and the symbiont tree includes 8 leaves.

SC - Seabirds & Chewing Lice. This dataset was extracted from the work of Paterson *et al.* [49]. The host tree includes 11 leaves and the symbiont tree includes 14 leaves.

SFC - Smut Fungi & Caryophillaceus plants. This dataset was extracted from the work of Refregier *et al.* [50]. The host tree includes 15 leaves and the symbiont tree includes 16 leaves.

SHA - Sigmodontinae Hantavirus & Arenaviridae. This dataset was extracted from the work of Jackson and Charleston [51]. The host tree includes 14 leaves and the symbiont tree includes 16 leaves.

SSA - Sigmodontinae Spumavirus & Arenaviridae. This dataset was extracted from the work of Jackson and Charleston [51] and is composed of a pair of host and symbiont trees which have each 10 leaves.

TC - Teleostei & Copepods. This dataset was extracted from the work of Paterson and Poulin [52]. The host tree includes 8 leaves and the symbiont tree includes 9 leaves.

TD - Tephritidae & Bacteria. This dataset was extracted from the work of Viale *et al.* [53]. The host tree includes 26 leaves and the symbiont tree includes 22 leaves.

The choice of these datasets 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. We call attention here to the fact that only 15 of these datasets present multiple associations (namely AP, AS, CA, CP, FA, FE, GM, MF, MP, RM, SBL, SFC, SHA, TC, TD) and are the ones used to study the robustness of the method in the case of multiple associations. Let us recall that whenever a symbiont inhabits more than one host, we have multiple associations. For a leaf $s \in L(S)$ (where $L(S)$ is the set of leaves of the symbiont tree S), we denote by $\phi(s)$ the set of host leaves to which it is associated. Given a dataset (H,S,ϕ), the number of multiple associations M for the dataset is:

$$M(H, S, \phi) = \sum\nolimits_{s \in L(S)} |\phi(s)| - 1. \quad (1)$$

Table 1 shows the number of multiple associations in the datasets where it is non null.

2.1.2 Simulated Datasets for Multiple Associations

To study the multiple associations, we generated simulated datasets with a variable amount of multiple associations, using a method developed by Drinkwater *et al.* [26]. The simulated datasets were generated using the 15 biological datasets that present multiple associations as follows.

For each of them, we simulated a number of multiple associations, as defined in 1, equal to $x\%$ of the total number of host tree leaves, with $x \in \{10, 15, 20, 25, 30, 35, 40, 45, 50\}$. We thus constructed 9 simulated datasets per original real dataset, by adding or removing multiple associations and keeping the host and symbiont trees fixed. More precisely, consider a dataset D with M multiple associations and an integer M^* (equal to the integer part of $x\%|L(H)|$). Whenever $M^* > M$, we randomly choose $M^* - M$ different pairs $\langle s, h\rangle \in L(S) \times L(H)$ such that we do not already have $h \in \phi(s)$ and we associate them (*i.e.* $h \in \phi(s)$). If $M^* < M$, we randomly choose $M - M^*$ different pairs $\langle s, h\rangle \in L(S) \times L(H)$, for which $h \in \phi(s)$ and $|\phi(s)| \geq 2$ and delete their association.

For each real dataset D, we denote by $D_{x\%}$ the dataset simulated from D with $x\%$ of multiple associations.

2.1.3 Simulated Datasets for Re-Rooting a Symbiont Tree

To study the re-rooting, we generated simulated datasets using a method that we previously developed, called COALA [21], and the 28 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 (ABC) 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.* [21] for more details) to obtain, for each vector γ, a simulated symbiont tree S' whose evolution follows that of the host tree H (under the parameter value γ). Each dataset (H, S, ϕ) and probability vector γ thus led to a simulated dataset (H, S', ϕ'). In total, we created $28 \times 50 = 1400$ such datasets. For each real dataset D, we denote by D-sim the 50 simulated datasets (generated using the parameter estimated on D).

2.2 Methods

2.2.1 Generating All the Optimal Solutions

We used EUCALYPT [8], which for a given dataset (H, S, ϕ) and a 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. Only time-feasible reconciliations are retained.

2.2.2 Choosing Among Multiple Associations

Fifteen of the real datasets we selected present multiple associations. For each dataset $D = (H, S, \phi)$, we considered all the datasets that can 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. In the end we have a set of datasets $\{D_1, ..., D_t\}$ with simple associations. For instance, in the SBL dataset, 5 of 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.

