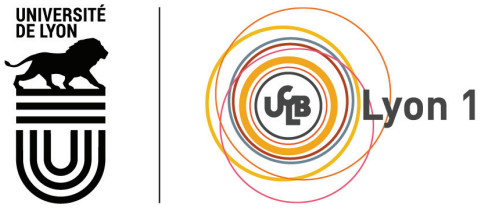


N° d'ordre NNT : 2017LYSE1214

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

École Doctorale E2M2
ED-341

Spécialité de doctorat : BioInformatique

Soutenue publiquement le 23/10/2017, par :

Laura Urbini

Models and algorithms to study the common evolutionary history of hosts and symbionts.

Devant le jury composé de :

Jean Lobry, Professeur des universités, UCBL	Examinateur
Martin Middendorf, Professeur des universités, Leipzig	Rapporteur
Paola Bonizzoni, Professeur des universités, UniMiB	Examinatrice
Yves Desdevises, Professeur des universités, UPMC	Rapporteur
Catherine Matias, Directrice de Recherche, CNRS	Directrice de thèse
Marie-France Sagot, Directrice de Recherche, INRIA	Directrice de thèse
Blerina Sinaimeri, Chargé de recherche, INRIA	Co-encadrant

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Remerciements

Contrairement à ce que on puisse penser, les remerciements sont l'une des parties les plus difficiles à écrire. Réaliser un doctorat n'implique pas de faire seulement de la recherche et de l'enseignement, mais permet aussi de vivre des nouvelles experiences et des moments de partage. Ces trois années passées au sein du LBBE m'ont fait grandir humainement et professionnellement. C'est pourquoi nommer ici tout le monde sera impossible, mais je vous remercie tous du plus profond de mon cœur.

Tout d'abord, je voulair remercier mes rapporteurs **Yves DESDEVISES** et **Martin MIDDENDORF**, qui on pris le temps de lire et améliorer ce travail de recherche. Je aussi voudrais remercier mes directrices de thèse, qui ont toujours été présentes, qui ne m'ont jamais abandonnée et qui m'ont toujours encouragées à donner le meilleur de moi-même. Merci **Marie-France** pour les discussions autour d'une tasse de thè, quand on parlais pas seulement du travail, mais aussi de nos vies. Merci à **Catherine**, pour avoir été toujour disponible pour mes appels desespéres par Skype quand je ne savais plus quoi faire. Je voudrais aussi remercier une autre personne qui m'a suivi pendant cette thèse, merci **Blerina**, tu sais qu'avant d'être une collègue tu es une amie pour moi, tu as été la personne la plus proche de moi tout au long de ce voyage.

Je ne peux pas oublier de remercier mes collègues et amis de l'équipe Erable
Baobab, avec vous j'ai vécu des moments incroyables: **Alex**, **Arnaud**, **Audric**, **Camille M.**, **Camille S.**, **Carol**, **Catherine M.**,**Christian**, **Clara**, **Delphine**, **Florence**, **Gustavo**, **Irene**, **Hélène**, **Laurent**, **Leandro**, **Marina**, **Martin**, **Mattia**, **Nina**, **Scheila**, **Susan**, **Taneli**, **Vincent L.** et je remercie spécialement : **Alice**, **Mariana** et **Ricardo**, qui ont été particulièrement présents dans ma vie au laboratoire.

Il y a d'autres personnes importantes que je ne peux pas oublier, mes amis, qui m'ont fait rire et passer de bons moments : **Antonio**, **Claudia**, **Élisa**, **Emanuela**, **Eleonora**, **Giovanna**, **Melissa**, **Paolo** et **Roberta**. Ces derniers mois, j'ai vécu avec les meilleurs colocataires du monde: **Émilie** et **Max**, merci à vous pour ces mois ensemble. Evidemment je ne peux pas oublier mes merveilleuses compagnes du Master, combien de fois nous nous sommes appelées, pour parler de nos problèmes de thèse et nous avons toujours été proches les unes des autres, vraiment, merci à : **Choumouss**, **Inès**, **Samia** et ma **Sarah**.

Je voudrais également remercier une personne qui heureusement, depuis un certain temps, fait partie de ma vie. Merci **Vincent** pour tes encouragements, pour m'avoir fait rêver et cru en moi. Tu m'as toujours aidé quand j'en avais besoin, écouté et conseillé quand j'avais des problèmes et tu m'as fait rire quand j'étais un peu triste. T'avoir à mes côtés est une immense chance.

Infine, un ringraziamento speciale a mia **madre** a mio **padre** che sono sempre stati i miei più grandi sostenitori.

Résumé en français

Presque chaque organisme terrestre interagit avec d'autres espèces. Quand cette interaction est proche et durable, elle est appelé *symbiose*. Les interactions symbiotiques impliquent en général une dépendance entre espèces et sont indispensables pour le fonctionnement de l'écosystème. Plusieurs types d'associations symbiotiques entre espèces existent en nature, tels que : le mutualisme (association bénéfique entre deux espèces vivantes) et le parasitisme (association étroite entre deux espèces vivantes dont l'une appelée l'hôte héberge la seconde qui vit à ses dépens), avec des situations intermédiaires où une espèce n'est pas affectée mais les autres espèces bénéficient de l'interaction (le commensalisme) ou elle en sont blessées (ici on parle de amensalisme). Certaines interactions peuvent devenir obligatoires dans le cas où aucune des espèces impliquées ne peut survivre sans l'autre. Cela peut par exemple se vérifier quand une des espèces vit à l'intérieur des cellules de l'autre. Nous porlons alors de *endosymbiose*. Les espèces qui hébergent les autres, que ce soit à l'intérieur ou à l'extérieur, s'appellent *hôtes*, tandis que les espèces hébergées s'appellent *symbiotes*. Puisque il est possible d'avoir des interactions symbiotiques pendant une très longue période, les espèces impliquées influencent mutuellement l'évolution l'une de l'autre. Ceci est connu sur le nom de *coévolution*.

L'étude de la coévolution et de l'influence évolutive réciproque entre les hôtes et leurs symbiotes est de plus en plus utilisé, notamment parce que ceux-ci peuvent aider à améliorer la production agricole, mais aussi aider à étudier des maladies dangereuses. Par exemple un dernier étude [66], envisage d'arrêter ou de limiter l'impact des infections, répandues par certains insectes, par une manipulation des endosymbiotes qui les habitent.

La toujours plus grande disponibilité de données phylogénétiques a permis de mieux comprendre les associations historiques qui existent entre différents hôtes et symbiotes. Cela a comme but d'essayer de comprendre: l'ancienneté de ces associations, si l'acquisition du symbiote est récente ou ancienne et si le symbiote est spécifique (c'est-à-dire si le symbiote a la capacité d'infecter un seul hôte ou une grande gamme des hôtes). Dans le premier cas, le symbiote est appelé spécialiste (avec une spécificité élevée), tandis que dans le second, le symbiote est appelé généraliste (avec une faible spécificité).

L'évolution d'un groupe d'espèces est généralement représentée avec un *arbre phylogénétique*, où les arcs correspondent aux lignées des espèces, les sommets internes représentent le moment de spéciation et les feuilles correspondent aux espèces qui vivent actuellement sur Terre. L'histoire évolutive commune entre les deux ensembles d'espèces, pour les hôtes et pour les symbiotes, peut être étudiée en comparant les arbres phylogénétiques respectifs, cela seulement si on connait quelles sont les interactions actuelles, c'est-à-dire si on connait les symbiotes qui actuellement habitent un hôte. Si deux arbres ont la même topologie, on parle de congruence entre phylogénies. D'habitude, la congruence est lié a la *cospeciation*. Cet evènemet se présente quand la spéciation de l'hôte et du symbiote sont strictement dépendantes l'un de l'autre. Cependant, il est très

rare de trouver deux arbres phylogénétiques des hôtes et des symbiotes qui sont exactement les mêmes. D'habitude, des incongruences sont présents, cela est dû au fait que les symbiotes peuvent changer de lignée (c'est appelé *saut*), specier indépendamment de l'hôte (c'est appelé *duplication*) disparaître ou ne pas coloniser tous les hôtes (ceux-ci sont appelés *perte*). Il faut remarquer que avoir des incongruences ne signifient pas nécessairement que les deux groupes d'espèces n'ont pas coévolué. Pas tout le monde est d'accord pour dire que deux espèces peuvent coévoluer sans cospecicer. Par exemple Brooks and Lennan [8] assument que la cospeciation et la coévolution représentent un même événement, ou plus précisément, ils assument que au niveau macro-évolutif la cospeciation est égale à la coévolution. Cependant, la majorité des auteurs chercheurs fait une distinction entre la cospeciation et la coévolution, et considèrent cette dernière comme l'adaptation réciproque des hôtes et des symbiotes [79] tandis que la cospeciation est vue comme un processus distinct [57]. Dans cette thèse, nous considérons la cospeciation comme une possible conséquence à la coévolution.

La méthode de réconciliation cophylogénétique que nous utilisons implique de associer (ou lier) un arbre, d'habitude celui des symbiotes, à l'autre, en utilisant un modèle appelé *basées sur des evènements*. Les evènement le plus utilsés sont la cospéciation, la duplication le saut et la perte. Les phylogénies des l'hôtes et des symbiotes sont généralement considérés comme correcte, ou sans aucune erreur. Ceci est une hypothèse forte qui est rarement vrai. Une autre assomption forte est que en général il est supposé qu'un symbiote peut habiter au plus un hôte. Cela est l'hypothèse dite un-hôte-par-symbiote [68]. Cette dernière rend le problème mathématiquement plus simple et est largement adopté, même si elle est biologiquement fausse. En effet, il existent de nombreux cas d'espèces multiples, pathogènes ou non, infectant un plusieurs hôtes.

Ces deux assomptions, qui sont rarement correctes sont les principaux études de cette thèse. Nous avons essayé d'affiner les modèles de réconciliation existants et de les rendre plus réalistes. À cette fin, nous avons présenté certains evènements peu ou pas formellement considérés dans la littérature. L'un d'entre eux est le "spread", qui correspond à l'invasion de différents hôtes par un même symbiote.

Cependant, avant de développer un tel modèle, nous nous sommes demandés quel est la robustesse de la méthode adoptée. C'est à dire que, compte tenu d'un modèle, si la méthode peut trouver toutes les réconciliations optimales même en présence des petites perturbations ou des erreurs dans les données en entrée.

Ce doctorat s'est concentré sur ces deux aspects :

- Robustesse du modèle : L'objectif était d'analyser la robustesse des méthodes de réconciliation des arbres phylogénétiques, qui sont très utilisés dans ce type d'étude.Nous avons essayé de comprendre les forces et les faiblesses du modèle parcimonieux utilisé et comprendre comment les résultats finaux peuvent être influencés en présence des petites perturbations ou des erreurs dans les données en entrée. Ici deux cas sont considérés, le premier est le choix erroné d'une association entre les feuilles des hôtes et des symbiotes dans le cas où plusieurs d'autres existent, le deuxième est lié au mauvais choix de l'enracinement de l'arbre des symbiotes. Nos résultats montrent que le choix des associations entre feuilles et le choix de l'enracinement peuvent avoir un fort impact sur la variabilité de la réconciliation obtenue. Nous avons également remarqués que l'evènement appelé "saut" joue un rôle important dans l'étude de la robustesse surtout pour le problème de l'enracinement.

- Modèle de cophylogénie plus réaliste : L'objectif est d'introduire certains evènements

peu ou pas formellement considérés en littérature. L'un d'entre eux est le "spread", qui correspond à l'invasion de différents hôtes par un même symbiote. Dans ce cas, quand les spreads ne sont pas pris en compte, les réconciliations optimales sont obtenus en tenant compte seulement des coûts des evènements classiques (cospeciation, duplication, saut, perte). La nécessité de développer des méthodes statistiques pour assigner les coûts le plus appropriées est toujours d'actualité. Deux types de spread sont introduits: verticale et horizontale. Le premier type correspond à ce qu'on pourrait appeler aussi une gèlé, c'est à dire que l'évolution du symbiote s'arrête et "gèle " alors que le symbiote continue d'être associé à un hôte et aux nouvelles espèces qui descendent de cet hôte. Le second comprend à la fois une invasion, du symbiote qui reste associé à l'hôte initial, mais au même temps il est associé (il "envahit") un symbiote incomparable avec le premier, ici nous avons que l'évolution du symbiotes " gèle " par rapport l'évolution des deux l'hôtes, celui auquel il était associé au début et celui qu'il a envahi. Nos résultats montrent que l'introduction de ces evènement rend le modèle plus réalise, mais aussi que désormais il est possible d'utiliser directement des jeux de données avec un symbiote qui habite plusieurs hôtes au même temps, ce qui n'était pas faisable auparavant.

Contents

Chapter 1

Introduction

Contents

1.1 Motivation

Although initial concepts related to evolution have been elaborated since Ancient Greece, the first more general theories were developed much later. J. B. Lamark's model, presented in his book "Philosophie Zoologique" dating from 1809, may be considered as one of the first examples of this. The real revolution came however a few years later, in 1859, when Charles Darwin's book "The origin of species" was published. In it, Darwin introduced his theory of the process through which organisms of the same species gradually evolve over time. Figure 1.1 presents the famous sketch of an evolutionary tree that he drew a few years before, in 1837-1838, and which appears in his "Notebook B: Transmutation of species". Darwin's main idea, that he called "natural selection", is that, in a world with stable populations where each individual has to struggle to remain alive, only those with the "best" features will have more chance to survive and to transmit those favourable traits to their descendants. These beneficial characteristics will thus gradually become the dominant ones in the population. According to Darwin, if the process of natural selection takes place over a long enough period of time, it produces changes in a population, possibly leading to the formation of new organisms. The latter phenomenon is called *speciation*.

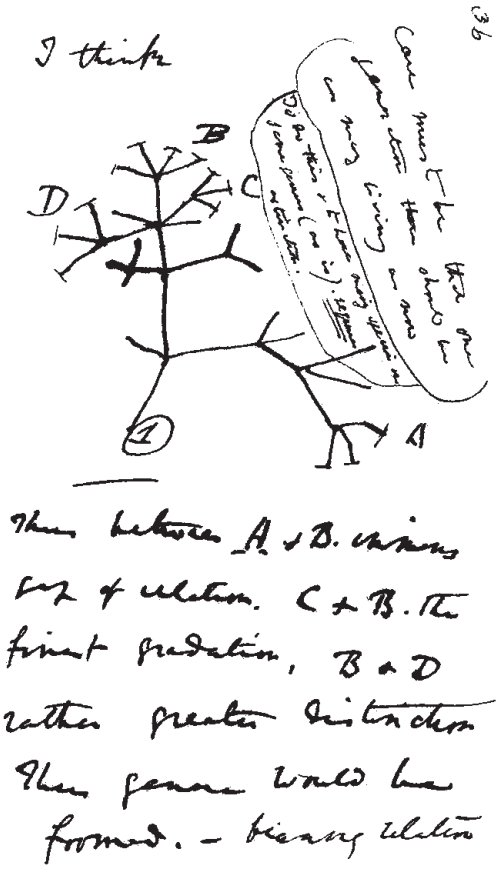


Figure 1.1: Darwin's first sketch of an evolutionary tree, dated around July 1837. Interpretation of the handwriting: "I think case must be that one generation should have as many living as now. To do this and to have as many species in same genus (as is) requires extinction. Thus between A + B the immense gap of relation. C + B the finest gradation. B+D rather greater distinction. Thus genera would be formed. – bearing relation".

In his book of 1859, Darwin mentioned also the existence of evolutionary interactions as these

may be important in particular for analysing the rate of evolution of the species involved. One example he gave concerned flowering plants and insects. Instances of such interactions have since been discovered to be very extensive. Indeed, it is nowadays believed that almost every organism is involved in what has been initially called (by Heinrich Anton de Bary in 1879) a *symbiotic* interaction with other biological species, or *symbiosis* for short, that is, in an interaction which, according to the initial meaning given to it by de Bary, is close and often long term. Both characteristics will in general imply in a dependency on the presence of the other species to survive, and are indispensable for the functioning of the ecosystem.

Symbiosis can involve two different species, or more than two. It can also be of various types, ranging from mutualism (when both species benefit) to parasitism (when one benefits to the detriment of the other), with intermediate situations where one species is unaffected but the other species either benefits from the interaction (this is called commensalism) or is harmed (one then speaks of amensalism). Some interactions may become obligatory in the sense that none of the species involved is able to survive anymore without the other. This may be the case in particular when one of the species lives inside the cells of the other. We speak then of *endosymbiosis*. An example is the insect *Acyrthosiphon pisum* and the bacterium *Buchnera aphidicola*. It is however important to notice that not all endosymbioses are mandatory. This is for instance the case of the bacteria from the *Wolbachia* genus which infect a range of arthropods and of nematodes, and where the interaction is sometimes obligatory, and at other times not.

The species harbouring others, whether inside or outside, is called a *host*, while the species being hosted is called *parasite* when it benefits from the interaction to the detriment of the host. As such interactions may not always be detrimental to the host, the term of *symbiont* has also been used in the literature [81] and is the one that will be adopted in this Ph.D.

Since symbiotic interactions may continue over very long periods of time, the species involved can affect each other's evolution. This is known as *coevolution*. There have been increasingly more studies of coevolution and of the reciprocal evolutionary influence between hosts and their symbionts, notably because these may help improve agricultural production [86] or even control for devastating diseases. As an example of the latter, one may consider recent attempts to stop or to limit the impact of infections that are spread by some insects through a manipulation of the endosymbionts that inhabit them [66].

A growing availability of phylogenetic data has allowed to better study the historical associations that exist between different types of hosts and symbionts [56] in order to try to understand how old such associations are, if the acquisition of the symbiont is recent or ancient, and what is the host specificity of the symbiont, that is whether the symbiont has the ability to infect only one or a large range of hosts. In the first case, one speaks of a specialist symbiont and of a high specificity, while in the second, one talks of a generalist symbiont with a low specificity.

The evolution of a set of species is usually represented by a so-called *phylogenetic tree*, where the arcs correspond to the species lineages, the internal vertices are the moment of speciation and the leaves correspond to the species that live at present time. The common evolutionary history between two sets of species, one for the hosts and one for the symbionts, can then be studied by comparing their respective phylogenetic trees if one further knows which are the present-day interactions, that is, which currently living symbiont closely interacts with (or inhabits in the case of endosymbiosis) which hosts. If the two trees have the same topology, there is congruence between the phylogenies. Usually, congruence is related to *cospeciation* (some authors prefer the term *codivergence*; we will be using both terms on this Ph.D. manuscript). This event occurs

when the speciation of hosts and of symbionts strictly depends on one another. An example is given in Figure 1.2(a). It is however very rare to find two phylogenetic trees of hosts and of symbionts that are exactly the same. Most often, there are incongruences that are due to the fact that the symbionts can switch lineage (Figure 1.2(b) – this is called *host switch*), speciate independently of the host (Figure 1.2(c) – this is called *duplication*), go extinct (Figure 1.2(d)), or not colonise all the hosts (Figure 1.2(e) – these are called a *loss* (of the symbiont)). It is important to notice that such incongruences do not necessarily mean that the two sets of species have not coevolved.

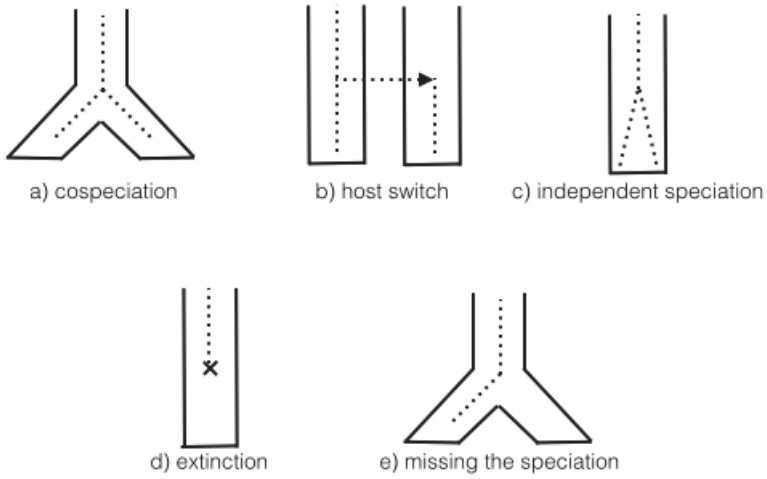


Figure 1.2: Host-symbiont associations. The tube represents the host tree and the dotted lines the symbiont tree. (a) Hosts and symbionts speciate at the same time – this is called cospeciation or codivergence. (b) Symbionts speciate and one of the new species switches from one host lineage to another – this is called host switch. (c) Symbionts speciate independently from the host – this is called in the literature duplication (of the symbiont). (d) Absence of a symbiont from a host which may be due to extinction of the symbiont (notice that other explanations are possible such as failure to detect the symbiont) – this is called loss (of the symbiont). (e) The ancestor of the host lineage may not have inherited the ancestral symbiont (also called loss).

Not everyone agrees that two species can coevolve without cospeciating. Brooks and Lennan [8] for instance assumed that cospeciation and coevolution represent a same event, or more precisely, they assumed that cospeciation equals coevolution in macro-evolutionary time. The majority of authors however distinguish between cospeciation and coevolution, and consider the latter as the reciprocal adaptation of hosts and of symbionts [79] while cospeciation is perceived as a separate process [57]. That is will be our case also in this Ph.D., where cospeciation will be seen as only one possible consequence of coevolution.

One must however be conscious of the fact that establishing where the line stands can often be difficult. As an example, Ramsden *et al.* (2009, Figure 2) considered the evolution of hantaviruses in relation to their rodent hosts. The two phylogenetic trees exhibited enough topological similarities for many [33, 38, 53, 62] to believe that the two sets of species coevolved. By more finely analysing the similarities between the two trees through what is called a *cophylogenetic reconciliation*, Ramsden *et al.* [64] showed that more than 20 cospeciations were however non significant. To then explain the observed congruence between the two trees, the authors suggested that this was the consequence of a number of symbionts switching lineages. Based

on this study, the authors concluded that the evolution between rodents and hantaviruses is the result of a recent history of preferential host switching and of local adaptation rather than of coevolution.

The cophylogenetic reconciliation method that Ramsden *et al.* [64] used, involves mapping one tree, most often the symbiont's, to the other using a so-called *event-based* model. The events considered are those that were mentioned above and that appear illustrated in Figure 1.2. The host and the symbiont phylogenies are usually considered as given and without any errors. This is clearly a strong assumption that is seldom if ever correct. It is further assumed in general that a symbiont can inhabit at most one host. This is the so-called one-host-per-symbiont assumption [68]. This latter in particular makes the problem mathematically simpler and is widely adopted even though it is biologically wrong. Indeed, there exist numerous cases of multiple species, pathogen or not, infecting a same host. Two examples of such are provided for instance in [12]. The first concern HIV viruses, more particularly HIV-1 and HIV-2 and their evolutionary relationship with humans. HIV is a human virus that causes failure of the immune system. One of the questions that had been addressed by the authors and by others was whether this virus had been recently acquired from monkeys or if the human species were hosting the virus since a much longer time. The second example concerns malaria parasites which include *Plasmodium* and related genera. Malaria is a disease present in different animals including humans. Perkins and Schall [61] found that the four types of parasites that cause malaria in humans are polyphyletic, which means that they do not share an immediate common ancestor. The question here was if humans acquired the parasite from monkeys or the opposite.

In both cases, what Charleston and Perkins argued [12] is that the results obtained and the conclusions that were then reached depend strongly on whether the two assumptions above – *i.e.* that the trees given as input are correct and that each symbiont is associated to at most one host – are correct.

These two examples make us understand the importance of trying to address such problems, or in the case of possible errors in the input, of trying to evaluate its potential impact on the results obtained. These are the main aims of this thesis.

Before we introduce more in detail each of the above problems in turn (in Sections 1.4.1 and 1.4.2), we provide a few basic concepts below on phylogeny (Section 1.2) and on cophylogeny (Section 1.3) that will be important to understand the remaining of the thesis.

1.2 Phylogeny

A phylogenetic tree is a directed graph with labelled leaves that represents the lines of evolutionary descent between the taxonomic groups of different organisms. The leaves represent the current taxa, the internal vertices correspond to the inferred speciation events and the arcs of the tree represent the life of a single species. Although a number of phylogenetic tree reconstruction algorithms produce unrooted trees [52, 70, 76], most provide phylogenetic trees that are rooted, where the root is the common ancestor of all the species in the tree. In this case, a direction is thus intrinsically assumed that corresponds to the one of increasing evolutionary time. Rooting of the tree is in general obtained using the so-called *outgroup method*. A correct indication of the root position therefore strongly depends on the availability of a proper outgroup [29, 63, 73].

1.2.1 Basic Definitions

In this work, we will consider always rooted trees. The inner vertices of such trees have in-degree 1 and out-degree 2 (except for the root that has in-degree 0), while the leaf vertices have in degree 1 and out-degree 0.

Given a tree T, the set of its vertices will be denoted by $V(T)$, the set of its arcs by $A(T)$, and the set of its labelled leaves by $L(T)$. The root of T is denoted by $r(T)$. Given an arc $a = (v, w) \in A(T)$ going from vertex v to vertex w, we denote the head of a by $h(a)$ and its tail by $t(a)$. We thus have that $v = t(a)$ and $w = h(a)$. In terms of evolutionary relationship, v is the parent of w. We denote this by $v = par(w)$. The only vertex without parent is the root $r(T)$. We assume that a parent and its children do not coexist, meaning that enough time has passed between the two to distinguish them.

Given a vertex $v \in V(T)$, we denote by T_v the subtree of T rooted at v (including v). We define the set of descendants of v, denoted by $Des(v)$, as the set of vertices in T_v. Similarly, the set of ancestors of v, denoted by $Anc(v)$, is the set of vertices in the unique path from the root of T to v (including the end-points). We denote by $lrca(v, w)$ the last recent common ancestor of v and w in T. We denote by $\geq$ the partial order induced by the ancestry relation in the tree. Formally, for $x, y \in V(T)$, we say that $x \geq y$ if $x \in Anc(y)$. If neither $x \in Anc(y)$ nor $y \in Anc(x)$, the vertices are said to be *incomparable*. All operations that may be performed on phylogenetic trees are followed by a cleaning of the vertices of out-degree 1 if any were created (see Figure 1.3) in order to obtain once again a phylogenetic tree.

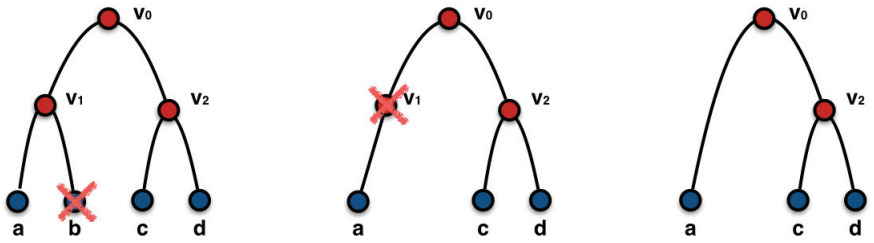


Figure 1.3: Example of an operation of cleaning a phylogenetic tree. Given such a tree, suppose we want to eliminate leaf b. In a second step, we then need to clean the phylogenetic tree, taking out the vertex v_1 that is of out-degree 1. At the end, we obtain a cleaned phylogenetic tree.

1.2.2 Rooting a Phylogenetic Tree

Phylogenetic methods reconstruct trees using as information the differences observed between taxa, but often they cannot orient the trees. This is why many phylogenetic tree reconstruction algorithms produce unrooted trees [52, 70, 76]. In this case, there are as many potential roots as the number of the arcs in the tree (see an example in Figure 1.4).

There are different methods to orient two phylogenetic trees. We indicate two of the main ones below:

- *Midpoint rooting.* This approach is used when all the species are believed to have evolved at the same velocity (this is called the molecular clock hypothesis). The evolutionary distance between each leaf and the root is therefore the same. The root is then positioned

in the tree equidistantly from all the leaves. The midpoint rooting works well if the tree is balanced with a long branch in the middle separating the groups of organisms. The main problem with midpoint rooting is that it is very susceptible to large deviations from a constant evolutionary rate, notably when these are not balanced. The rooting can also be wrong when it places the root amongst a dense set of short branches. In this case, a small deviation will place the root on the wrong branch.

- *Outgroup rooting.* This is the most widely used method. Here a species (or set of species), exterior to those analysed, is included in the input. The root is positioned at the vertex that binds the outgroup to the studied tree. The main problem with outgroup rooting is that it is very sensitive to the choice of the outgroup as a distant one can lead to a wrong positioning of the root.

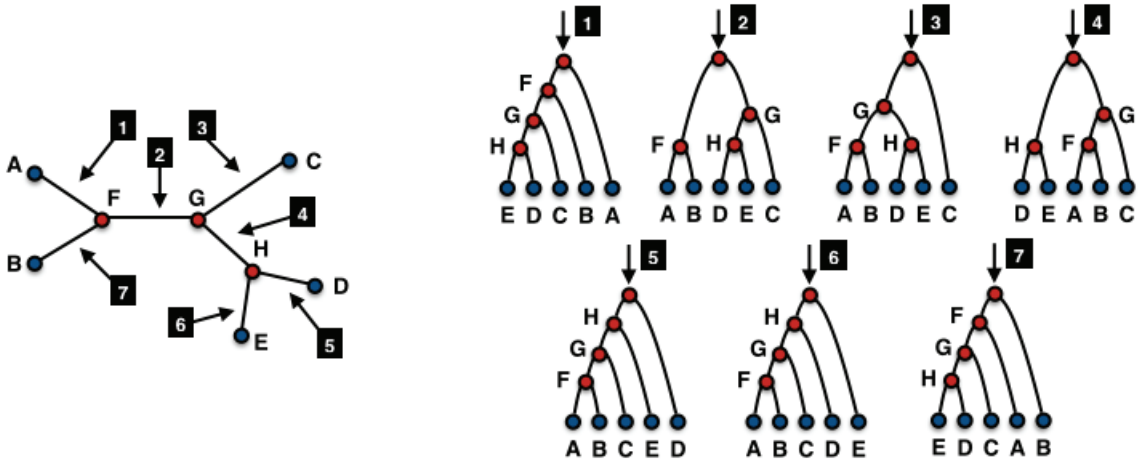


Figure 1.4: All possible roots for an unrooted phylogenetic tree. On the right, all the rooted trees obtained from the left unrooted phylogenetic tree. Tree 1 is obtained by placing the root in the arc 1 of the unrooted phylogenetic tree. The same for all the other rooted trees, leading to 7 possible rooted phylogenetic trees.

1.2.3 Dated and Undated Trees - Timing information

Estimating the timescales for a phylogenetic tree can be important to understand the evolutionary history of species. Using molecular clocks to estimate divergence dates depends on other methods of dating that most often rely on molecular data. In this case, the length of an arc in a tree represents the rate at which a stretch of DNA changes. The molecular clock hypothesis asserts that, over time and among different organisms, DNA and protein sequences evolve at a rate that is relatively constant. A direct consequence of this is that the difference observed between two species is proportional to the time since these species shared a common ancestor. This hypothesis is in particular used to study organisms that have left few fossil traces in their biological history, as is the case of viruses.

In many phylogenetic trees, timing information is not available. The trees are thus undated and any time information is given by the topology only. The ancestor species is the root and

the leaves are those species that live on earth today. It is assumed that an ancestor and its descendants cannot coexist. However, two incomparable vertices may have coexisted.

In this Ph.D., we will work with trees that are rooted and are undated.

1.3 Cophylogeny

Phylogenetic trees can be used to study the evolutionary history between two sets of species, in what is known under the name of *cophylogeny*. In this case, we need two trees (more may also be considered) and must also know the associations between the leaves, that is, we must have as additional information which present-day symbiont has a close and long term interaction (parasitic, mutualistic, commensalist, amensalist) with which present-day host (in the case of endosymbiosis, inhabits the host).

The corresponding dataset will be denoted by the triple (H, S, ϕ) where H is the host tree, S is the symbiont tree and *phi* is a function which indicates the association between the leaves of the host tree and the leaves of the symbiont tree. An example is given in Figure 1.5. In this model, the one-host-per-symbiont assumption is made, meaning that each present-day symbiont is associated to one and only one present-day host. A host on the other hand can have more than one symbiont associated to it.

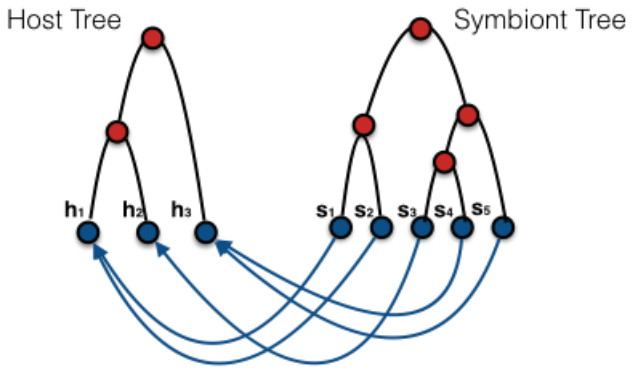


Figure 1.5: Example of a cophylogeny dataset (H, S, ϕ), with the host tree H at the right, the symbiont tree S at the left and the blue arrows representing the associations between the leaves of the two trees.

Cophylogeny may be addressed in different ways, we briefly mention two main approaches that were used in the literature before entering more in detail into the reconciliation method which is the one we adopted in this work.

1.3.1 Comparison, or Congruence Test Method

In earlier studies of coevolution, the objective was to just evaluate whether two sets of species could be said to have coevolved, not to infer the ancestral associations from present-day ones. Most such methods therefore performed what was called a *congruence test*. Such studies thus relied on the assumption that two sets of species have coevolved if their trees are measured as being "congruent enough", meaning topologically "similar enough". If hosts and symbionts often speciated at the same time, some dependence between these two species could be assumed. The notion of congruent phylogenies thus implies a high number of cospeciations, while incongruence

implies host switching [8]. Many tests of tree congruence have been proposed in the literature [14, 15, 30, 32, 44]. In most, the main idea is to fit a host and a symbiont tree together by maximising the number of cospeciations. Even though this method is easily applicable, it cannot say anything about events other than cospeciation.

1.3.2 Brooks Parsimony Analysis

One of the earliest methods that attempted to go further than a congruence test was the so-called Brooks Parsimony Analysis (BPA for short) developed by Brooks and McLennan [8]. The method takes two phylogenetic trees as input, each corresponding to a set of species, and uses a parsimonious analysis to reconstruct the coevolutionary relationships between the two sets. Each leaf of a phylogenetic tree (of the host or of the symbiont) is coded, using binary characters, in a manner that indicates non only the identity of the species, but also the common ancestors of each species. This code is then represented in a binary cost matrix which is optimised on the other tree. The results of BPA can be interpreted *a posteriori* in terms of events but it has proved difficult to properly formalise this translation. It is important to note that this analysis is not based on a model with associated event-cost assignments.

1.3.3 Reconciliation Model

A more widely used model for studying the evolutionary history of two sets of species has been called *phylogenetic tree reconciliation* [9, 10, 49, 55].

The main idea of such model is, given a triple, (H, S, ϕ) to find a reconciliation λ that associates each vertex of the symbiont tree S to a vertex of the host tree H. The function λ must be an extension of the function ϕ that associates each leaf $l_S \in L(S)$ of the symbiont tree to a leaf $l_H \in L(H)$ of the host tree.

Thanks to this λ function, it is then possible to unambiguously identify a series of events that explain the coevolution of the two sets of organisms [37].

