Cophylogeny Reconstruction Allowing for Multiple Associations Through Approximate Bayesian Computation

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

Cophylogeny Reconstruction Allowing for Multiple Associations Through Approximate Bayesian Computation

Blerina Sinaimeri[1,2], Laura Urbini[2*], Marie-France Sagot[2] and Catherine Matias[3]

[1] *LUISS University, Rome, Italy*

[2] *Inria Lyon, 56 Bd Niels Bohr, 69100 Villeurbanne, France, and Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558; 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France*

[3] *Sorbonne Université, Université de Paris Cité, Centre National de la Recherche Scientifique, Laboratoire de Probabilités, Statistique et Modélisation, Paris, France*

Corresponding author: Blerina Sinaimeri, LUISS University, Rome, Italy; E-mail: bsinaimeri@luiss.it.

Abstract

Nowadays, the most used method in studies of the coevolution of hosts and symbionts is phylogenetic tree reconciliation. A crucial issue in this method is that from a biological point of view, reasonable cost values for an event-based parsimonious reconciliation are not easily chosen. Different approaches have been developed (Alcala et al., 2017; Baudet et al., 2015; Szöllősi et al., 2013) to infer such cost values for a given pair of host and symbiont trees. However, a major limitation of these approaches is their inability to model the *invasion* of different host species by a same symbiont species (referred to as a spread event), which is thought to happen in symbiotic relations. To mention one example, the same species of insects may pollinate different species of plants. This results in multiple associations observed between the symbionts and their hosts (meaning that a symbiont is no longer specific to a host), that are not compatible with the current methods of coevolution.

In this paper, we propose a method, called AmoCoala (a more realistic version of a previous tool called Coala) which for a given pair

*First co-authors.

of host and symbiont trees, estimates the probabilities of the cophylogeny events, in presence of spread events, relying on an approximate Bayesian computation (ABC) approach.

The algorithm that we propose, by including spread events, enables the multiple associations to be taken into account in a more accurate way, inducing more confidence in the estimated sets of costs and thus in the reconciliation of a given pair of host and symbiont trees. Its rooting in the tool Coala allows it to estimate the probabilities of the events even in the case of large datasets. We evaluate our method on synthetic and real datasets. The software is available at `https://team.inria.fr/erable/en/software/amocoala/`

Keywords: reconciliation, cophylogeny, ABC method, spread.

1 Introduction

A powerful framework for modelling host-symbiont coevolution is provided by cophylogeny, a method which allows to infer combined evolutionary scenarios for a pair of phylogenetic trees of hosts and symbionts. The cophylogeny problem is usually envisioned as a problem of mapping the phylogenetic tree of the symbionts into the one of the hosts (see *e.g.* Charleston, 2003; Merkle and Middendorf, 2005; Page, 1994; Donati et al., 2015). Such mapping, called a *reconciliation*, allows the identification of (up to) four types of biological events: (a) cospeciation, when the symbiont diverges in correspondence to the divergence of a host species; (b) duplication, when the symbiont diverges but not the host; (c) host switch, when a symbiont switches from one host species to another independently of any host divergence; and (d) loss, which can describe for instance speciation of the host species independently of the symbiont, which then follows just one of the new host species.

The reconciliation method is abstract enough that it may actually be applied to different types of data, of which a common one is gene-species associations (Bansal et al., 2012; Doyon et al., 2011; Hallett and Lagergren, 2001; Stolzer et al., 2012; Tofigh et al., 2011). In fact, the trees that are compared do not even need to be representations of phylogenies. For instance in Becerra (1997), the phylogenetic tree of the beetle genus *Blepharida* is compared to a tree of the host plants (genus *Bursera*) whose construction is based on chemical similarity. Such generality may be seen as an advantage since the methods developed for host-symbiont associations (Conow et al.,

2010; Merkle et al., 2010; Baudet et al., 2015; Donati et al., 2015) could be applicable to other situations (such as the gene-species context). However, this also shows that these models do not fully capture the specificity of the host-symbiont context. Among the most important aspects that have been only partially addressed is the fact that a same symbiotic species can interact, and therefore be associated with more than one host species; we refer to this as a *multiple association*. This is in sharp contrast with the gene-species context where a gene (sequence) is naturally associated to only one species (the one it is extracted from) (Stolzer et al., 2012; Bansal et al., 2018).

In host-symbiont systems, a multiple association can result from a combination of biologically different situations. Following Banks and Paterson (2005), such association can indeed be explained by: (i) cryptic symbiont species (that is, different symbiont species that are morphologically indistinguishable); (ii) misclassified (over-split) hosts (if the apparently different host species to which the symbiont is related represent in fact a same single species); (iii) recent host switches (when the symbiont has recently colonised a new host species and in the newly established population, there is very limited genetic diversity compared to the original symbiont population); (iv) failure to speciate by the symbiont population despite the fact that the host diverged (which might happen if the symbiont populations maintain genetic contact despite the host speciation); and (v) incomplete host switching (if a symbiont colonised a sister taxon of its original host, and maintained genetic contact with the source population).

While in the cases (i)-(ii) the multiple associations are due to errors in defining the real input, in the cases (iii)-(v) those are caused by the ability of the symbiont to be associated to more than one host species and hence require the introduction of an additional biological event that has been called *spread* in the literature. The first use of such term seems to be in Brooks and McLennan (1991). Several methods in the literature deal with multiple associations in a more or less ad-hoc way but to the best of our knowledge none of them fully considers spread events. As multiple associations can be caused by spread events, any method that deals with multiple associations without considering spread events is not satisfying. Below, we briefly review the state of the art of cophylogenetic methods that consider multiple associations.

The methods presented in the literature on host and symbiont trees reconciliation can be divided into two main groups: parsimony methods and statistical methods of inference. The first group is based on an optimisation problem where, given a cost for each of the events, an optimal reconciliation

is found by minimising its total cost. An advantage of parsimony methods is that they allow not only to estimate the frequencies of each of the events but also to infer the past associations. However, a major problem with these methods is that the solutions obtained are strongly dependent on the costs that have to be chosen *a priori*. Indeed, costs are inversely proportional to the obtained frequencies: the larger an event cost, the smaller the corresponding frequency of this event. Statistical approaches can then be used in addition to or as an alternative as they remove the subjective step of cost parameter choice and rely instead on a simultaneous inference of parameter values (*i.e.* event probabilities) and events.

To the best of our knowledge, the parsimony-based methods that address multiple associations are the following: CoRe-Pa (Merkle et al., 2010), Jane 4 (Conow et al., 2010) and WiSPA (unpublished, see Drinkwater et al. (2016)). The tool CoRe-Pa (Merkle et al., 2010) deals only with the case of cryptic species and solves the multiple associations locally in a parsimonious way. Jane 4 (Conow et al., 2010) and WiSPA (Drinkwater et al., 2016) consider multiple associations as resulting only from recent host switches (case (iii) above, see Brooks et al. (2004)).