2.2.3 Re-Rooting of the Symbiont Tree

Most phylogenetic reconstruction algorithms produce unrooted trees, or rooted ones that have an unreliable root [19]. 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 in a first step all the optimal reconciliations for the pair H, S' where S' is obtained by positioning the root of S in an edge of S. With these re-rooted trees, we explore the plateau property (see below). In a second step, we want to study the robustness from a slightly different perspective, taking into account the distance from the new root to the original one. We then focus on the subset of re-rooted datasets, where the root is positioned in an edge of S at distance at most k to the original root. More precisely, given a dataset (H, S, ϕ), let $k = \max(5\%|V(S)|, 3)$. We focus on 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 vertex and the edge endpoints. The variable k captures the "closeness" of the new root to the original one.

2.2.4 The Plateau Property

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.* [54] 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 symbiont tree a value indicating the cost of an optimal reconciliation when considering the symbiont tree rooted in that edge, the edges with minimum value form a connected subtree in the symbiont 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 the biological and the simulated datasets, we use the sets of all optimal reconciliations of the datasets with all possible symbiont tree rootings to count the number of plateaux (*i.e.* subtrees of the symbiont tree where the rootings in their edges lead to a minimum cost), and we further keep track whether the original root belongs to a plateau.

2.2.5 Comparing Two Sets of Reconciliations

To evaluate the similarity of the outputs of two different runs of the tree reconciliation algorithm, we need a measure to compare two sets of tree reconciliations. A first step is to compare the respective optimal costs obtained at each run (note that this makes sense only when tree topologies and cost vectors are fixed). When these optimal costs are equal, we need to keep more information on the sets of optimal 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 need 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 [55]), 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 multisets of patterns Λ_1 and Λ_2, we define a *dissimilarity measure* $d(\Lambda_1, \Lambda_2)$ as follows:

$$d(\Lambda_1, \Lambda_2) = \frac{||v_{\Lambda_1} - v_{\Lambda_2}||}{(|\Lambda_1| + |\Lambda_2|) \max_{\pi \in \Lambda_1 \cup \Lambda_2} ||\pi||} \qquad (2)$$

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. Notice 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 summarise 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).

2.2.6 Dissimilarities in the Case of Multiple Associations

As already explained, for each dataset D, we have extracted a set of datasets $\{D_1, \ldots, D_t\}$ each with simple associations. We fixed a cost vector c and for each $1 \leq j \leq t$, we computed all the optimal reconciliations for D_j. We denoted by $\Lambda_{D_j,c}$ the multiset of patterns (as defined above) obtained for these optimal reconciliations and $opt(D_j, c)$ their optimal cost. In most of the cases, the set $\{opt(D_j, c); 1 \leq j \leq t\}$ will contain many different values (this is a first observation that the corresponding multisets of reconciliations are different). Then, to further analyse the diversity of these different optimal reconciliations, we focused on the most frequent optimal cost $opt^*(D_j, c)$ and on the subset $\mathcal{D}^* \subseteq \{D_1, \ldots, D_t\}$ of datasets that exhibit this most frequent optimal cost. For any pair of datasets $D, D' \in \mathcal{D}^*$, the optimal reconciliations for D and D' have same cost (by construction) and we further analyse how different they are by computing the dissimilarity between these sets. Given $\Lambda_{D,c}$ and $\Lambda_{D',c}$ the sets of optimal reconciliations for D and D' respectively, we thus compute $d(\Lambda_{D,c}, \Lambda_{D',c})$ for any pair $D, D' \in \mathcal{D}^*$.

2.2.7 Dissimilarities in Case of Re-Rooting at Distance k

In order to study the robustness of the parsimonious tree reconciliation method with respect to the position of the root in the symbiont tree, we explore "small perturbations" of the rooting by varying the distance k of the position of the new root with respect to the original one. We then compare the sets of reconciliations obtained with the true positioning of the root and with the positioning at distance k using our dissimilarity measure defined in Eq. (2). Notice that here we are interested in the variation of dissimilarity at distance less than k from the original root. Thus, we are not necessarily inside a plateau. For this reason, we use our dissimilarity measure to compare sets of reconciliations where the optimum cost may not be the same.

2.2.8 Empirical Distribution of Dissimilarity

It is important to understand what values of the dissimilarity measure correspond to low/high values between two multisets of patterns. To answer this question, we studied the behaviour of the dissimilarity under the null hypothesis $\mathcal{H}_0$ that there is a random association between H and S. More precisely, the empirical distribution of the dissimilarity between two multisets of patterns is computed in the following way: we fix the topologies of H and S as well as the association ϕ between their leaves, and we randomly permute the labels of the leaves of H and S to obtain permuted datasets.