Four major macro-evolutionary events are in general considered in the literature: a) cospeciation, when the divergence of a symbiont is in correspondence to the divergence of a host; (b) duplication, when the divergence of a symbiont is independent of the divergence of a host; (c) host switch, when after a divergence of the symbiont, one symbiont jumps from one host species to another that is incomparable with the first; and (d) loss, which can describe three different and undistinguishable situations: (i) when the divergence of the host is independent of the divergence of a symbiont, which then follows just one of the new host species due to factors such as, for instance, geographical isolation; (ii) when there is a cospeciation of host and symbiont, followed by the extinction of one of the new symbiont species, and (iii) when there is a cospeciation of host and symbiont, with failure to detect the symbiont in one of the two new host species [9, 10]. These events are depicted in Figure 1.6.

A cost is associated with each of the four types and the best reconciliation (the optimal one) is chosen according to a parsimony model [55]. The final objective is to find the λ function associating the vertices of the symbiont tree to the vertices of the host tree which has minimum total cost.

Biologically, in a host switch event, a symbiont can only jump from one host species to another that is contemporaneous to the first. If timing information (*i.e.* the order in which the speciation

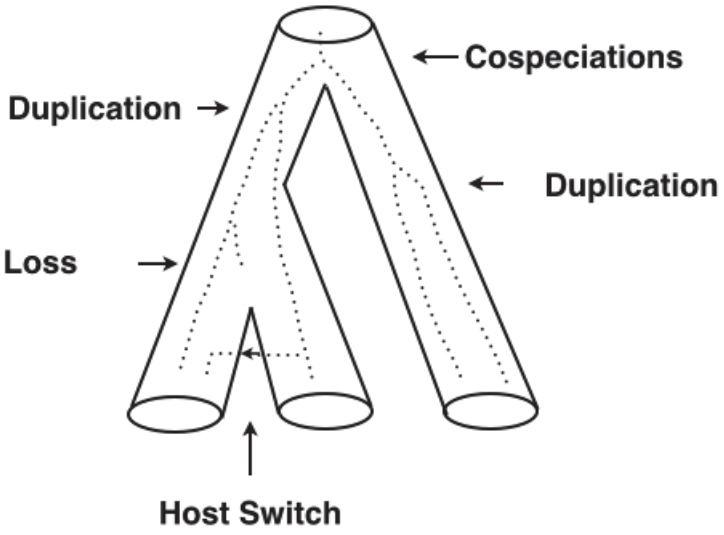


Figure 1.6: Example of reconciliation. The tube represents the host tree and the dotted lines represent the symbiont tree.

events occurred along the host phylogeny) is not available, some reconciliations proposed may be biologically impossible in the sense that some of the switches induce a contradictory time ordering for the internal vertices of the host tree.

An example is given in Figure 1.7B where a cycle is present. We have here two host switches, one involving a symbiont and another one involving of its descendants, where the latter jumps to a host that is an ancestor of the host to which the former jumped. This is biologically unfeasible because it is assumed that an ancestor and a descendant cannot coexist.

We will henceforth call a reconciliation without cycles (as in Figure 1.7A), a *time feasible reconciliation* and one with cycles (as in Figure 1.7B) a *time unfeasible reconciliation*. In reality, things are slightly more complicated than this, as indicated in [74].

If timing information is not known, as is usually the case, the problem of reconciling two phylogenetic trees is NP-hard [54, 80]. 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 vertices of the host tree. In this case, the problem can be solved in polynomial time [4, 18, 20, 50, 74]. In most situations, as shown in [18], among the many optimal solutions, at least some will be time-feasible.

Another option is to rely on heuristics [13], that is on approaches that are not guaranteed to be optimal.

In both cases, providing a single optimal solution is not a good option as it may either be not optimal among those that are time-feasible, or it may be biologically unfeasible. Observe however that this is what the majority of the existing reconciliation algorithms do.

Moreover, the common evolutionary history of the hosts and of the symbionts is only partially captured by the reconciliation model. The most desirable solution may thus depend on biological quality criteria that are not taken into consideration in the model. Indeed, it is frequent that given a host and a symbiont tree, there may be many optimal solutions which, although having the same total cost, can be quite different among them (*i.e.* can correspond to a different set of

events). For all these reasons, it is better not to rely on only one optimal solution and instead to output all solutions. Once such a list of candidates is generated, one can then rely on more sophisticated biological criteria to choose among or to classify them. Enumeration, also called listing algorithms are therefore of particular interest.

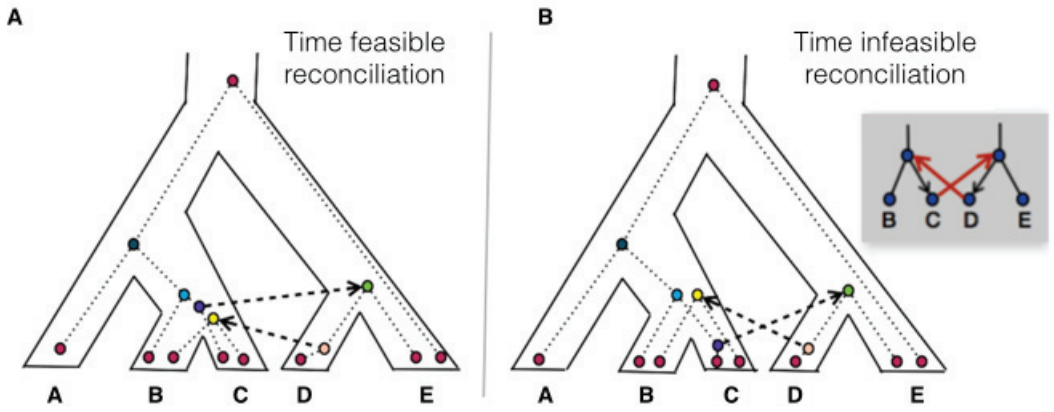


Figure 1.7: Two examples of reconciliation. The tube represents the host tree and the dotted line represents the symbiont tree. The dotted arrows represent a symbiont that jump to from an host to an other. (A) Example of time feasible reconciliation. (B) Example of time unfeasible reconciliation, in this case a cycle is present.

Notice that, in the context of gene-species associations, the reconciliation model presented here is known as the *DTL* (for "Duplication, Transfer, and Loss") model and has been extensively studied (see, for example, [4, 19, 27, 75, 80]).

1.3.4 Enumeration Algorithm

It is important to observe that the complexity of enumeration algorithms is different from traditional complexity theory as usually the number of solutions is exponential. It is therefore meaningless to measure efficiency by a polynomial running time in the input size ignoring the output size.

New notions of efficiency have thus been developed for enumeration problems. Examples are provided by Johnson *et al.* in [41]. For instance, an enumeration algorithm is said to be *polynomial delay* if the time between the output of any one solution and the next one is bounded by a polynomial function of the input size [41]. Eucalypt (which will be introduced below and then more in detail in Chapter 2) is such a polynomial delay algorithm, that, given a cost model for the events, generates *all* the optimal solutions for the reconciliation problem [18]. Observe that in order to guarantee such polynomial delay, the optimal solutions are enumerated without requiring that the reconciliations are time-feasible. This is because even producing one optimal time-feasible solution may take exponential time as mentioned before [54, 80]. A post-processing of the solutions obtained may be done to filter out for unfeasible solutions.

Other methods that generate more than one solution include CoRe-Pa [50], Mowgli [20], Jane 4 [13], Notung [75], Ranger-DTL [4] and ecceTERA [39]. For a survey on the

different features each method presents, and their possible limitations, see *e.g.* [18, 39]. More recent methods, namely CoRe-ILP [87] and ILPEACE (see the arXiv file at: https://arxiv.org/abs/1410.7004), find optimal solutions using integer linear programming.

Among the above methods, three have been recently extended and one newly developed to address the problem in the case where the input datasets include symbionts that are associated to more than one host. These methods are CoRe-Pa [50], Jane 4 [13], and WiSPA (see the arXiv file at: https://arxiv.org/abs/1603.09415). CoRe-Pa solves the multiple associations locally, by starting from the leaves that are already mapped and choosing for a parent vertex the unique associations of its children that give the best cost. Jane 4 uses a heuristic approach based on a genetic algorithm to recover the best solutions. Finally, WiSPA is a new model for reconciling trees where the symbionts are permitted to be associated with more than one host that includes additional evolutionary events, which the authors call *spread events*.

Considering spread events, or spreads for short, as in WiSPA was one of the objectives of this Ph.D. The term of spread was first used (to the best of our knowledge) by Brooks and McLennan in [8].

Before explaining what was done on this question, I briefly introduce below a method, Eucalypt, previously developed in the team where I did my Ph.D that was used in the work that I did.

Eucalypt (EnUmerator of Co-evolutionary Associations in PolYnomial-Time delay) is an algorithm that, given a triple (H, S, ϕ) and a vector of costs $\langle c_c, c_d, c_s, c_l \rangle$, generates all reconciliations with minimum cost (both time-feasible and time-unfeasible) [17].

Eucalypt adopts, as in [4], a dynamic programming approach to enumerate all optimal reconciliations. This uses a matrix D of dimension m by n where m is the number of symbionts and n the number of hosts. Each cell $c = D(s, h)$ of the matrix D indicates the cost of an optimal reconciliation that maps s to h. The complexity to find one optimal solution is $O(nm)$. A pseudo-code for finding the cost of one optimal reconciliation is given in Algorithm 1 in Appendix A. Actually two matrices are used. The first corresponds to D. The second, denoted by D_{ST} and of size also m by n, contains the optimal solutions of the subtrees. Formally, $D_{ST}(s, h)$ indicates the cost of an optimal solution with s mapped to some vertex i in the host subtree rooted in h.

In the case where we want to enumerate all solutions, each cell $c = D(s, h)$ of D must contain a list of $O(n^2)$ pointers, one to each of the mappings of the children s_1 and s_2 of s having led to the cost of an optimal sub-solution that mapped s to h. Each list has in the worst case a size of n^2, meaning that the total space required becomes $O(n^3 m)$.

Figure 1.8 (left side) shows the representation of a cell $c = D(s, h)$. In Figure 1.8 (right side), the information is visualised in the form of a local tree, with a parent vertex c as the root which corresponds to the mapping of vertex s in the symbiont tree to vertex h in the host tree (denoted in Figure 1.8 by $s : h$) and one child for each alternative mapping solution (denoted in the figure by rectangle vertices). Each alternative mapping solution corresponds to a pair of pointers that have optimal cost (denoted in Figure 1.8 by circle vertices). Accordingly, these vertices correspond to other cells of the matrix D which contain a similar local tree.

Finally, a time-feasibility test, following the approach used in [75], was implemented in Eucalypt to enumerate only time-feasible reconciliations. It has a time complexity of $O(n^2)$.

To enumerate all the optimal reconciliations, a pseudo-code is given in Algorithm 2 in Appendix A. A stack M is used to select which sub-solutions to add to the reconciliation that is currently being built. This stack is filled with couples of the form $\langle cell, index \rangle$. The function

$M(cell)$ returns, in constant time, the couple $\langle cell, index \rangle$ at the top of M, if M is not empty.

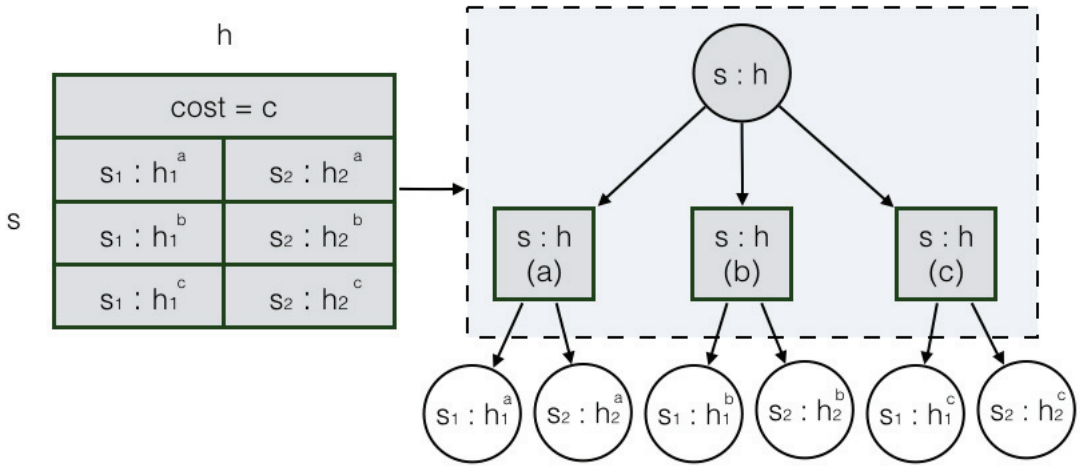


Figure 1.8: Representation of the content of the cell $c = D(s, h)$ in the dynamic programming matrix D: Suppose the cell is related to the association $s : h$ and let s_1, s_2 be the two children of s. One single cell-root vertex is created to represent the association $s : h$. This association has a local minimum cost that can be obtained in different ways, meaning by this, choosing different associations for $s1$ and $s2$. Each equivalent alternative is represented by a vertex (squared vertex in the picture). The number of alternatives is variable. Each squared vertex has exactly two children corresponding to the associations of s_1 and s_2 respectively (circle vertices in the picture)

1.3.5 Cost Inference Problem

Another crucial issue related to the parsimonious framework is that, from a biological point of view, reasonable cost values for a reconciliation are not easy to establish. Some approaches [11, 45] attempt to choose the costs of the events by adopting some minimisation constraints. Others, such as CoRe-Pa [50], propose to find cost minimal reconstructions using a parameter adaptive approach. The space of cost vectors is explored either by sampling such vectors at random assuming a uniform distribution model or by using more sophisticated approaches. In both cases, no costs have to be assigned in advance to the coevolutionary events. The method instead seeks for the optimal reconstruction in which the used costs are inversely related to the relative frequency of the corresponding events.

As indicated in [67], if each event is associated with a cost that is inversely related to its likelihood (the more likely is the event, the smaller is its cost), then the most parsimonious reconstruction will also, in some sense, be the most likely explanation of the observed data.

Likelihood-based approaches are on the other hand sometimes preferred to parsimony-based methods. A number of works have been done along these lines, for instance in [31, 32]. However in [32], the authors tested restricted hypotheses and mainly focused on the congruence between the trees, while in [31] the authors excluded duplications and tended to over-estimate the number of host switches. In [77] instead, all four types of events are considered, but the method was developed with the objective of inferring reconciled gene trees from a species tree (see also [78]). The aim is similar in [2] but the type of approach is different and the model again incomplete as in [31], this time not allowing for host switches. The Monte Carlo Markov chain approach adopted in [31, 77] moreover presents the inconvenience of being computationally intensive.

In the teams where this Ph.D. was done, an algorithm called Coala (for "CO-evolution Assessment by a Likelihood-free Approach") was then introduced to deal with the problem of cost inference [5]. For a given pair of host and symbiont trees, Coala estimates the frequency of the events based on an approximate Bayesian computation (ABC for short) approach. Indeed, in complex models where the likelihood calculation is often unfeasible or computationally prohibitive, classical Monte Carlo methods and their variants are being replaced by ABC, a set of more efficient statistical techniques [7].

In [5], starting with a probability distribution associated with the events, Coala simulates accordingly the temporal evolution of a set of species (the symbionts) following the evolution of another set (the hosts) for which a phylogenetic tree is already available. During an evolution simulation, Coala thus generates symbiont trees which are compared to the "known" symbiont tree. The set of probabilities that generates trees "closer" to the real one are in some sense the most likely explanation of the observed data. The approach thus consists in selecting parameter values (*i.e.* event probabilities) giving rise to symbiont trees that are "similar" to the known one. In this way, starting from a prior distribution on the parameter values, the approximate posterior probability for the events that best explains the observed data is deduced. The algorithm proposed, on one hand provides some confidence in the set of costs to be used for a given pair of host and symbiont trees, while on the other hand it allows to estimate the frequency of the events in cases where the dataset consists of trees with a large number of taxa.

Coala includes two main parts: the first (corresponding to Algorithm 1 below) simulates the evolutionary history of symbionts while the second (Algorithm 2) uses ABC in order to select the most probable frequency of the four events: cospeciation, duplication, host switch and loss.

Given a host tree H and a vector of four probabilities $\theta = \langle p_c, p_d, p_s, p_l \rangle$, a simulated tree $\tilde{S}$ is created using Algorithm 1 explained below. Using a distance between $\tilde{S}$ and S, it is possible to compare the simulated symbiont tree to the real one. The more similar are the two trees in terms of size and of topology, the lower will be the distance. At the end, the best vector θ is the one that creates a simulated tree $\tilde{S}$ that is most similar to the real tree S.

Algorithm 1: Simulation of the evolutionary history of the symbionts. The parameter vector of the model is composed of the probabilities of each one of the four events: cospeciation, duplication, host switch and loss. We thus have that θ stands for a vector of four probabilities $\langle p_c, p_d, p_s, p_l \rangle$. Following the topology of H and the vector θ, a simulated symbiont tree $\tilde{S}$ is created. At the same time as the simulation of $\tilde{S}$, a function λ that associates each vertex of $\tilde{S}$ to a vertex of H is created. If a host vertex does not match any symbiont vertex, we have a loss event. For this reason, θ is constrained such that $p_c + p_d + p_s + p_l = 1$.

Algorithm 2: Approximate Bayesian Computation – Sequential Monte Carlo procedure (ABC-SMC). In this case, N vectors in the space $[0, 1]^3$ are randomly chosen under some prior distribution (usually uniform). The two main steps of the Algorithm 2 consist in:

Step 1 For each vector θ, Algorithm 1 is used to generate M simulated trees $\tilde{S}$. A distance value, obtained computing the difference between S and $\tilde{S}$, is associated to each parameter vector. The vectors are then ordered respecting the distance value, in ascending order. The choice of the best vectors is made by taking into account two values which are:

- Tolerance: This value determines the percentage of parameter vectors to be accepted. If for instance, the tolerance value is 0.2, then 20% of the vectors with lower distance will be accepted.

- Threshold: This corresponds to the largest value of the summary statistics associated with the accepted parameter vectors.

Step 2 Only the best vectors previously selected are used in this step, and are perturbed. The perturbation is performed by adding to each coordinate of the vector a randomly chosen value in the interval $[-0.01, +0.01]$ and by doing a normalisation. A new distance value is calculated for the perturbed vector. If the new distance is lower than the threshold, the perturbed vector is conserved.

The ABC-SMC procedure is composed of $R > 1$ rounds. For each of round, a tolerance value is defined which determines the percentage of parameter vectors to be accepted.

The final set of accepted parameter vectors is the result of the ABC-SMC procedure and characterises the list of vectors that may explain the evolution of the pair of host and symbiont trees given as input. Observe that, since in all experiments a uniform prior distribution is assumed and also the perturbations are performed in a uniform way, the weights induced by the proposals will also appear to be uniform [6]. However, in the case of a different prior, weights should be used in the process in order to correct the posterior distribution according to the perturbation made.

Very recently, Alcala *et al.* [1] developed a new ABC framework to infer the rates of host switch and cospeciation. The authors indicate that their method allows to consider that a symbiont may be associated to more than one host. This is called parasite speciation as a generalist (Figure 1 in the paper). The ABC method itself resembles what was done in Coala. The summary statistics used is not a metric between trees as in Coala but corresponds instead to the summary classically adopted in ABC methods. The host and symbiont trees together with the leaf associations are represented as a network and 32 ecological measures (such as modularity, connectivity, etc.) are used for such summary to compare the simulated networks to the real one.

1.4 Questions Addressed in this Ph.D.

The topic of my Ph.D. was the development of models and algorithms to study the evolutionary history between hosts and symbionts. The main goal was to refine the existent reconciliation models and to make them more realistic. To this purpose, it was important to introduce some events that were little or not formally considered in the literature. One of them is the spread, which corresponds to the invasion of different hosts by a same symbiont [8]. In this case, as when spreads are not considered, the optimal reconciliations obtained will depend on the choice made for the costs of the events. The need to develop statistical methods to assign the most appropriate ones therefore remains of actuality.

Before considering such extended models however, one may question the robustness of any adopted method for, given a model, find all optimal reconciliations, under editing or small perturbations of the input or when some types of errors may be present in the input.

This Ph.D focused on these two aspects starting with the second:

- Robustness of the model: The objective here was to understand the strengths and weaknesses of the parsimonious model, and how the final results may be influenced when small errors are present, or are introduced in the input datasets. This may correspond either to

a wrong choice of present-day symbiont-host association in the case where multiple ones exist, or to small errors related to a wrong rooting of the symbiont tree.

- More realistic cophylogenetic model: The objective in this case was to consider spread events that are observed in nature but that are not taken into account in most models, or are done so in limited ways.

1.4.1 Robustness of the model

In EUCALYPT, a parsimonious model is used. An important issue related to this is that it makes a strong assumption on the input data which may not be verified in practice. The robustness of an algorithm is its ability not to be affected by "small" changes in the input. We examine two cases where this situation happens. The first is related to a limitation in most of the currently available methods for tree reconciliation where the association ϕ of the leaves is in general required to be a function. A leaf s of the symbiont tree can therefore be mapped to at most one leaf of the host tree. This is clearly not realistic as a single symbiont species can infect more than one host. We henceforth use the term *multiple association* to refer to this phenomenon. For each present-day symbiont involved in a multiple association, one is thus often forced to choose a single one. Clearly, this may have an influence on the solutions obtained.

The second case addresses a different type of problem related to the phylogenetic trees of hosts and symbionts. These indeed are assumed to be correct, which may not be the case already for the hosts even though these are in general eukaryotes for which relatively accurate trees can be inferred, and can become really problematic for the symbionts which most often are prokaryotes and can recombine among them [52, 73, 76]. We decided to deal for now with one single type of error, that corresponds to the problem of correctly rooting a phylogenetic tree. Many phylogenetic tree reconstruction algorithms in fact produce unrooted trees [52, 70, 76]. As we saw, 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 [29, 63, 73]. A wrong rooting of the trees given as input may lead to an incorrect output.

The goal of this study was, 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.* [5, 10]).

This work is described in Chapter 2. It was initially presented in a paper [81] that was accepted at the *International Conference on Algorithms for Computational Biology* (*AlCoB* 2016) whose proceedings appeared in LNBI/LNCS. A full version of the paper is currently under revision in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (*TCBB*) [82].

1.4.2 More realistic cophylogeny model

Most often, the coevolutionary models considered in the literature address only cases where a symbiont can be associated with at most one host, following the one-host-per-symbiont assumption [68]. In nature, however, a same symbiont can be associated to different hosts.

In order to deal with this problem, we modified both the model and the algorithm used in Coala [5]. In this case, the relation ϕ from $L(S)$ to $L(H)$ need no longer to be a function. It can be one-to-many, or many-to-many and not only one-to-one or many-to-one. It may thus be viewed as a bipartite network on the set $L(S) \cup L(H)$ and encoded as a binary adjacency matrix ϕ with size $|L(S)| \times |L(H)|$. In this matrix, the entry ϕ_{sh} is 1 whenever the symbiont s is associated to the host h, and zero otherwise.

This required introducing the new event of *spread*. Two kinds of spreads are considered: vertical and horizontal.

We recall that Coala includes two main steps, that were indicated above in Section 1.3.5.

We changed the first step to simulate the evolution of the symbiont, where the frequency of the events (cospeciation, duplication, host switch and loss) are calculated (Figure 1.2 shows the different events). At the beginning of this step, the probability of a vertical or of a horizontal spread is calculated for each vertex of H. Thus $p_{vert-spread}$ (probability of a vertical spread) and $p_{hor-spread}$ (probability of a horizontal spread) are estimated based on the topology of H and the association ϕ of the leaves. We changed also the second step, introducing a new distance that we called MAS and denoted by d_{MAS} to compare the simulated to the real symbiont trees even in this case of multiple associations.

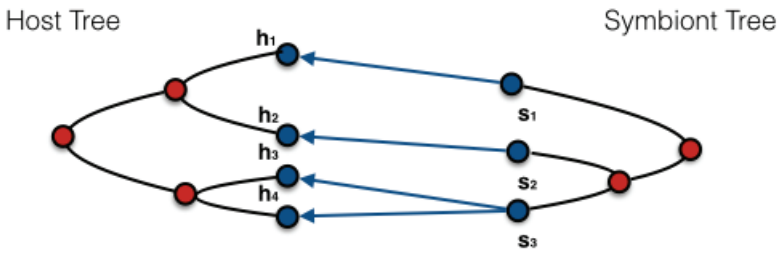


Figure 1.9: A tanglegram with a host-symbiont system including multiple associations. The host tree is on the left, the symbiont tree on the right and the associations are represented with blue arrows.

This work is described in Chapter 3. A paper is in preparation that will be submitted to a journal before the end of 2017.

Chapter 2

Robustness of the Model

Contents

2.1 Overview

A phylogenetic tree reconciliation makes strong assumptions on the input data which may not be verified in practice. In this chapter, we examine two cases where this situation happens.

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

The second case addresses a different type of problem related to the phylogenetic trees of hosts and symbionts. These indeed are assumed to be correct, which may not be the case already for the hosts even though these are in general eukaryotes for which relatively accurate trees can be inferred, and can become really problematic for the symbionts which most often are prokaryotes and can recombine among them [52, 73, 76]. This is the problem of correctly rooting a phylogenetic tree. Many phylogenetic tree reconstruction algorithms in fact produce unrooted trees [52, 70, 76]. 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 [29, 63, 73]. A wrong rooting of the trees given as input may lead to an incorrect output.

The organisation of the chapter 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.

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 [5, 10]. 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

$$\mathcal{C} = \{\langle -1, 1, 1, 1\rangle, \langle 0, 1, 1, 1\rangle, \langle 0, 1, 2, 1\rangle, \langle 0, 2, 3, 1\rangle, \\ \langle 1, 1, 1, 1\rangle, \langle 1, 1, 3, 1\rangle\}.$$

The implemented methods are included in the software previously developed in the team, Eucalypt [18], that is available at this web page: `http://eucalypt.gforge.inria.fr/`.

2.2 Material

2.2.1 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.* [24]. 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.* [71, 72] 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 [22]. 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.* [48]. 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.* [83]. 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.* [16]. 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 [47]. 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.* [3]. 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.* [51]. 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 [26]. 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.* [46] 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.* [69] 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.* [42]. 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.* [40] 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.* [34] 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.* [34] 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 [35]. 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.* [64]. 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 [85]. 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 [36] 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.* [57]. 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.* [60]. 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.* [65]. 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 [38]. 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 [38] 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 [58]. 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.* [84]. 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:

$$M(H, S, \phi) = \sum\nolimits_{s \in L(S)} |\phi(s)| - 1. \tag{2.1}$$

Table 2.1 shows the number of multiple associations in the datasets where it is non null.

Figure 2.1 shows the multiple associations for the MP dataset. MP has five symbiont leaves $s \in L(S)$ with multiple associations *{Phengaris_ atroguttata,rebeli,alcon,nausithous,teleius}*. If we apply Equation 2.1, the value of M is 8. For example, the symbiont leaf *Phengaris_ atroguttata* is associated to the host leaves *formosae* and *arisana*, and $\phi(Phengaris\_atroguttata) = \{formosae, arisana\}$. There are therefore 2 simple association or 1 multiple associations for *Phengaris_ atroguttata*.

Table 2.1: List of datasets exhibiting multiple associations, the number $M(H, S, \phi)$ of such multiple associations as in Equation 2.1 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
$M/\|L(H)\|$ (%)	244	21	27	83	29	75	36	109	100	55	94	27	7	13	15

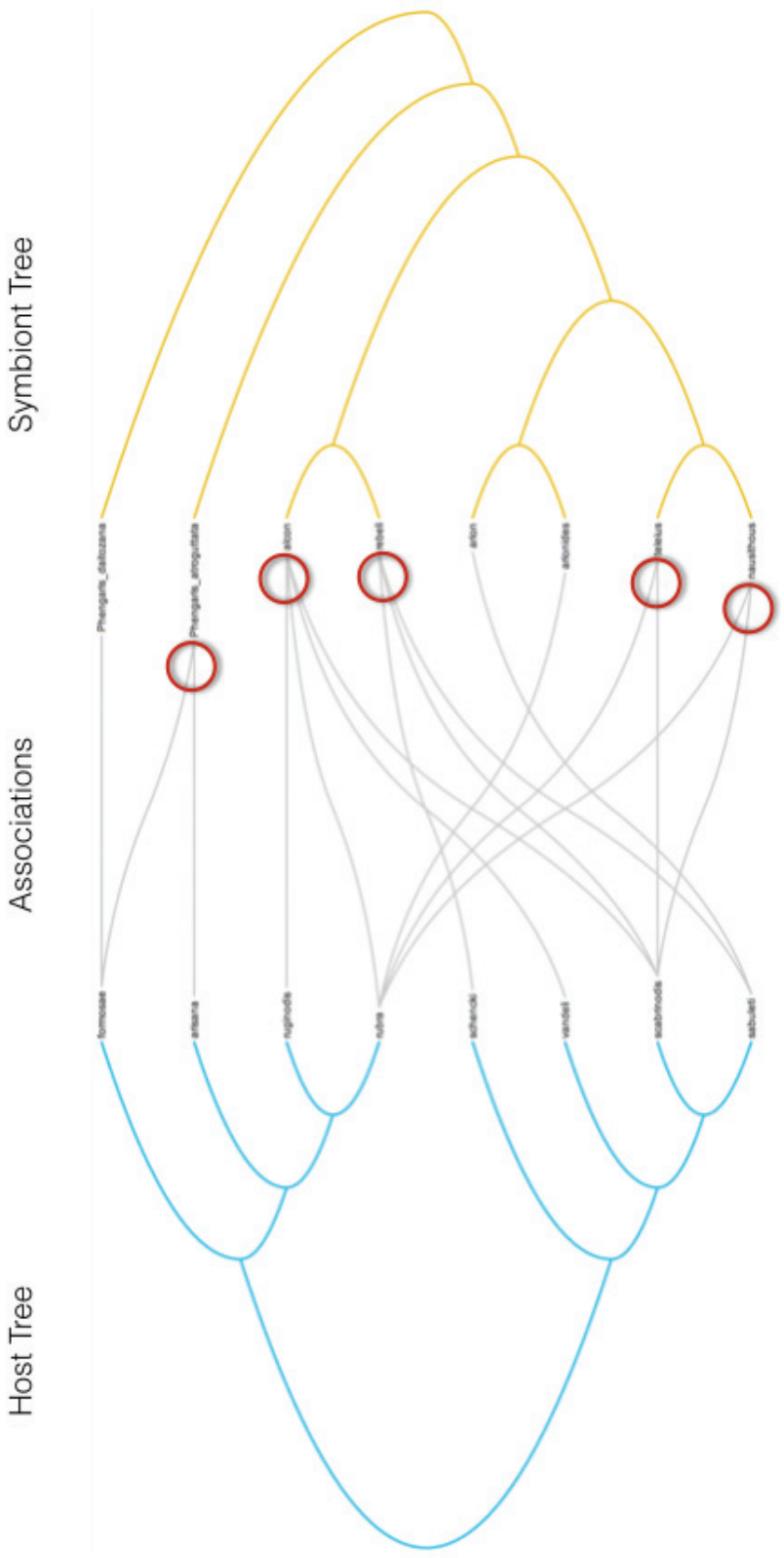


Figure 2.1: Dataset MP. Figure obtained with TreeMap 3 (https://sites.google.com/site/cophylogeny/treemap/using-treemap). In blue the host tree, in yellow the symbiont tree, in grey the associations between leaves. The red circles show the symbiont leaves that have multiple associations.

2.2.2 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.* [21]. 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 2.1, 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.

2.2.3 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 [5], 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.* [5] 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).

2.3 Methods

2.3.1 Generating All the Optimal Solutions

We used Eucalypt [18], 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.

2.3.2 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 (an example is shown in Figure 2.2). 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 Figure B.1 in the Appendix B. By choosing in all possible ways among the multiple associations, we thus obtain 560 datasets.

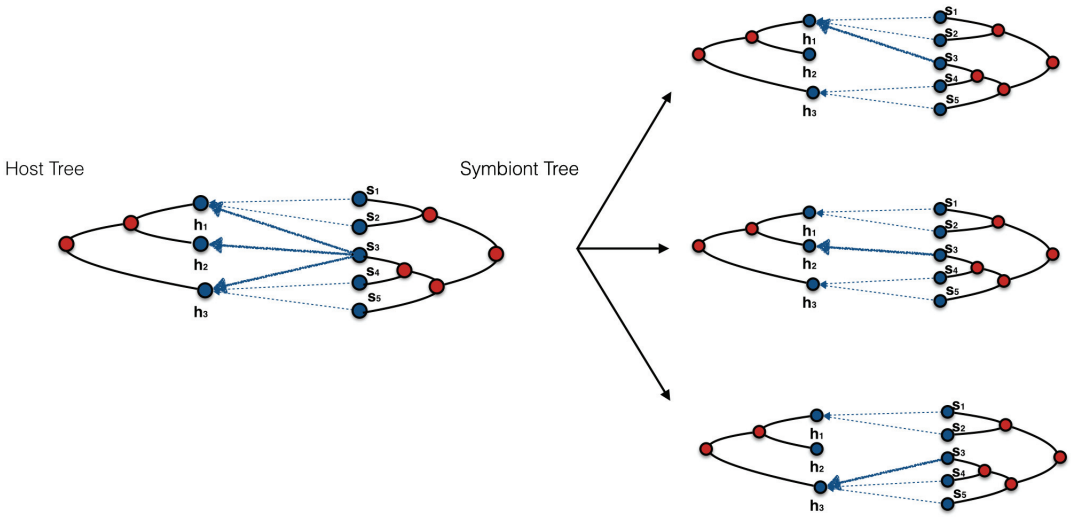


Figure 2.2: On the left, a dataset with multiple associations. There are three multiple associations. On the right, the datasets obtained by resolving multiple associations in all possible ways.

2.3.3 Re-Rooting of the Symbiont Tree

Most phylogenetic reconstruction algorithms produce unrooted trees, or rooted ones that have an unreliable root [29]. 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. In Figure 2.3, an example is shown of the re-rooting of a given dataset. We have on the left a host tree H and a symbiont tree S with 5 leaves; in this case $k = 3$. We create all possible datasets obtained with the different root (shown on the right-hand side).

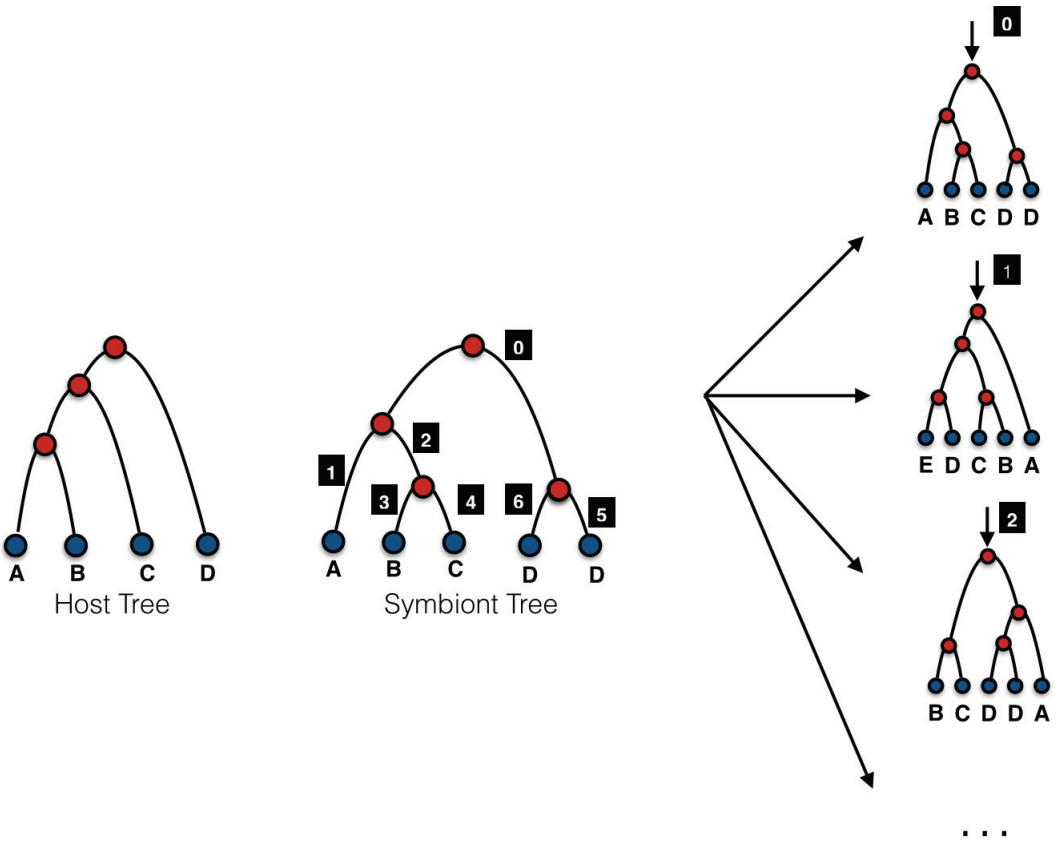


Figure 2.3: On the left, the host and symbiont trees. On the right, some of the re-rooted symbiont trees.