For what concerns the statistical approaches, only Alcala et al. (2017) proposed a method of inference addressing multiple associations. The authors develop an approximate Bayesian computation (ABC) method to infer the rates of only two events: host switch and cospeciation. Their approach is different from the current literature on tree reconciliation in many ways. First, their method relies on symbiont genomic sequences to produce sets of dated phylogenies instead of relying on a single symbiont tree. Moreover, they pre-estimate extinction and speciation rates from the set of reconstructed symbiont phylogenies. As cospeciation occurs independently from the speciation process in their cophylogeny model, one might expect that the symbiont trees obtained with this method exhibit more speciations than expected. Finally, their method outputs only a host-shift rate and a cospeciation probability but no quantification of duplication or loss events. Note that a very recent work addresses multiple associations in host-parasite systems, by modelling host repertoire evolution along the branches of a parasite tree (Braga et al., 2020). However, this method is far from the reconciliation approach and uses the host tree only through host pairwise distances.

In this paper, we introduce spread as a fifth event in the method called Coala (for *COevolution Assessment by a Likelihood-free Approach*) originally proposed in Baudet et al. (2015) which to our knowledge was the first

method to rely on ABC in the context of tree reconciliation. Coala infers a probability for each of the four cophylogeny events: cospeciation, duplication, host switch and loss but requires that the input has no multiple association. Introducing a spread event is a challenge and there is yet no canonical way to do this.

We choose to introduce two kinds of spread events, called vertical and horizontal spreads respectively. In this way, we capture the 2 different situations occurring in the cases (iii)-(v) above. The first event, called vertical spread, corresponds to a spread of a symbiont in the entire subtree below a host species. This event could also be called a *freeze* in the sense that the evolution of the symbiont *freezes* while the symbiont continues to be associated with a host and with the new species that descend from this host. As will be further detailed in Section Model and Method, this event covers case (iv) above and is related to what is known in the literature as *failure to diverge* (see for example Conow et al. (2010)). This also corresponds to the *speciation as a generalist* introduced in Alcala et al. (2017). The second event, called horizontal spread, informally corresponds to the combination of a "host switch" with 2 different vertical spreads, one occurring in the initial host subtree and the second in the new host subtree. Thus, this horizontal spread event includes both an *invasion*, of the symbiont which remains with the initial host but at the same time gets associated with (*invades*) another host that is not a descendant of the first, plus a *freeze*, actually a double freeze as the evolution of the symbiont *freezes* in relation to the evolution of the host to which it was initially associated and to the evolution of the second host it *invaded*. This event is useful to describe the cases (iii) and (v) from above. It allows to explain the case where two host clades that are phylogenetically distant are associated with the same symbiont species. Notice that a fundamental difference between host switch and horizontal spread is that in the former, the symbiont that switches hosts will further create 2 different symbionts, each one associated to the initial and to the new host respectively. In particular, a host switch never induces a multiple association, in sharp contrast with a horizontal spread. Notice also that cases (i) and (ii) above correspond to input errors rather than real biological events. Nonetheless, these situations are dealt with by our model. Indeed, case (i) is considered as a horizontal spread while case (ii) counts as a vertical spread. Our goal here is not to correct for these potential input errors but to provide a comprehensive framework that handles the diversity of biological situations.

In this paper, we propose a method, called AmoCoala, which for a given

pair of host and symbiont trees, first estimates the probabilities of spread events directly from the input and second estimates the probabilities of the remaining four classical cophylogeny events, relying on an ABC approach. In doing so, we also define a new distance to compare two symbiont trees that are associated with the same host tree in presence of multiple associations. Indeed, ABC methods heavily rely on the ability to compare observations with simulated datasets. In the cophylogeny context, this means comparing trees (as these are the most complete information on the data), a task that is far from trivial. Our new distance is an extension of the classical Maximum Agreement Subtree distance (*MAS*) (Ganapathy et al., 2005) to what we call *set-labelled* trees. We believe this new distance can be of independent interest (see Section Model and Method and also the Supplementary Material).

We test AmoCoala on both synthetic and real datasets and compare the results with Coala. We could not compare our approach with the tool Alcala et al. (2017) due to the, previously described, substantial differences both in the model and in the input. Our tests show that AmoCoala produces results that seem closer than those of Coala to what is expected from the judgment of a biological expert.

2 Model and method

2.1 Reconciliations and cophylogeny events

Similarly to Coala, AmoCoala is built on the event-based model presented in Charleston (2002); Tofigh et al. (2011). The *input* of AmoCoala consists of a triple (H, S, ϕ) where H and S correspond to the phylogenetic trees of the hosts and symbionts, respectively, and ϕ is a relation from the leaves of the symbiont tree $L(S)$ to the leaves of the host tree $L(H)$. The relation ϕ describes the existing associations between currently living symbiont species and their hosts. More precisely, ϕ is a function from the set of symbiont leaves to the set of all subsets of host leaves. Notice that a multiple association will correspond to a leaf in the symbiont tree that is associated to more than one leaf in the host tree. In Coala, as well as in all the models that do not allow for multiple associations, the relation ϕ assigns to each $s \in L(S)$ exactly one host leaf in $L(H)$ (notice however that one host can be associated to more than one symbiont). In AmoCoala, this constraint will be dropped and thus we have that each leaf $s \in L(S)$ in the symbiont tree is

associated to $\phi(s)$, a subset of $L(H)$.

A *reconciliation* λ is a function from the vertices of the symbiont species tree to the set of all subsets of vertices of the host tree that is an extension of ϕ, *i.e.* that is the same function as ϕ when restricted to the sets of leaves. In the classical setting, a reconciliation can be associated to a set of cospeciations, duplications, host switches and losses (the four classical cophylogeny events). For more details about the reconciliation model, we refer to Charleston (2002); Tofigh et al. (2011); Stolzer et al. (2012); Donati et al. (2015); Baudet et al. (2015) and our Supplementary Material. In this paper, we extend the classical reconciliation model to include other biological events.

Finally notice that here we focus on models that do not require the host tree to be dated. This is a clear advantage of the method as this information is rarely available and when it is available, is often not reliable. However, as we do not require the host tree to be dated some combinations of host switches can introduce an incompatibility due to the temporal constraints imposed by the host and symbiont trees, as well as by the reconciliation itself. We say that a reconciliation is *time-feasible* if it does not violate the time-feasibility constraints. The exact criterion we use to assess time-feasibility is the one defined in Stolzer et al. (2012) and that was already the one used in Coala.