In the multiple associations setup, for any original dataset $D = (H, S, \phi)$ and any cost vector c, we previously obtained a set of dissimilarities $\{d_i; 1 \leq i \leq K\}$

between all the pairs of datasets that have the same most frequent optimal cost. We generated 1000 permuted datasets $\{D^0, D^1, ..., D^{999}\}$, by permuting the labels of the leaves of H and S and keeping the associations between the leaves fixed, *i.e.* fix the topology of the tree H and consider the tree H' given by a permutation of the labels of its leaves (similarly for S). The associations ϕ between H', S' remains the same as in H, P. In other words, for a leaf s of the symbiont tree and a leaf h of the host tree, if $\phi(s) = h$, they are associated in the trees H', S'. For each D^j, we resolved the multiple associations into simple ones, extracted the subset $\mathcal{D}^{j,*}$ of datasets that exhibit the most frequent optimal cost and for all pairs of such datasets, computed the dissimilarity of their optimal reconciliation sets. We thus ended up with a set of dissimilarities $\{d_i^j; 1 \leq i \leq K_j\}$. We then plotted a histogram of the values $\{d_i^j; 1 \leq i \leq K_j, 0 \leq j \leq 999\}$. This is the empirical distribution of the dissimilarities under the null hypothesis of random associations between H and S. We computed the 10%–quantiles and the 90%-quantiles of this empirical distribution.

For the original dataset D, we denote by $freq_{dissim}(D)$, the most frequent non null dissimilarity. Whenever this value is less than the 10%-quantile, we are observing a value that is statistically significantly small. When this value is larger than the 90%-quantile, we are observing a value that is statistically significantly large.

3 Results and Discussions

For both the editing of the host-symbiont associations and the perturbations of the symbiont tree root, we present here only part of the results obtained in our analysis (in terms of datasets and/or of cost vectors). 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 the Supplementary Material. Notice that the time-unfeasible reconciliations have been filtered-out.

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 [48] 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 agreement with the TreeMap philosophy. Our results for this dataset with the other cost vectors together with the other datasets presenting multiple associations (AP, AS, CA, CP, FA, FE, GM, MF, MP, RM, SFC, SHA, TC and TD) are presented in Section 2.1 from the online Supplementary Material.

Figure 1 (top) shows the optimal reconciliation costs obtained for the 560 datasets that were generated 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 the 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 through the dissimilarity measure introduced in Eq. (2). Figure 1 (bottom) shows a density histogram of the pairwise dissimilarities between the reconciliation sets of the 392 datasets with same optimum cost of 7. Even if in many cases 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 82% of the cases this is not so, and the value instead ranges inside $[0.004, 0.6]$, the largest dissimilarity (value of 0.6) being observed in 8.5% of the cases.

In order to assess whether the values of the dissimilarity are (statistically) large or not, we plotted in Figure 2 the empirical distribution (under a null hypothesis of random association) of the dissimilarities between sets of reconciliations (for the cost vector $\langle 0, 1, 1, 1\rangle$) of datasets with same most frequent optimal cost, obtained by resolving in all possible ways permuted versions of the original SBL dataset (as explained in the paragraph "Empirical Distribution of Dissimilarity"). As already explained, we focus on $freq_{dissim}(SBL)$, the most frequent non null dissimilarity observed in the original dataset. In this case, it takes two different values (0.32 and 0.6), which appear to be the quantiles at levels 86.607% and 97.64% respectively of the empirical distribution. We then cannot conclude whether the dissimilarity value of 0.32 is statistically big or not. However, the dissimilarity value of 0.6 is bigger than the 90%-quantile, so that we can conclude that this is a statistically big dissimilarity. This result shows that even if two datasets have the same optimal cost, they may exhibit very different reconciliations.

Still considering the SBL dataset, now for the other cost vectors c (Section 2.2 in the Supplementary Material), the values of the most frequent non null dissimilarity $freq_{dissim}(SBL)$ are as follows. For the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$ and $\langle 1, 1, 1, 1\rangle$, the values are larger than the 90%-quantile and we conclude that they are statistically significantly large. For the cost vectors $\langle 0, 2, 3, 1\rangle$ and $\langle 1, 1, 3, 1\rangle$, the results are not conclusive. There are no cases with a value smaller than the 10%-quantile.

For the other datasets (Section 2.2 in the Supplementary Material), we observe that whenever the original datasets have less than three multiple associations, or if the multiple associations are in the same clade (AS, TC, TD), the value $freq_{dissim}(D)$ is smaller than the 10%-quantile of the empirical distribution. This means that if two datasets have the same optimum cost, they have very similar reconciliations (their dissimilarities are statistically significantly small). For some datasets (CP, FA, FE, MF and SFC), the value $freq_{dissim}(D)$ is between the 10%- and the 90%-quantiles. In these cases, we cannot conclude about the values of the dissimilarities of the reconciliations. In the other cases (GM, MP, SBL), there are some cost vectors such that $freq_{dissim}(D)$ is larger than the 90%-quantile while it

is never smaller than the 10%-quantile. For these three last datasets, even if the cost of the optimal solution is the same, we can thus obtain very different reconciliations. Indeed, if we have a tree with symbionts that inhabit different hosts which are topologically far, the way in which we choose the associations may have a big impact in terms of reconciliation. This means that the resulting dissimilarity is directly related to the leaves association ϕ.