2.3.4 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.* [25] 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.* of 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.

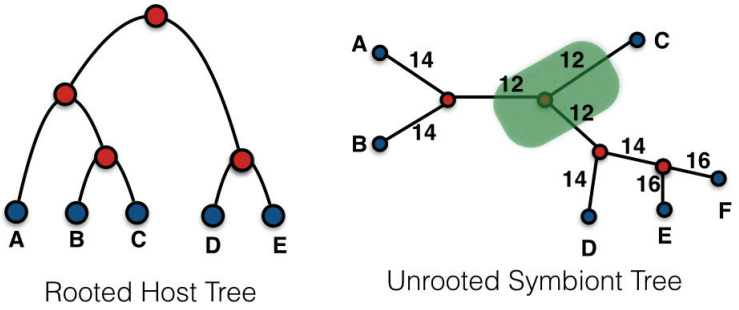


Figure 2.4: Example of plateau property. On the left, a rooted host tree. On the right, an unrooted symbiont tree. The number associated to each edge corresponds to the cost of an optimal reconciliation. The plateau is in green and shows the minimum cost values obtained by other re-rootings. The edges with minimum reconciliation cost are close each other. The edge values in any path from a plateau towards a leaf are monotonically increasing.

2.3.5 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 [43]), 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 by $v_\Lambda = \sum_{\pi \in \Lambda} \pi$. 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:

$$d(\Lambda_1, \Lambda_2) = \frac{||v_{\Lambda_1} - v_{\Lambda_2}||}{(|\Lambda_1| + |\Lambda_2|) \max_{\pi \in \Lambda_1 \cup \Lambda_2} ||\pi||} \tag{2.2}$$

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).

2.3.6 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 by $opt(D_j, c)$ 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 $D, D' \in \mathcal{D}^*$.

2.3.7 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. (2.2). 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.

2.3.8 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.* fixing the topology of the tree H and considering the tree H' given by a permutation of the labels of its leaves (similarly for S). The association 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.

2.4 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 of the results appears in the Appendix B. Notice that the time-unfeasible reconciliations have been filtered-out.

2.4.1 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 [57] tried to maximise the number of cospeciations between hosts and symbionts but found out that sometimes host switches must be postulated to maximise cospeciations. 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 B.2.1 in the Appendix B.

Figure 2.6 (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. (2.2). Figure 2.6 (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 2.5 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.

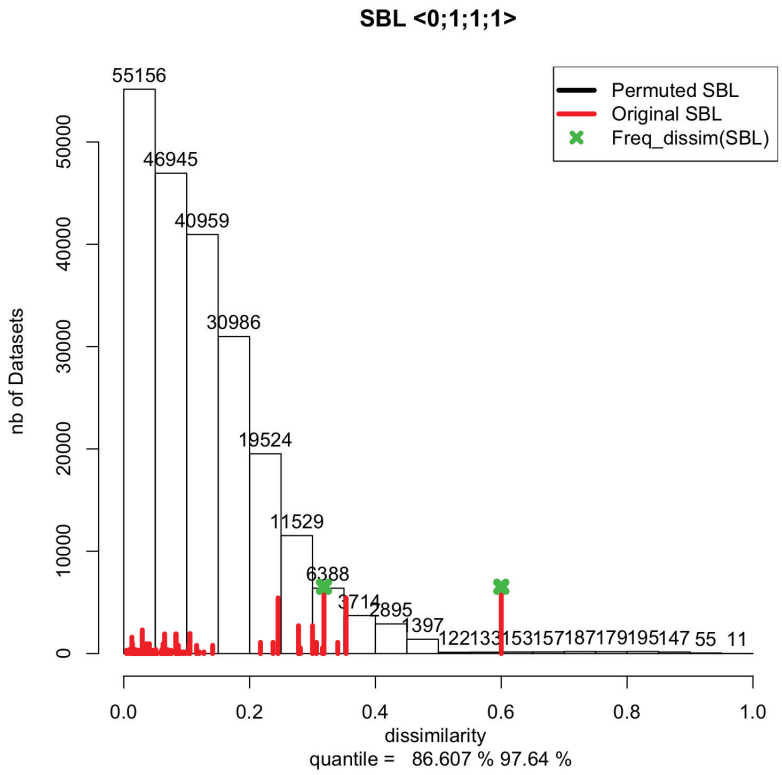


Figure 2.5: Histogram of dissimilarity derived from SBL dataset with the cost vector $\langle 0, 1, 1, 1\rangle$. The black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset SBL. The green crosses are the $freq_{dissim}(SBL)$.

Still considering the SBL dataset, now for the other cost vectors c (Figure B.87 in Sec-

tion B.2.2 in the Appendix B), 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 B.2.2 in the Appendix B), 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 2.2.

Table 2.2: 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)

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 2.7, 2.8 and 2.9 are similar to Figure 2.6 (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 B.2.3 in the Appendix B), 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 $SBL_{x\%}$ (Section B.2.3 in the Appendix B). 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.

2.4.2 Re-Rooting of the Symbiont Tree

Testing the Plateau Property

Table B.1 and Table B.2 in the Appendix B 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 B.1 in the Appendix B). 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 B.3 and Table B.4 in the Appendix B). 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).

Re-rooting at Distance k

We show in Figure 2.10 the results obtained with the biological dataset MP. Similar figures are presented with other biological datasets in Section B.2.5 in the Appendix B. Here the dissimilarity of the reconciliation globally increases as k also increases. The farther is the new root from the

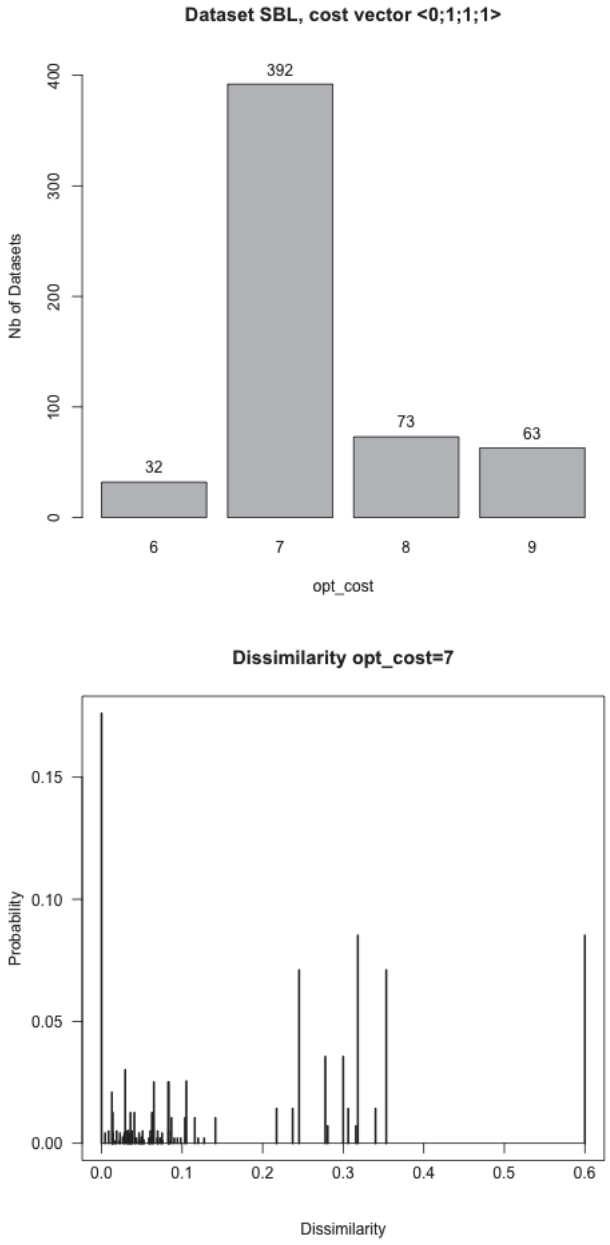


Figure 2.6: Barplots of optimum cost (top) and dissimilarity between pairs of reconciliations with optimum cost 7 (bottom) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$.

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 B.2.6 in the Appendix B). 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.

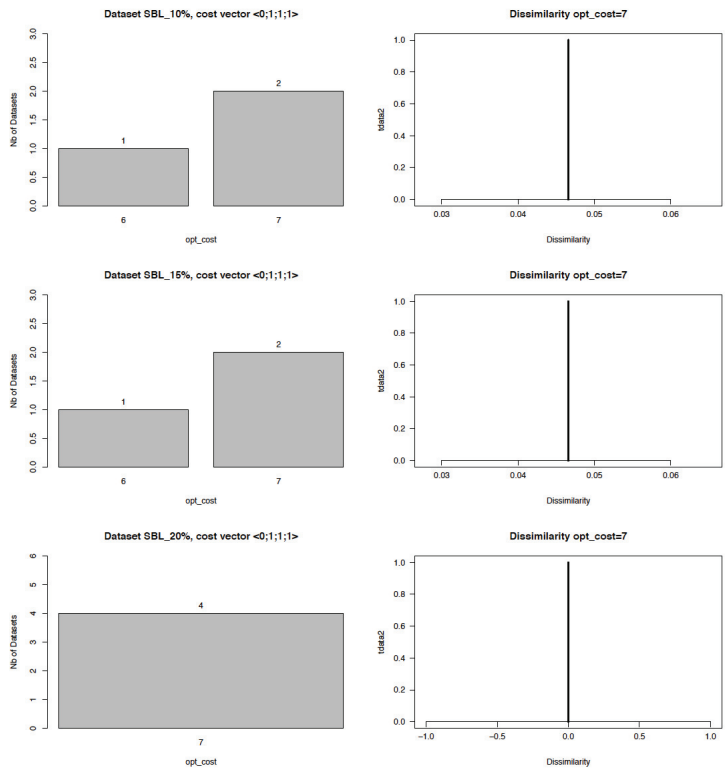


Figure 2.7: Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimum cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each lines is a different $SBL_{x\%}$ with $x = 10, 15, 20$.

2.5 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,

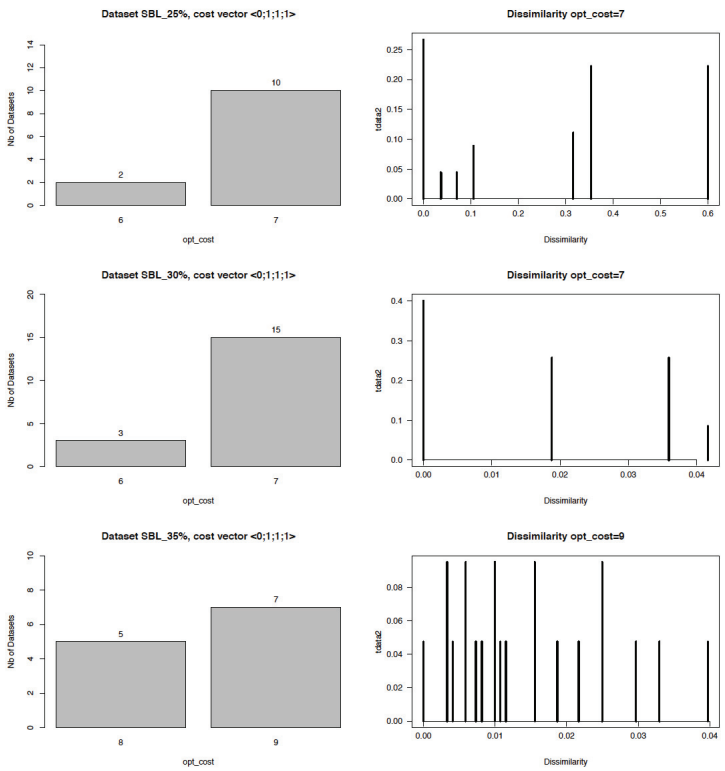


Figure 2.8: Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimum cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each lines is a different $SBL_{x\%}$ with $x = 25, 30, 35$.

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 [25]. 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

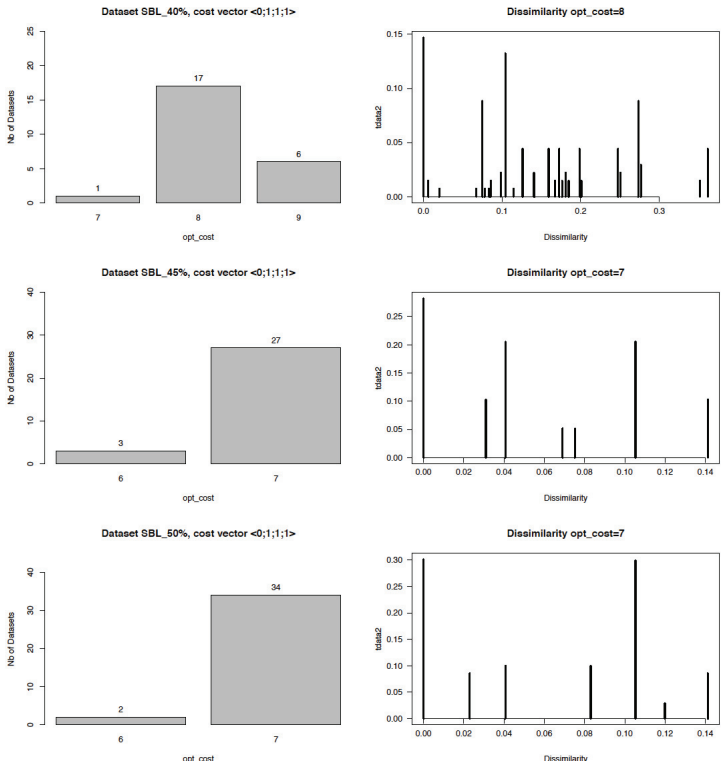


Figure 2.9: Barplots of optimum cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimum cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each lines is a different $SBL_{x\%}$ with $x = 40, 45, 50$.

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.

We end by noticing that since our paper [81], appeared, a more recent study was published in the context of gene-species reconciliation (this appeared from now as an Abstract in *ISBRA 2017* with the full version available at this address: `http://compbio.engr.uconn.edu/papers/Kundu_RootingUncertainty_2017.pdf`). By analyzing a data set of over 4500 gene families from 100 species the authors showed that an important fraction of gene trees have multiple optimal rootings, with the roots often, but not always, appearing clustered together in a same region of the gene tree with other aspects of the reconciliation also remaining conserved across the different rootings.

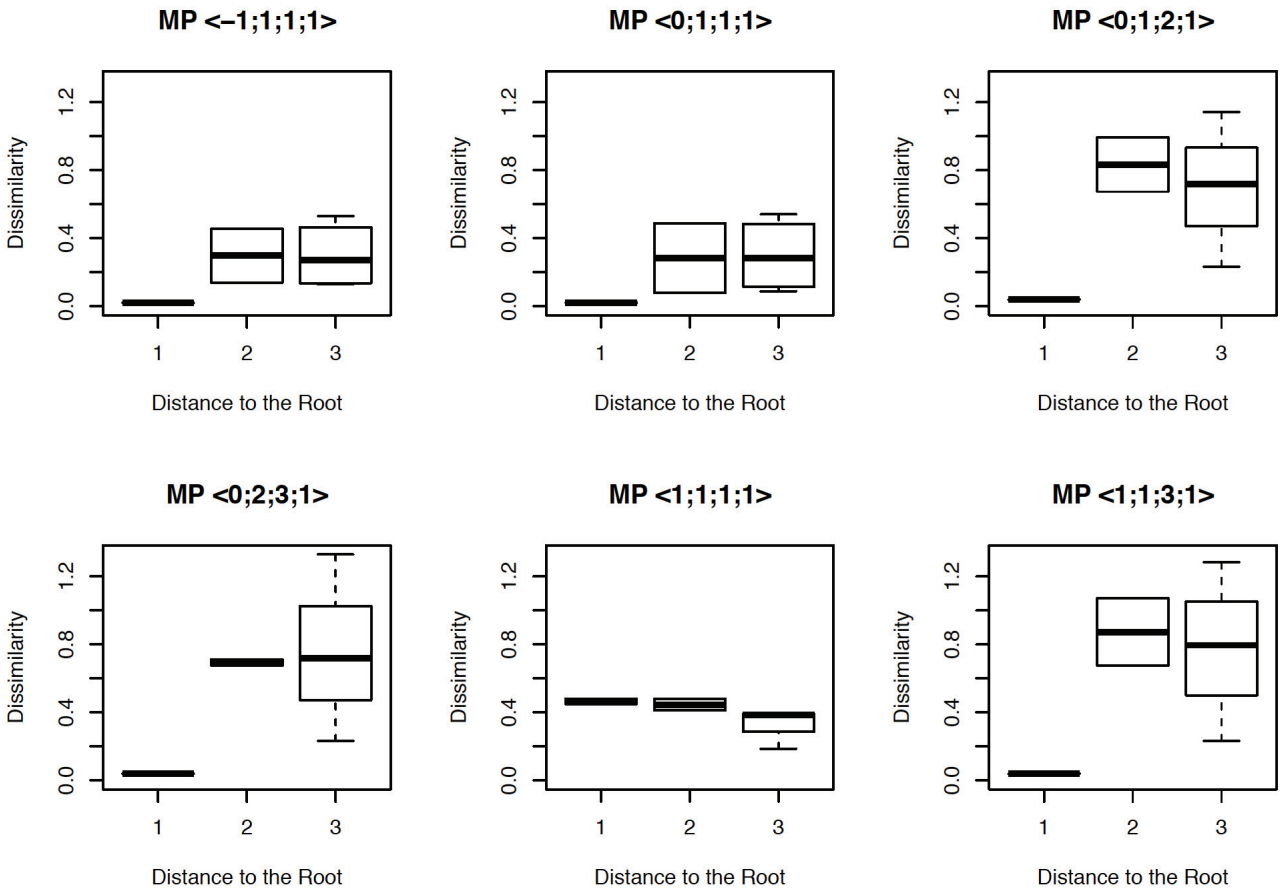


Figure 2.10: Boxplots of the dissimilarities between reconciliations obtained for the original dataset MP and all datasets obtained from MP by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between new and original root. The y-axis shows the value d of the dissimilarity of the reconciliation patterns.

Chapter 3

More Realistic Cophylogenetic Model

Contents

3.1 Overview

The cophylogeny methods present in the literature do not handle (or they do so only partially) the cases where a symbiont is associated to more than one host. In nature, however, such multiple associations are often observed. Such associations introduce a type of events that have been called *spread* in the literature. To the best of our knowledge, the first use of such term was by Brooks and McLennan in [8].

Among the methods for dealing with host-symbiont cophylogeny that enumerate all optimal reconciliations, four address the problem in the case where the input datasets include symbionts that are associated to more than one host. These are CoRe-Pa [50], Jane 4 [13], WiSPA (see the arXiv file at: https://arxiv.org/abs/1603.09415), and the approach of Alcala *et al.* [1]. CoRe-Pa solves the multiple associations locally, by starting from the leaves that are already mapped and choosing for a parent vertex the unique associations of its children that give the best cost. Jane 4 uses a heuristic approach based on a genetic algorithm to recover the best solutions. Furthermore, the spread events in this case appear to be allowed only just before the leaves. WiSPA is a new model for reconciling trees where the symbionts are permitted to be associated with more than one host which however apparently uses also a restricted form of spread events. Finally, very recently, Alcala *et al.* [1] developed a new ABC framework to infer the rates of host switch and cospeciation. The authors indicate that their method allows also to consider that a symbiont may be associated to more than one host. This is called parasite speciation as a generalist (Figure 1 in the paper). The part of the paper related to this is however not detailed in the paper.

In this chapter, we describe how we extended Coala to create a new model, that we called AmoCoala, to address the problem of multiple associations.

Coala (for "CO-evolution Assessment by a Likelihood-free Approach") [5] is an algorithm for inferring a rate for each of the four macro-evolutionary events most used in the literature, namely cospeciation, duplication, host switch and loss. For a given pair of host and symbiont trees, Coala estimates the frequency of the events based on an approximate Bayesian computation (ABC) approach that may be more efficient than a classical likelihood approach [7].

Coala includes two main parts: the first simulates the evolutionary history of symbionts while the second uses ABC in order to select the most probable frequency of the four events: cospeciation, duplication, host switch and loss.

In the case of AmoCoala, ϕ is no longer a function from $L(S)$ to $L(H)$ since a symbiont may now be associated with multiple hosts. Instead, ϕ can be viewed as a bipartite network on the set $L(S) \cup L(H)$ and is encoded as a binary adjacency matrix ϕ with size $|L(S)| \times |L(H)|$. In this matrix, the entry ϕ_{sh} is 1 whenever symbiont s is associated to host h, and zero otherwise.

Two kinds of spread events are considered: vertical and horizontal. The first case corresponds to what could be called also 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. The second includes both an *invasion*, of the symbiont which remains with the initial host but at the same time gets associated ("invades") another one incomparable with the first, and 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".

We changed the first step of Coala to simulate the evolution of the symbiont, where the

frequency of the events (cospeciation, duplication, host switch and loss) are calculated. At the beginning of this step, the probability of vertical and horizontal spreads, respectively denoted by $p_{\text{vert-spread}}$ and $p_{\text{hor-spread}}$, are computed for each vertex of H. These are estimated based on the topology of H and the association ϕ of the leaves. The second step was also modified by introducing a new distance, that we called *MAS*, to compare the simulated to the real symbiont trees in this case where we have multiple associations.

We start by briefly recalling how the algorithm Coala works before presenting the methods and distance introduced in AmoCoala for handling vertical and horizontal spread events. We then indicate the datasets that will be used to test the method, both synthetic and real ones, as well as in the first case, the method to generate them. We conclude by showing and discussing the results obtained by AmoCoala on these datasets.

This chapter presents an advanced draft of a paper that will be submitted before the end of 2017.

3.2 Model

In order to make this Ph.D. self-contained, we start by first describing the existing model, Coala, which served as starting point for our new one, AmoCoala, which is introduced next. The parts of this section related to Coala are based on the paper of Baudet *et al.* [5] and on the Ph.D. of Beatrice Donati [17].

3.2.1 Coala Model

Coala [5] is an algorithm that, given a host tree H, a symbiont tree S, and the associations between leaves $\phi : L(S) \mapsto L(H)$, estimates the vector of frequencies of the macro-evolutionary events of cospeciation, duplication, loss and host switch, using an approximate Bayesian computation (ABC) approach. We recall here what these events stand for: a) cospeciation is when the parasite diverges in correspondence to the divergence of a host species; (b) duplication is when the symbiont diverges "without the stimulus of host speciation" [59]; (c) host switch is when a symbiont switches, or jumps from one host species to another independent of any host divergence; and (d) loss 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.

The parameter vector θ of the model is thus composed of the probabilities $\langle p_c, p_d, p_s, p_l \rangle$ of each one of these four events.

Following the topology of H and a given vector θ, a simulated symbiont tree $\tilde{S}$ is created. At the same time as the simulation of $\tilde{S}$, a function λ that associates each vertex of $\tilde{S}$ to a vertex of H is also created. Once a symbiont tree $\tilde{S}$ is simulated, a distance summary between $\tilde{S}$ and S is computed.

Evolution of symbionts in Coala

The evolution of the symbionts is simulated by following the evolution of the hosts traversing the phylogenetic tree H from the root to the leaves, and progressively constructing the phylogenetic tree for the symbionts. The probabilities of the parameter vector θ are used to define the type of mapping chosen. In this process, a symbiont vertex can be in two different states: mapped or unmapped. At the moment of its creation, a new vertex v is unmapped and is assigned a temporary position on an arc a of the host tree H. We denote this position by $\langle v, a\rangle$. Vertex v is mapped to a vertex w of H (for cospeciation, duplication and host switch). For the three events except the loss case, v is always mapped to the vertex $h(a)$. This mapping is denoted by $[v : w]$ with $w = h(a)$.

In the cases of cospeciation, duplication, and host switch, the symbiont is supposed to speciate and two children are created for v, denoted by v_1 and v_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 v is created (at this step) since there is no symbiont speciation, and v is just moved to one of the two arcs outgoing from $h(a)$ chosen randomly. These choices, together with the general framework for the symbiont tree generation method, are provided next.

Evolutionary events

For any vertex s of $\tilde{S}$ that is not yet mapped and whose position is $\langle s, a\rangle$ (Figure 3.4 (b)), Coala chooses to apply one among the four allowed operations, depending on the probability of each event. In what follows, we denote by a_1, a_2 the arcs outgoing from the head $h(a)$ of the arc a.

- Cospeciation (Figure 3.4 e)): We apply the mapping $[v : h(a)]$ and we create the vertices v_1 and v_2 as children of v. We position them as follows: $\langle v_1, a_1\rangle$ and $\langle v_2, a_2\rangle$. This operation is executed with probability p_c.

- Duplication (Figure 3.4 f)): We apply the mapping $[v : h(a)]$ and we create the vertices v_1 and v_2 as children of v. Both v_1 and v_2 are positioned on a. This operation is executed with probability p_d.

- Host switch (Figure 3.4 g)): We apply the mapping $[v : h(a)]$ and we create the vertices v_1 and v_2 as children of v. We then randomly choose one of the two children and position it on a. Finally, we randomly choose an arc a' that does not violate the time feasibility of the reconstruction so far [75]. If such an arc does not exist, it is not possible for a host switch to take place. In this case, we choose between the three remaining events with probability $p_i/(p_c + p_d + p_l)$ with $i \in \{c, d, l\}$. Otherwise, we position v_2 on a'. This operation is executed with probability p_s.

- Loss (Figure 3.4 h)): This operation is executed with probability p_l and consists of randomly choosing an arc outgoing from the head $h(a)$ of a and positioning v on it. Observe that we are considering only losses resulting from lineage sorting. It would be interesting to incorporate extinction or failure to detect infection but this would require the addition of new parameters, thus making the model more complex to analyse. However, if v was created by a duplication event and is being processed for the first time, we have to verify

if its sibling vertex v' was already processed and also suffered a loss. In this case, v must be positioned on the same arc a' where v' was positioned. This procedure is adopted to avoid later mappings where a duplication followed by two losses would be confused with a cospeciation.

3.2.2 AmoCoala model

As mentioned at the start of the chapter, when a symbiont is associated to more than one host, the association ϕ is no longer a function from $L(S)$ to $L(H)$ but may rather be viewed as a bipartite network on the set $L(S) \cup L(H)$. It can then be encoded as a binary adjacency matrix ϕ with size $|L(S)| \times |L(H)|$. In this matrix, the entry ϕ_{sh} is 1 whenever symbiont s is associated to host h, and zero otherwise. Before explaining how we handle this situation and the event of spreads, vertical and horizontal (or if one prefers, of freeze and invasion followed by a double freeze), we introduce some basic definitions.

Basic definitions

Given a dataset (H,S,ϕ), the *number of multiple associations*, denoted by MA, for such dataset is given by:

$$MA(H, S, \phi) = \sum\nolimits_{s \in L(S)} |\phi(s)| - 1, \tag{3.1}$$

where $\phi(s) = \{h \in L(H); \phi_{sh} = 1\}$.

Given a set-labelled tree T, that is a tree whose leaves are labelled with a set of labels, we denote its *weight* by:

$$w(T) = \sum\nolimits_{v \in L} |l(v)|, \tag{3.2}$$

where $l(v)$ is the set of labels of leaf v. If T is a symbiont tree S, then $l(v)$ is equal to $\phi(v)$.

The χ^2 distance that we use to compare two vectors $\theta = \langle p_0, p_1, \ldots, p_n \rangle$ and $\theta' = \langle q_0, q_1, \ldots, q_n \rangle$ is a weighted Euclidean distance defined as follows:

$$d(\theta, \theta') = \sqrt{\sum_{i=1}^{n} 2 \times \frac{(p_i - q_i)^2}{(p_i + q_i)}}. \tag{3.3}$$

For a set-labelled tree T, we recall that we denote by T_v the subtree of T rooted in v. An operation performed on phylogenetic trees may be associated with the removal of some elements from a set, or followed by, a cleaning of the vertices of out-degree 1 if any were created in order to obtain once again a phylogenetic tree. An example is given in Figure 3.1.

Given a set-labelled tree T, we say that a set T^* is a *subtree* of T, if the following holds: (i) there exists a function $f : L(T^*) \to L(T)$ such that f is injective (*i.e.*, for $v_1, v_2 \in L(T^*)$, if $v_1 \neq v_2$ then $f(v_1) \neq f(v_2)$), (ii) for every $v \in L(T^*)$, $l(v) \subseteq l(f(v))$, and (iii) T^* can be obtained by cleaning the leaves of T that are not in the images of f.

Given two set-labelled trees T_1 and T_2, we say that T^* is *an agreement subtree* if T^* is a subtree of both T_1 and T_2. A *maximum agreement subtree* is an agreement subtree of maximum weight (an example is given in Figure 3.2).

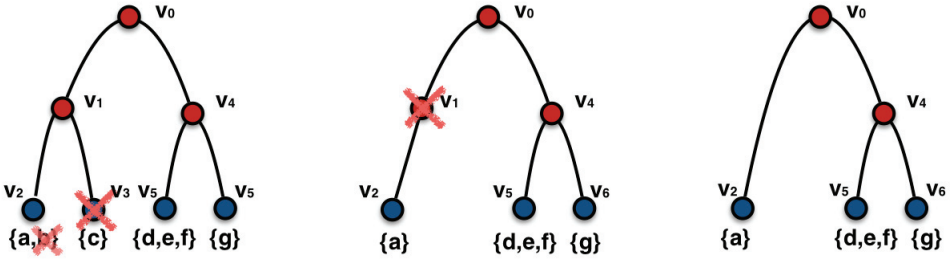


Figure 3.1: Example of operation of cleaning a tree. Following some operation, label b from the leaf v_2 was removed as was leaf v_3. The latter further necessitates that vertex v_1, which is of out-degree 1, is taken out.

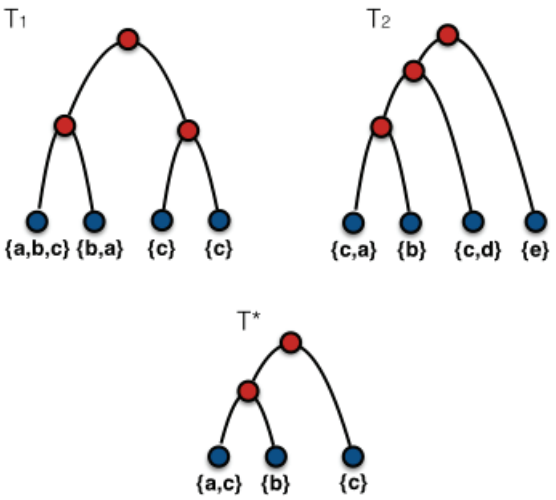


Figure 3.2: Example of a maximum agreement subtree for a set-labelled tree. T^* is the agreement subtree of the set-labelled trees $T1$ and $T2$ that has the maximum weight.

Spread events

To make the model more realistic by being able to handle multiple associations, we define two types of spreads: vertical and horizontal.

A vertical spread happens when a symbiont s is associated to a host h and to all the vertices in the subtree of h. The difference between cospeciation and a vertical spread event is that, in the first case the symbiont s speciates into two new species of symbionts s_1 and s_2, while in the second it does not.

A horizontal spread happens when a symbiont s is associated to two incomparable hosts h_i and h_j and all the vertices in the subtrees of h_i and h_j are also associated to the same symbiont s. The difference between host switch and a horizontal spread event is that, in the first case the symbiont s speciates into s_1 and s_2 and each is associated to a distinct host vertex with the two being incomparable, while in the second the same symbiont s is associated to two incomparable host vertices.

The probabilities of the "classical" events (cospeciation, duplication, host switch and loss) are parameters to be inferred. **On the contrary, the probabilities for the spread events**

are given *a priori*.

We now define the different probabilities that will be used in our new algorithm. Observe that each spread event has its own probability.

A probability $p_{\text{vert-spread}}(h)$ is associated to a vertical spread event at host h as follows. For each $h \in L(H)$, we set $p_{\text{vert-spread}}(h)$ to 1. For internal vertices h of the host tree H, the probability $p_{\text{vert-spread}}(h)$ is given by:

$$p_{\text{vert-spread}}(h) = \left(\frac{1}{|S^{L(h)}|}\right) \frac{\sum_{s \in S^{L(h)}} |\phi(s) \cap L(h)| - 1}{|L(h)| - 1}, \tag{3.4}$$

where $L(h)$ is the set of leaves in the subtree of H rooted in h, the set of symbionts in the leaves of this subtree is $S^{L(h)}$ and the number of leaves in this subtree infected by a symbiont s is $|\phi(s) \cap L(h)|$.

Observe that the "classical events" (of cospeciation, duplication, host switch and loss) have the same probability everywhere in the tree, while the probability of a vertical spread is specific to each vertex of the host tree. This probability $p_{\text{vert-spread}}(h)$ is large whenever most of the symbionts are associated to almost all the leaves $L(h)$ in the subtree rooted in h. On the contrary, the probability is low when most of those symbionts appear only in a few of the leaves. In particular, whenever all the symbionts associated to the leaves below h are associated to only one of those leaves, the probability $p_{\text{vert-spread}}(h)$ is zero.

For two incomparable vertices h_i and h_j, a probability $p_{\text{jump}}(h_i \to h_j)$ is computed as follows:

$$p_{\text{jump}}(h_i \to h_j) = \frac{|S^{L(h_i)} \cap S^{L(h_j)}|}{|S^{L(h_i)} \cup S^{L(h_j)}|}. \tag{3.5}$$

Here again, the probability of a jump (we remind that jump and host switch are synonym) is specific to each pair of vertices of the host tree. This quantity is symmetric. The probability $p_{\text{jump}}(h_i \to h_j)$ is high whenever the leaves of the subtrees below h_i and h_j have many associated symbionts in common. In particular, it is zero when they do not share any associated symbiont, and 1 when they have exactly the same set of associated symbionts.