Spread events. In AmoCoala, we introduce two new additional cophylogeny events: vertical and horizontal spreads. We now define and illustrate both of them.

Vertical Spread. When for a symbiont s, a vertical spread happens at a host h, the evolution of the symbiont s *freezes* in h, *i.e.* s continues to be associated with h and with the new species that descend from h. In the toy example depicted in Figure 1(a), we see that the symbionts s_1, s_2 are both related to all the hosts h_3, h_4, h_5. One possible explanation is that the symbiont s_5 (the most recent common ancestor of s_1, s_2) was present in all the clade of h_8 (which is the most recent common ancestor of h_3, h_4, h_5). In that case, we say that h_8 is the ancestral host of s_5 and the two clades S_{s_5} (which denotes the symbiont clade rooted in s_5) and H_{h_8} (the host clade rooted in h_8) are related. We say that a vertical spread has happened at symbiont s_5 and we associate s_5 with all the vertices in the subtree rooted in h_8 (see Figure 1(b)).

Horizontal Spread. In some datasets, we see the occurrence of a same

symbiont in two different clades of the host tree. Such a situation cannot occur when relying only on cospeciation, duplication, loss, host switch or vertical spread events. Indeed, as already underlined, the 4 initial events never produce multiple associations, while the vertical spread produces them only within clades. For this reason, we introduce a horizontal spread event. In the horizontal spread event, the symbiont remains with the initial host but at the same time gets associated with (*invades*) another host incomparable with the first, and undergoes a freeze, actually a double freeze as the evolution of the symbiont *freezes* in relation to the evolution of the host to which it was initially associated and in relation to the evolution of the second one it *invaded*. For illustrative purposes only, we show in Figure 2 an example of a reconciliation involving a horizontal spread event. The horizontal spread event happens in vertex s_5 as it is associated to two subtrees of the host tree, rooted in h_6 and h_7, respectively.

2.2 General framework of AmoCoala

The method we propose is based on the approximate Bayesian computation (ABC) method that was already used in Coala (Baudet et al., 2015). We briefly recall it here for the sake of completeness. ABC methods belong to a family of likelihood-free Bayesian inference algorithms that attempt to estimate posterior densities for problems where the likelihood is unknown or may not be easily computed. ABC only requires that simulations under the statistical model at stake are possible. We recall that the likelihood function expresses the probability of the observed data under a particular statistical model. More specifically, given a set of observed data D_0 (in our case the input (H, S, ϕ)) and starting with a prior distribution π on the space of the parameters of the model (here, the probabilities of the four classical cophylogeny events), the objective is to estimate the parameter values θ that could lead to the observed data using a Bayesian framework. Formally, we are interested in the posterior distribution $p(\theta|D_0) = p(D_0|\theta)\pi(\theta)/p(D_0)$.

For simple models, the likelihood function $p(D_0|\theta)$ can typically be derived. However, for more complex models the likelihood function might be computationally very costly to evaluate. In these cases, ABC methods approximate the posterior distribution by simulations, the outcomes of which are compared with the observed data. First, a population of N parameter values θ_i is sampled from the prior distribution. Then, for each sampled parameter θ, a dataset $\tilde{D}_\theta$ is simulated. It consists in a simulated symbiont tree

$\tilde{S}_\theta$ together with a reconciliation $\tilde{\lambda}$ from $\tilde{S}_\theta$ to H. This dataset $\tilde{D}_\theta$ is then compared with the real dataset D_0 through a discrepancy measure which is used as a *quality measure* to accept or reject the candidate parameter value θ. In many cases when it is believed that the prior and posterior densities are very different, the acceptance rate is very low. To deal with that issue, we can rely more specifically on a likelihood-free Sequential Monte Carlo (SMC) search that involves many iterations of the simulation procedure, each iteration targeting more precisely good candidate parameter values.

Given an input dataset (H, S, ϕ), an ABC-SMC method was developed in Coala (Baudet et al., 2015) to infer the posterior density of the probability of each of the four classical events, namely cospeciation, duplication, host switch and loss. Coala includes two main parts. The first consists in a simulation algorithm of the coevolutionary history of symbionts and their hosts. More specifically, given the host tree H and a vector $\theta = \langle p_c, p_d, p_s, p_l \rangle$ specifying the probability of each of the classical cophylogeny events, the model generates a symbiont tree $\tilde{S}_\theta$ together with a reconciliation from $\tilde{S}_\theta$ to H describing the ancient host-symbiont associations. In AmoCoala, this first part is improved by introducing spread events whose probabilities of occurrence are fixed throughout all the simulations, while being heterogeneous along the host tree and specific to the original dataset. The second part concerns a method to select the most likely probability vectors based on an ABC-SMC variant. It relies on the main idea that the most likely vectors θ will generate trees $\tilde{S}_\theta$ together with reconciliations $\tilde{\lambda}$ from $\tilde{S}_\theta$ to H that are similar to the original input (H, S, ϕ).

In Coala, the symbiont trees together with their leaf associations were summarised through labelled trees and this step thus relied on a phylogenetic distance between labelled trees. In AmoCoala, this part is improved by the introduction of a new distance that accounts for the possibility of multiple associations between $\tilde{S}_\theta$ and H. Indeed, the symbiont trees together with their leaves association may now be summarised through set-labelled trees (*i.e.* trees with leaves labelled by subsets of $L(H)$). We thus provide and rely here on a new phylogenetic distance between set-labelled trees. To the best of our knowledge, distances between set-labelled trees have not been considered in the literature and our proposal for such may be of independent interest.

In a nutshell, to deal with multiple associations coming from spread events, we thus extend Coala as follows: (i) we first propose estimators of vertical and horizontal spread probabilities given the input (H, S, ϕ); (ii)

we introduce a new method to simulate the cophylogeny of the symbiont tree, along the host tree and given a candidate probability for each of the 4 classical cophylogeny events (cospeciation, duplication, host switch and loss) which also takes into account the probabilities of vertical and horizontal spread; (iii) we introduce a new distance to compare the simulated to the real symbiont trees in presence of multiple associations to the host tree.