In order to better understand what may be happening, if there is a relation between the number of multiple associations and the dissimilarity observed, we considered the simulated datasets $D_{x\%}$ constructed with different values of multiple associations. The SBL dataset has originally 94% of multiple associations. This means that in order to create a dataset $SBL_{x\%}$, we deleted some associations. The structure of the 9 datasets $SBL_{x\%}$ is shown in Table 2.

It is important to note that the number of datasets obtained by resolving the multiple associations into simple ones is not related to the percentage $x\%$, but rather to the combinatorial way to solve it. For example $SBL_{30\%}$ and $SBL_{35\%}$ have the same number of multiple associations (this is due to the fact that the integer parts of the values $x\%|L(H)|$ are the same in this case). However, in $SBL_{30\%}$ the multiple associations are spread among more leaves than for $SBL_{35\%}$. This is why $SBL_{30\%}$ is resolved with more datasets than $SBL_{35\%}$. Currently we are not able to create datasets with multiple associations that would lead to a fixed number of resolutions (*i.e.* datasets with simple associations obtained from the original dataset). Figures 3, 4 and 5 are similar to Figure 1 (which concerns the original dataset SBL) but now for the simulated datasets $SBL_{x\%}$, with cost vector $\langle 0, 1, 1, 1\rangle$. We see that in general the number of optimal reconciliations and the dissimilarity increase with the value of x. A particular case is $SBL_{25\%}$ that presents the largest most frequent non null dissimilarity. If we look at this dataset for the other cost vectors (Section 2.3 in the Supplementary Material), we observe that when we consider low costs for the host switch, namely for the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$ and $\langle 1, 1, 1, 1\rangle$, this dataset exhibits a value of $freq_{dissim}(SBL_{25\%})$ larger than what is obtained for other values of x. We believe that this is due to a high number of host switches in the reconciliations.

The other simulated datasets present similar results as $SBL_{x\%}$ (Supplementary Material). In general, the number of optimal reconciliations and the dissimilarity increase with the value of x. However, it is important to note that the results are related to the combinatorial way in which datasets with multiple-associations are resolved into datasets of simple associations.

3.2 Re-Rooting of the Symbiont Tree

3.2.1 Testing the Plateau Property

Table 1 and 2 in the Supplementary Material in the online Supplementary Material present the results for the 28 biological datasets evaluated with the 6 cost vectors in $\mathcal{C}$. Most of the datasets present only 1 plateau, 3 datasets (CA, CT and EC) present 2 plateaux and 1 dataset (CT) present 3 plateaux. Moreover for 5 out of the 6 cost vectors tested, there is always a biological dataset for which 2 plateaux are observed. The cost vector $\langle 1, 1, 1, 1\rangle$ is the one that gives, for the CT dataset, 3 plateaux.

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 28 biological datasets (except for TC, with cost vector $\langle 1, 1, 3, 1\rangle$), there were never more than 2 plateaux. This may be due to the relatively small size of the trees.

We also note that in 53% 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 the online Supplementary Material). Among these 53%, in addition, for the datasets AW, CP, FD, GM, MF, RH, SFC, SHA, TC, TD, 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 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 3 and 4 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 CA-sim and FE-sim exhibit up to 5 plateaux. In 25% of the simulations, the original root does not belong to a plateau (data not shown).

3.2.2 Re-rooting at Distance k

We show in Figure 6 the results obtained with the biological dataset MP. Similar figures are presented with other biological datasets in Section 2.5 in 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 27 of the remaining biological datasets. 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 the patterns with larger values taken by the dissimilarities (see Section 2.5 in the online Supplementary Material). This might be due to the fact that there are many more datasets (50 simulated datasets corresponding to one biological dataset). The trend of a global increase of the values and of 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 concerns the problem of the rooting, we were able to show that allowing for host switches invalidates the plateau property that had been previously 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 at most of 3). Moreover, having more than one plateau 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 at least 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 6 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). In 10 real datasets among the 28, the original root is never in the plateau. We hypothesised that for the biological 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 ones, 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 (17 out of 28 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 [54]. 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 an estimated evolutionary time.

TABLE 1
List of datasets exhibiting multiple associations, the number $M(H, S, \phi)$ of such multiple associations as in Equation (1) and the ratio (in percentage) of this number to the number of host leaves.