From these probabilities $p_{\text{jump}}(h_i \to h_j)$ of jumps, we construct a probability of horizontal spread at each vertex h_i. The associated probability depends on all the vertices that are incomparable with h_i. Indeed, such vertices are all those that may be reached from h_i. In practice, a horizontal spread corresponds to a jump combined with two vertical spreads. We thus associate a probability of spread $p_{\text{hor-spread}}(h)$ to each vertex h of the host tree that takes into account both horizontal and vertical spreads and is defined as:

$$p_{\text{hor-spread}}(h) = \min\{1, p^*(h)\}, \tag{3.6}$$

where

$$p^*(h) = p_{\text{vert-spread}}(h) \sum_{h';h,h' \text{ incomparable}} p_{\text{vert-spread}}(h') p_{\text{jump}}(h \to h').$$

This probability of horizontal spread $p_{\text{hor-spread}}(h)$ is high whenever $p_{\text{vert-spread}}(h)$ is high (so most of the symbionts associated below h are spread all over the leaves) and there are vertices h' incomparable to h with large $p_{\text{vert-spread}}(h)$ and large value $p_{\text{jump}}(h \to h')$ (so that the leaves below h and h' share a lot of associated symbionts). Observe that $p^*(h)$ is not a probability but

a value, that in particular may be bigger than 1.

Finally, if a horizontal spread happens at vertex h, we sample an incomparable vertex h' where the symbiont s has to jump to. We associate a probability $p_{\text{invasion}}(h \to h', \lambda)$ to thus be invaded to h and every incomparable vertex h'. For a current mapping λ of the vertices of S to the vertices of H, the probability of a vertex h' to be invaded from a symbiont s mapped in h is:

$$p_{\text{invasion}}(h \to h', \lambda) = \frac{p_{\text{jump}}(h \to h') 1\{E_{h,h',\lambda}\} p_{\text{vert-spread}}(h) p_{\text{vert-spread}}(h')}{p_{\text{vert-spread}}(h) \sum_{h'} p_{\text{vert-spread}}(h') p_{\text{jump}}(h \to h') 1\{E_{h,h',\lambda}\}}, \tag{3.7}$$

where $1\{E_{h,h',\lambda}\} = 1$ whenever the event induces a time feasible reconciliation, and the sum on the denominator is restricted to the vertices h' that are incomparable to h. If no vertex induces a time feasible reconciliation (namely $p_{\text{invasion}}(h \to h', \lambda) = 0$), the horizontal spread is not applied and another event is sampled.

Figure 3.3 shows an example of a dataset with horizontal and vertical spread probabilities. Here the host tree H has 5 leaves, the symbiont tree S has 4 leaves and there are 3 multiple associations.

Computing the spread probabilities

The spread probabilities are calculated at the beginning of the algorithm. These values depend only on the host tree H, the symbiont tree S and the associations ϕ. In a first step, we start by setting to 1 the probabilities $p_{\text{vert-spread}}$ for the leaves. Then, for the internal vertices h, these probabilities are computed as in Equation (3.4). In a second step, the probabilities of a jump are calculated for each pair of incomparable vertices h and h' as in Equation (3.5). In the last step, the probabilities of a horizontal spread for vertex h are computed as in Equation (3.6). Observe that the probabilities of invasion (Equation (3.7)) depend on the current simulation (one has to take into account acyclicity to choose the target h' of a horizontal spread). These are computed during the simulated algorithm, each time a horizontal spread is selected.

Evolution of the symbionts in AmoCoala

The evolution of the symbionts in AmoCoala enables to construct the simulated symbiont tree $\tilde{S}$. We follow a process similar to the one used in Coala, but we introduce the mapping of vertical and horizontal spread events.

As in Coala, at the moment of its creation, a new vertex v of $\tilde{S}$ is unmapped and is assigned a temporary position on an arc a of the host tree H. We first sample whether we have a horizontal spread event. If yes, then we map v to the vertex $h(a)$ that is the head of the arc a, and to all the descendants of $h(a)$. We denote this mapping by $\{v, H_{h(a)}\}$. We then choose an incomparable vertex $h(a')$ based on the probabilities given in Equation (3.7) and we map v to it and to all its descendants.

If we do not have a horizontal spread event, we sample whether a vertical spread occurs. If yes, we map v to the vertex $h(a)$ and to all its descendants as before.

In both cases, of vertical or of horizontal spread, the evolution of v stops and v becomes a leaf.

If we do not have a spread, we map v to a vertex w of H (cospeciation, duplication or host switch) or, in the case of a loss, we move v to another position. This last part is done exactly as

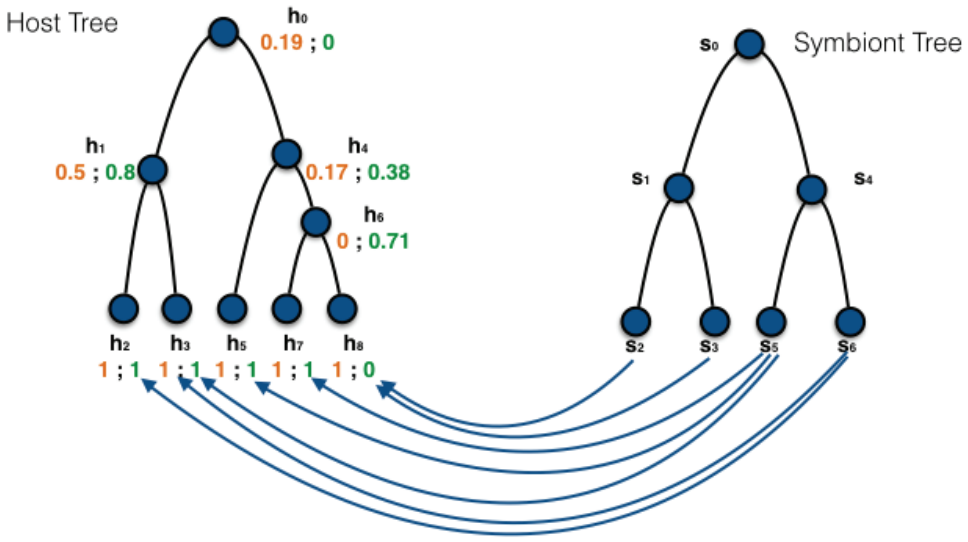


Figure 3.3: Probabilities of horizontal and vertical spreads. The host tree is to the left, the symbiont tree to the right and the associations are represented by the blue arrows. For each host vertex in the host tree are shown: in black the vertex name, in orange the value of $p_{\text{vert-spread}}(h)$ and in green the value of $p_{\text{hor-spread}}(h)$.

in Coala. In the first case, v is always mapped to the vertex $h(a)$. We denote this mapping by $[v : w]$ with $w = h(a)$.

If a spread event has been sampled, we construct a subtree below v (see details in the paragraph "Algorithm for simulating the evolution of the symbiont in the case of horizontal and vertical spreads" below) together with its mapping in H, and we stop the evolution of $\tilde{S}$ in this direction.

These choices, together with the general framework for our symbiont tree generation method, are provided next.

Starting the generation in AmoCoala

The generation of the simulated symbiont tree $\tilde{S}$ starts with the creation of its root vertex $\tilde{S}_{root}$. This vertex is positioned before the root of H on the arc $a = (\rho, H_{root})$. This allows the simulation of events that happened in the symbiont tree before the most recent common ancestor of all host species in H. Figure 3.4a) depicts this initial configuration. For any vertex v of $\tilde{S}$ that is not yet mapped and whose position is $\langle v, a \rangle$ (Figure 3.4b)), we choose an event according to the following procedure:

1. If $h(a)$ is a leaf, *STOP*.
2. With probability $p_{\text{hor-spread}}(h(a))$ do a *horizontal spread* event $(h(a) \to h(a'))$. The incomparable vertex $h(a')$ is chosen with a multinomial distribution, according to Equation 3.7. Construct a subtree below v together with its mapping in H and stop the evolution of $\tilde{S}$ in this direction.
3. If we do not draw a horizontal spread event, then with probability $p_{\text{vert-spread}}(h(a))$ we do a *vertical spread* event. Construct a subtree below v together with its mapping in H and stop the evolution of $\tilde{S}$ in this direction.

4. Otherwise, sample with a multinomial distribution one of the four events: cospeciation, duplication, host switch or loss event, and apply the same procedure as in Coala.

The evolutionary events in AmoCoala

As in Coala, for any vertex v of $\tilde{S}$ that is not yet mapped and whose position is $\langle v, a\rangle$ (see Figure 3.4b)), we choose to apply either a vertical spread, a horizontal one, or one of the "classical events", depending on the probability of each event. We do not repeat how the "classical events" are handled because it is the same as in Coala. Here we will just describe how to apply the vertical and horizontal spreads. In what follows, we denote by a_1, a_2 the arcs outgoing from the head $h(a)$ of the arc a.

- Vertical Spread (Figure 3.4c)): We apply the mapping $\{v, H_{h(a)}\}$. This operation is executed with probability $p_{\text{vert-spread}}(h_a)$.
- Horizontal Spread (Figure 3.4d)): We apply the mapping $\{v, H_{h(a)}\}$ and the mapping $\{v, H_{h(a')}\}$. This operation is executed with probability $p_{\text{hor-spread}}(h_a)$. The choice of the incomparable vertex $h(a')$ can change due to the need to preserve time feasibility: thus the probabilities described in Equation (3.7) are updated accordingly to the new set of incomparable vertices.

Algorithm for simulating the evolution of the symbiont in the case of horizontal and vertical spreads

During the simulation, if a spread event is chosen (horizontal or vertical), the symbiont is associated to a set of hosts. We need to choose a way to simulate the topology of the symbiont tree $\tilde{S}$ below the symbiont vertex $\tilde{s}$ that undergoes a spread event. With the passing of time, both the host and the symbiont have evolved and in addition, it is possible that some hosts have lost their symbiont. Reconstructing all the possible evolutions of the symbiont is practically impossible. Trying all the possible topologies is hard. Therefore, for computational reasons, we decide to promote the more realistic situations, that is those present in the real symbiont. In this context, we choose the topology and the leaves associations that are identical to those observed in the original symbiont tree. In the following, for any tree T and any set of taxa $t_1, \ldots, t_n$, we let $T_{|\{t_1,\ldots,t_n\}}$ denote the subtree of T with leaves $t_1, \ldots, t_n$.

Given a dataset (H, S, ϕ), the procedure to follow in case of a vertical spread in h_i, during the simulation of $\tilde{S}$ is:

- List the symbionts $\{s_1, \ldots s_n\}$ associated with $L(h)$, the leaves of the subtree under h.
- Create a subtree with the leaves $\{s_1, \ldots, s_n\}$ identical to $S_{|\{s_1,\ldots,s_n\}}$, and attach the root of this subtree to $\tilde{s}$ in $\tilde{S}$.
- The associations $\phi(s_1, ..., s_n)$ of $\tilde{S}$ will be like in the original symbiont tree.

Figure 3.5 shows an example of a mapping that involves a vertical spread. For this, we used the dataset of Figure 3.3 with their probabilities of horizontal and vertical spreads. At first, we have a cospeciation (Figure 3.5(A)), next there is a vertical spread in h_4, the symbiont's

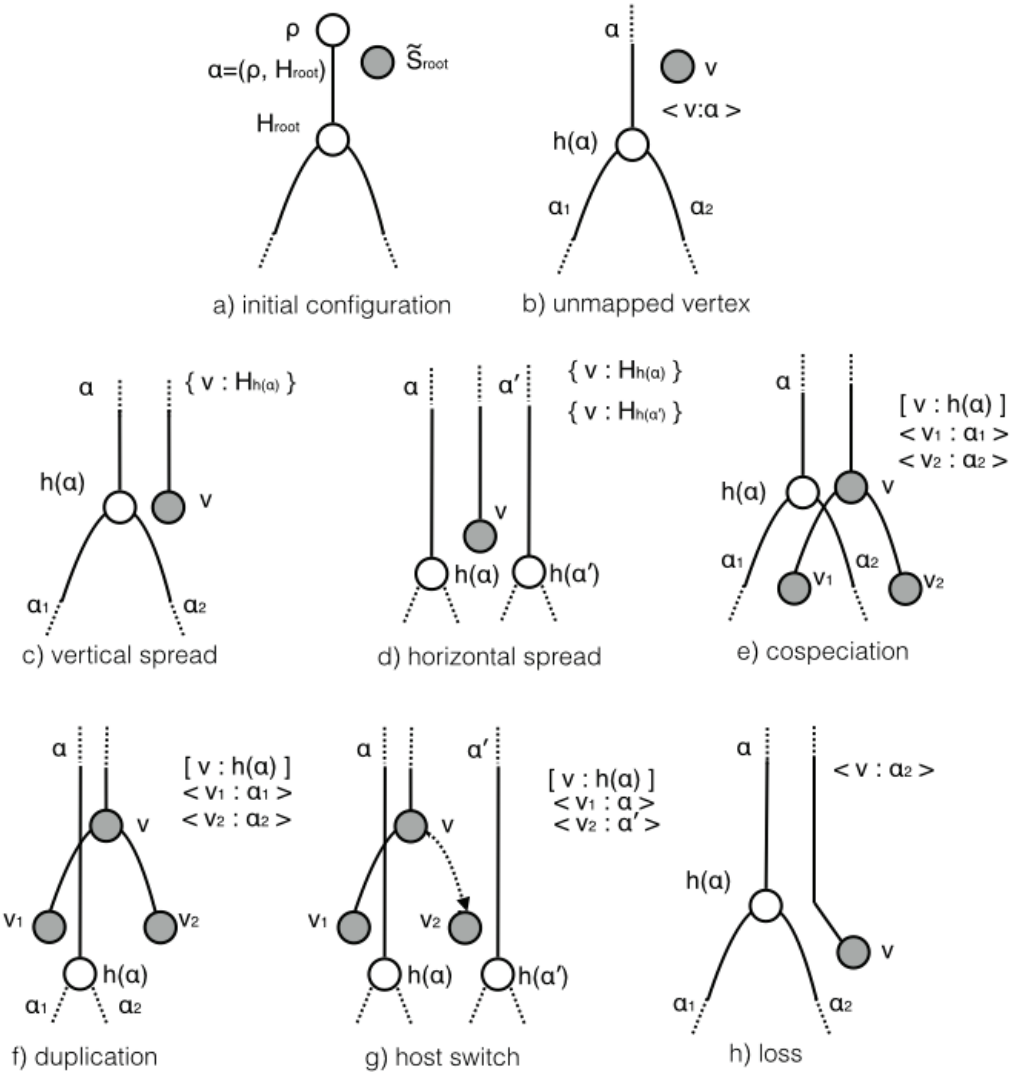


Figure 3.4: Events during the generation of the symbiont tree $\tilde{S}$. The host tree has white vertices and the symbiont tree grey vertices. The association $\{v, H_{h(a)}\}$ indicates that we map v to the vertex $h(a)$ that is the head of the arc a and to all the descendants of $h(a)$. The association $\langle v : a \rangle$ indicates that an unmapped symbiont vertex v is positioned on the arc a of the host tree. The association $[v : w]$ indicates that the symbiont vertex v is mapped to the host vertex w.

simulation for the vertical spread starts as described above. The list of the symbionts associated to h_4 are $\{s_2, s_3, s_5\}$. A phylogenetic tree identical to $S_{|\{s_2,s_3,s_5\}}$ is attached to $\tilde{s}_2$. Finally the associations are copied, as in the original: we thus have that $\phi(\tilde{s_4}) = \{h_8\}$, $\phi(\tilde{s_5}) = \{h_8\}$ and $\phi(\tilde{s_6}) = \{h_5, h_7\}$.

Given a dataset (H, S, ϕ), the procedure to follow in the case of a horizontal spread in $h_i \to h_j$ is:

- List the symbionts $\{s_1, \ldots, s_n\}$ associated with $L(h)$, the leaves of the subtrees under h_i and h_j.
- Create a subtree with the leaves $\{s_1, \ldots, s_n\}$ identical to $S_{|\{s_1,\ldots,s_n\}}$, and attach the root of this subtree to $\tilde{s}$ in $\tilde{S}$.
- The associations $\phi(s_1, \ldots, s_n)$ of $\tilde{S}$ will be like in the original symbiont tree.

Figure 3.6 shows an example of mapping that involves a horizontal spread. For this example, we used the dataset of Figure 3.3 with their probabilities of horizontal and vertical spreads. The choice of the incomparable vertex h_j is made using Table 3.1. At first, we have a cospeciation (Figure 3.6(A)), next there is a horizontal spread $h_1 \to h_5$ of $\tilde{s}$, then the symbiont's simulation starts for the horizontal spread as described above. The list of the symbionts associated to h_1 and h_5 are $\{s_5, s_6\}$. A phylogenetic tree identical to $S_{|\{s_5,s_6\}}$ is attached to $\tilde{s}$. Finally the associations are copied, as in the original: we thus have that $\phi(\tilde{s_3}) = \{h_3, h_5\}$ and $\phi(\tilde{s_4}) = \{h_2, h_3\}$.

Table 3.1: Probabilities of a jump between vertices for the example shown in Figure 3.3

$p_{\text{jump}}(h_i \to h_j)$	h_0	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
h_0	0	0	0	0	0	0	0	0	0
h_1	0	0	0	0	0.25	0.5	0.25	0.5	0
h_2	0	0	0	1	0	0	0	0	0
h_3	0	0	1	0	0.25	0.5	0.25	0.5	0
h_4	0	0.25	0	0.25	0	0	0	0	0
h_5	0	0.5	0	0.5	0	0	0.33	1	0
h_6	0	0.25	0	0.25	0	0.33	0	0	0
h_7	0	0.5	0	0.5	0	1	0	0	0
h_8	0	0	0	0	0	0	0	0	0

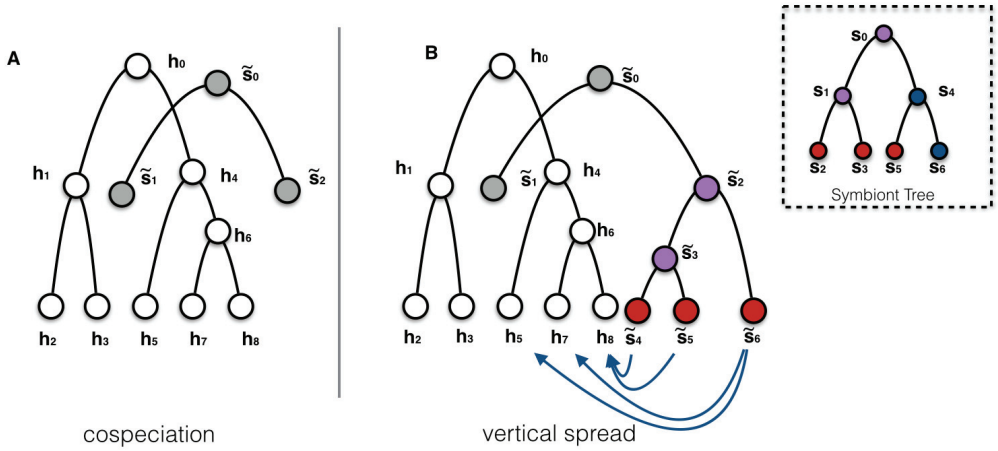


Figure 3.5: Example of a simulation of the evolution of a symbiont tree $\tilde{S}$ in the case of a vertical spread. The vertices of the host tree (resp. simulated tree) are in white (resp. in colour) while the real symbiont tree is shown in the top right rectangle (the original associations are shown in Figure 3.3). (A) A cospeciation event with the symbiont tree root $\tilde{s}_0$ mapped in h_0 (namely $[\tilde{s}_0 : h_0]$) while the children $\tilde{s}_1$, $\tilde{s}_2$ are positioned on the arcs $a_1 = (h_0, h_1)$ and $a_2 = (h_0, h_4)$ (namely $\langle \tilde{s}_1 : a_1 \rangle$ and $\langle \tilde{s}_2 : a_2 \rangle$). (B) A vertical spread event for the symbiont $\tilde{s}_2$ positioned on the arc a_2. We thus first map $\tilde{s}_2$ to h_4 and all its descendants (namely $\{\tilde{s}_2 : h(a_2)\}$). We then attach a subtree to $\tilde{s}_2$ constructed as explained in the text.

3.2.3 ABC-SMC Inference Method

In this section, we describe the Bayesian computation (ABC method) that was developed in Coala (this section is based on [5, 17]).

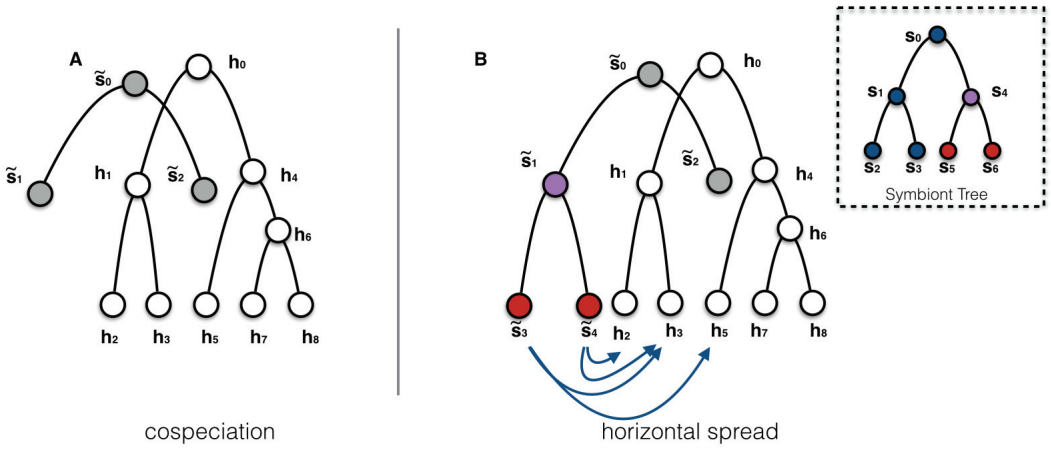


Figure 3.6: Example of a simulation of the evolution of a symbiont tree $\tilde{S}$ in the case of a horizontal spread. The vertices of the host tree (resp. simulated tree) are in white (resp. in colour) while the real symbiont tree is shown in the top right rectangle (the original associations are shown in Figure 3.3). (A) A cospeciation event with the symbiont tree root $\tilde{s}_0$ mapped in h_0 (namely $[\tilde{s}_0 : h_0]$) while the children $\tilde{s}_1$, $\tilde{s}_2$ are positioned on the arcs $a_1 = (h_0, h_1)$ and $a_2 = (h_0, h_2)$ (namely $\langle \tilde{s}_1 : a_1 \rangle$ and $\langle \tilde{s}_2 : a_2 \rangle$). (B) A horizontal spread event for the symbiont $\tilde{s}_2$ in $h_1 \to h_5$. We thus first map $\tilde{s}_1$ to h_1 and all its descendants (namely $\{\tilde{s}_1 : h(a_1)\}$). We then map $\tilde{s}_1$ to h_5 and all its descendants (namely $\{\tilde{s}_1 : h(a_5)\}$), where $a_5 = (h_4, h_5)$. Note that h_5 does not have a descendant. We then attach a subtree to $\tilde{s}_1$ constructed as explained in the text.

In general, a likelihood-free computation involves a chain of parameter proposals and only accepts a set of parameter values on the condition that the model with these values generates data that satisfy a performance criterion with respect to the observed data. In this case, N parameter vectors in the space $[0, 1]^3$ are randomly chosen under some prior distribution (usually uniform). The two main steps of the ABC-SMC procedure consist in:

Step 1 For each vector θ, generate M simulated trees $\tilde{S}$. A distance value, obtained computing the difference between S and $\tilde{S}$, is associated to each parameter vector. The vectors are then ordered respecting the distance value, in ascending order. The choice of the best vectors is made by taking into account two values which are:

- Tolerance: This value determines the percentage of parameter vectors to be accepted. If for instance, the tolerance value is 0.2, then 20% of the vectors with lower distance will be accepted.
- Threshold: This corresponds to the largest value of the summary statistics associated with the accepted parameter vectors.

Step 2 Only the best vectors previously selected are used in this step, and are perturbed. The perturbation is performed by adding to each coordinate of the vector a randomly chosen value in the interval $[-0.01, +0.01]$ and by doing a normalisation. A new distance value is calculated for the perturbed vector. If the new distance is lower than the threshold, the perturbed vector is conserved.

The ABC-SMC procedure is composed of $R > 1$ rounds. For each round, a tolerance value is defined which determines the percentage of parameter vectors to be accepted.

The final set of accepted parameter vectors is the result of the ABC-SMC procedure and characterises the list of vectors that may explain the evolution of the pair of host and symbiont trees given as input. Observe that, since in all experiments a uniform prior distribution is assumed and also the perturbations are performed in a uniform way, the weights induced by the proposals will also appear to be uniform [6]. However, in the case of a different prior, weights should be used in the process in order to correct the posterior distribution according to the perturbation made.

Distance for comparing trees

We now propose an extension of MAAC (Maximum Agreement Area Cladogram) [23] for comparing two phylogenetic trees whose leaves are labelled by a subset of a set X. We remind that in Coala, one host can be associated to more than one symbiont, but one symbiont cannot be associated to more than one host, so that $\tilde{S}$ is a *multi-labelled* tree, meaning that it is a tree whose leaf labels need not be unique.

Instead in AmoCoala, one host can be associated to more than one symbiont and one symbiont can be associated to more than one host. In this case, $\tilde{S}$ is a *set-labelled* tree (that is, the leaves are labelled with sets and not singletons). We can extend the concept of maximum agreement subtree in order to handle the case of *set-labelled* trees. We recall that the maximum agreement subtree of two trees T_1 and T_2 is an agreement subtree with the maximum weight. We will denote such weight by $MAS(T_1, T_2)$. The MAS distance, denoted by d_{MAS}, between two set-labelled phylogenetic trees T_1,T_2 is defined as:

$$d_{MAS}(T_1, T_2) = \max\{w(T_1), w(T_2)\} - w(MAS(T_1, T_2)) \tag{3.8}$$

where we remind that $w(T)$ is the weight of a tree T. The MAS distance is thus the difference between the maximum weight of T_1 and T_2, and the weight of a maximum agreement subtree of T_1 and T_2. Note that the maximum agreement subtree does not necessarily have the maximum number of vertices as showed in Figure 3.7.

We show that the distance d_{MAS} is a metric. For this, we check that d_{MAS} satisfies the following properties:

1. $d_{MAS}(T_1, T_2) \geq 0$ for all T_1,T_2: this is trivial.

2. $d_{MAS}(T_1, T_2) = 0$ if and only if $T_1 = T_2$. Clearly if $T_1 = T_2$ then $d_{MAS}(T_1, T_2) = 0$. Otherwise, let $d_{MAS}(T_1, T_2) = 0$. Then $\max\{w(T_1), w(T_2)\} = MAS(T_1, T_2)$. The proof follows by observing that in general if T^* is a subtree of T such that $w(T^*) = w(T)$ then $T^* = T$.

3. $d_{MAS}(T_1, T_2) = d_{MAS}(T_2, T_1)$: this is trivial.

4. For any triplet of trees T_1, T_2, T_3, it holds that $d_{MAS}(T_1, T_2)+d_{MAS}(T_2, T_3) \geq d_{MAS}(T_1, T_3)$: For simplicity, we set $w_i = w(T_i)$ and $m_{i,j} = MAST(T_i, T_j)$. Hence $d_{MAS}(T_i, T_j) = \max\{w_i, w_j\} - m_{i,j}$. Furthermore, we denote by $m_{1,2,3}$ the size of the maximum agreement

subtree that is common to the three trees T_1, T_2, T_3. We then have:

$$\begin{aligned} d_{MAS}(T_1,T_2)+d_{MAS}(T_2,T_3) &= \max\{w_1,w_2\}-m_{1,2}+\max\{w_2,w_3\}-m_{2,3} \\ &= \max\{w_1,w_2\}+\max\{w_2,w_3\}-(m_{1,2}+m_{2,3}-m_{1,2,3}+m_{1,2,3}) \\ &\geq \max\{w_1,w_2,w_3\}+w_2-(w_2-m_{1,2,3}) \\ &\geq \max\{w_1,w_3\}-m_{1,3}, \end{aligned}$$

where for the first inequality, we use the fact that $\max\{w_1,w_2\}+\max\{w_2,w_3\} \geq \max\{w_1,w_2,w_3\}+w_2$ and $m_{1,2}+m_{2,3}-m_{1,2,3}$ is at most w_2. This concludes the proof.

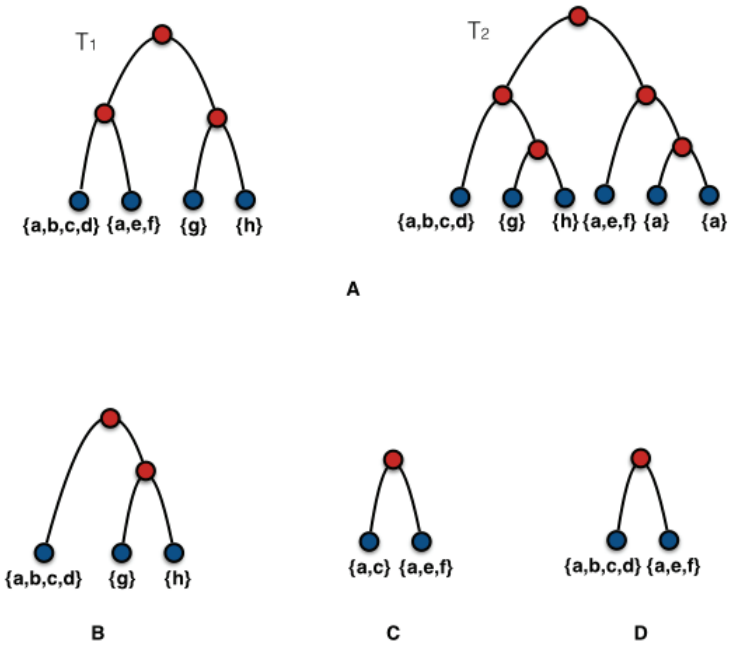


Figure 3.7: (A) Two set-labelled phylogenetic trees. T_1 has weight 9 and T_2 has weight 11. In (B), (C), (D), three different agreement subtrees of weight 6, 5 and 7 respectively. The maximum agreement subtree is the one depicted in (D).

The previous proof and comments show that the MAS distance d_{MAS} is very similar to the $MAAC$ one [23] for *multi-labelled* trees. Thus, it is natural to ask whether comparing two set-labelled trees can be reduced to comparing two multi-labelled trees. One idea is to transform a set-labelled tree into a multi-labelled tree. However, the straightforward transformation seems not to work well for our purpose. For instance, we can transform each set-labelled tree into a multi-labelled tree by substituting each set-labelled leaf by a subtree with a fixed topology (say a complete binary tree, or a multifurcating vertex) as in Figure 3.8. However, in these cases the two trees in Figure 3.8A-B would be considered equivalent, but in our context they are different. In fact, the set-labelled tree in Figure 3.8 indicates that there is a symbiont that infects 4 different hosts $\{a,b,c,d\}$, while in the multi-labelled trees, we will have 4 different symbionts infecting each a different host.

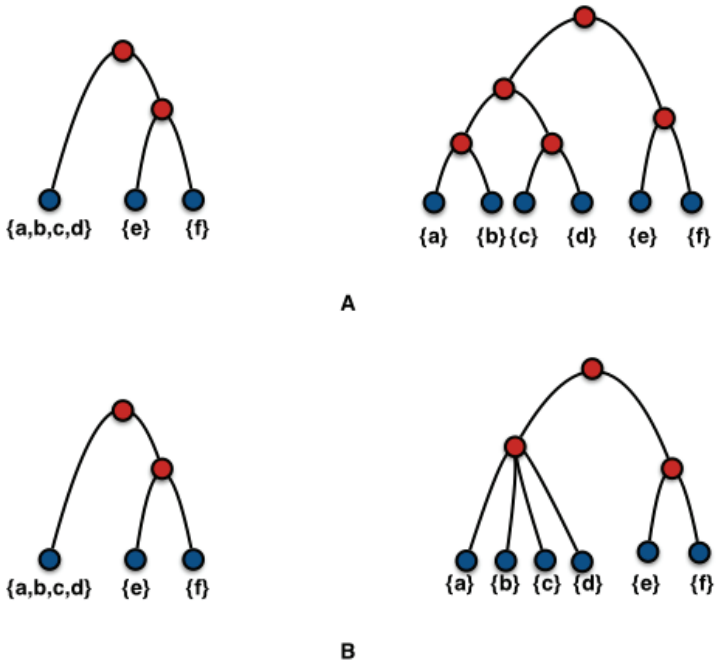


Figure 3.8: (A) The two phylogenetic trees will be considered at distance 0 if we substitute the vertex labelled by the set $\{a, b, c, d\}$ by: (A) a balanced binary tree, (B) a multifurcated vertex.

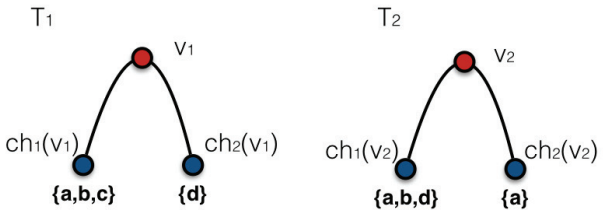
A polynomial time algorithm for computing the MAS distance d_{MAS}

We now show that it is possible to calculate the distance $d_{MAS}(T_1, T_2)$ in polynomial time with respect to the size of the trees. The algorithm is based on dynamic programming and extends quite straightforwardly the algorithm for calculating the MAAC distance [23]. We denote by $MAS(v_1, v_2)$ the maximum agreement subtree between the two trees T_1 and T_2 rooted in v_1 and v_2, respectively. For a leaf v, we denote by $l(v)$ the set of labels associated with it. Finally, for an internal vertex v, we denote by $ch_1(v)$ and $ch_2(v)$ the two children of v.

The dynamic programming algorithm starts from the leaves and ends in the roots of T_1 and T_2 following a recursion. We have that $MAS(v_1, v_2)$ is given by:

- If v_1 and v_2 are both leaves then $MAS(v_1, v_2) = |L(v_1) \cup L(v_2)|$
- If v_1 or v_2 (could be both) are internal vertices, $MAS(v_1, v_2)$ is the maximum value among the following three quantities:
  1. $\max\{MAS(ch_1(v_1), v_2), MAS(ch_2(v_1), v_2)\}$
  2. $\max\{MAS(v_1, ch_1(v_2)), MAS(v_1, ch_2(v_2)\}$
  3. $\max\{MAS(ch_1(v_1), ch_1(v_2)) + MAS(ch_2(v_1), ch_2(v_2)),\ MAS(ch_1(v_1), ch_2(v_2))$ $+MAS(ch_2(v_1), ch_1(v_2))\}$

Figure 3.9 shows an example of dynamic programming matrix for the set-labelled trees T_1 and T_2 rooted in v_1 and v_2 respectively. In this example, $MAS(v_1, v_2)$ is equal to 2.



Dynamic Programming Matrix

	$ch_1(v_2)$	$ch_2(v_2)$	v_2
$ch_1(v_1)$	2	1	2
$ch_2(v_1)$	1	0	1
v_1	2	1	2

Figure 3.9: Example of dynamic programming matrix for the set-labelled trees T_1 and T_2 rooted in v_1 and v_2 respectively.

3.3 Datasets

We evaluated our method AmoCoala using synthetic and real datasets. We describe below each one of these datasets.

3.3.1 Synthetic datasets

We first evaluated the model using synthetic datasets. The idea is, given a dataset (H, S, ϕ) and a specific probability vector θ, to produce a simulated symbiont tree $\tilde{S}_\theta$. Observe that θ stands for a vector of four probabilities $\langle p_c, p_d, d_s, p_l \rangle$.

In this way, we create a synthetic dataset $(H,\ \tilde{S}_\theta,\ \phi')$ for which we know the "truth", that is the real parameter vector θ associated to it. Hence, to test AmoCoala with this dataset, we want to check whether the vector θ (or a vector very "similar" to it) is found among those accepted in the last round of the procedure.

To test the algorithm, we used the 5 biological datasets described in Section 3.3.2 below, and we generated the synthetic ones using the method described below.