2.3 Estimation of the probabilities of the events

In AmoCoala, the probabilities $\langle p_c, p_d, p_s, p_l \rangle$ of the classical cophylogeny events (cospeciation, duplication, host switch and loss) are parameters inferred relying on the ABC-SMC approach, namely they are first sampled from a prior distribution and then later selected according to some criteria that are specified later. **On the contrary, the probabilities p_{vs} and p_{hs} for the (vertical and horizontal) spread events are not estimated within the ABC-SMC method but rather in a preliminary step, directly from the input**. This choice is mainly driven by the fact that in a realistic model the spread probabilities are not constant throughout the host tree. For instance, a spread event appearing near to the root is less likely to happen than one close to the leaves. Indeed, spread events were introduced partly to account for recent host switches (see point (iii) in the introduction) and more generally they are motivated by the fact that symbionts may not diversify immediately, which happens less likely close to the root. Then, as the probability of a spread event is specific to each vertex of the host tree, sampling the spread events will increase significantly the size of the parameter space and thus the size of the space of the generated symbiont trees. Hence, in this framework the spread probabilities cannot be inferred in the procedure. Nevertheless, these probabilities are clearly related to the shape of the host and symbiont trees and to the associations between their leaves. For this reason, we exploit the signal from the input to calculate the probabilities of the spread events. These probabilities are used in the generation of the putative symbiont trees and are not parameters to be inferred through the ABC-SMC method. Details are given in the Supplementary Material.

2.4 Simulation of a symbiont tree in AmoCoala

We now describe the procedure of generation of simulated symbiont trees in AmoCoala. Similarly to Coala, our algorithm takes as input (H, S, ϕ)

and the probabilities of each of the events, and simulates the evolution of the symbionts by following the evolution of the hosts, *i.e.* by traversing H from the root to the leaves, and progressively constructing the phylogenetic tree $\tilde{S}$ for the symbionts and at the same time mapping them to subsets of vertices of the host tree, *i.e.* constructing $\tilde{\lambda}$. In this process, a symbiont vertex can be in two different states: mapped or unmapped. At the moment of its creation, a new vertex $\tilde{s}$ is unmapped and is assigned a temporary position on an arc a of the host tree H. We denote this situation by $\langle \tilde{s} : a \rangle$. We let $h(a)$ denote the head of the arc a (*i.e.* the vertex at the endpoint of a that is farthest from the root). Then vertex $\tilde{s}$ is mapped to either vertex $h(a)$ of H (*i.e.* $\tilde{\lambda}(\tilde{s}) = \{h(a)\}$ for cospeciation, duplication and host switch) or to a subset $\mathcal{H}$ of vertices of H (*i.e.* $\tilde{\lambda}(\tilde{s}) = \mathcal{H}$ for vertical and horizontal spread). Notice that for the vertical spread, the subset of vertices $\mathcal{H}$ corresponds to a clade in H, while for the horizontal spread it corresponds to the union of two clades in H.

In the cases of cospeciation, duplication, and host switch, a speciation has occurred in the symbiont tree and hence two children are created for $\tilde{s}$, denoted by $\tilde{s}_1$ and $\tilde{s}_2$. Their positioning along the arcs of the host then depends on which of the three events took place. In the case of a loss, no child for $\tilde{s}$ is created (at this step) since there is no symbiont speciation, and $\tilde{s}$ is just moved to one of the two arcs outgoing from $h(a)$ chosen randomly.

The case of a spread event is different. Consider for instance the example in Figure 3. A vertical spread occurs at the symbiont s_6 on the host h_8 and thus s_6 is associated to all the subtree H_{h_8} (the host clade rooted in h_8). Moreover, we choose that all the symbionts descendent from s_6 are associated to the same clade as s_6 (see Definition 2, part 3c in the Supplementary Material). We now need to choose a realistic way of continuing the simulation of the symbiont subtree below s_6. We call the subtree of the symbiont tree rooted at a vertex associated to a spread event (vertical or horizontal) a *ghost subtree*. In Figure 3, the subtree S_{s_6} is a ghost subtree. Then during the generation of the symbiont tree $\tilde{S}$ when a symbiont $\tilde{s}$ undergoes a spread event, we need to simulate the ghost subtree rooted in $\tilde{s}$ up to its leaves, in order to end up the simulation in this part of the tree. After a spread event, with the passing of time, both the host and the symbiont have evolved and in addition, it could be that some hosts have lost some of their symbionts. Taking into account all the possible evolutions of the symbiont is computationally unfeasible in practice. Therefore, for computational reasons, we decide to promote the simplest situation. In particular, no other

event takes place after a spread event and we mimic in this part of the simulated symbiont tree the evolution occurring in the real symbiont tree. To this purpose, we choose a topology and leaf associations that are identical to those present in S. More formally, if a vertical spread occurs at $\tilde{s}$ on the host h, we consider the set of host leaves descendent from h, namely $L = L(H_h)$. Let L' be the set of symbiont leaves that are associated to the leaves in L, *i.e.* $L' = \phi^{-1}(L) \cap L(S)$. The ghost subtree $\tilde{S}_{\tilde{s}}$ is then set equal to $S_{|L'}$, the subtree of the *real* symbiont tree induced by L'. The case of horizontal spread is analogous, except that the set of leaves L is given by the union of $L(H_h)$ and $L(H_{h'})$ where h, h' are the two host vertices involved in the horizontal spread. Once the ghost tree is set, the simulation ends in this part of the tree. Notice that as already mentioned, the spread events are more likely to occur far from the root, so that the loss of variability in the simulated tree $\tilde{S}$ induced by this choice is counterbalanced by the fact that it should affect a small part of the tree. More details are given in the Supplementary Material.

2.5 ABC-SMC inference method

AmoCoala is based on the same ABC-SMC method presented in Coala. It is an iterative method with many rounds, and it involves a summary distance that describes the quality of any candidate vector θ (*i.e.* how much it is susceptible to have generated the observed dataset). For the sake of completeness, we include the details of the method in the Supplementary Material and report here only the differences. In particular, the main difference consists in the summary distance measure used to quantify the quality of the vectors θ. The discrepancy between the simulated dataset (the generated symbiont trees and their host associations) and the observed one (the real symbiont tree and its host associations) is measured through a distance between phylogenetic trees which can be calculated in polynomial time. Similarly as in Coala, this discrepancy is constructed from two components: (i) d_1, that describes how much the simulated tree $\tilde{S}_\theta$ is representative of the vector θ, and ii) d_2 that measures how much is $\tilde{S}_\theta$ (and its labels) topologically similar to S (and its labels). The value of d_1 is computed identically as in Coala. As concerns point (ii), the metric used here is different from the one used in Coala and we detail its definition and motivation in the next paragraph.

2.6 A distance between set-labelled trees

In AmoCoala, the leaves of both the observed and the simulated symbiont trees (S and $\tilde{S}$ respectively) are labelled by the host leaves to which they are associated. Thus, due to possible multiple associations in AmoCoala, those symbiont trees are what we call *set-labelled* trees, that is, their leaves are labelled with sets and not with singletons. To the best of our knowledge, distances for set-labelled trees have not been considered in the literature and we believe our proposal for such is thus of independent interest.