Dataset	AP	AS	CA	CP	FA	FE	GM	MF	MP	RM	SBL	SFC	SHA	TC	TD
$M(H,S,\phi)$	22	4	11	5	2	3	5	12	8	6	15	4	1	1	4
$M/\lvert L(H)\rvert$ (%)	244	21	27	83	29	75	36	109	100	55	94	27	7	13	15

TABLE 2
Table showing some details for the $SBL_{x\%}$ datasets. Each line shows a summary of $SBL_{x\%}$. Column A indicates the number of multiple associations; column B shows the number of datasets obtained resolving those multiple associations into simple ones; column C describes how many leaves in the symbiont tree S have multiple associations (and the cardinality of their image $|\phi(s)|$ in the host tree H).

$SBL_{x\%}$	A	B	C
$SBL_{10\%}$	2	3	1 leaf (3 associations)
$SBL_{15\%}$	2	3	1 leaf (3 associations)
$SBL_{20\%}$	3	4	1 leaf (4 associations)
$SBL_{25\%}$	4	12	3 leaves (2, 2 and 3 associations)
$SBL_{30\%}$	5	18	3 leaves (3, 3 and 2 associations)
$SBL_{35\%}$	5	12	2 leaves (3 and 4 associations)
$SBL_{40\%}$	6	24	3 leaves (2, 3 and 4 associations)
$SBL_{45\%}$	7	30	3 leaves (2, 3 and 5 associations)
$SBL_{50\%}$	8	36	3 leaves (2, 3 and 6 associations)

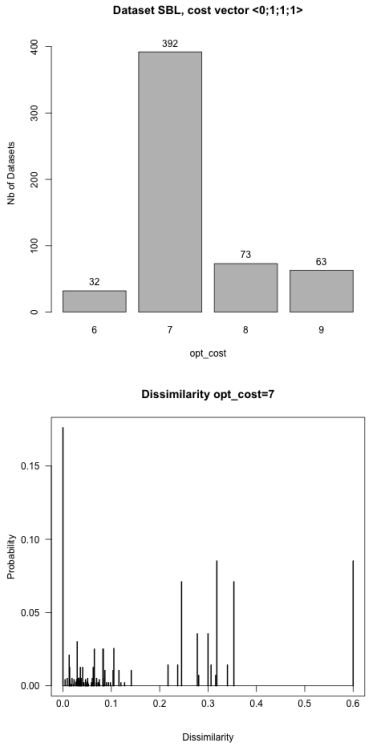

Fig. 1. Barplots of optimum cost (top) and dissimilarity between pairs of reconciliations with optimum cost 7 (bottom) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$.

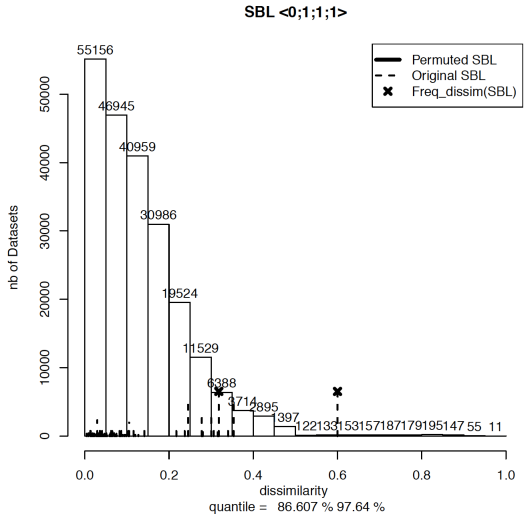

Fig. 2. Histogram of dissimilarity derived from SBL dataset with the cost vector $\langle 0, 1, 1, 1\rangle$. The black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The dotted lines are obtained by resolving the multiple associations in all possible ways for the original dataset SBL. The crosses are the $freq_{dissim}(SBL)$.

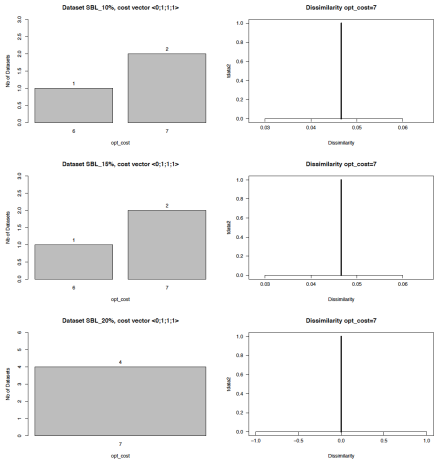

Fig. 3. Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimum cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each lines is a different $SBL_{x\%}$ with $x = 10, 15, 20$.

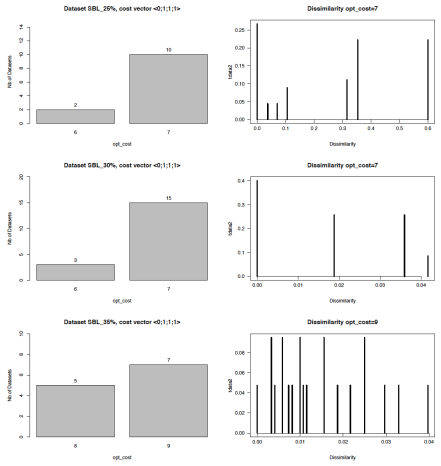

Fig. 4. Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimum cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each lines is a different $SBL_{x\%}$ with $x = 25, 30, 35$.