For any such dataset, we generated 8 datasets (H,S) associated with the following 8 probability vectors: $\theta_1 = \langle 0.70, 0.10, 0.10, 0.10 \rangle$, $\theta_2 = \langle 0.80, 0.15, 0.01, 0.04 \rangle$, $\theta_3 = \langle 0.75, 0.01, 0.16, 0.08 \rangle$, $\theta_4 = \langle 0.70, 0.05, 0.02, 0.23 \rangle$, $\theta_5 = \langle 0.60, 0.20, 0.00, 0.20 \rangle$, $\theta_6 = \langle 0.55, 0.00, 0.20, 0.25 \rangle$, $\theta_7 = \langle 0.45, 0.10, 0.15, 0.30 \rangle$ and $\theta_8 = \langle 0.40, 0.20, 0.10, 0.30 \rangle$. The choice of vectors was done with the aim to cover different patterns of probability. All datasets were generated with the same host tree H and symbiont tree S.

We used these vectors and the symbiont tree generation algorithm in Coala (see [5] for more details) to obtain, for each vector θ, a simulated symbiont tree $\tilde{S}_\theta$ whose evolution follows that of the host tree H under the parameter value of θ. It is important to note that during the simulation of $\tilde{S}_\theta$, we can have vertical or horizontal spreads (so also multiple associations created as explained in "Algorithm for simulating the evolution of the symbiont in the case of horizontal and vertical spreads"). Each dataset (H, S, ϕ) and probability vector γ thus led to a simulated dataset (H, S', ϕ').

Due to the high variability of the symbiont trees which can be simulated given a host tree H and a vector θ, the task of choosing the most "typical" tree can be hard. To simplify this task and select a typical tree, we impose two conditions which must be observed by the simulated tree. The first one requires that the candidate tree should have a size close to the median for all the trees which are simulated using H and θ. The second condition requires that the observed number of events of a candidate tree should be very close to the expected number given θ. For

this reason, the number of synthetic datasets created is not 5×8=40, because there are some vectors θ that do not produce enough simulated trees (for reasons of size or number of events). In this case, the vector θ is not used, and this is why in the end, we have only 36 synthetic datasets.

In practical terms, in order to simulate a realistic symbiont tree, we choose a real host tree H and a probability vector θ. We then generate 2000 symbiont trees, imposing constraints on the size of the simulated trees $\tilde{S}_\theta$ can be at most two times bigger than S). We then compute the median size of all the generated trees and filter out those whose size is far from this value (difference greater than 1 or 2 leaves from the median value). Finally, we select as typical tree $\tilde{S}_\theta$ the one that shows the smallest χ^2 distance between the vector θ and the vector of observed frequencies of events.

3.3.2 Biological datasets

To test our method, we selected 5 biological datasets from the literature:

AP - *Acacia & Pseudomyrmex.* This dataset was extracted from the work of Gómez-Acevedo *et al.* [24]. The host tree includes 9 leaves and the symbiont tree includes 7 leaves. The dataset has 22 multiple-associations.

MP - *Myrmica & Phengaris.* This dataset was extracted from the work of Jansen *et al.* [40] and is composed of a pair of host and symbiont trees which have each 8 leaves. The dataset has 8 multiple-associations.

SBL - *Seabirds & Lice*, This dataset was extracted from the work of Paterson *et al.* [57]. The host tree includes 15 leaves and the symbiont tree includes 8 leaves. The dataset has 15 multiple-associations.

SFC - *Smut Fungi & Caryophillaceus plants.* This dataset was extracted from the work of Refrégier *et al.* [65]. The host tree includes 15 leaves and the symbiont tree includes 16 leaves. The dataset has 4 multiple-associations.

SFCsimple - *Smut Fungi & Caryophillaceus plants.* This is the same dataset as the previous one, except that it has no multiple associations. The authors in [5] created the SFCsimple dataset because in Coala it was impossible to use datasets with multiple associations. With AmoCoala we can study not only SFCsimple , but also the SFC dataset where there are four symbiont species that are associated each to two host species.

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, and (3) the possibility to compare the results with those presented in [5].

3.4 Results and Discussion

We tested the model on both the synthetic and the datasets. We present here the results for all datasets, however giving a particular attention to the SFC dataset (Figure 3.10) and to the SFCsimple dataset (Figure 3.11).

3.4.1 Experimental setting

Parameter setting using AmoCoala

All datasets were processed using AmoCoala configured with the same parameters. For each dataset, we generated $N = 2000$ vectors in the first round. For each vector, $M = 1000$ symbiont trees were generated using the method described in Section *"Evolution of symbionts in* AmoCoala*"* . These trees have a size at most twice the one of the real symbiont tree, otherwise the tree was discarded as being too different from the original one. The tolerance value used in the first round was $\tau_1 = 0.1$. We ran $R = 3$ rounds and we defined $\tau_i = 0.25$. Notice that $\tau_1 \times N = 200$ defines the size Q of the quantile set which must be produced in each new round. Thus, after the last round, we have $\tau_3 \times Q = 50$ accepted vectors.

Behaviour of the algorithm

In this section, we check if the algorithm is able to produce trees $\tilde{S}$ similar in terms of number of multiple associations and of size to the original symbiont tree S. We expect that in the first round, where the vectors are sampled uniformly at random, the simulated trees differ from S. However, after each round, the method tends to select vectors producing trees near to the original one. Observe that we expect the generated tree to be slightly bigger in terms of number of multiple associations and of size compared to the original symbiont S. This is due to the fact that during the simulations, if we have a horizontal or a vertical spread, we simulate the symbiont tree choosing the topology and the leaf associations, identical to those observed in the original symbiont tree. For each generated vector, we simulated $M = 1000$ symbiont trees, we calculated the average number of multiple associations and the average size, and we plotted the frequencies of each value. We produce these plots for the 3 rounds.

Self-test

Here we controlled if, given a dataset (H, S, ϕ) and a specific probability vector θ, we produce a simulated symbiont tree $\tilde{S}_\theta$ as explained in the Section *Evolution of symbionts in* AmoCoala. We want to know if, running AmoCoala on a host tree H, a symbiont tree $\tilde{S}_{\theta'}$ and the associations between leaves ϕ, among the vectors accepted on the last round there is one θ' that is close to θ. The distance that we use between the vectors $\theta = \langle p_c, p_d, p_s, p_l \rangle$ and $\theta' = \langle q_c, q_d, q_s, q_l \rangle$ is the χ^2.

3.4.2 Results on the Synthetic Datasets

We present now the results for the five datasets. We gave more importance to the results for the datasets SFC and SFCsimple because this were the ones that were analysed in [5] and thus it is possible to compare the results of the two models, Coala and AmoCoala.

Behaviour of the algorithm

We discuss here the results for the SFC dataset. Figure 3.12 shows for each round, given a vector, the average number of multiple associations of the simulated symbiont trees. The original dataset has 4 multiple associations. We see that during the simulation this value is observed. Notice that

in the second and third rounds, there is a tendency to have more than 4 multiple associations: most of the vectors have an average of multiple associations between 5 and 6.

Figure 3.13 shows for each round, given a vector, the average size of the simulated symbiont trees. The original dataset has 31 vertices, we see that during the simulation this value is observed. Notice that in second and third rounds, there is a tendency to have a size bigger than 31: indeed, most of the vectors have an average size between 35 and 40. As before, this is probably due to the fact that during the simulations, if we have a horizontal or a vertical spread, we simulate the symbiont tree choosing the topology and the leaf associations identical to those observed in the original symbiont tree. Thus, if a spread happens more than once in a host vertex, the subtree of the symbiont tree will be copied more than once. However, this is a rare event in practice as shown empirically by our experiments where we have symbiont trees very near to the real tree.

In general, we find the same characteristics (as concerns both number of multiple associations and size) between the original and the simulated symbiont trees, in all the datasets with multiple associations used (Figures 3.14, 3.16 and 3.18 for the number of multiple associations for the datasets AP, MP and MP respectively, Figures 3.15, 3.17 and 3.19 for the size of the trees for the datasets AP, MP and MP respectively). The number of multiple associations expected is present in the histogram, but it is not the mode, probably due to the method to simulate $\tilde{S}$ with horizontal and vertical spreads.

The dataset SFCsimple does not present this behaviour. The number of multiple associations is equal to zero (Figure 3.20) and the average size of the simulated symbiont trees is close to the size of the original tree (Figure 3.21). This is probably due to the fact that SFCsimple does not have multiple associations and the method to simulate $\tilde{S}$ with horizontal and vertical spreads is not used.

Self-test

We applied the self-test to the synthetic datasets obtained from the real ones. For those obtained from the SFC dataset, we see that running AmoCoala under the parameter value $\tilde{S}_\theta$, we can find, among the vectors accepted in the last round, a vector close to θ. Table 3.2 shows for each setting: the probability vector θ, the probability vector θ', the χ^2 distance $d(\theta, \theta')$, the cluster where θ' appears and how many vectors this cluster has. The results show that if we produce a tree from a vector θ, then this vector is close to those accepted at the end of the algorithm.

We obtain the two bigger χ^2 distances with the vectors that present the lowest probability of cospeciations. This is due to the fact that, in the case of coevolution, usually there is a big probability of cospeciations. When this value is low, it is difficult to simulate good symbiont trees, as the symbiont is not following the host anymore and potentially every phylogenetic tree would be possible for the symbiont. We found the same type of results for all the used datasets: AP(Table 3.3), MP (Table 3.4), SBL (Table 3.5) and SFCsimple (Table 3.6).

3.4.3 Results with the Biological Datasets

In the case of the biological datasets, to see if the algorithm really simulates horizontal and vertical spreads but also if the new distance (d_{MAS}) accepts the good vectors, we list the histograms of distances, event probabilities and number of vertical and horizontal spreads obtained at the end of each one of the 3 rounds (Figures 3.22, 3.23 and 3.24). As we expect, in the case of

Table 3.2: Result of the self-test for the SFC synthetic datasets

$\theta = \langle p_c, p_d, p_s, p_l \rangle$	$\theta' = \langle q_c, q_d, q_s, q_l \rangle$	χ^2	*cluster*	*#vectors*
$\langle 0.7, 0.1, 0.1, 0.1 \rangle$	$\langle 0.779, 0.130, 0.079, 0.012 \rangle$	0.399	2	16
$\langle 0.8, 0.15, 0.01, 0.04 \rangle$	$\langle 0.82, 0.114, 0.028, 0.04 \rangle$	0.167	4	6
$\langle 0.75, 0.01, 0.16, 0.08 \rangle$	$\langle 0.834, 0.004, 0.115, 0.0474 \rangle$	0.214	1	31
$\langle 0.7, 0.05, 0.02, 0.23 \rangle$	$\langle 0.752, 0.017, 0.004, 0.227 \rangle$	0.239	2	11
$\langle 0.6, 0.2, 0, 0.2 \rangle$	$\langle 0.519, 0.268, 0.002, 0.211 \rangle$	0.192	2	12
$\langle 0.55, 0, 0.2, 0, 25 \rangle$	$\langle 0.729, 0.006, 0.262, 0.004 \rangle$	0.292	1	18
$\langle 0.45, 0.1, 0.15, 0.3 \rangle$	$\langle 0.672, 0.000, 0.157, 0.170 \rangle$	0.599	3	9
$\langle 0.4, 0.2, 0.1, 0.3 \rangle$	$\langle 0.478, 0.358, 0.005, 0.159 \rangle$	0.602	5	3

Table 3.3: Result of the self-test for the AP synthetic datasets

$\theta = \langle p_c, p_d, p_s, p_l \rangle$	$\theta' = \langle q_c, q_d, q_s, q_l \rangle$	χ^2	*cluster*	*#vectors*
$\langle 0.7, 0.05, 0.02, 0.23 \rangle$	$\langle 0.757, 0.0292, 0.0211, 0.1926 \rangle$	0.260	3	14
$\langle 0.6, 0.2, 0, 0.2 \rangle$	$\langle 0.7075, 0.0308, 0.0055, 0.2563 \rangle$	0.854	1	17
$\langle 0.55, 0, 0.2, 0, 25 \rangle$	$\langle 0.8931, 0.0374, 0.0409, 0.0285 \rangle$	1.375	3	2
$\langle 0.45, 0.1, 0.15, 0.3 \rangle$	$\langle 0.7799, 0.0314, 0.0136, 0.1751 \rangle$	1.421	2	12
$\langle 0.4, 0.2, 0.1, 0.3 \rangle$	$\langle 0.4844, 0.0232, 0.0076, 0.4848 \rangle$	1.35	5	4

Table 3.4: Result of the self-test for the MP synthetic datasets

$\theta = \langle p_c, p_d, p_s, p_l \rangle$	$\theta' = \langle q_c, q_d, q_s, q_l \rangle$	χ^2	*cluster*	*#vectors*
$\langle 0.7, 0.1, 0.1, 0.1 \rangle$	$\langle 0.834, 0.035, 0.02, 0.111 \rangle$	0.7629	1	25
$\langle 0.8, 0.15, 0.01, 0.04 \rangle$	$\langle 0.198, 0.418, 0.015, 0.37 \rangle$	2.126	2	9
$\langle 0.7, 0.05, 0.02, 0.23 \rangle$	$\langle 0.607, 0.019, 0.237, 0.137 \rangle$	1.106	2	12
$\langle 0.6, 0.2, 0, 0.2 \rangle$	$\langle 0.734, 0.025, 0.031, 0.21 \rangle$	0.958	3	11
$\langle 0.55, 0, 0.2, 0, 25 \rangle$	$\langle 0.6, 0.00, 0.325, 0.074 \rangle$	0.7163	4	9
$\langle 0.45, 0.1, 0.15, 0.3 \rangle$	$\langle 0.409, 0.015, 0.426, 0.15 \rangle$	1.251	2	16
$\langle 0.4, 0.2, 0.1, 0.3 \rangle$	$\langle 0.702, 0.0403, 0.034, 0.223 \rangle$	1.272	1	35

Table 3.5: Result of the self-test for the SBL synthetic datasets

$\theta = \langle p_c, p_d, p_s, p_l \rangle$	$\theta' = \langle q_c, q_d, q_s, q_l \rangle$	χ^2	*cluster*	*#vectors*
$\langle 0.7, 0.1, 0.1, 0.1 \rangle$	$\langle 0.858, 0.022, 0.026, 0.094 \rangle$	0.811	2	22
$\langle 0.8, 0.15, 0.01, 0.04 \rangle$	$\langle 0.79, 0.168, 0.028, 0.014 \rangle$	0.34	3	10
$\langle 0.75, 0.01, 0.16, 0.08 \rangle$	$\langle 0.737, 0.013, 0.167, 0.083 \rangle$	0.0731	1	20
$\langle 0.7, 0.05, 0.02, 0.23 \rangle$	$\langle 0.752, 0.008, 0.017, 0.224 \rangle$	0.346	2	17
$\langle 0.6, 0.2, 0, 0.2 \rangle$	$\langle 0.671, 0.034, 0.024, 0.272 \rangle$	0.939	1	29
$\langle 0.55, 0, 0.2, 0, 25 \rangle$	$\langle 0.667, 0.003, 0.213, 0.117 \rangle$	0.74	1	22
$\langle 0.45, 0.1, 0.15, 0.3 \rangle$	$\langle 0.439, 0.001, 0.298, 0.262 \rangle$	0.839	4	2
$\langle 0.4, 0.2, 0.1, 0.3 \rangle$	$\langle 0.411, 0.006, 0.337, 0.246 \rangle$	1.233	1	24

Table 3.6: Result of the self-test for the SFCsimple synthetic datasets

$\theta = \langle p_c, p_d, p_s, p_l \rangle$	$\theta' = \langle q_c, q_d, q_s, q_l \rangle$	χ^2	*cluster*	*#vectors*
$\langle 0.7, 0.1, 0.1, 0.1 \rangle$	$\langle 0.723, 0.152, 0.113, 0.012 \rangle$	0.593	1	23
$\langle 0.8, 0.15, 0.01, 0.04 \rangle$	$\langle 0.794, 0.159, 0.007, 0.04 \rangle$	0.065	2	20
$\langle 0.75, 0.01, 0.16, 0.08 \rangle$	$\langle 0.706, 0.006, 0.15, 0.137 \rangle$	0.292	3	15
$\langle 0.7, 0.05, 0.02, 0.23 \rangle$	$\langle 0.657, 0.007, 0.01, 0.326 \rangle$	0.579	4	6
$\langle 0.6, 0.2, 0, 0.2 \rangle$	$\langle 0.56, 0.206, 0.00, 0.234 \rangle$	0.164	3	10
$\langle 0.55, 0, 0.2, 0, 25 \rangle$	$\langle 0.76, 0.001, 0.229, 0.01 \rangle$	0.528	1	25
$\langle 0.45, 0.1, 0.15, 0.3 \rangle$	$\langle 0.59, 0.006, 0.207, 0.197 \rangle$	0.945	4	9
$\langle 0.4, 0.2, 0.1, 0.3 \rangle$	$\langle 0.346, 0.27, 0.001, 0.383 \rangle$	0.815	5	3

a dataset with multiple associations, the histograms of the number of vertical and horizontal spreads contain also values bigger than zero. We see further that the representative distance histograms for the accepted parameter vectors decrease at every round. This means that our method that relies on the distance d_{MAS} accepts vectors of simulated trees that are increasingly more similar to the original symbiont tree.

At the end of the third round, AMOCOALA performs a hierarchical clustering procedure to group the final list of accepted parameter vectors. Table 3.7 shows the cluster for the SFC and SFCsimple datasets. In the study of Baudet *et al.* [5], COALA was tested with the SFCsimple dataset. One question is whether using the same dataset with AMOCOALA, the cluster patterns are the same or not.

We compared our results in Table 3.7 with the results in Table 3 of the Supplementary Material of [5]. We observe that the results are very similar. This means that if a dataset does not have multiple associations, the results obtained with AMOCOALA are very close of those obtained with COALA. Notice that both methods are stochastic. Thus the results cannot be identical.

Next, we considered the real dataset SFC with multiple associations proposed in [65], where the 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 we do not have for now a parsimonious algorithm for computing optimal reconciliations in the case of a dataset with multiple associations (we are working on one). However, we know that the sum of 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 SFCsimple) has 15 internal vertices. Based on the Refrégier's data, we expect to have events with the following probabilities: 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). In Table 3.7, the most similar result to the expected one according to [65] is the cluster 2 for the SFC dataset, because it exhibits zero probability of duplication and high probability of host switch. It is also important to note that the number of vectors that are part of this cluster is high. The results obtained with AMOCOALA which allows to consider the whole dataset SFC with multiple associations, are thus closer to the result presented in [65] than those that were obtained by COALA which ignores such multiple associations. This shows again the

importance of taking into account the latter.

Table 3.7: Representative probability vectors produced by AmoCoala at round 3 while processing the biological datasets SFC and SFCsimple .

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
SFCsimple	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

3.5 Conclusion and Open Problems

In this work, we extended the algorithm of Coala to make the model more realistic. The new model, that we called AmoCoala, allows for multiple associations, meaning that a symbiont can be associated to more than one host. In AmoCoala, it is possible to estimate the probabilities of the "classical events" (cospeciation, duplication, host switch and loss) and also the number of times that the new events, horizontal and vertical spreads, are present. These two events allow to study datasets that contain multiple associations. The model uses set-labelled trees and a new distance, called MAS (which is an extension of the $MAAC$ distance [23]), to compare two trees set-labelled trees. We tested AmoCoala on both synthetic and real datasets. We saw that AmoCoala simulated symbiont trees that are slightly bigger in terms of number of multiple associations and of size than the original symbiont S. This is due to the fact that during the simulations, if we have a horizontal or a vertical spread, we simulate the symbiont tree choosing the topology and the leaf associations identical to those observed in the original symbiont tree. We also used a self-test to check if, for a dataset and its evolutionary history given by the θ parameter, AmoCoala predicts well the expected vector θ. Finally, we applied AmoCoala to the SFC dataset with multiple associations [65]. We checked if the algorithm really simulated horizontal and vertical spreads but also if the new distance (d_{MAS}) accepted the good vectors. As expected, AmoCoala simulated a number of vertical and horizontal spreads bigger than zero. We saw also that the representative distance values for the accepted parameter vectors decrease at every round of the algorithm. This means that our method, that relies on the MAS distance, accepts vectors for the simulated trees that are increasingly more similar to those of original symbiont tree. Based on Refrégier's data, we expected to have zero probability of duplication and high probability of host-switch. We indeed find this. The results obtained with AmoCoala using the SFC dataset with multiple associations are thus closer to the expected value (based on Refrégier's data) than what Coala can get. This demonstrates the importance of having a method that takes into account multiple associations.

The accuracy of the results obtained depends on the choice of the metric used for comparing trees. Designing a new metric that takes in account also if the number of vertical and horizontal spreads observed are similar to the number of spread expected would be one future direction to follow. Another interesting direction would be to change the model and the probability of spread depending on the type of symbiotic association. Indeed, in the case of endosymbiosis, we expect the probability of spread to be lower.

Figure 3.10: Dataset SFC: Host (blue) and symbiont (yellow) trees together with their leaf associations (grey lines). The figure was created using TreeMap3 [11].

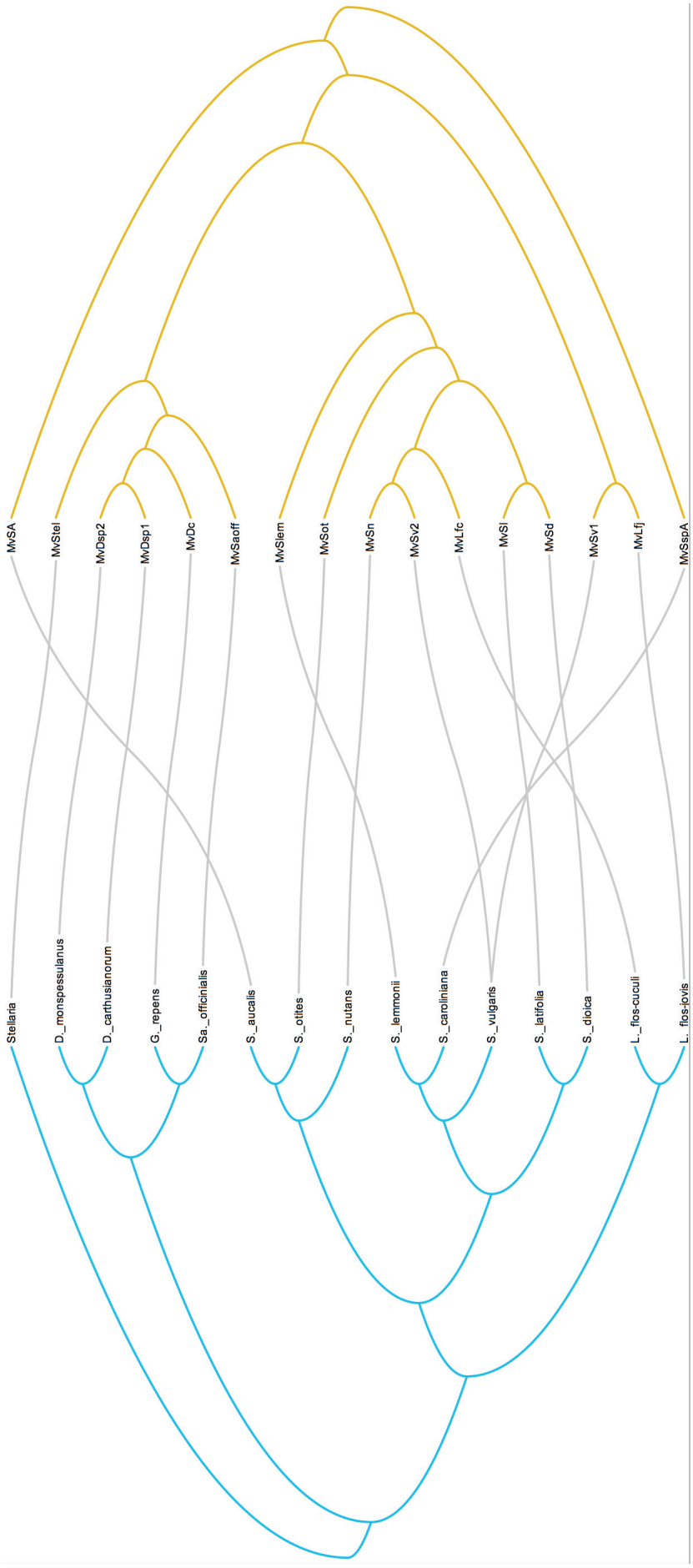


Figure 3.11: Dataset SFCsimple : Host (blue) and symbiont (yellow) trees together with their leaf associations (grey lines). The figure was created using TreeMap3 [11].

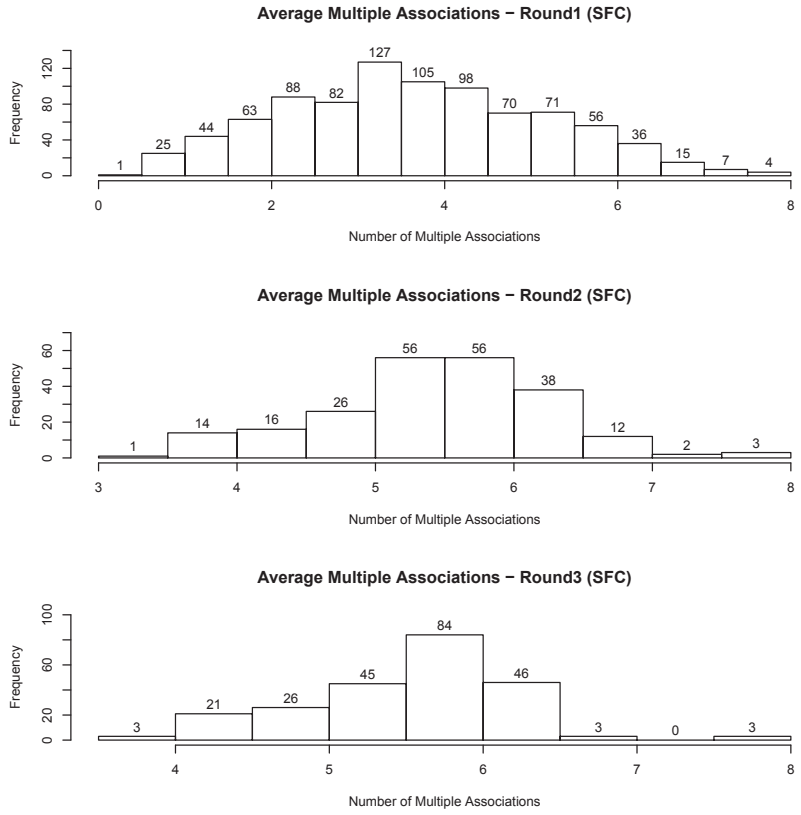


Figure 3.12: Histogram of the average number of multiple associations obtained on the simulated datasets derived from the SFC dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average number of multiple associations of the simulated symbiont trees, given a cost vector.

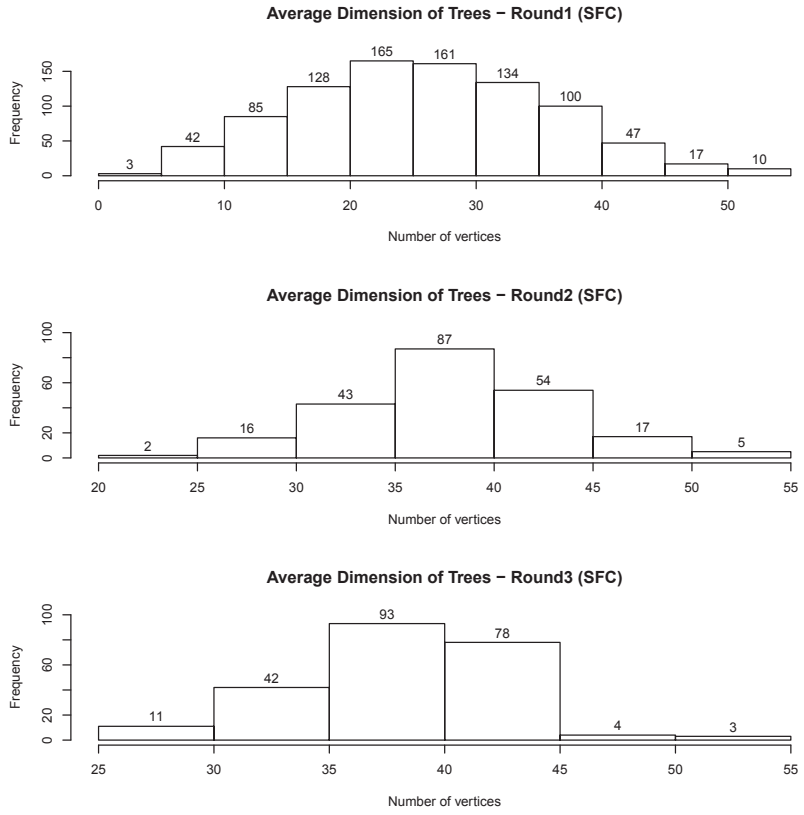


Figure 3.13: Histogram of the average number of tree size obtained on the simulated datasets derived from the SFC dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average tree size of the simulated symbiont trees, given a cost vector.

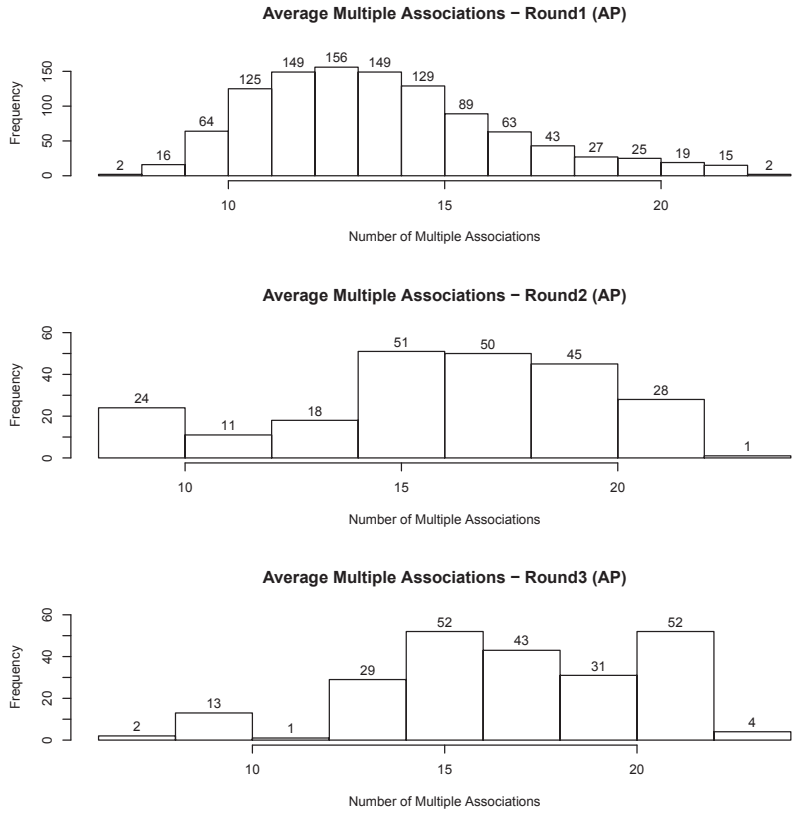


Figure 3.14: Histogram of the average number of multiple associations obtained on the simulated datasets derived from the AP dataset.For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average number of the multiple associations of the simulated symbiont trees, given a cost vector.

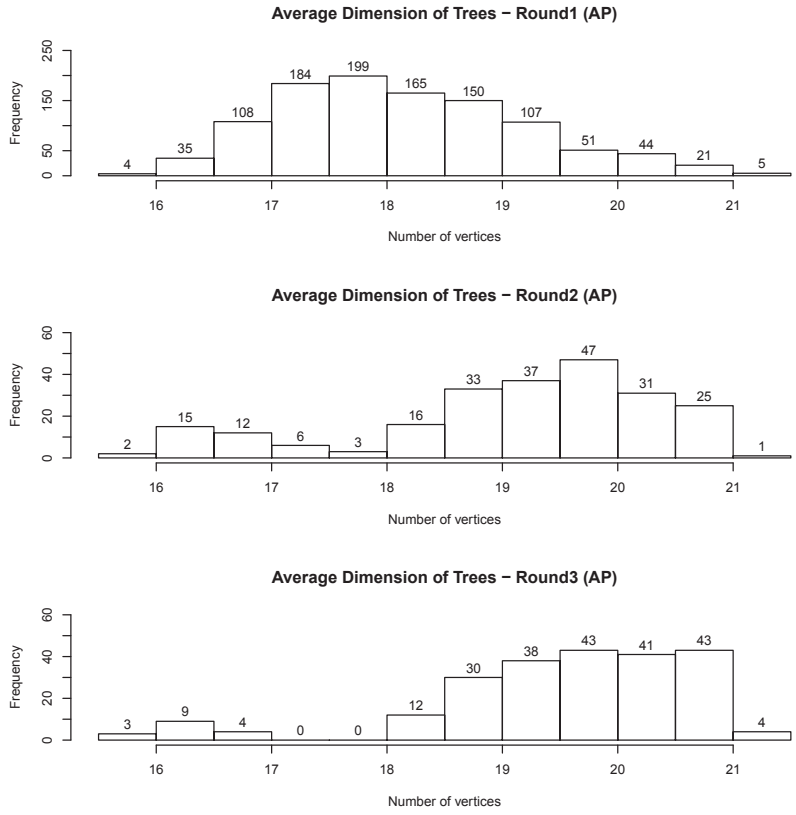


Figure 3.15: Histogram of the average number of tree size obtained on the simulated datasets derived from the AP dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average tree size of the simulated symbiont trees, given a cost vector.

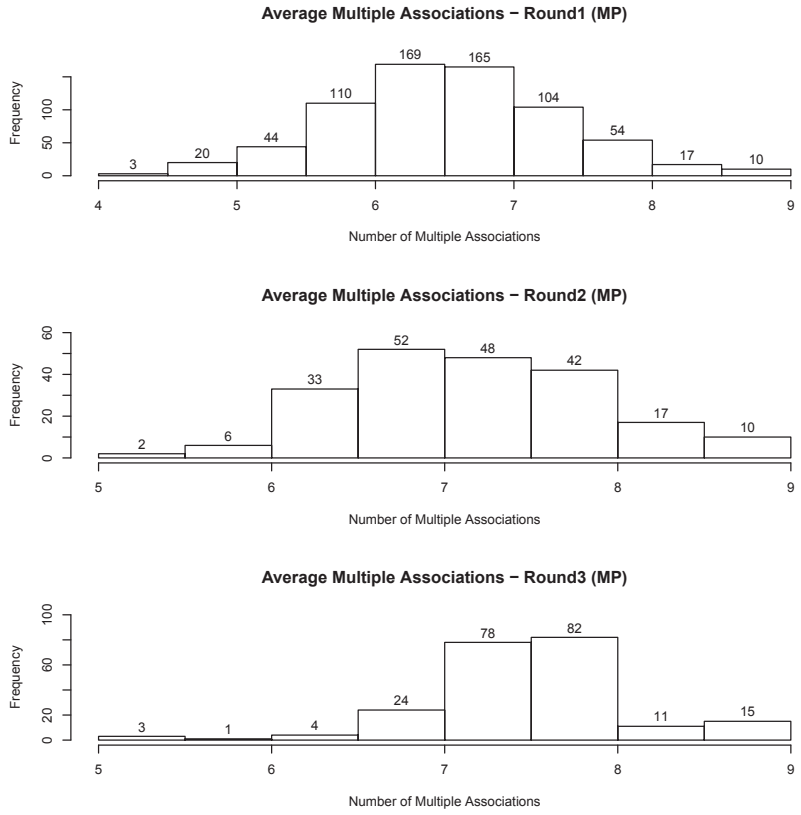


Figure 3.16: Histogram of the average number of multiple associations obtained on the simulated datasets derived from the MP dataset.For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average number of multiple associations of the simulated symbiont trees, given a cost vector.