We first recall that the MAST distance of two phylogenetic trees T_1 and T_2 corresponds to the number of leaves in the largest isomorphic subtree that is common to the two trees (subtrees common to the two trees are called agreement subtrees and we look for the one with the largest number of leaves). Clearly this isomorphism takes into account the labels of the trees. The MAST distance can be calculated in $O(n^2)$ time where n is the size of the largest input tree (Ganapathy et al., 2006). For set-labelled trees, we need to take into account the sizes of the sets of labels in the possible agreement subtrees.

Thus, given a set-labelled tree T, we denote its *weight* by $w(T) = \sum_{v \in L(T)} |l(v)|$, where $l(v)$ is the set of labels associated to the leaf v. Now, a *maximum agreement set-labelled subtree*, denoted by $MAS(T_1, T_2)$, is a set-labelled subtree that is common to the two trees T_1, T_2 and which has largest weight. Notice that the MAS of two trees does not necessarily have the maximum number of leaves among the set-labelled agreement subtrees, as shown in Figure 4. In the same way as the MAST distance is defined, we introduce the *maximum agreement set* distance, denoted by d_{MAS}, between two set-labelled phylogenetic trees T_1,T_2 as well as a normalized related quantity d_2, respectively defined as

$$d_{MAS}(T_1, T_2) = \max\{w(T_1), w(T_2)\} - w(MAS(T_1, T_2))$$
$$d_2(T_1, T_2) = \frac{d_{MAS}(T_1, T_2)}{\max\{w(T_1), w(T_2)\}} = 1 - \frac{w(MAS(T_1, T_2))}{\max\{w(T_1), w(T_2)\}}.$$

We can prove that d_{MAS} is a metric and that it can be calculated in polynomial time using a dynamic programming algorithm. Note that the normalized quantity d_2 has the advantage of lying in $[0, 1]$ and is computed with the same complexity as d_{MAS}. It is only a pseudo-distance (as it does not satisfy the triangular inequality). However, the resulting $d(\theta)$ defined relying on d_2 will,

with some abuse of notation, be called a *summary distance* (see details in the Supplementary Material).

3 Experimental results and discussion

To evaluate our tool, we used both synthetic and real datasets. The synthetic datasets were generated in a similar way as in Baudet et al. (2015). The results are much the same to the ones of Coala and do not bring any new insights. These are a proof of concept of AmoCoala and can be found in the Supplementary Material.

3.1 Biological datasets

To test our method, we selected 4 biological datasets from the literature. The choice of these datasets was dictated by: (1) the availability of the data in public databases, (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. The phylogenetic trees of each dataset can be found in the Supplementary Material. As already mentioned, any dataset D containing multiple associations cannot be analysed with Coala. Thus, in order to compare the results with those obtained by Coala (Baudet et al., 2015), for each real dataset D we generated a dataset D_{Coala} which is obtained from D by randomly choosing exactly one association (among existing ones and whenever there are more than 2 such associations) for each symbiont leaf. Notice that this is what is usually done in the literature when analysing such datasets with a method that does not allow for multiple associations. We detail here the results obtained for only two datasets, the reader can find the remaining ones in the Supplementary Material.

Dataset 1*: AP - Acacia & Pseudomyrmex.* This dataset was extracted from Gómez-Acevedo et al. (2010) and displays the interaction between *Acacia* plants and *Pseudomyrmex*, a genus of ants. Although the authors did not use a cophylogeny reconstruction tool to analyse the dataset, this is considered as a typical example of mutualism between ants and plants, and the authors show that their relationship originated in Mesoamerica between the late Miocene to the middle Pliocene, with eventual diversification of both groups in Mexico. The host and symbiont trees include 9 and 7 leaves, respectively. The dataset has 22 multiple-associations. The corresponding

dataset with no multiple association is called AP_{Coala}.

Dataset 2*: SFC - Smut Fungi & Caryophillaceus plants.* This dataset was extracted from Refrégier et al. (2008). The host and symbiont trees include 15 and 16 leaves, respectively. The dataset has 4 multiple associations. The corresponding dataset with no multiple association is called SFC_{Coala}. Notice that this is the same dataset used in Baudet et al. (2015).

In Figures 5 and 6, we present for each of the cophylogeny events, the distribution of the inferred probabilities obtained by running AmoCoala and Coala. First notice that the results change substantially when we consider the complete dataset instead of the one obtained by removing the multiple associations. Indeed, from the graphics in the third row of Figure 5, we see that if we ignore multiple associations, then Coala explains the dataset using a very low cospeciation frequency and a high number of switches and losses. In general, we can say that Coala detects a high incongruence between the trees which cannot be explained by cospeciations. However, if the complete dataset is considered, *i.e.* the one including all the multiple associations, we see from the first two rows of Figure 5 that the dataset can be explained by only 2-3 horizontal spreads, a high number of cospeciations, a very low number of duplications and switches and also a significantly lower number of losses. Thus, the incongruence between the two phylogenetic trees can be explained by approximately 3 horizontal spreads and then most of the events correspond to cospeciations, which is an indication of coevolution. This is in accordance with what is expected for this dataset, which, as already mentioned in the previous paragraph, is considered as a typical example of mutualism between ants and plants.

Next, we considered the dataset SFC with multiple associations proposed in Refrégier et al. (2008). From Figure 6, we can see that both methods show similar results concerning cospeciations, duplications and host switches while AmoCoala outputs a smaller number of losses (less then 25%) compared to Coala (less then 40%). In Refrégier et al. (2008), the different analyses performed indicated that the most plausible reconciliations presented for the SFC dataset have from 0 to 3 cospeciations, no duplication, 12 to 15 host switches and 0 to 2 losses. It is impossible for us to calculate the number of events in a parsimony framework because there is no parsimonious algorithm for computing optimal reconciliations in the presence of vertical and horizontal spreads. Nonetheless, we have access to estimated frequencies of the reconstructed events. Moreover, from the definition of the model (see the Supplementary Material) we know that the sum of the classical events

(cospeciation, duplication and host switch), excluding the loss event, is equal to the number of internal vertices of the symbiont tree. The symbiont tree (that is the same for SFC and SFC_{Coala}) has 15 internal vertices. Based on the analyses presented in Refrégier et al. (2008), we expect to have events with the following frequencies: between 0% and 20% for cospeciations (from 0 to 3 events), 0% for duplications (no duplications), between 80% and 100% for host switches (from 12 to 15 events) and between 0% and 13% for losses (from 0 to 2 events). To compare the results output by the two methods (Coala and AmoCoala) with those expected from the analyses of Refrégier et al. (2008), we cluster the parameter vectors output by the methods. Indeed, both Coala and AmoCoala perform a hierarchical clustering procedure to group the final list of accepted parameter vectors. We then compared the cluster patterns found by the two methods. Table 1 shows the representative vectors of each of the clusters output by AmoCoala (for the SFC dataset) and by Coala (for the SFC_{Coala} dataset). Notice that as already mentioned in Baudet et al. (2015), a vector with a high frequency of host switches can generate a large space of simulated trees, many of which can have a high distance from the real symbiont tree. Thus, it is clear that such vectors are more difficult to be output by both Coala and AmoCoala.