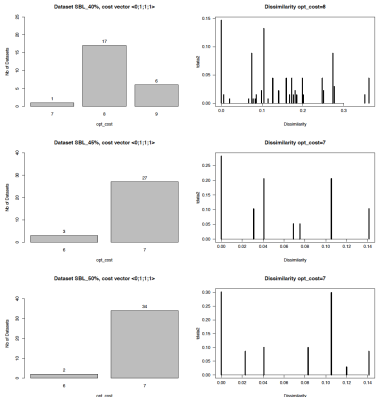

Fig. 5. Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimum cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each lines is a different $SBL_{x\%}$ with $x = 40, 45, 50$.

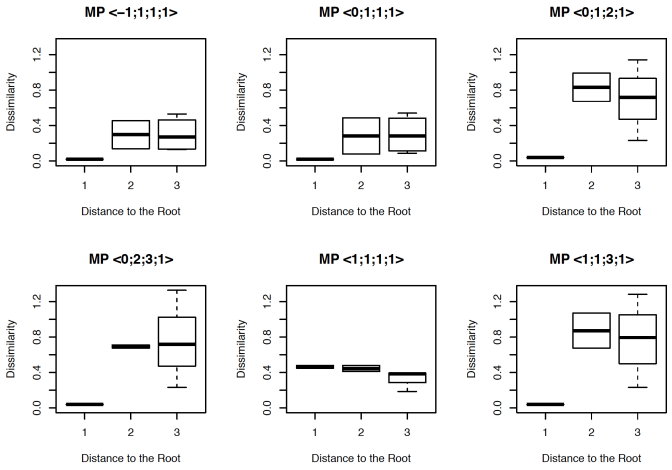

Fig. 6. Boxplots of the dissimilarities between reconciliations obtained for the original dataset MP and all datasets obtained from MP by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k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

REFERENCES

[1] M. A. Charleston, "Jungles: a new solution to the host/parasite phylogeny reconciliation problem," *Math. Biosci.*, vol. 149, no. 2, pp. 191–223, 1998.

[2] ——, "Recent results in cophylogeny mapping," *Adv. Parasitol.*, vol. 54, pp. 303–330, 2003.

[3] D. Merkle and M. Middendorf, "Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information," *Theory Biosci.*, vol. 123, no. 4, pp. 277–299, 2005.

[4] R. D. M. Page, "Parallel phylogenies: reconstructing the history of host-parasite assemblages," *Cladistics*, vol. 10, no. 2, pp. 155–173, 1994.

[5] Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas, "The cophylogeny reconstruction problem is NP-complete," *J. Comput. Biol.*, vol. 18, no. 1, pp. 59–65, 2011.

[6] A. Tofigh, M. Hallett, and J. Lagergren, "Simultaneous identification of duplications and lateral gene transfers," *IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB)*, vol. 8, no. 2, pp. 517–535, 2011.

[7] M. S. Bansal, E. J. Alm, and M. Kellis, "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss," *Bioinf.*, vol. 28, no. 12, pp. i283–i291, 2012.

[8] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, and M.-F. Sagot, "EUCALYPT: efficient tree reconciliation enumerator." *Algo. Mol. Biol.*, vol. 10, no. 1, p. 3, 2014.

[9] 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 *Proceedings of the 8th annual RECOMB Satellite Workshop on Comparative Genomics (RECOMB-CG 2010)*, ser. LNB, E. Tannier, Ed., vol. 6398. Spring-Verlag Berlin Heidelberg, 2011, pp. 93–108.

[10] D. Merkle, M. Middendorf, and N. Wieseke, "A parameter-adaptive dynamic programming approach for inferring cophylogenies," *BMC Bioinf.*, vol. 11, no. Suppl 1, p. S60, 2010.

[11] 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.*, vol. 28, no. 18, pp. i409–i415, 2012.

[12] C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas, "Jane: a new tool for the cophylogeny reconstruction problem," *Algo. Mol. Biol.*, vol. 5, no. 16, pp. 1–10, 2010.

[13] N. Wieseke, T. Hartmann, M. Bernt, and M. Middendorf, "Cophylogenetic reconciliation with ilp," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 6, pp. 1227–1235, Nov 2015.

[14] N. Wieseke, M. Bernt, and M. Middendorf, *Unifying Parsimonious Tree Reconciliation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 200–214.

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

[16] J. Stavrinides and D. S. Guttman, "Mosaic evolution of the severe acute respiratory syndrome coronavirus," *J. Virol.*, vol. 78, no. 1, pp. 76–82, 2004.