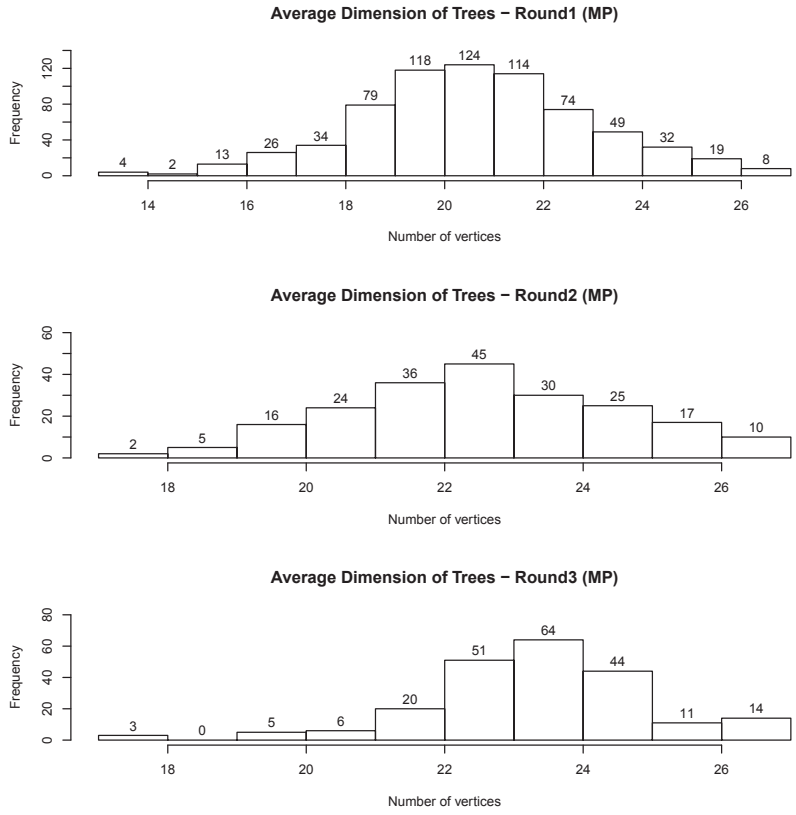


Figure 3.17: Histogram of the average number of tree size obtained on the simulated datasets derived from the MP dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average tree size of the simulated symbiont trees, given a cost vector.

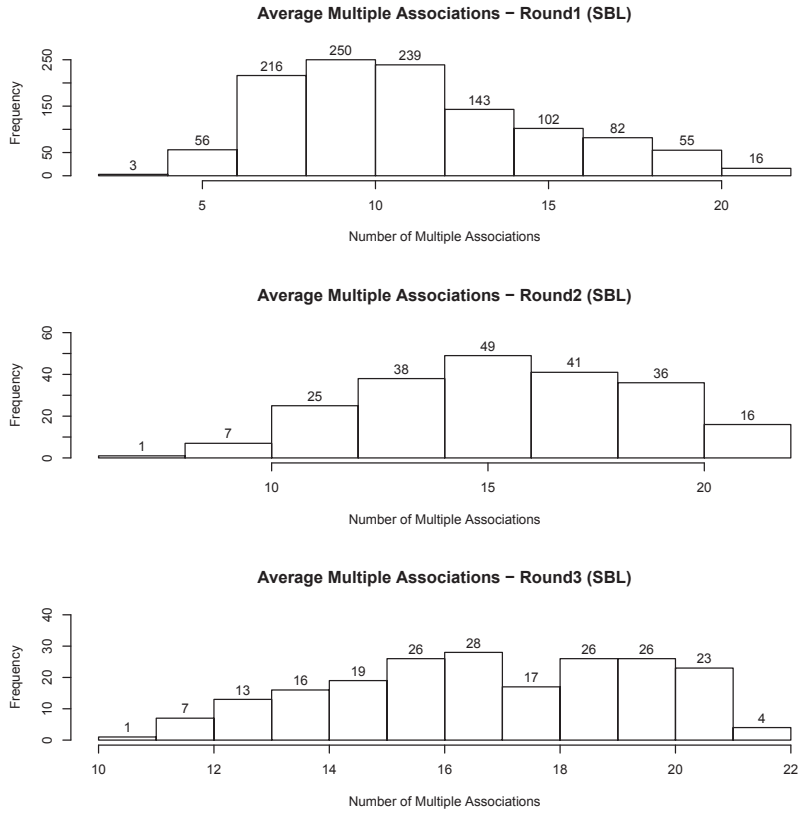


Figure 3.18: Histogram of the average number of multiple associations obtained on the simulated datasets derived from the SBL dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average number of multiple associations of the simulated symbiont trees, given a cost vector.

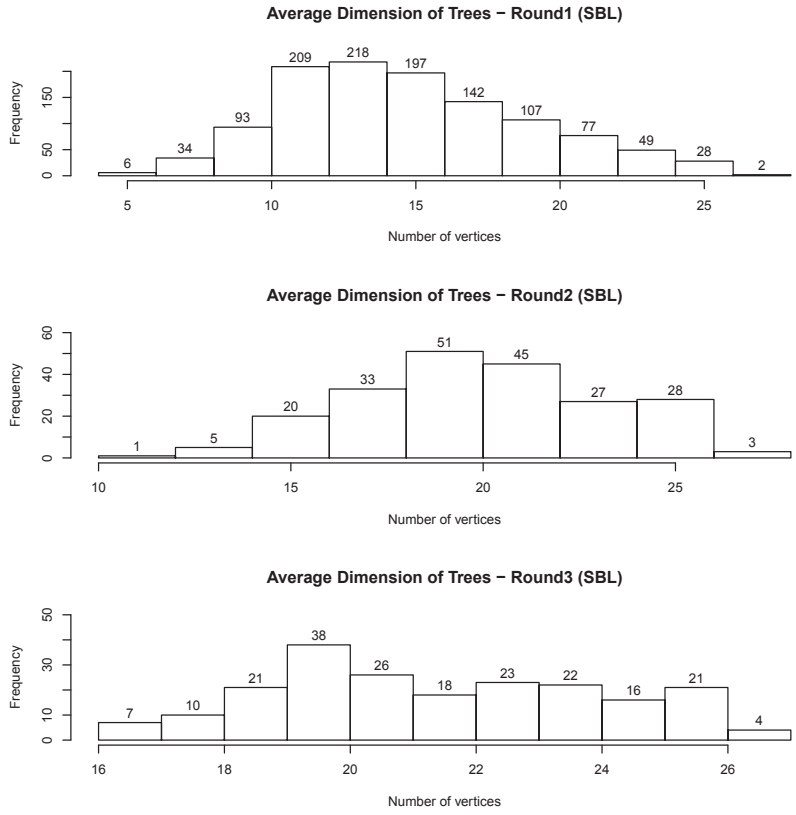


Figure 3.19: Histogram of the average number of tree size obtained on the simulated datasets derived from the SBL dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average tree size of the simulated symbiont trees, given a cost vector.

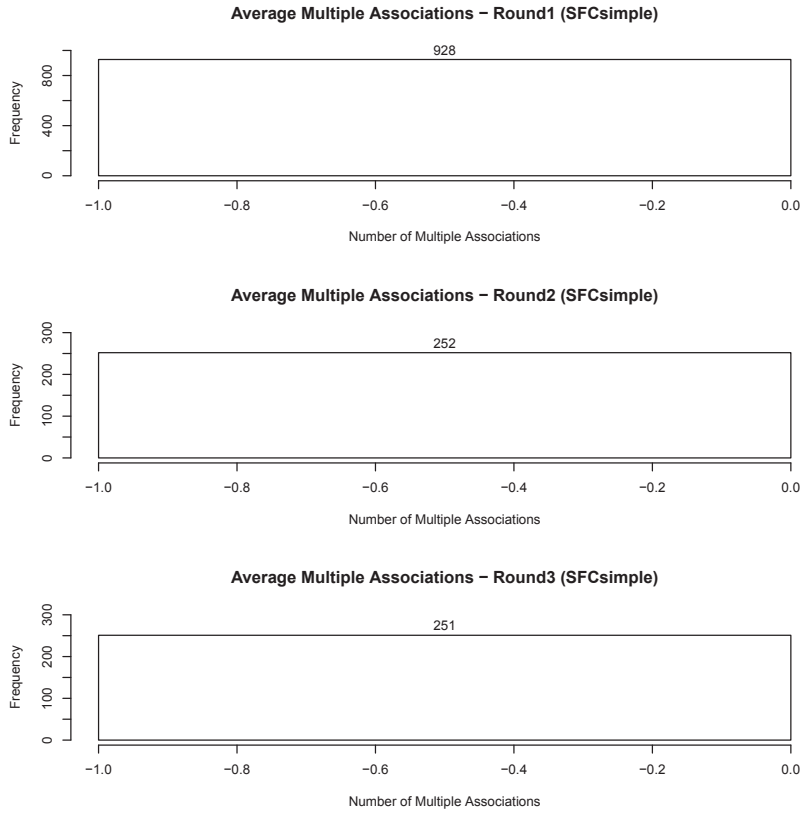


Figure 3.20: Histogram of the average number of multiple associations obtained on the simulated datasets derived from the SFCsimple dataset.For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average number of multiple associations of the simulated symbiont trees, given a cost vector.

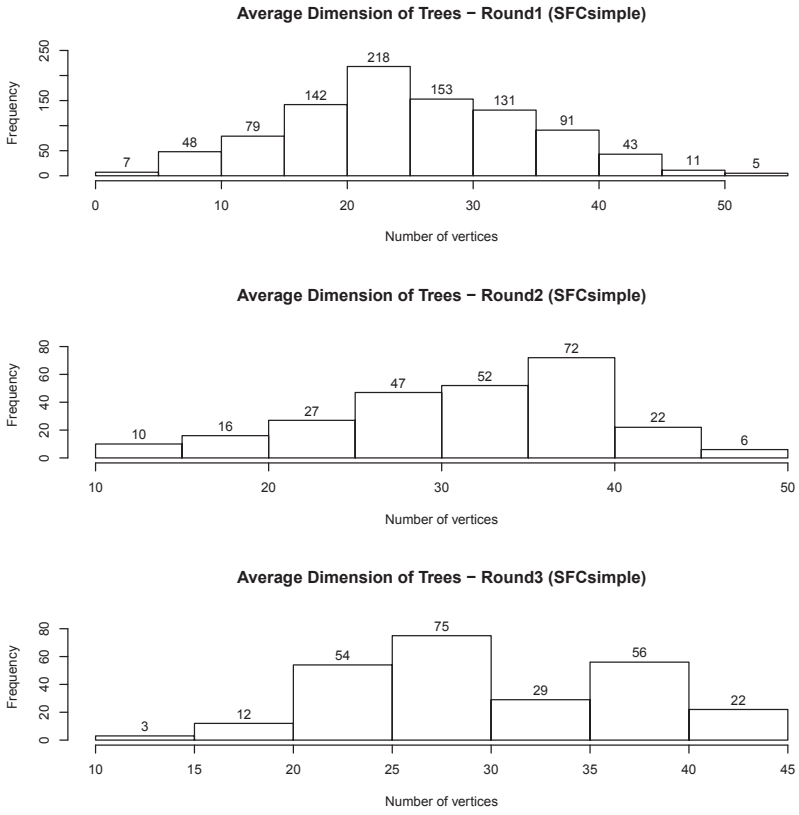


Figure 3.21: Histogram of the average number of tree size obtained on the simulated datasets derived from the SFCsimple dataset. For each cost vector, $M = 1000$ symbiont trees are generated. Each line represents a round of AmoCoala. The histogram shows the average tree size of the simulated symbiont trees, given a cost vector.

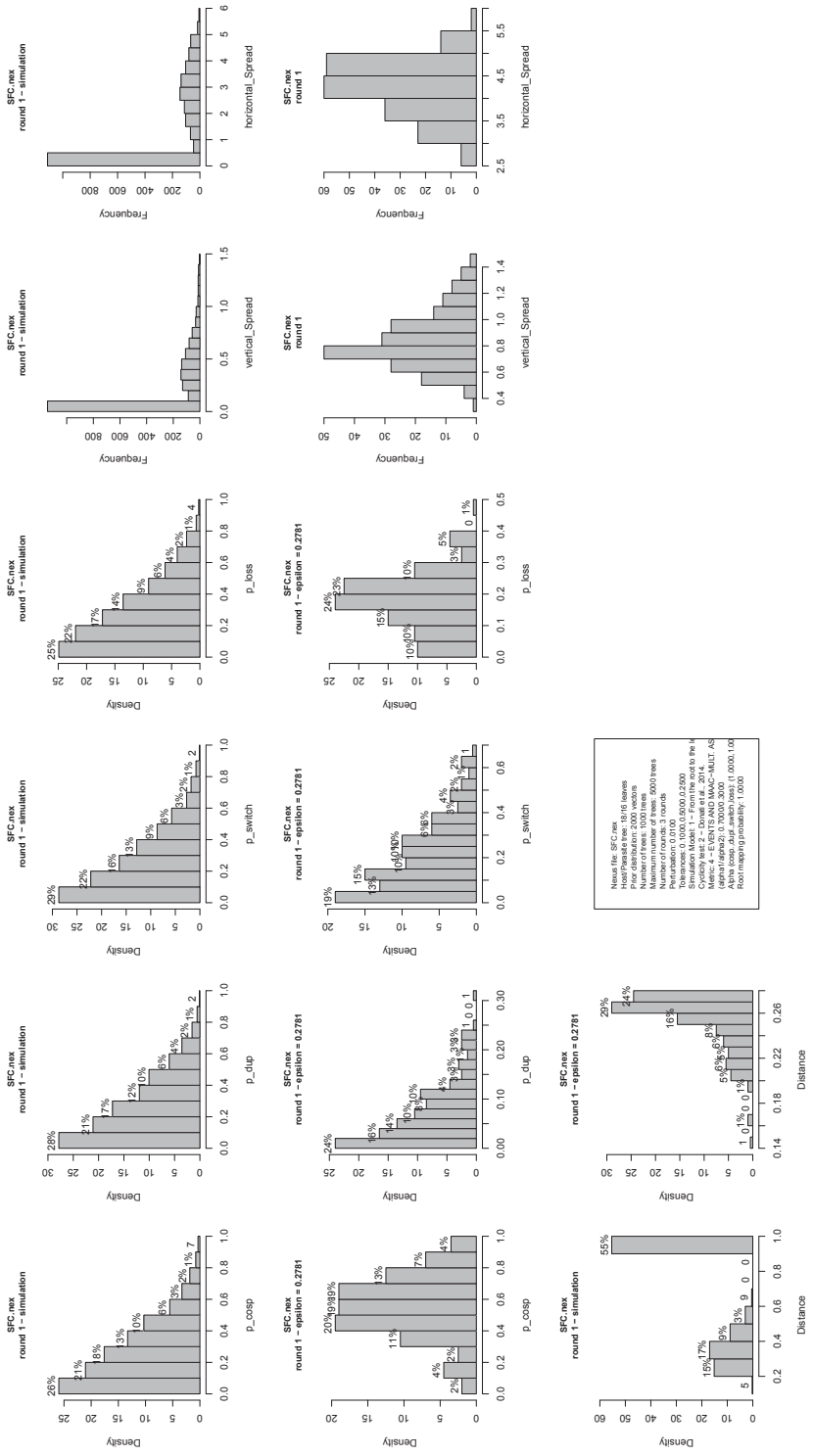


Figure 3.22: Results for the SFC dataset at round 1: First row: 4 event probability histograms (for cospeciation, duplication, host switch and loss) and 2 number of event histograms (for vertical and horizontal spreads) of all the simulated parameter vectors at round 1. Second row: 4 event probability histograms (for cospeciation, duplication, host switch and loss) and 2 number of event histograms (for vertical and horizontal spreads) of all the accepted parameter vectors at round 1. Third row: Representative distance histograms for all the simulated parameter vectors (first column) and accepted parameter vectors (second column).

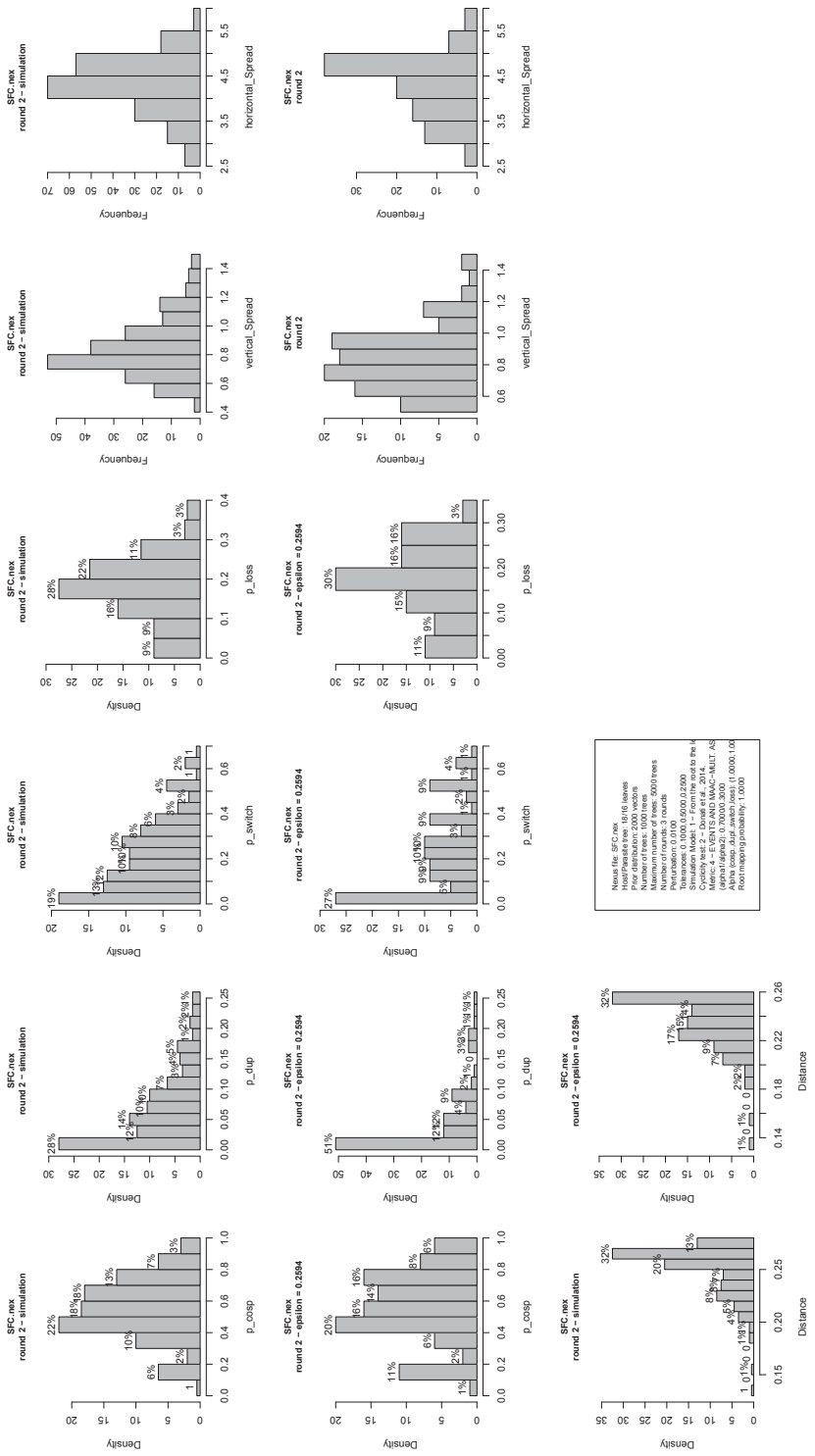


Figure 3.23: Results for the SFC dataset at round 2: First row: 4 event probability histograms (for cospeciation, duplication, host switch and loss) and 2 number of event histograms (for vertical and horizontal spreads) of all the simulated parameter vectors at round 2. Second row: 4 event probability histograms (for cospeciation, duplication, host switch and loss) and 2 number of event histograms (for vertical and horizontal spreads) of all the accepted parameter vectors at round 2. Third row: Representative distance histograms for all the simulated parameter vectors (first column) and accepted parameter vectors (second column).

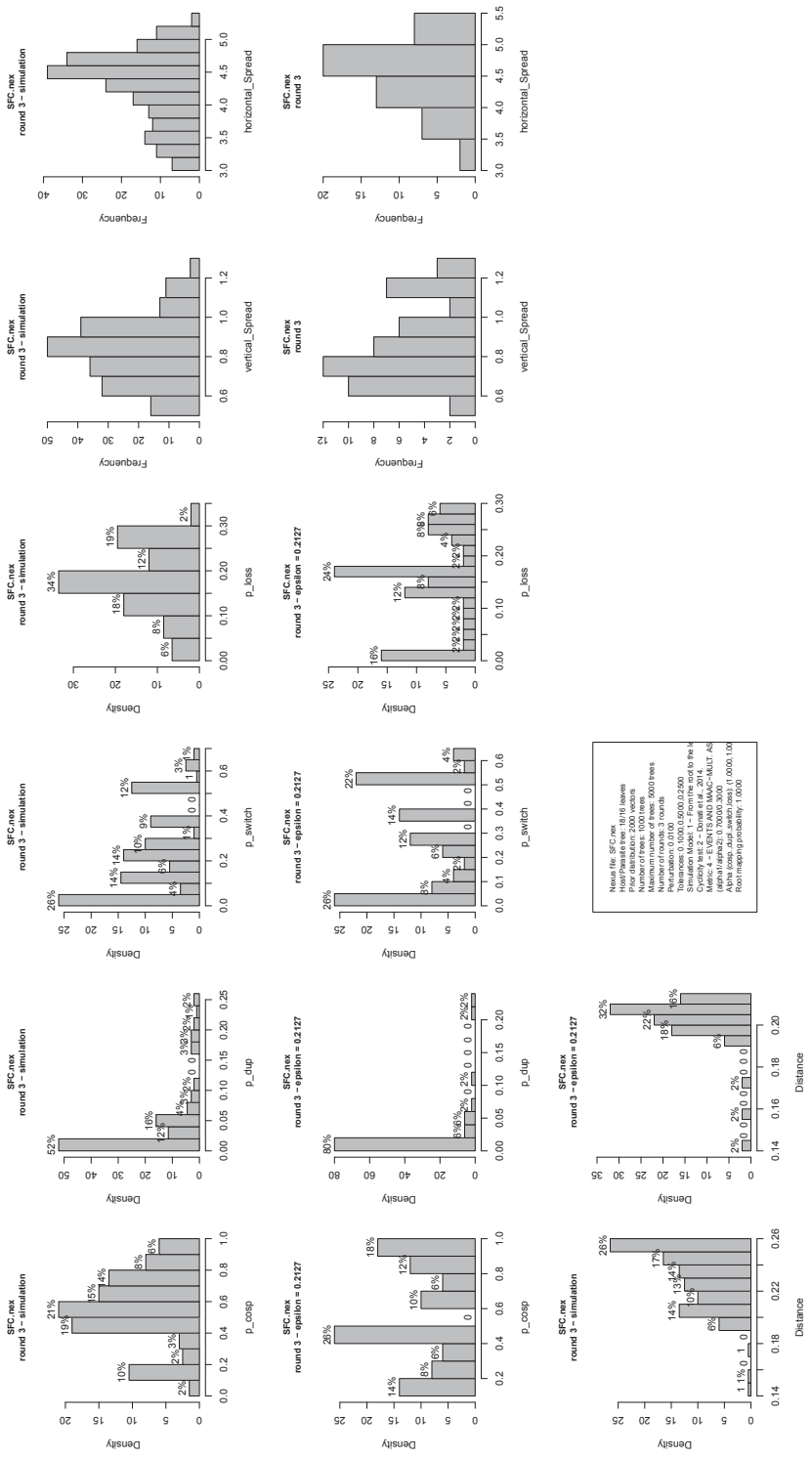


Figure 3.24: Results for the SFC dataset at round 3: First row: 4 event probability histograms (for cospeciation, duplication, host switch and loss) and 2 number of event histograms (for vertical and horizontal spreads) of all the simulated parameter vectors at round 3. Second row: 4 event probability histograms (for cospeciation, duplication, host switch and loss) and 2 number of event histograms (for vertical and horizontal spreads) of all the accepted parameter vectors at round 3. Third row: Representative distance histograms for all the simulated parameter vectors (first column) and accepted parameter vectors (second column).

Chapter 4

General Conclusion and Perspectives

In this Ph.D., we aimed at understanding and creating models and algorithms to study the common evolutionary history of host and symbionts.

This study was focused on two main aspects. The first is the robustness of the model. The objective in this case was to better understand the strengths and weaknesses of the parsimonious reconciliation method. We analysed the robustness in the two cases: of "editing" (multiple associations) and of "small perturbations" of the input (rooting problem). Notice that the first case is in general due to the fact that in the parsimonious method, it is difficult 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 considered involved, 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 explored the influence of six cost vectors that are commonly used in the literature.

We observed that the choice of leaf associations and the re-rooting may have a strong impact on the variability of the reconciliation output. We also noticed that the host switch event has an important role in particular for the rooting problem. Indeed, we showed that if we use a model that contains host switches, the plateau property is no longer verified. It will be interesting to better understand which role this event plays on the robustness of the model. For instance, it would be interesting to know if, in the case where we have a reconciliation with more than one plateau, the number of host switches in the optimal reconciliations is always the same. More work must be done in this direction.

Given a same cost vector, in order to evaluate the similarity of the output of two different optimal reconciliations, we could use the *pattern* of integers $\pi = \langle n_c, n_d, n_s, n_l \rangle$ representing the number of each event that it contains. However, different optimal reconciliations may induce such same pattern. A future work can be to study the robustness using a method that can better distinguish two reconciliations.

Another important point is that the correctness of the input datasets depend not only on a correct rooting. More in general, it is not easy to be sure of the correctness of a dataset and a wrong phylogeny will affect the final result. For the future, it would be interesting to be able to infer the phylogenetic information directly from the sequence data.

Finally, we noticed that using Eucalypt, we can enumerate all possible optimal reconciliations, but it can happen that these reconciliations are time unfeasible. Another important

work could thus be to find the suboptimal reconciliations, meaning reconciliations with a slightly larger cost that are time feasible.

The second goal of the Ph.D. was to refine the existent reconciliation model and make it more realistic. In the literature, there are different datasets with symbiont leaves associated to more than one host leaf (multiple association case). A few recent algorithms treat this case, but in limited ways. Most often, only the four "classical events" are used (cospeciation, duplication, host switch and loss). In this study, we thus considered also the *spread* event. The latter corresponds to the invasion of different hosts by a same symbiont. We developed a statistical method adding such event to infer the probabilities of the the four "classical events".

To this purpose, we modified the model of Coala and we created a new one, AmoCoala. This model allows multiple associations, meaning that a symbiont leaf can be associated to more than one host leaf. Here the dataset is composed by a host tree H, a symbiont tree S and the association between leaves ϕ that in this case is no longer a function from $L(S) \rightarrow L(H)$. It can rather be viewed as a bipartite network on the set $L(S) \cup L(H)$ and encoded as a binary adjacency matrix ϕ with size $|L(S)| \times |L(H)|$. If the symbiont s is associated to the host h, the value in the matrix ϕ_{sh} is 1, it is zero otherwise.

We tested AmoCoala on synthetic and on real datasets. Horizontal and vertical spread events are simulated based on the probabilities associated to each host event. We controlled if AmoCoala is able to produce trees $\tilde{S}$ similar in terms of number of multiple associations and of size to the original symbiont tree S. We showed that, after each round, the method tends to select vectors producing trees near to the original one, although the generated trees tend to be slightly bigger in terms of both factors. This is due to the fact that, during the simulations, if we have a horizontal or a vertical spread, we simulate the symbiont tree choosing the topology and the leaf associations identical to those observed in the original symbiont tree. We also tested AmoCoala using the real datasets SFC and SFCsimple . The two datasets are identical except for the associations (SFCsimple has no multiple associations). Based on the study of Refrégier *et al.*, we expected that the reconciliation would have zero probability of duplication and high probability of host switch. We showed that the results obtained with AmoCoala which allows to consider the whole dataset SFC with multiple associations, are closer to the result presented in [65] than those that were obtained by Coala which ignores such multiple associations. This shows again the importance of taking into account the latter.

With AmoCoala it is now possible to directly study datasets where a symbiont is associated to more than one host. The accuracy of the results obtained by our model depend on the distance used to capture the similarity between trees. The accepted parameter vectors are obtained by the ABC-SMC procedure and characterise the list of vectors that may explain the evolution of the pair of host and symbiont trees given as input. In AmoCoala, we also developed a new distance, namely MAS and denoted by d_{MAS}, for comparing two set-labelled trees. The number of events obtained during the simulation of a tree is important for choosing the accepted parameters: we expected that such number would be similar to the number of expected events. This calculation is not easy to do in the case of spread events. In fact, "classical events" (cospeciation, duplication, host switch and loss) have the same probability everywhere in the tree, while the probability of a vertical spread is specific to each vertex of the host tree. It would be interesting to be able to infer a probability of spread that is general for the tree, that could then be used to choose the accepted parameters.

One future work will be to include spread events in Eucalypt which would allow to reconcile

trees where the symbionts are associated to more than one host. A more efficient exploration of the parameter space allowing to handle larger trees and increase the efficiency of the procedure is another important future issue. It will also be important to consider the case where the phylogeny is not fully solved, for example when phylogenetic trees are not binary. In this study, we used trees that are not too big (maximum 34 leaves). It would be important to increase the efficiency of algorithm and allow to handle larger trees. Finally, we do not have to treat all symbionts in the same way. For example, we could have a symbiont that lives inside the host, what is called an endosymbiont. In this case, it may be more difficult for it to be involved in a spread event. An interesting work would thus be to adapt the probability of spread according to the type of symbiont.

Appendix A

Eucalypt

A.1 EUCALYPT - Algorithm 1

Algorithm 1: Finding the cost of an optimal solution

$\langle H, S, \phi \rangle$ **and a cost vector** $\langle c_c, c_d, c_s, c_l \rangle$
Output: Optimal cost
for $s \in V(S)$ *and* $h \in V(H)$ **do**
    Initialise $D(s,h)$,$D_{ST}(s,h)$ to ∞
for $l \in L(S)$ **do**
    Initialise $D(l, \phi(l)) = 0$
    for $a \in Anc(\phi(l))$ **do**
        $D_{ST}(l,a) = c_l * d(a, \phi(l))$
for $s \in V(S)$ *in post order with children* s_1,s_2 **do**
    for $h \in V(H)$ *in post order with children* h_1,h_2 **do**
        if $h \in L(H)$ **then**
            $\delta_d \rightarrow c_d + c(s_1,h) + c(s_2,h)$
            $D(s,h) = \min\{\delta_d, \delta_s\}$
            $D_{ST}(s,h) = D(s,h)$
        else
            $\delta_c \rightarrow \min\{(c_c + D_{ST}(s_1,h_1) + D_{ST}(p2,h2)), (cc + DST(p1,h2) + DST(p2,h1))\}$
            $\delta_d \rightarrow \min\{D(s_1,h) + D(s_2,h), D(s_1,h) + D_{ST}(s_2,h_1) + c_l, D(s_1,h) + DST(s_2,h_2) + c_l, D(s_2,h) + D_{ST}(s_1,h_1) + c_l, D(s_2,h) + D_{ST}(s_1,h_2) + c_l, D_{ST}(s_1,h_1) + D_{ST}(p_2,h_1) + 2c_l, D_{ST}(p_1,h_2) + D_{ST}(p_2,h_2) + 2c_l\}$
            $D(s,h) = \min\{\delta_c, \delta_d, \delta_s\}$
            $D_{ST}(s,h) = \min\{D(s,h), c_l + D_{ST}(s,h_1), c_l + D_{ST}(s,h_2)\}$
return $\min\{D(r(S),h) | h \in V(H)\}$

A.2 EUCALYPT - Algorithm 2

Algorithm 2: Enumerating all optimal solutions

```
1  Input: The dynamic programming matrix D
2  Output: All optimal solutions
3  for All cells root in D containing an optimal mapping of r(S) (or the unique cell mapping
   r(s) to r(H)) do
4      currentCell → root
5      A stack M → ∅
6      do
7          while currentCell! = null do
8              if |List(current)| ≥ 1 then
9                  //There are different sub-solutions for this mapping if M(currentCell) is
                   not in M then
10                     Push(⟨currentCell, 0⟩) in M
11                     currentSubsolution → 0^th-element of M(currentCell)
12                 else if M(currentCell) is the last element of M then
13                     //In the final part of the solution I pass to consider the next option
14                     Pop(⟨currentCell, i⟩) from M
15                     Push(⟨currentCell, i + 1⟩) in M
16                     currentSubsolution → (i + 1)^th-element of M(currentCell)
17                 else
18                     //In the first part of the current solution, the mappings are the same
                       as for the previous one
19                     ⟨cell, index⟩ → M(currentCell)
20                     currentSubsolution → index^th-element of M(currentCell)
21             else
22                 //There is a unique possible sub-solution
23                 Add to the solution the mapping relative to currentCell
24                 currentSubsolution → 0^th-element of List(currentCell)
25                 //currentSubsolution is unique (or null if the vertex is a leaf)
26                 currentCell = the next vertex following the pointers of
                   currentSubsolution (in post-order)
27             Output the solution
28             Pop from M until the first couple ⟨s, i⟩ is found for which i < |M(s)| − 1 and
               the stack is not empty
29     while M is not empty;
```

Appendix B

Robustness – Supplementary Material

B.1 Supplementary figures

Figure B.1 represents the SBL dataset (host and symbiont trees together with their leaf associations).

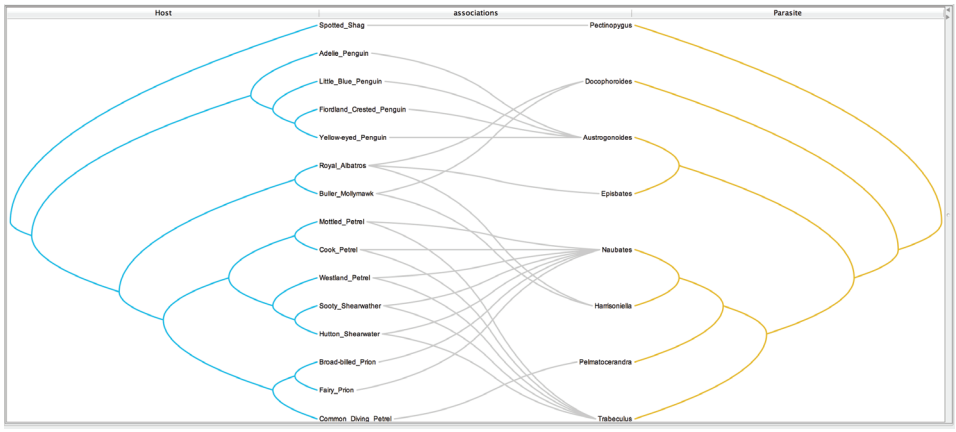


Figure B.1: Dataset SBL: host (left) and symbiont (parasite, on the right) trees together with their leaf associations (middle). The figure was created using TreeMap3 [11].

B.2 Additional results

B.2.1 Changing associations for real datasets

Figures B.2 to B.6 contain the results for the SBL dataset with the cost vectors in $\mathcal{C}$ (except for $c = \langle 0, 1, 1, 1\rangle$ that appeared in the paper). Figures B.7 to B.79 then present the results for the fourteen remaining datasets (AP, AS, CA, CP, FA, FE, GM, MF, MP, RM, SFC, SHA, TC and TD) with all cost vectors in $\mathcal{C}$. Only time-feasible reconciliations are retained.