From the results in Table 1, we have that the event vector that is most similar to the expected one according to Refrégier et al. (2008) is Cluster 2 for AmoCoala run on the SFC dataset (22.4% for cospeciations, 0.4% for duplications, 54.3% for host switches and 22.8% for losses). It is also important to note that the number of vectors that are part of this cluster is high (12 out of 50 vectors accepted in the third round). Notice that Cluster 5 of Coala run on SFC_{Coala} is also close to these values, however this cluster is supported by only 2 of the accepted vectors. Moreover, all the representative vectors of the clusters output by AmoCoala have a frequency of duplication close to 0, which is in agreement with what is expected from Refrégier et al. (2008).

Overall the results obtained with AmoCoala are closer to the result presented in Refrégier et al. (2008) than those that were obtained by Coala which ignores such multiple associations. This shows again the importance of taking into account the latter.

4 Concluding comments

In this paper, we propose a method, called AmoCoala, which for a given pair of host and symbiont trees, estimates the probabilities of the co-phylogeny events, in presence of spread events, relying on an approximate Bayesian computation (ABC) approach. In AmoCoala, it is possible to estimate the probabilities of the classical cophylogeny events (cospeciation, duplication, host switch and loss) and also the probabilities of horizontal and vertical spreads. These two latter events allow to study datasets that contain multiple associations. The model uses set-labelled trees and to compare them we introduced a new distance, called MAS, which we believe can be of independent interest.

AmoCoala on one hand provides more confidence in the set of costs to be used for the reconciliation of a given pair of host and symbiont trees, while on the other hand, it allows to estimate the probabilities of the events even in the case of large datasets. We evaluate our method on synthetic and real datasets.

This work leads to different research directions. First, it would be interesting to define better distances for set-labelled trees. To the best of our knowledge, these types of trees have not been considered in the literature and it would be interesting to generalise (if possible) some of the well-known phylogenetic distances to set-labelled trees. Another direction is to include the vertical and horizontal spreads in the parsimonious reconciliation framework. Thus, a perspective to this work is to design a reconciliation procedure that includes these switches.

5 Software and Supplementary Material

The software is available at `https://team.inria.fr/erable/en/software/amocoala/`. Supplementary Material is available at `http://team.inria.fr/erable/files/2022/05/AmoCoala_Supplementary.pdf`.

References

N. Alcala, T. Jenkins, P. Christe, and S. Vuilleumier. Host shift and cospeciation rate estimation from co-phylogenies. *Ecology Letters*, 20:1014–1024, 2017.

J. C. Banks and A. M. Paterson. Multi-host parasite species in cophylogenetic studies. *International Journal for Parasitology*, 35(7):741 – 746, 2005.

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

M. S. Bansal, M. Kellis, M. Kordi, and S. Kundu. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216, 2018.

C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C. Gautier, C. Matias, and M.-F. Sagot. Cophylogeny reconstruction via an Approximate Bayesian Computation. *Systematic Biology*, 64(3):416–31, 2015.

J. X. Becerra. Insects on plants: Macroevolutionary chemical trends in host use. *Science*, 276(5310):253–256, 1997.

M. P. Braga, M. J. Landis, S. Nylin, N. Janz, and F. Ronquist. Bayesian inference of ancestral host-parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic biology*, 69(6):1149–1162, 2020.

D. Brooks, A. P. G. Dowling, M. van Veller, and E. P. Hoberg. Ending a decade of deception: a valiant failure, a not-so-valiant failure and a success story. *Cladistics*, 20:32–46, 2004.

D. R. Brooks and D. A. McLennan. *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology*. University of Chicago press, 1991.

M. A. Charleston. *Biological Evolution and Statistical Physics*, volume 585 of *Lecture Notes in Physics*, chapter Principles of cophylogenetic maps, pages 122–147. Springer Berlin Heidelberg, 2002.

M. A. Charleston. Recent results in cophylogeny mapping. *Advances in Parasitology*, 54:303–330, December 2003.

C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. Jane: A new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(16):10 pages, 2010.

B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, and M. Sagot. EUCALYPT: efficient tree reconciliation enumerator. *Algorithms for Molecular Biology*, 10(1):3, 2015. URL `http://www.almob.org/content/10/1/3`.

J.-P. Doyon, S. Hamel, and C. Chauve. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):26–39, April 2011.

B. Drinkwater, A. Qiao, and M. A. Charleston. WiSPA: A new approach for dealing with widespread parasitism, 2016. URL `https://arxiv.org/abs/1603.09415`. arXiv:1603.09415.

G. Ganapathy, B. Goodson, R. Jansen, V. Ramachandran, and T. Warnow. Pattern Identification in Biogeography. In R. Casadio and G. Myers, editors, *Algorithms in Bioinformatics*, volume 3692 of *Lecture Notes in Computer Science*, pages 116–127. Springer Berlin Heidelberg, 2005.

G. Ganapathy, B. Goodson, R. Jansen, H. Le, V. Ramachandran, and T. Warnow. Pattern identification in biogeography. *IEEE/ACM Trans. on Comput. Biol. Bioinf.*, 3(4):334–346, 2006.

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*, 56(1):393–408, 2010.

M. T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. *In Lengauer, T. (ed), Proceedings of the fifth Annual International Conference on Research in Computational Molecular Biology (RECOMB), ACM (New York)*, pages 149–156, 2001.

D. Merkle and M. Middendorf. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences*, 123:277–299, 2005.

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

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

G. Refrégier, M. Le Gac, F. Jabbour, A. Widmer, J. A. Shykoff, R. Yockteng, M. E. 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 Evolutionary Biology*, 8(1):100, 2008.

M. L. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415, 2012. doi: 10.1093/bioinformatics/bts386.

G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.*, 62 (6):901–912, 2013.