[17] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis, "Phylogenetic inference," in *Molecular systematics*, D. M. Hillis, C. Moritz, and B. K. Mable, Eds. Sinauer Associates, Inc., 1996, pp. 407–514.

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

[19] B. Holland, D. Penny, and M. Hendy, "Outgroup misplacement and phylogenetic inaccuracy under a molecular clock: A simulation study," *Syst. Biol.*, vol. 52, no. 2, pp. 229–238, 2003.

[20] 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.*, vol. 18, no. 9, pp. 1745–1753, 2001.

[21] 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.*, vol. 64, no. 3, pp. 416–431, 2015.

[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.*, vol. 28, no. 18, pp. i409–i415, 2012.

[23] E. Jacox, C. Chauve, G. J. Szll?si, Y. Ponty, and C. Scornavacca, "eccetera: comprehensive gene tree-species tree reconciliation using parsimony," *Bioinformatics*, vol. 32, no. 13, pp. 2056–2058, 2016.

[24] S. Keller-Schmidt, N. Wieseke, K. Klemm, and M. Middendorf, "Evaluation of host parasite reconciliation methods using a new approach for cophylogeny generation," Univ. of Leipzig, Tech. Rep., 2011. [Online]. Available: https://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/11-013.pdf

[25] L. Urbini, B. Sinaimeri, C. Matias, and M.-F. Sagot, "Robustness of the parsimonious reconciliation method in cophylogeny," in *Algorithms for Computational Biology: Third International Conference, AlCoB 2016, Trujillo, Spain, June 21-22, 2016, Proceedings*, M. Botón-Fernández, C. Martín-Vide, S. Santander-Jiménez, and A. M. Vega-Rodríguez, Eds. Cham: Springer International Publishing, 2016, pp. 119–130.

[26] B. Drinkwater, A. Qiao, and M. Charleston, "WiSPA: A new approach for dealing with widespread parasitism," arXiv:1603.09415, Tech. Rep., 2016.

[27] S. Gómez-Acevedo, L. Rico-Arce, A. Delgado-Salinas, S. Magallón, and L. E. Eguiarte, "Neotropical mutualism between acacia and pseudomyrmex: phylogeny and divergence times," *Molecular Phylogenetics and Evolution*, vol. 56, no. 1, pp. 393–408, 2010.

[28] S. A. Hendricks, M. E. Flannery, and G. S. Spicer, "Cophylogeny of quill mites from the genus syringophilopsis (acari: Syringophilidae) and their north american passerine hosts," *The Journal of parasitology*, vol. 99, no. 5, pp. 827–834, 2013.

[29] P. M. Simões, "Diversity and dynamics of Wolbachia-host associations in arthropods from the Society archipelago, French Polynesia," Thesis, Univ. Lyon I, 2012. [Online]. Available: https://tel.archives-ouvertes.fr/tel-00850707

[30] P. Simões, G. Mialdea, D. Reiss, M.-F. Sagot, and S. Charlat, "Wolbachia detection: an assessment of standard PCR protocols," *Mol. Ecol. Resour.*, vol. 11, no. 3, pp. 567–572, 2011.

[31] M. Escudero, "Phylogenetic congruence of parasitic smut fungi (anthracoidea, anthracoideaceae) and their host plants (carex, cyperaceae): Cospeciation or host-shift speciation?" *American journal of botany*, vol. 102, no. 7, pp. 1108–1114, 2015.

[32] M. Mendlová, Y. Desdevises, K. Civáňová, A. Pariselle, and A. Šimková, "Monogeneans of west african cichlid fish: evolution and cophylogenetic interactions," *PLoS One*, vol. 7, no. 5, p. e37268, 2012.

[33] M. P. M. Vanhove, A. Pariselle, M. Van Steenberge, J. A. M. Raeymaekers, P. I. Habltzel, C. Gillardin, B. Hellemans, F. C. Breman, S. Koblmller, C. Sturmbauer, J. Snoeks, F. A. M. Volckaert, and T. Huyse, "Hidden biodiversity in an ancient lake: phylogenetic congruence between lake tanganyika tropheine cichlids and their monogenean flatworm parasites," *Scientific Reports*, vol. 5, 2015.

[34] J. Deng, F. Yu, H.-B. Li, M. Gebiola, Y. Desdevises, S.-A. Wu, and Y.-Z. Zhang, "Cophylogenetic relationships between anicetus parasitoids (hymenoptera: Encyrtidae) and their scale insect hosts

(hemiptera: Coccidae)," *BMC Evol. Biol.*, vol. 13, no. 1, pp. 1–11, 2013.

[35] M. J. McLeish and S. V. Noort, "Codivergence and multiple host species use by fig wasp populations of the ficus pollination mutualism," *BMC evolutionary biology*, vol. 12, no. 1, p. 1, 2012.