Notice that in the case of the dataset CP, for the cost vectors $c = \langle -1, 1, 1, 1\rangle$, $c = \langle 0, 1, 1, 1\rangle$, $c = \langle 0, 1, 2, 1\rangle$ and $c = \langle 0, 2, 3, 1\rangle$, all the optimal reconciliations are cyclic, so that no time-feasible reconciliation is retained. In the case of SHA, for all cost vectors c, there is only one pattern for the most frequent optimal cost, so that there are no results to show. In the case

of TD, we did not obtain any results for the cost vector $c = \langle 1, 1, 1, 1 \rangle$. Indeed, the number of reconciliations was too large and our program could not handle all these solutions (due to memory capacities).

Let us now comment some of the results. There are some cases where the dissimilarity for a cost vector is always zero, and we checked that this is because the multisets of patterns are equal. This is the case of AS for cost vectors $c = \langle -1, 1, 1, 1 \rangle$, $c = \langle 0, 1, 1, 1 \rangle$, $c = \langle 0, 1, 2, 1 \rangle$ and $c = \langle 0, 2, 3, 1 \rangle$ (Figures B.13 to B.16); FA for cost vectors $c = \langle 0, 1, 2, 1 \rangle$, $c = \langle 1, 1, 3, 1 \rangle$ (Figures B.29 and B.32); FE for cost vector $c = \langle 0, 2, 3, 1 \rangle$ (Figure B.36); SFC for cost vector $c = \langle -1, 1, 1, 1 \rangle$ (Figure B.63); TC for all cost vectors in $\mathcal{C}$ (Figures B.69 to B.74) and TD for cost vectors $c = \langle -1, 1, 1, 1 \rangle$, $c = \langle 0, 1, 1, 1 \rangle$, $c = \langle 0, 1, 2, 1 \rangle$, $c = \langle 0, 2, 3, 1 \rangle$ and $c = \langle 1, 1, 3, 1 \rangle$ (Figures B.75 to B.79).

There are cases where there is only one optimal cost and there is only one non-zero dissimilarity. We checked that this is because there are exactly two multisets of patterns among all the possible reconciliations. This is the case of AS for cost vectors $c = \langle 1, 1, 1, 1 \rangle$ and $c = \langle 1, 1, 3, 1 \rangle$ (Figures B.17 and B.18); CA for cost vector $c = \langle 1, 1, 1, 1 \rangle$ (Figure B.23).

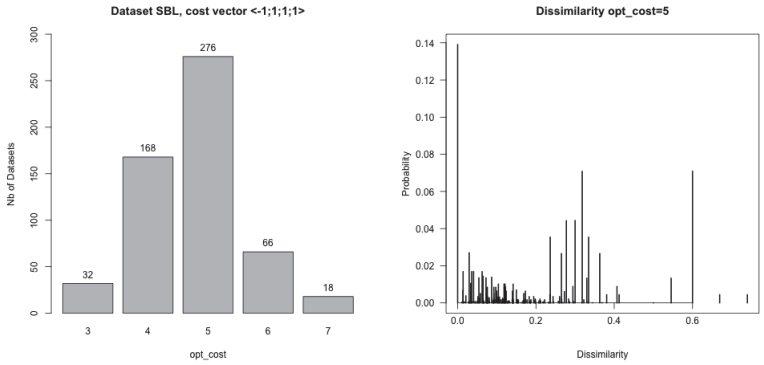


Figure B.2: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 5 (right) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

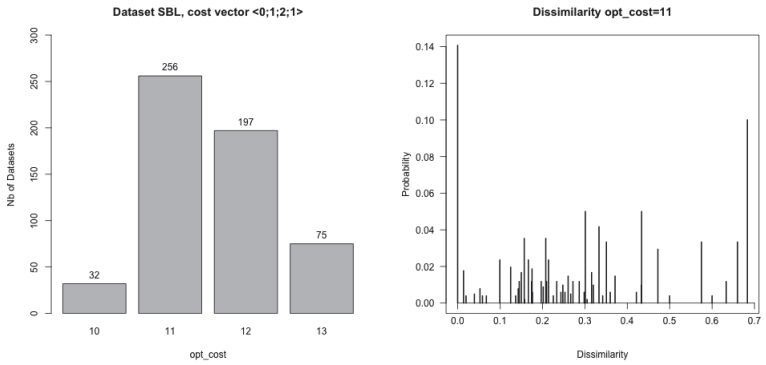


Figure B.3: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 11 (right) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

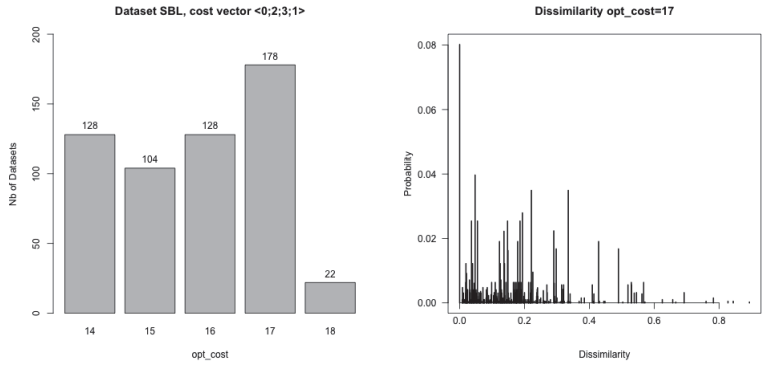


Figure B.4: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 17 (right) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

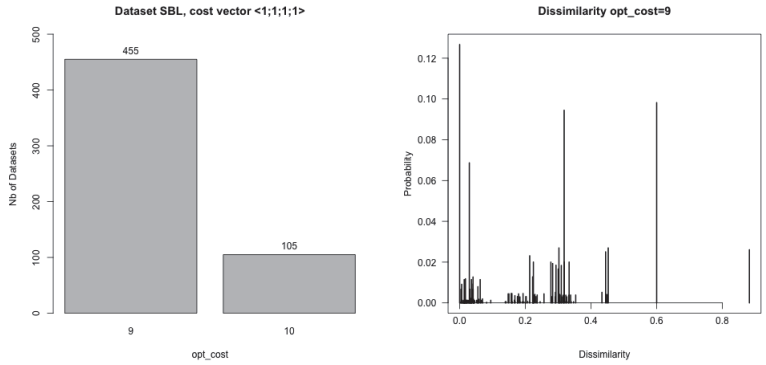


Figure B.5: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

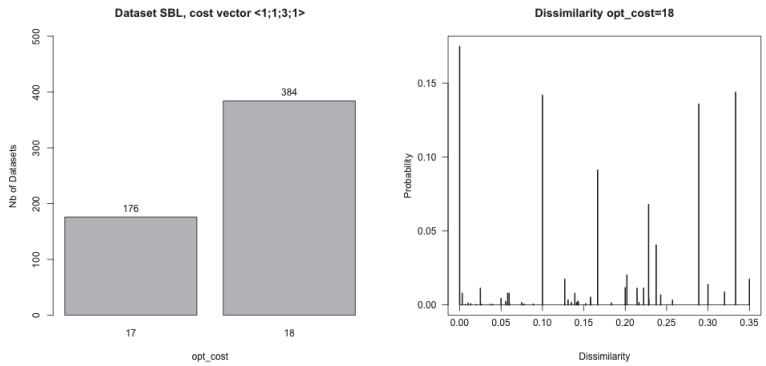


Figure B.6: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 18 (right) obtained on the datasets derived from the SBL dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

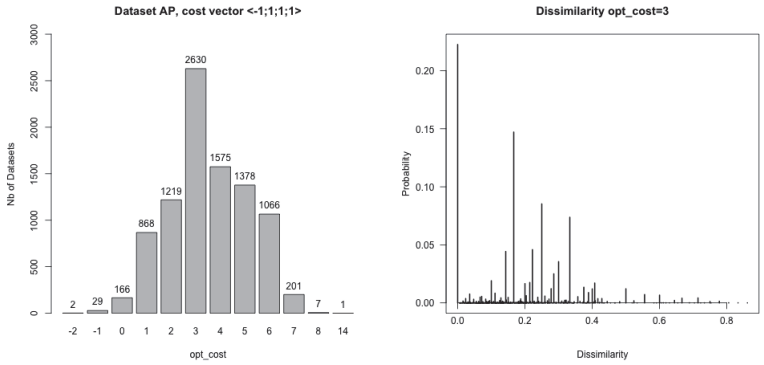


Figure B.7: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 3 (right) obtained on the datasets derived from the AP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

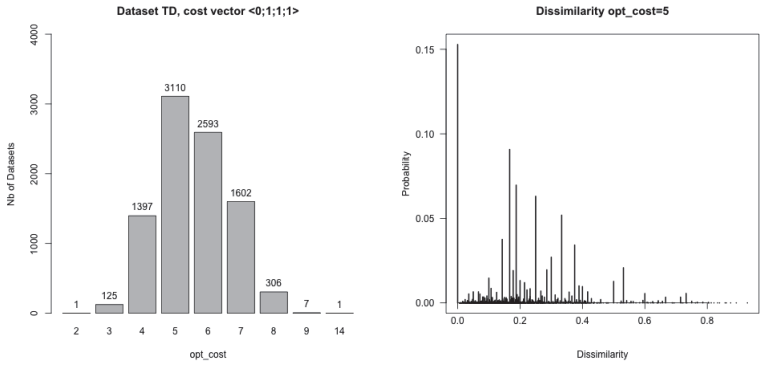


Figure B.8: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 5 (right) obtained on the datasets derived from the AP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

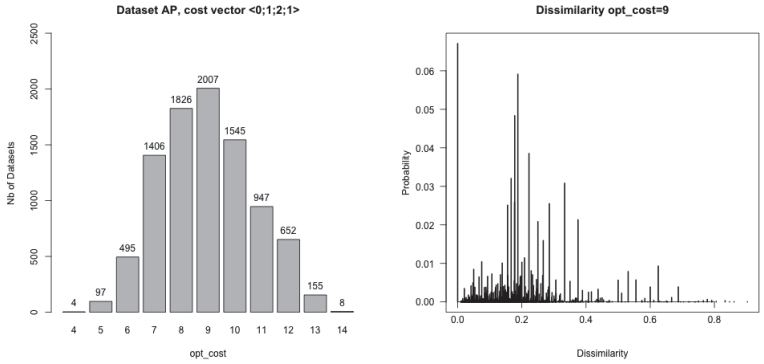


Figure B.9: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the AP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

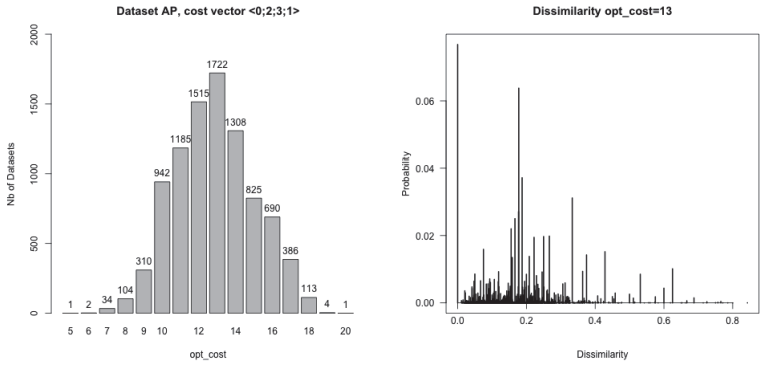


Figure B.10: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 13 (right) obtained on the datasets derived from the AP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

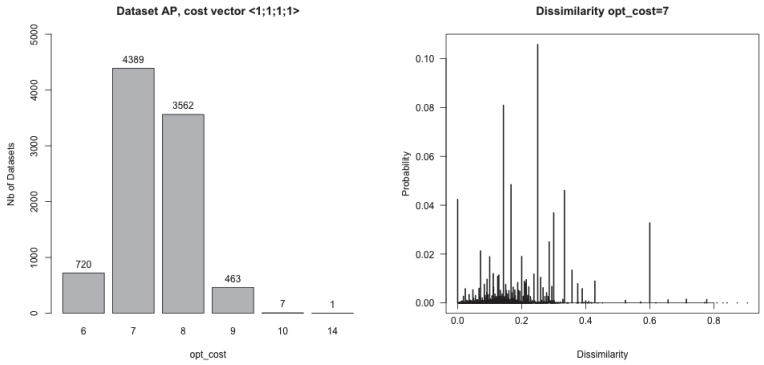


Figure B.11: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 7 (right) obtained on the datasets derived from the AP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

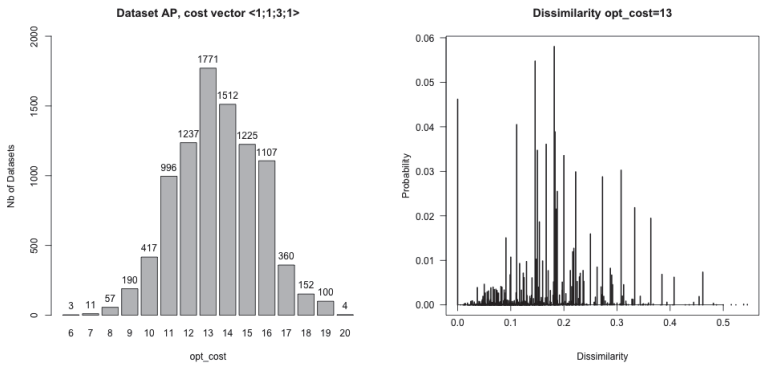


Figure B.12: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 13 (right) obtained on the datasets derived from the AP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

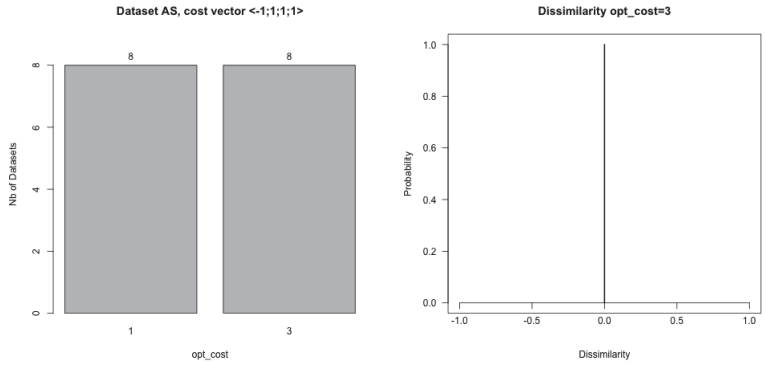


Figure B.13: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 3 (right) obtained on the datasets derived from the AS dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

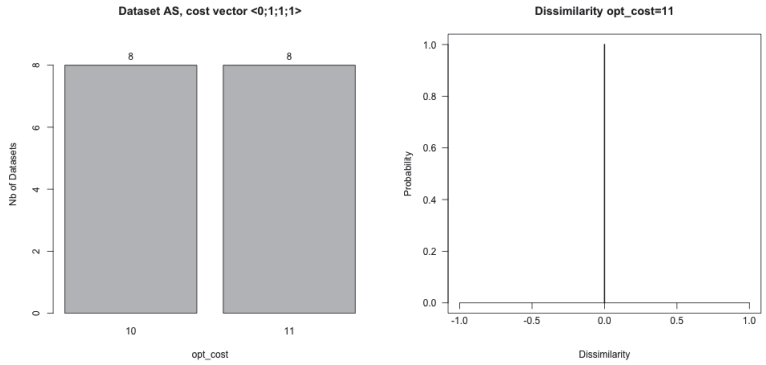


Figure B.14: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 11 (right) obtained on the datasets derived from the AS dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

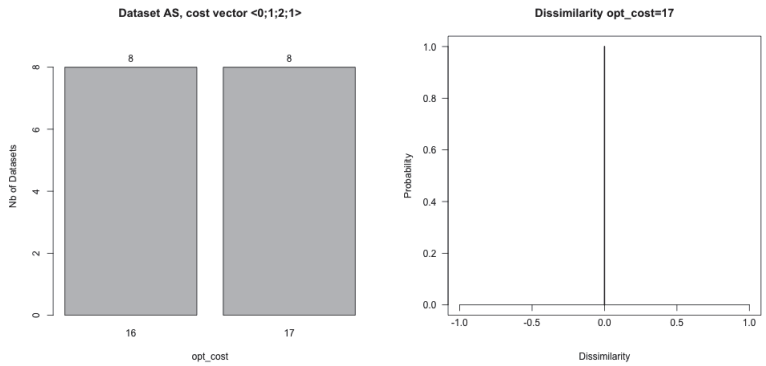


Figure B.15: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 17 (right) obtained on the datasets derived from the AS dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

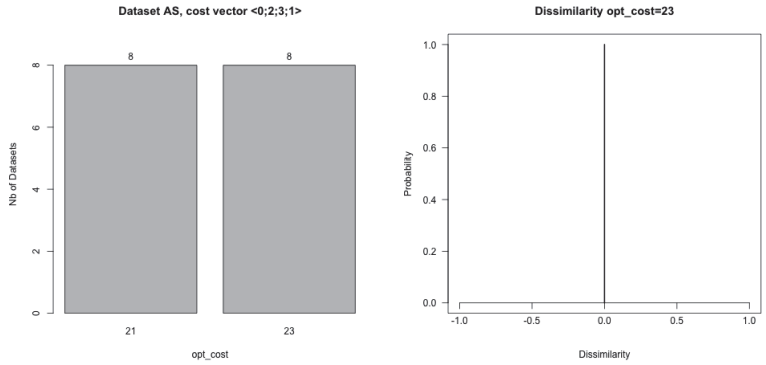


Figure B.16: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 23 (right) obtained on the datasets derived from the AS dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

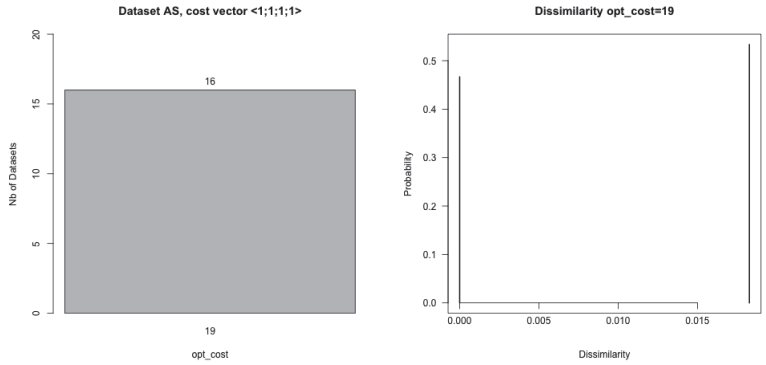


Figure B.17: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 19 (right) obtained on the datasets derived from the AS dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

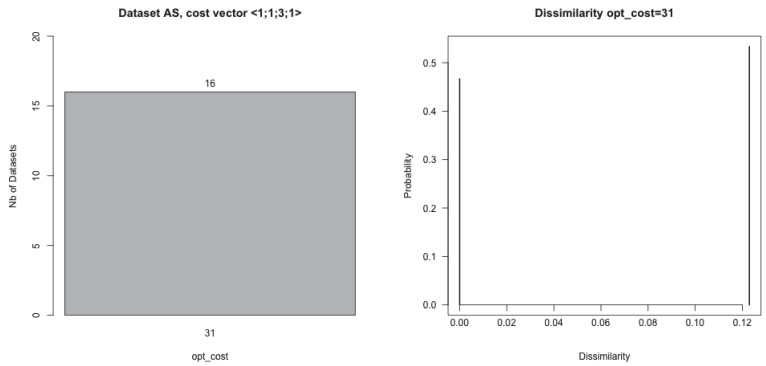


Figure B.18: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 31 (right) obtained on the datasets derived from the AS dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

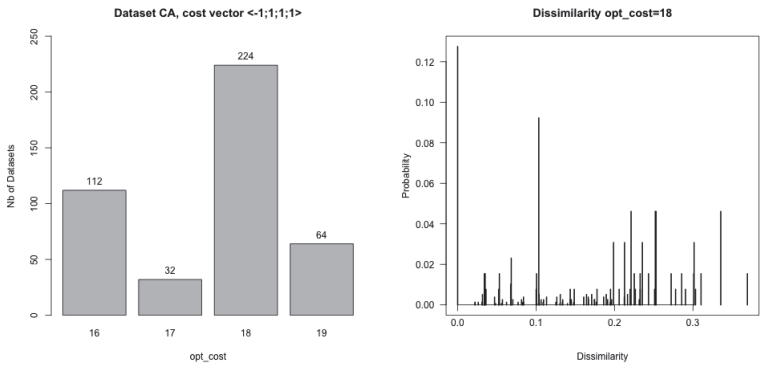


Figure B.19: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 18 (right) obtained on the datasets derived from the CA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

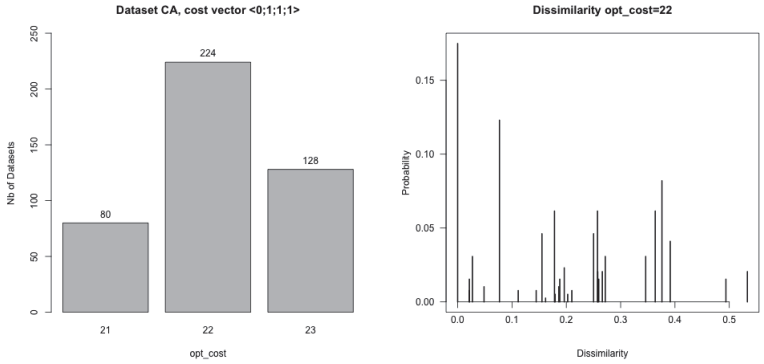


Figure B.20: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 22 (right) obtained on the datasets derived from the CA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

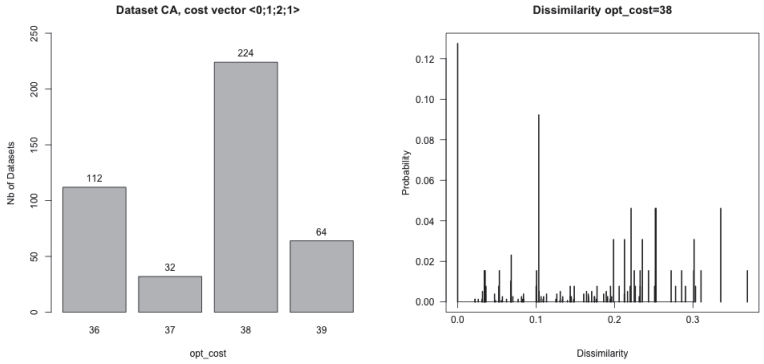


Figure B.21: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 38 (right) obtained on the datasets derived from the CA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

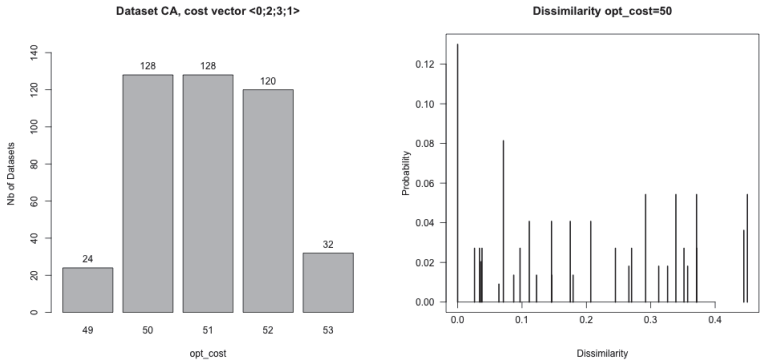


Figure B.22: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 50 (right) obtained on the datasets derived from the CA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

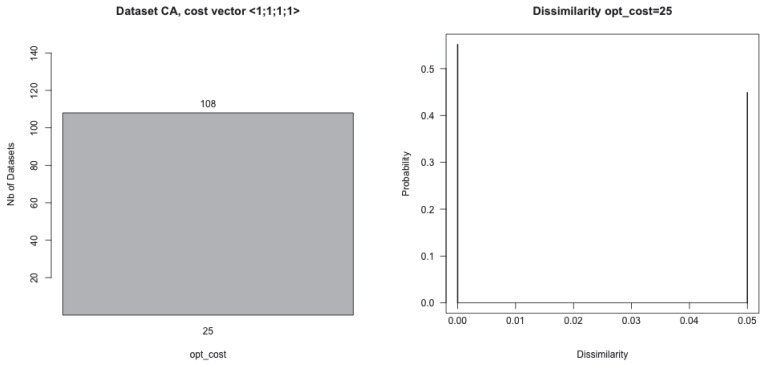


Figure B.23: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 25 (right) obtained on the datasets derived from the CA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

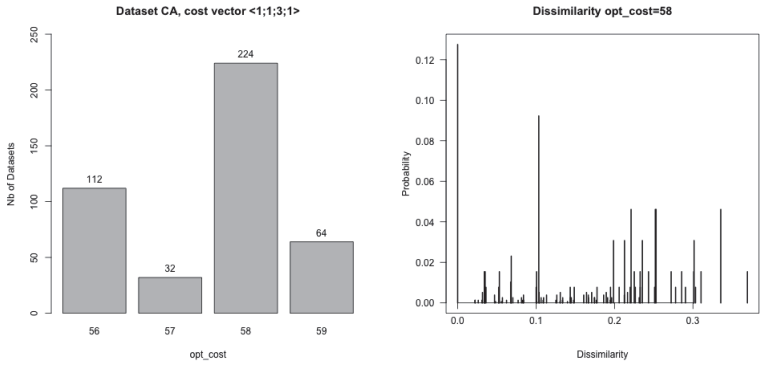


Figure B.24: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 58 (right) obtained on the datasets derived from the CA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

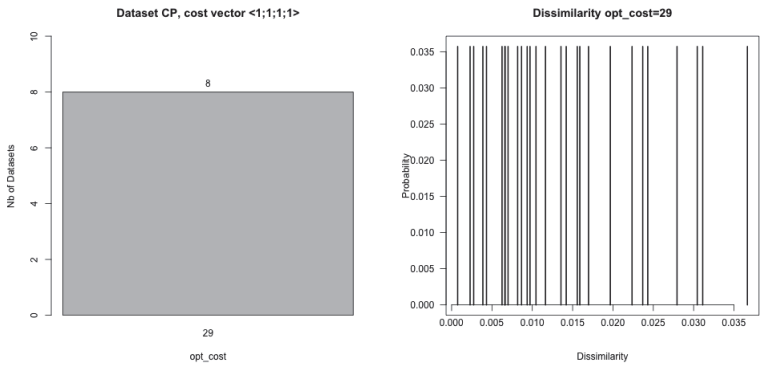


Figure B.25: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 29 (right) obtained on the datasets derived from the CP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

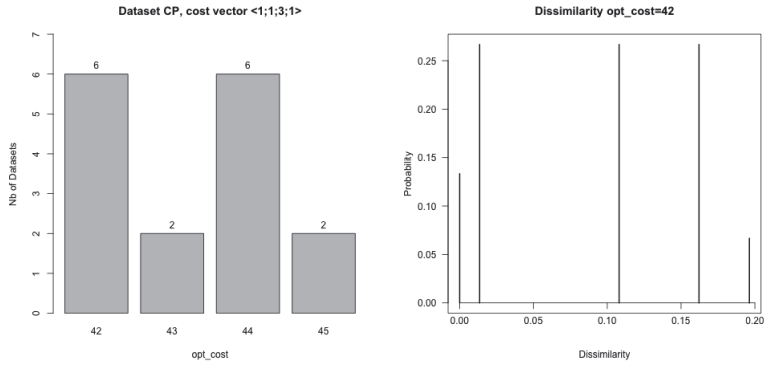


Figure B.26: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 42 (right) obtained on the datasets derived from the CP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

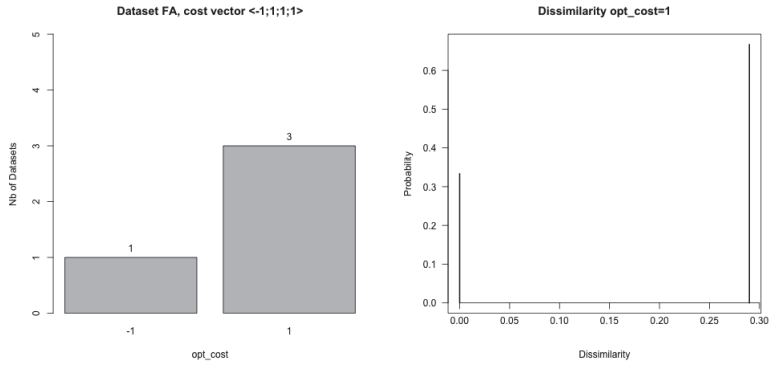


Figure B.27: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 1 (right) obtained on the datasets derived from the FA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

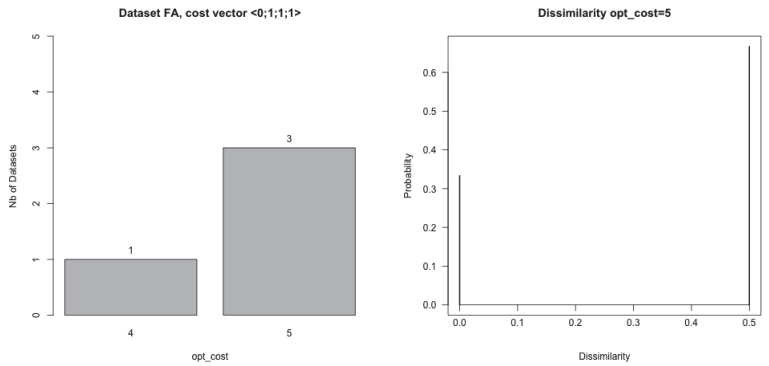


Figure B.28: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 5 (right) obtained on the datasets derived from the FA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

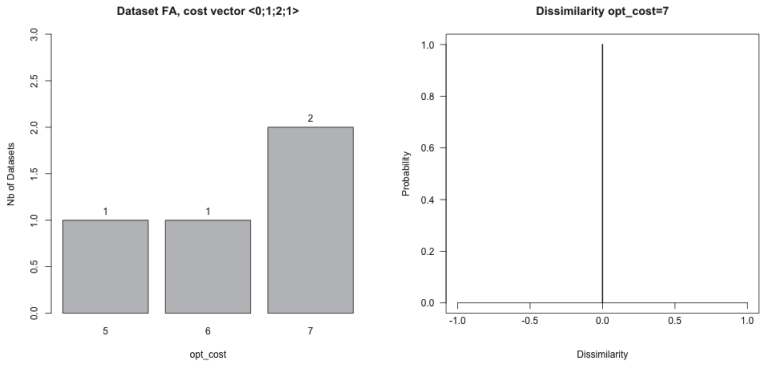


Figure B.29: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 7 (right) obtained on the datasets derived from the FA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

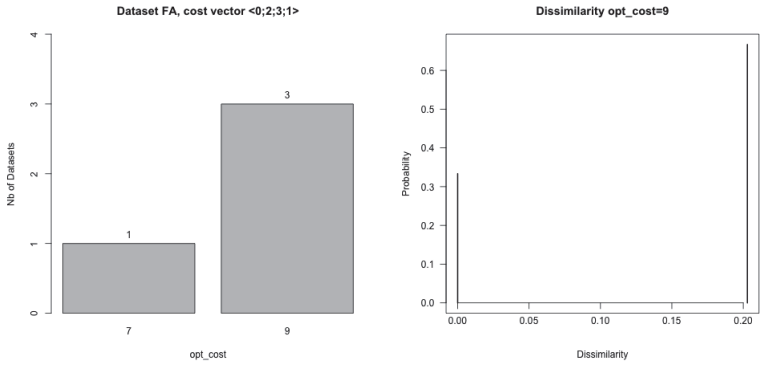


Figure B.30: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the FA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

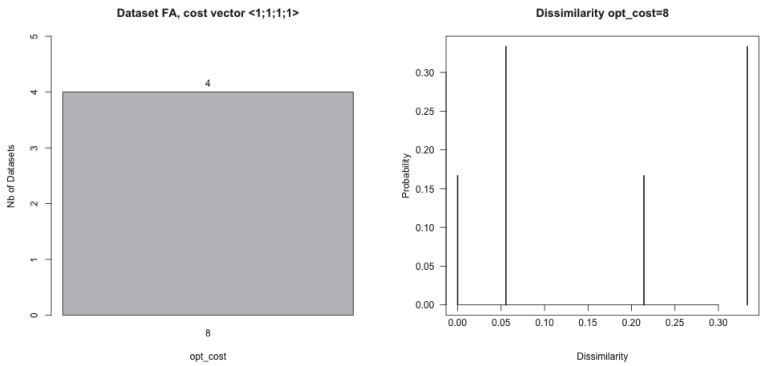


Figure B.31: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 8 (right) obtained on the datasets derived from the FA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

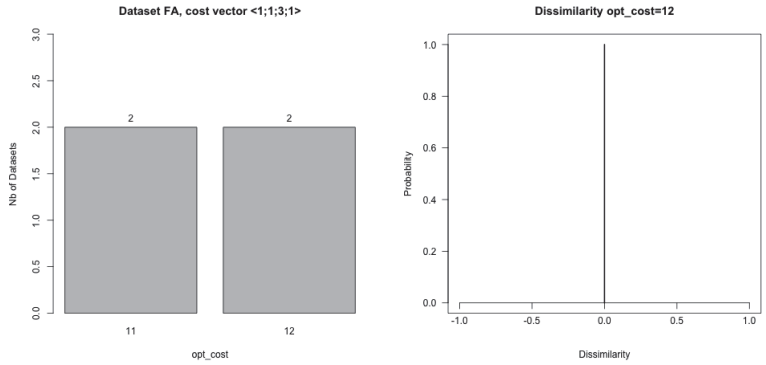


Figure B.32: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 12 (right) obtained on the datasets derived from the FA dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1\rangle$.

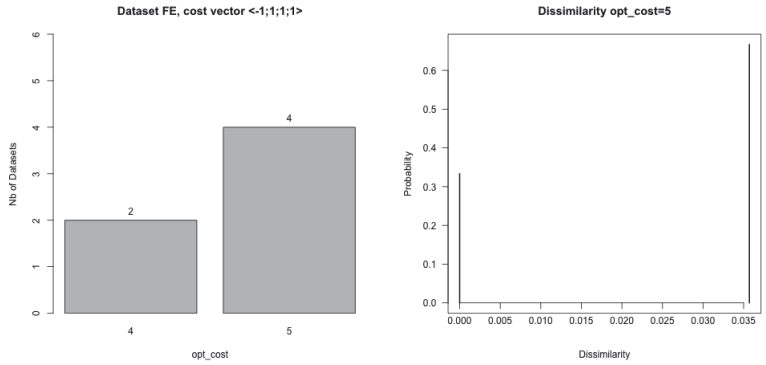


Figure B.33: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 5 (right) obtained on the datasets derived from the FE dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1\rangle$.

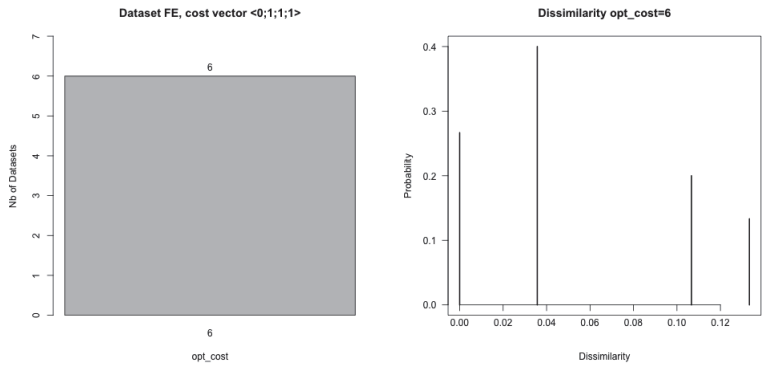


Figure B.34: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 6 (right) obtained on the datasets derived from the FE dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$.