A. Tofigh, M. T. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 8(2):517–535, 2011.

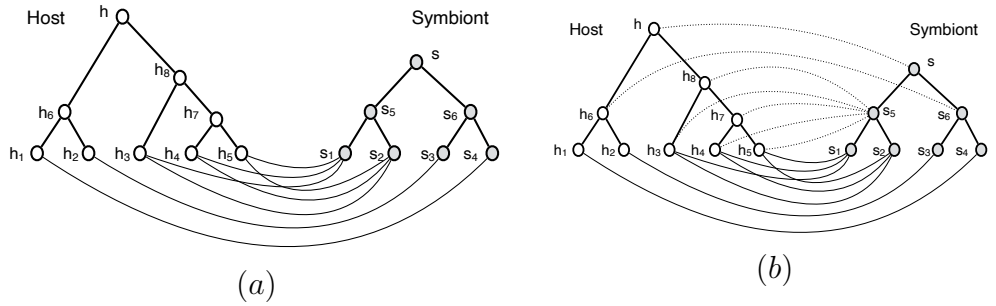

Figure 1: (a) Example of a dataset with multiple associations. The leaf associations are represented by plain lines and given by $\phi(s_1) = \{h_3, h_4, h_5\}; \phi(s_2) = \{h_3, h_4, h_5\}; \phi(s_3) = \{h_2\}; \phi(s_4) = \{h_1\}$. (b) In dotted lines, a reconciliation involving 2 cospeciations in s and s_6 and 1 vertical spread in s_5. More precisely, the reconciliation is given by $\lambda(s) = \{h\}; \lambda(s_6) = \{h_6\}$ and $\lambda(s_5) = \{h_3, h_4, h_5, h_7, h_8\}$ (on the symbiont leaves, we have $\lambda = \phi$).

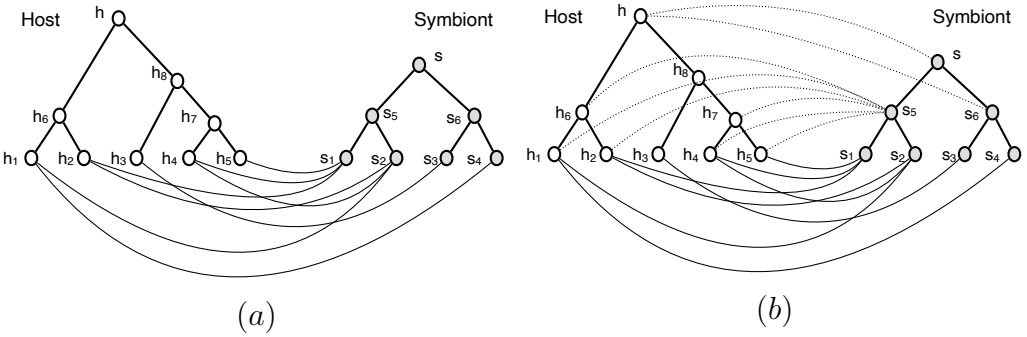

Figure 2: (a) Example of a dataset with multiple associations. The leaf associations are represented by plain lines and given by $\phi(s_1) = \{h_2, h_4, h_5\}; \phi(s_2) = \{h_1, h_2, h_4\}; \phi(s_3) = \{h_3\}; \phi(s_4) = \{h_1\}$. (b) In dotted lines, a reconciliation involving a horizontal spread event is shown. The symbiont s_5 makes a horizontal spread from h_6 to h_7 (or from h_7 to h_6) and thus is associated to the two subtrees H_{h_6} and H_{h_7} (*i.e.* $\lambda(s_5) = H_{h_6} \cup H_{h_7}$). The symbiont s is associated to a duplication (and $\lambda(s) = \{h\}$) and the symbiont s_6 to a cospeciation (and $\lambda(s_6) = \{h\}$).

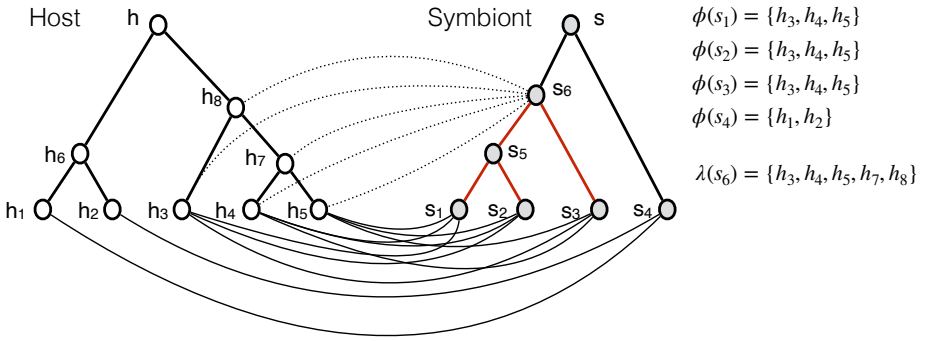

Figure 3: The symbiont s_6 is associated to a vertical spread on the host h_8 and thus is associated to all the subtree H_{h_8}. As we do not know exactly how the symbiont s_5 is associated, we symbolically associate it to all the vertices in H_{h_8}. The subtree of S in red corresponds to a ghost subtree.

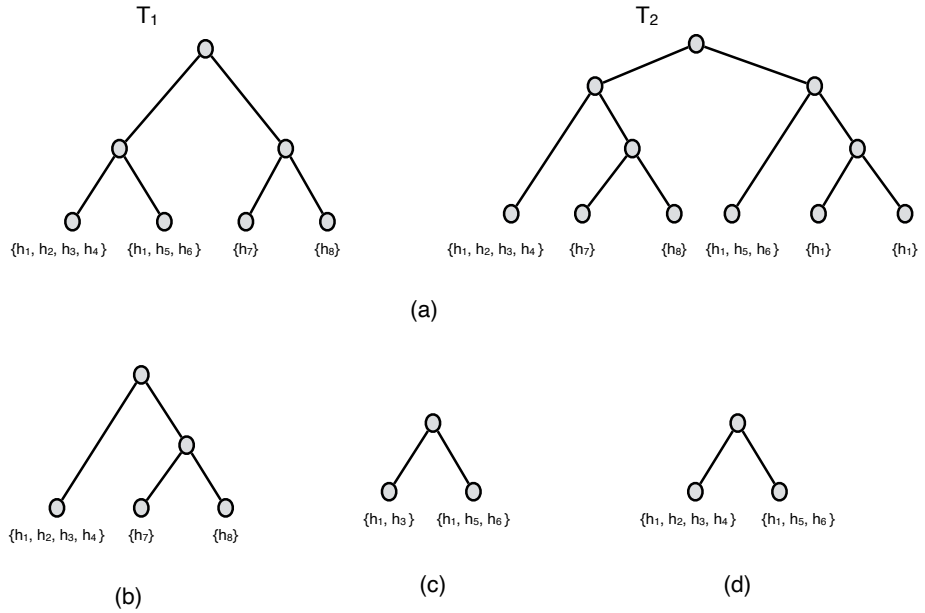

Figure 4: (a) Two set-labelled phylogenetic trees. T_1 has weight 9 and T_2 has weight 11. In (b), (c), (d), three different agreement set-labelled subtrees of weights 6, 5 and 7 respectively. The maximum agreement set-labelled subtree is the one depicted in (d) and notice that it does not have the maximum number of leaves.