[36] I. B.-C. Juan A. Balbuena, Raúl Míguez-Lozano, "Paco: A novel procrustes application to cophylogenetic analysis," *PLoS ONE*, vol. 8, no. 4, 2013.

[37] E. A. Murray, A. E. Carmichael, and J. M. Heraty, "Ancient host shifts followed by host conservatism in a group of ant parasitoids," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 280, no. 1759, p. 20130495, 2013.

[38] N. S. Hafner M, "Phylogenetic trees support the coevolution of parasites and their hosts," *Nature*, vol. 332, p. 258259, 1988.

[39] A. Martínez-Aquino, F. S. Ceccarelli, L. E. Eguiarte, E. Vázquez-Domínguez, and G. P.-P. de León, "Do the historical biogeography and evolutionary history of the digenean margotrema spp. across central mexico mirror those of their freshwater fish hosts (goodeinae)?" *PloS one*, vol. 9, no. 7, p. e101700, 2014.

[40] M. Rosenblueth, L. Sayavedra, H. Sámano-Sánchez, A. Roth, and E. Martínez-Romero, "Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (hemiptera: Coccoidea)," *J. Evol. Biol.*, vol. 25, pp. 2357–2368, 2012.

[41] K. Kellner, H. Fernández-Marín, H. Ishak, R. Sen, T. Linksvaye, and U. Mueller, "Co-evolutionary patterns and diversification of ant–fungus associations in the asexual fungus-farming ant mycocepurus smithii in panama," *Journal of evolutionary biology*, vol. 26, no. 6, pp. 1353–1362, 2013.

[42] G. Jansen, K. Vepsäläinen, and R. Savolainen, "A phylogenetic test of the parasite-host associations between maculinea butterflies(lepidoptera: Lycaenidae) and myrmica ants(hymenoptera: Formicidae)," *Eur. J. Entomol.*, vol. 108, no. 1, pp. 53–62, 2011.

[43] J. Hughes, M. Kennedy, K. P. Johnson, R. L. Palma, and R. D. Page, "Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and pectinopygus lice," *Syst. Biol.*, vol. 56, no. 2, pp. 232–251, 2007.

[44] J. Hugot, "Primates and their pinworm parasites: the cameron hypothesis revisited," *Syst. Biol.*, vol. 48, no. 3, p. 523546, 1999.

[45] C. Ramsden, E. C. Holmes, and M. A. Charleston, "Hantavirus evolution in relation to its rodent and insectivore hosts: No evidence for codivergence," *Mol. Biol. Evol.*, vol. 26, no. 1, pp. 143–153, 2009.

[46] J. D. Weckstein, "Biogeography explains cophylogenetic patterns in toucan chewing lice," *Systematic Biology*, vol. 53, no. 1, pp. 154–164, 2004.

[47] J. Hugot, "New evidence for hystricognath rodent monophyly from the phylogeny of their pinworms," in *Tangled trees: Phylogeny, cospeciation, and coevolution*, R. D. M. Page, Ed. Univ. Chicago Press, 2003, pp. 144–173.

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

[49] A. M. Paterson, R. L. Palma, and R. D. Gray, "Drowning on arrival, missing the boat, and x-events: how likely are sorting events?" in *Tangled trees: Phylogeny, cospeciation, and coevolution*, R. D. M. Page, Ed. Univ. Chicago Press, 2003, pp. 287–309.

[50] G. Refregier, M. Le Gac, F. Jabbour, A. Widmer, J. Shykoff, R. Yockteng, M. Hood, and T. Giraud, "Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: Prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation," *BMC Evol. Biol.*, vol. 8, no. 1, p. 100, 2008.

[51] A. P. Jackson and M. Charleston, "A cophylogenetic perspective of rna–virus evolution," *Molecular biology and evolution*, vol. 21, no. 1, pp. 45–57, 2004.

[52] A. Paterson and R. Poulin, "Have chondracanthid copepods co-speciated with their teleost hosts?" *Systematic Parasitology*, vol. 44, no. 2, pp. 79–85, 1999.

[53] E. Viale, I. Martinez-Sanudo, J. Brown, M. Simonato, V. Girolami, A. Squartini, A. Bressan, M. Faccol, and L. Mazzon, "Pattern of association between endemic hawaiian fruit flies (diptera, tephritidae) and their symbiotic bacteria: Evidence of cospeciation events and proposal of candidatus stammerula trupaneae," *Molecular phylogenetics and evolution*, vol. 90, pp. 67–79, 2015.

[54] P. Górecki, O. Eulenstein, and J. Tiuryn, "Unrooted tree reconciliation: A unified approach," *IEEE/ACM Trans. Comput. Biology Bioinf.*, vol. 10, no. 2, pp. 522–536, 2013.

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