Table 1: Representative vectors of the clusters produced by AmoCoala (for the SFC dataset) and by Coala (for the SFC_{Coala} dataset). The column $\#vectors$ indicates the number of vectors in the cluster.

Dataset	*Cluster*	p_c	p_d	p_s	p_l	$\#vectors$
SFC	1	0.531	0.004	0.282	0.183	19
	2	0.226	0.004	0.543	0.228	14
	3	0.898	0.020	0.040	0.042	12
	4	0.859	0.062	0.002	0.077	5
SFC_{Coala}	1	0.437	0.002	0.357	0.204	20
	2	0.417	0.274	0.003	0.306	19
	3	0.850	0.002	0.005	0.144	5
	4	0.005	0.418	0.003	0.575	4
	5	0.144	0.001	0.548	0.308	2

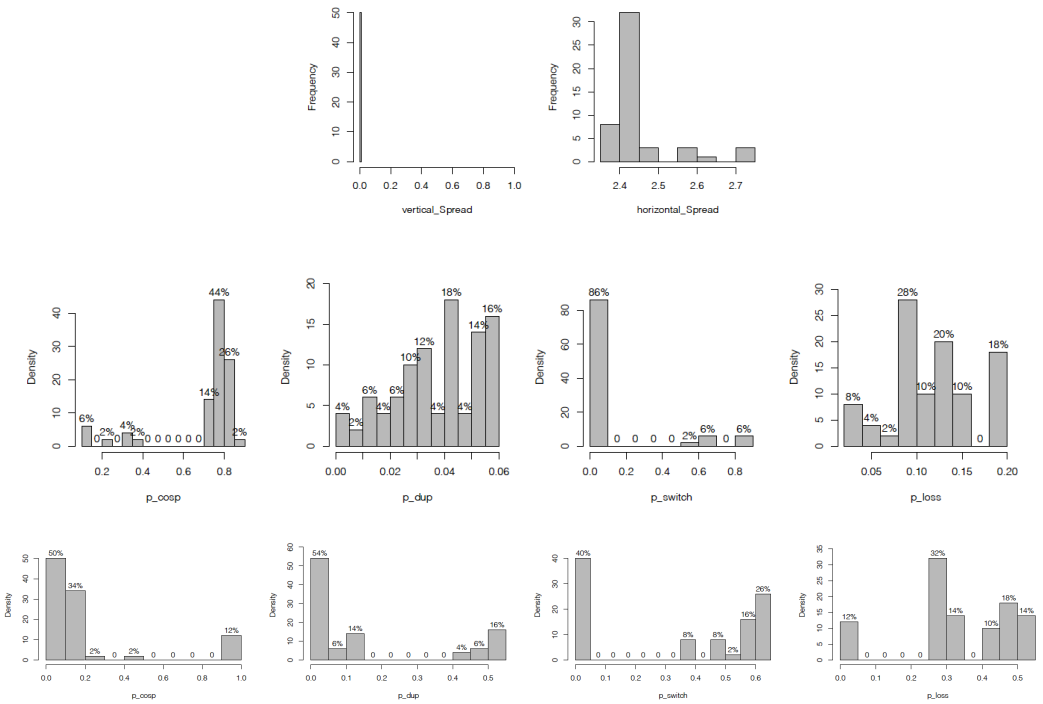

Figure 5: Comparison of the results obtained with AmoCoala and Coala for the dataset AP. In each graphic, we show for each event type, the distribution of the parameter values. In the first two rows, the results provided by AmoCoala and in the third row, the ones provided by Coala.

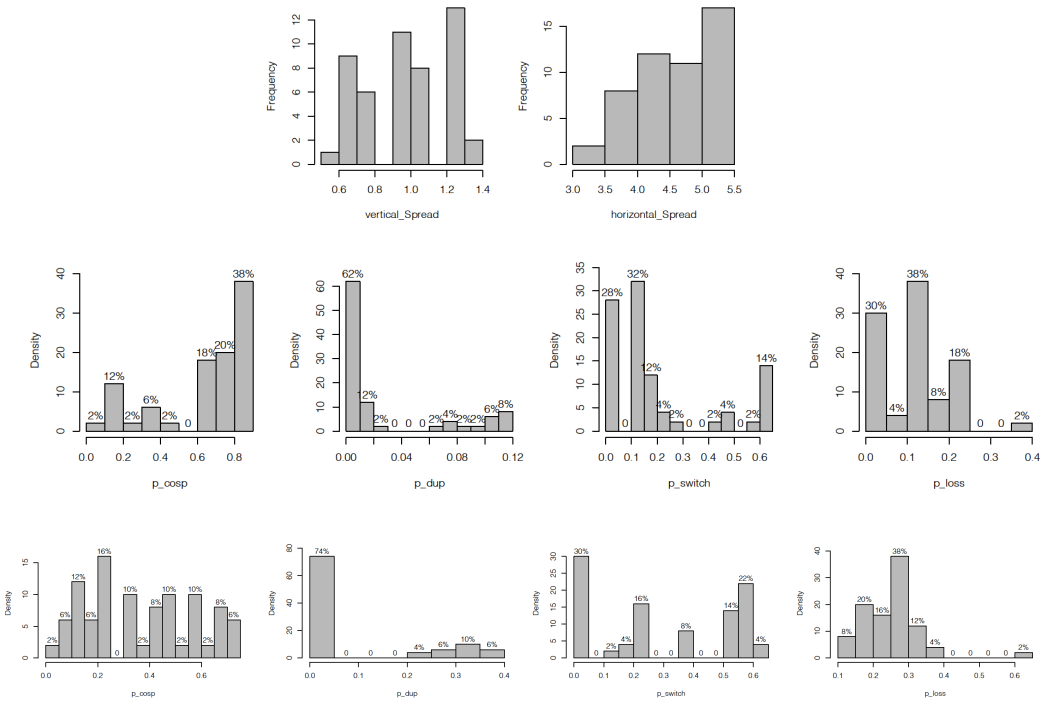

Figure 6: Comparison of the results obtained with AmoCoala and Coala for the dataset SFC. In each graphic, we show for each event type, the distribution of the parameter values. In the first two rows, the results provided by AmoCoala and in the third row, the ones provided by Coala.