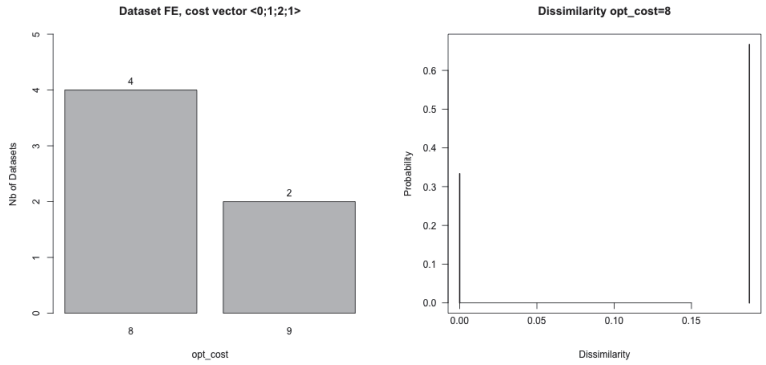


Figure B.35: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 8 (right) obtained on the datasets derived from the FE dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

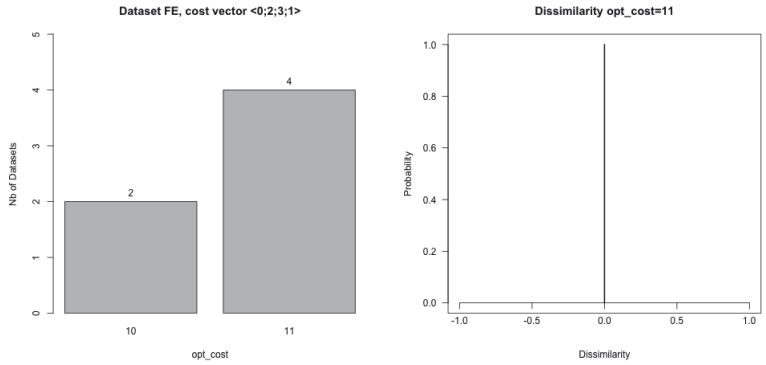


Figure B.36: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 11 (right) obtained on the datasets derived from the FE dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

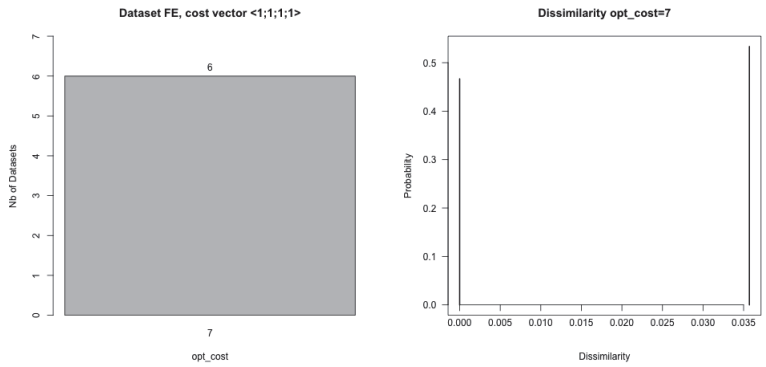


Figure B.37: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 7 (right) obtained on the datasets derived from the FE dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

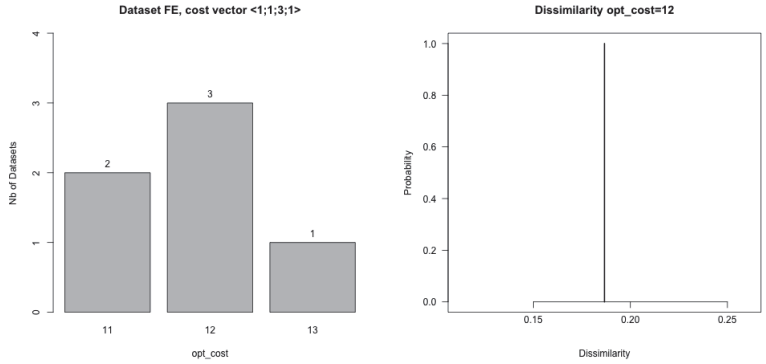


Figure B.38: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 12 (right) obtained on the datasets derived from the FE dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

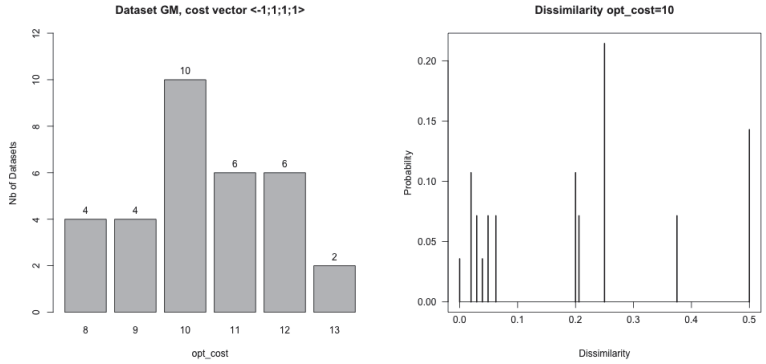


Figure B.39: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 10 (right) obtained on the datasets derived from the GM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

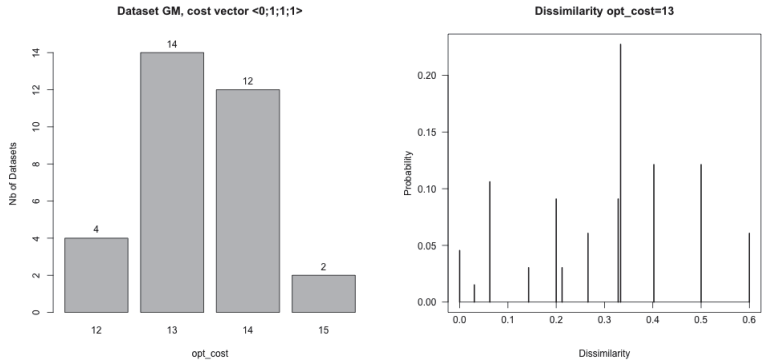


Figure B.40: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 13 (right) obtained on the datasets derived from the GM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

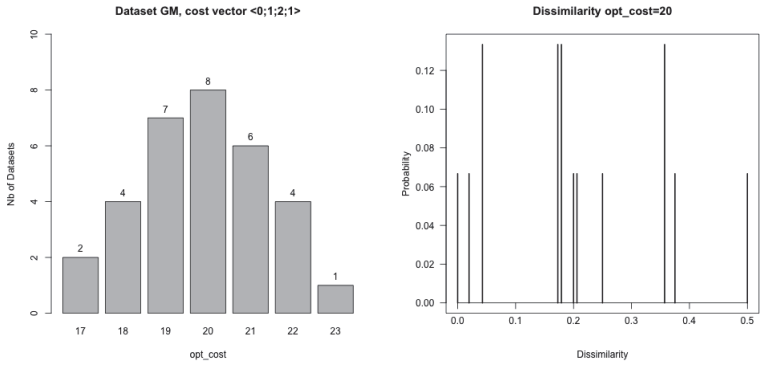


Figure B.41: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 20 (right) obtained on the datasets derived from the GM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

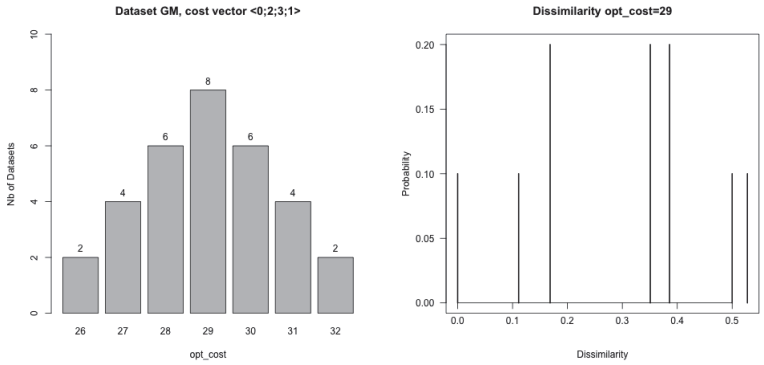


Figure B.42: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 29 (right) obtained on the datasets derived from the GM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

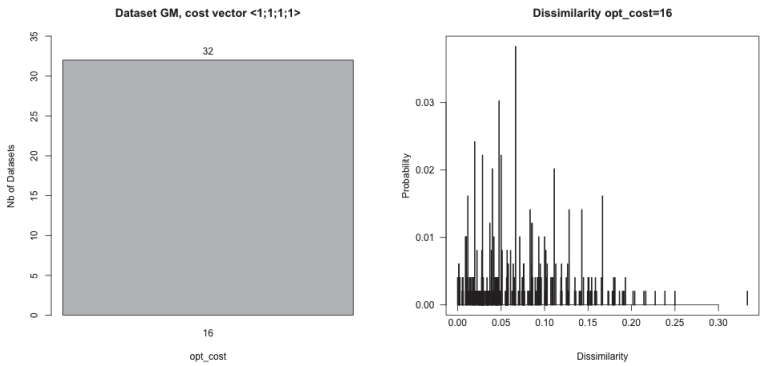


Figure B.43: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 16 (right) obtained on the datasets derived from the GM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

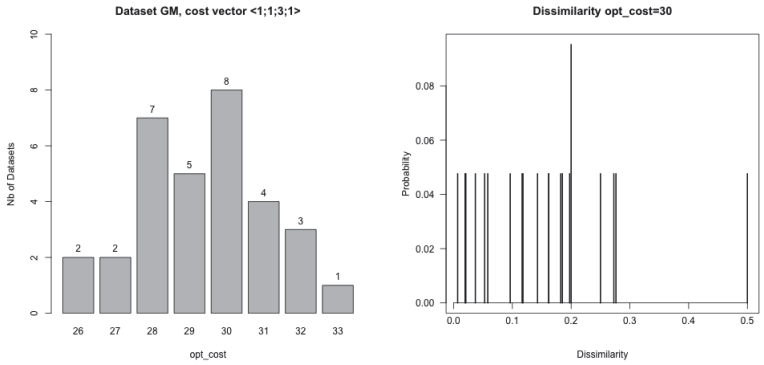


Figure B.44: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 30 (right) obtained on the datasets derived from the GM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

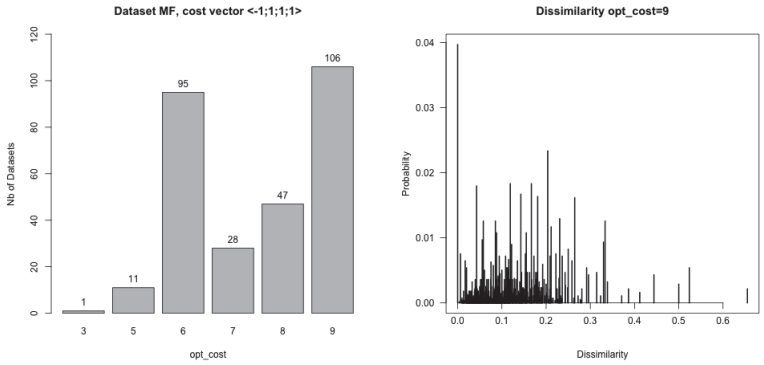


Figure B.45: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the MF dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

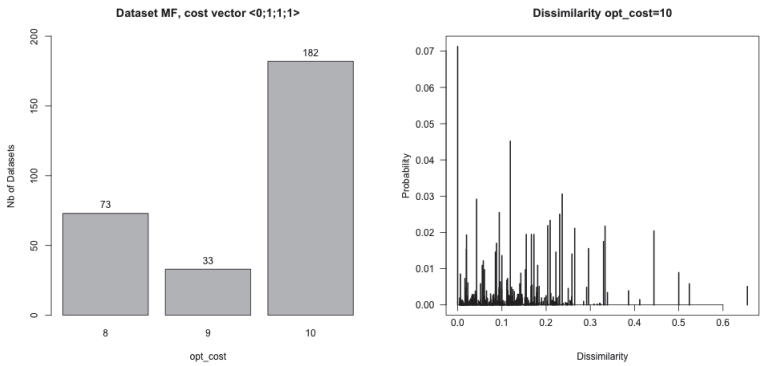


Figure B.46: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 10 (right) obtained on the datasets derived from the MF dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

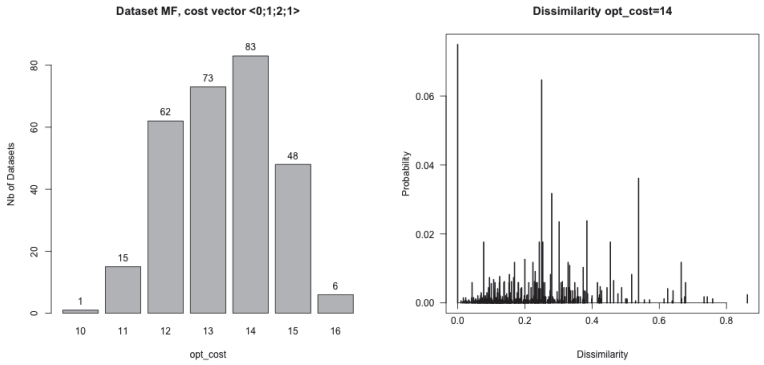


Figure B.47: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 14 (right) obtained on the datasets derived from the MF dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

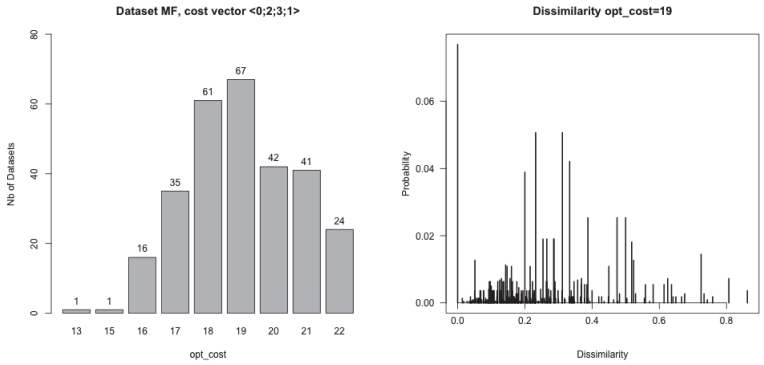


Figure B.48: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 19 (right) obtained on the datasets derived from the MF dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

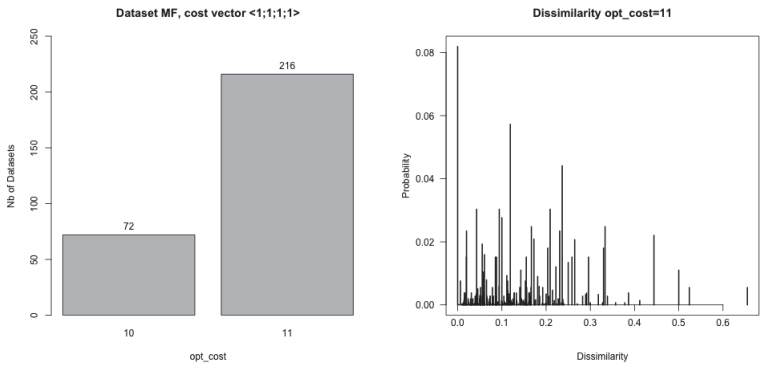


Figure B.49: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 11 (right) obtained on the datasets derived from the MF dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

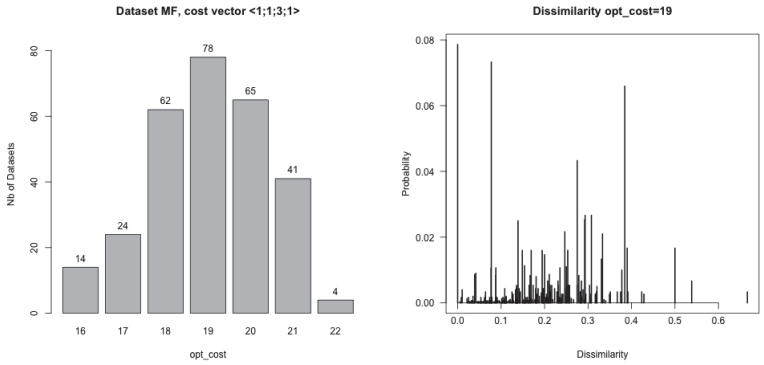


Figure B.50: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 19(right) obtained on the datasets derived from the MF dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

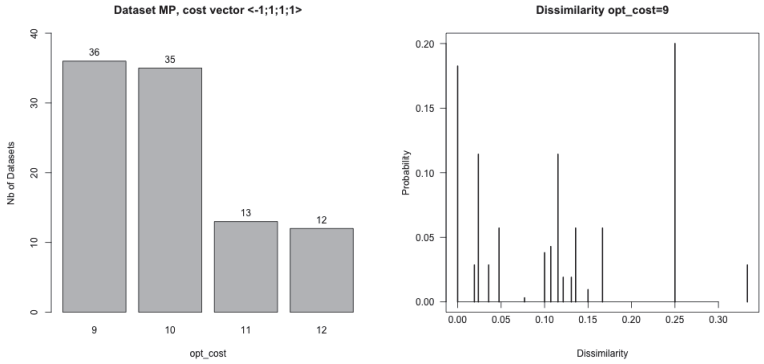


Figure B.51: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the MP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1\rangle$.

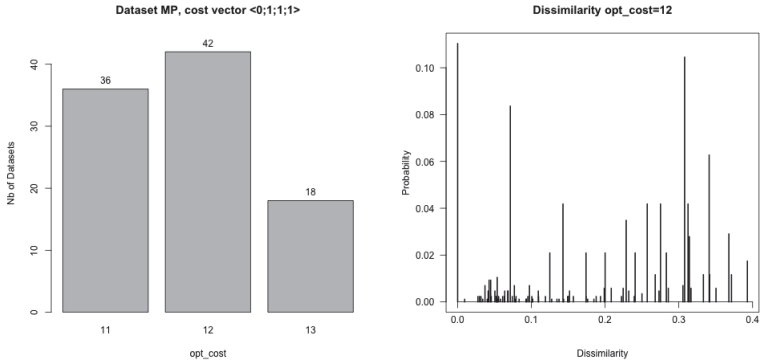


Figure B.52: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 12 (right) obtained on the datasets derived from the MP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$.

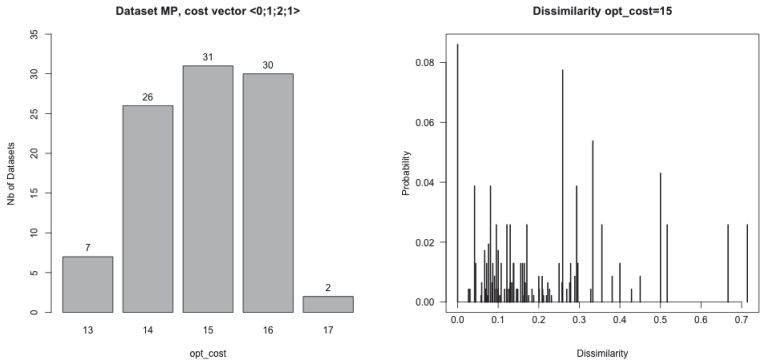


Figure B.53: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 15 (right) obtained on the datasets derived from the MP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$.

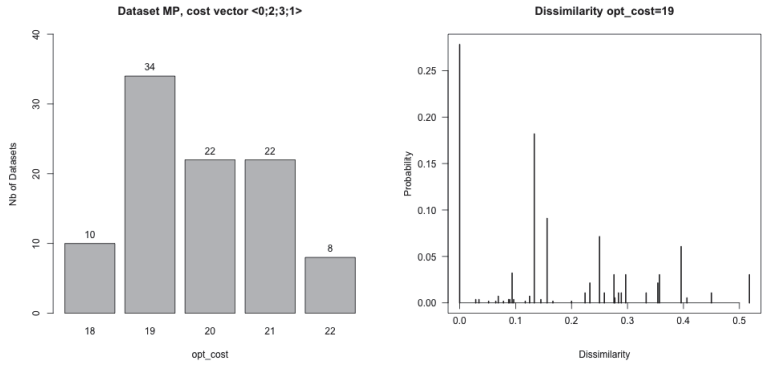


Figure B.54: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 19 (right) obtained on the datasets derived from the MP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

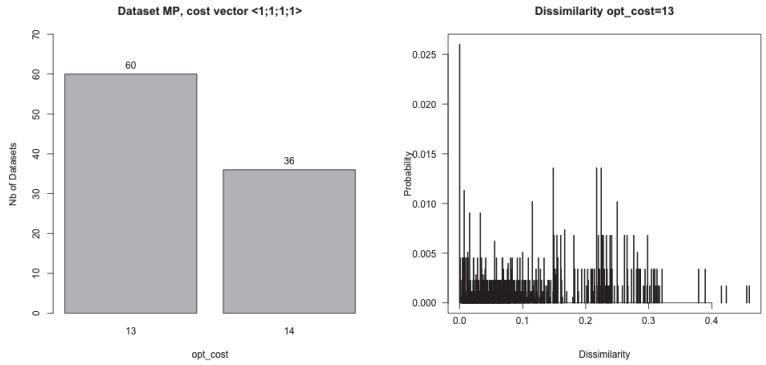


Figure B.55: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 13 (right) obtained on the datasets derived from the MP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

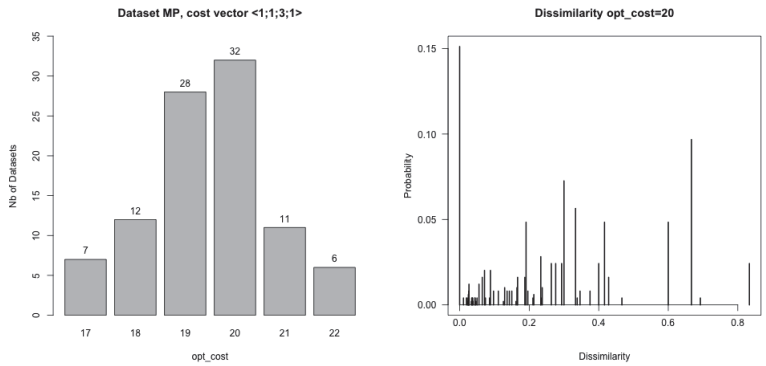


Figure B.56: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 20 (right) obtained on the datasets derived from the MP dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

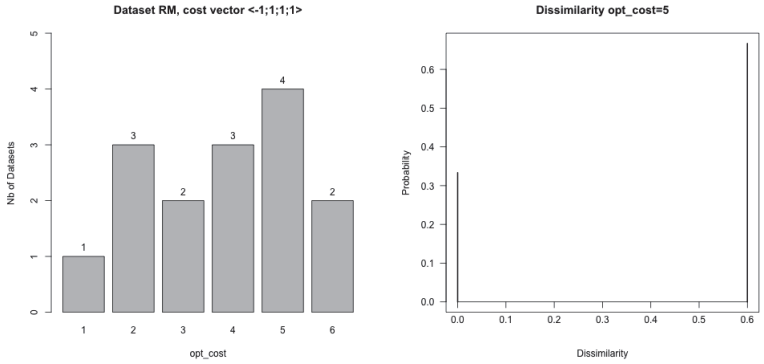


Figure B.57: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 5 (right) obtained on the datasets derived from the RM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

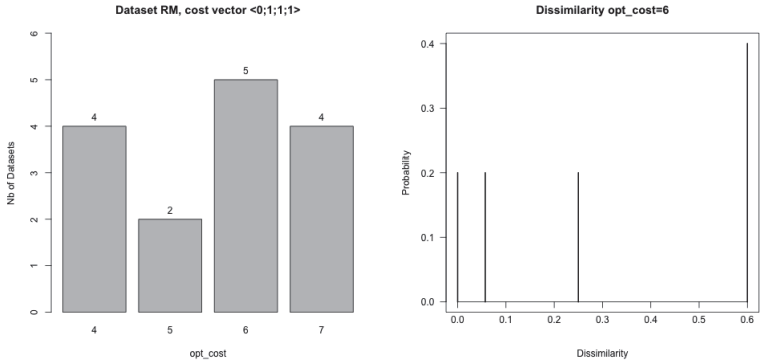


Figure B.58: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 6 (right) obtained on the datasets derived from the RM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

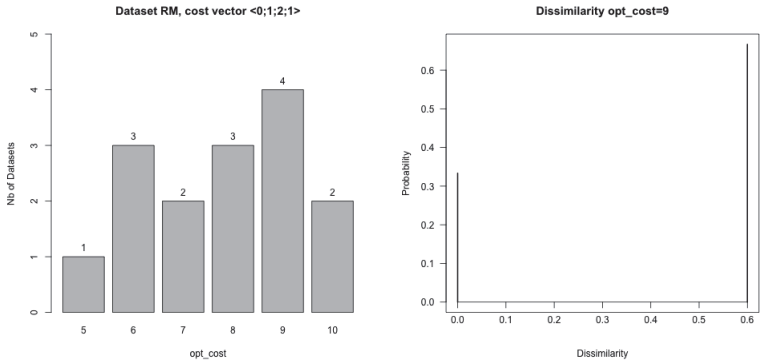


Figure B.59: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the RM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

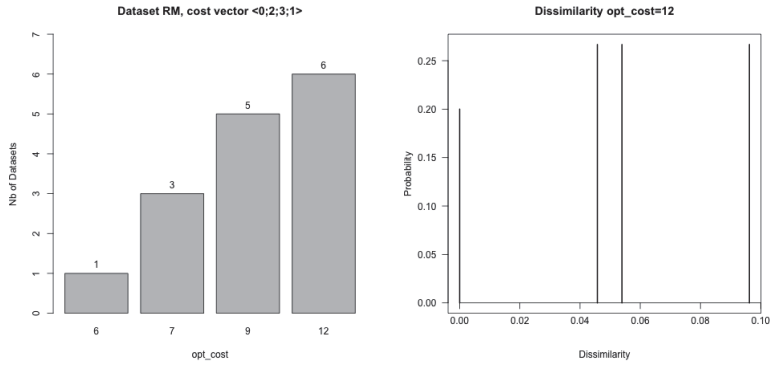


Figure B.60: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 12 (right) obtained on the datasets derived from the RM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

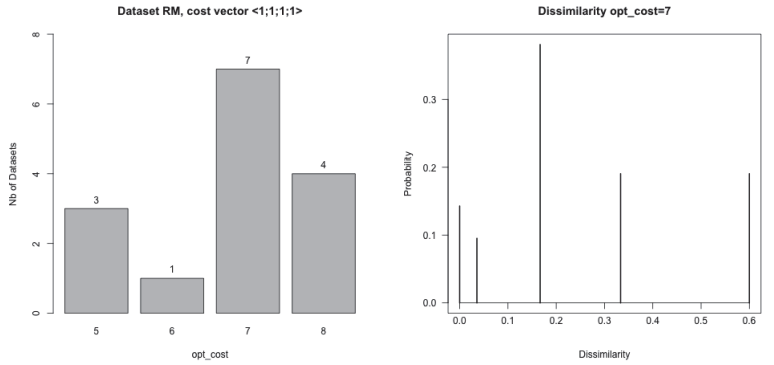


Figure B.61: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 7 (right) obtained on the datasets derived from the RM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

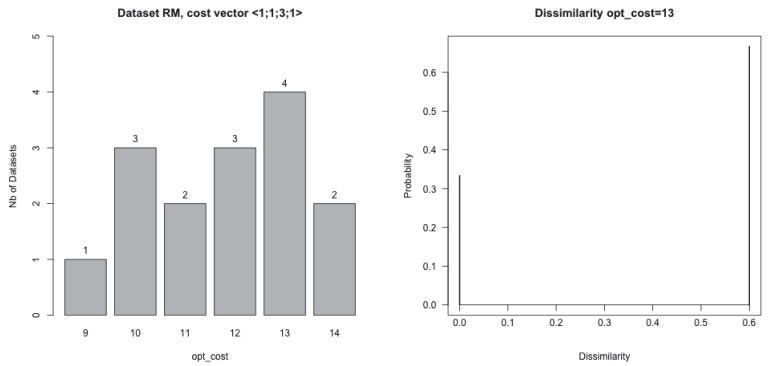


Figure B.62: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 13 (right) obtained on the datasets derived from the RM dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

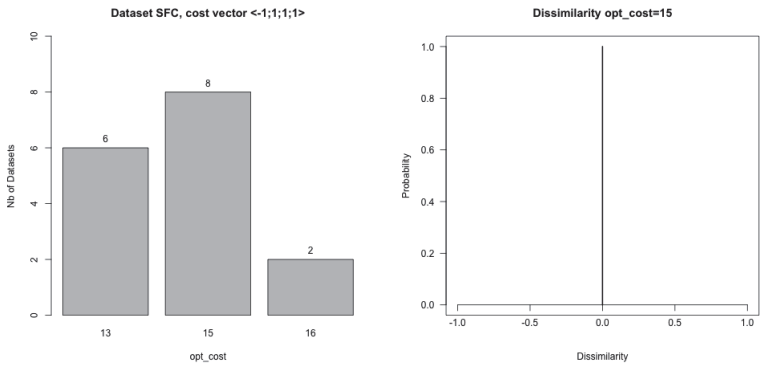


Figure B.63: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 15 (right) obtained on the datasets derived from the SFC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1\rangle$.

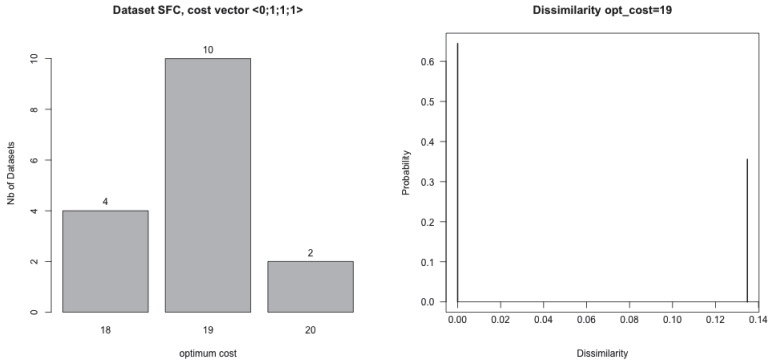


Figure B.64: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 19 (right) obtained on the datasets derived from the SFC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$.

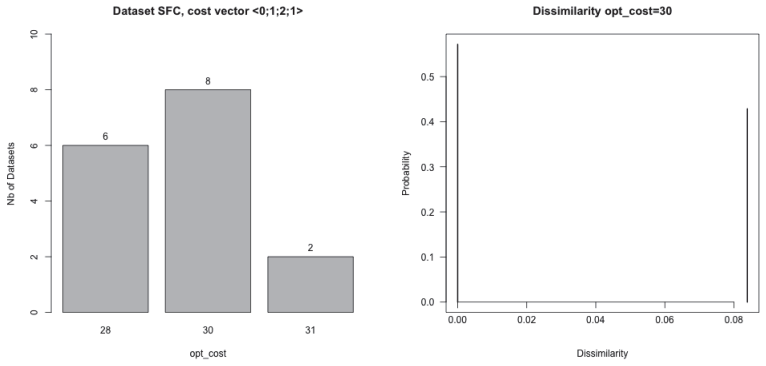


Figure B.65: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 30 (right) obtained on the datasets derived from the SFC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$.

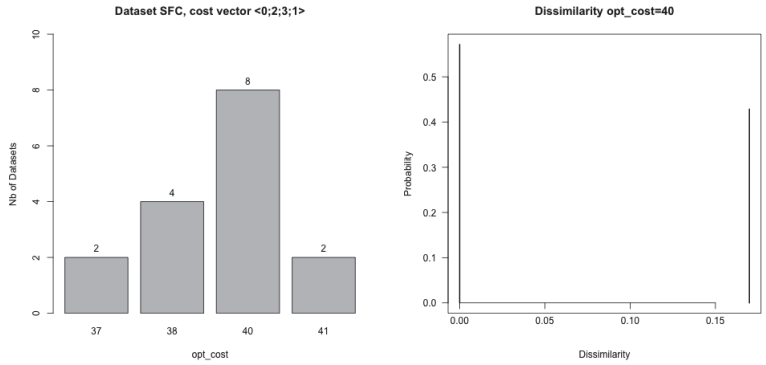


Figure B.66: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 40 (right) obtained on the datasets derived from the SFC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

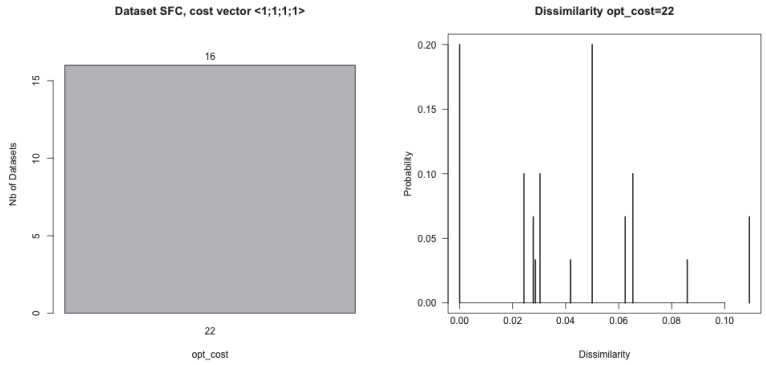


Figure B.67: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 22 (right) obtained on the datasets derived from the SFC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

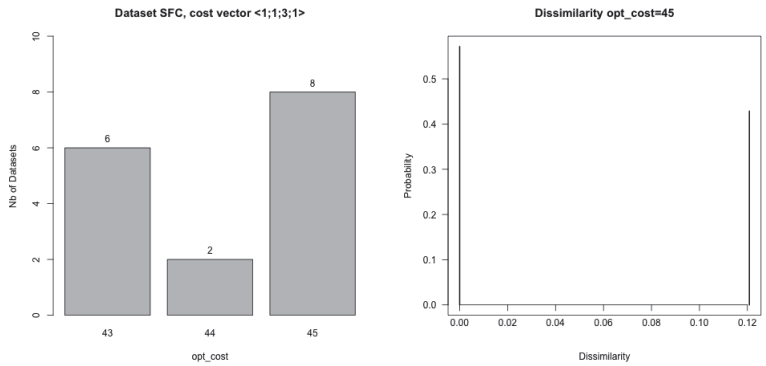


Figure B.68: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 45 (right) obtained on the datasets derived from the SFC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

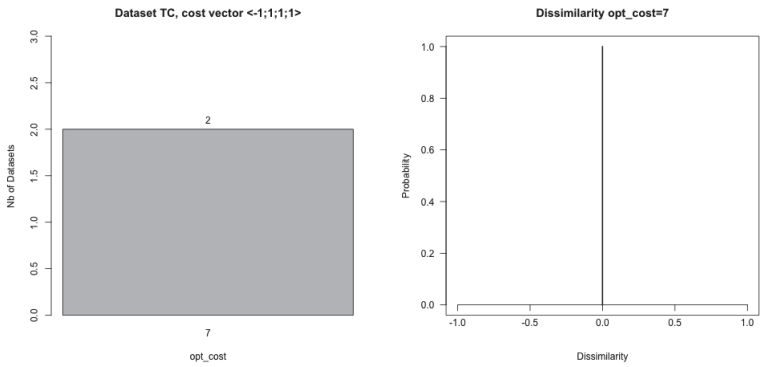


Figure B.69: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 7 (right) obtained on the datasets derived from the TC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

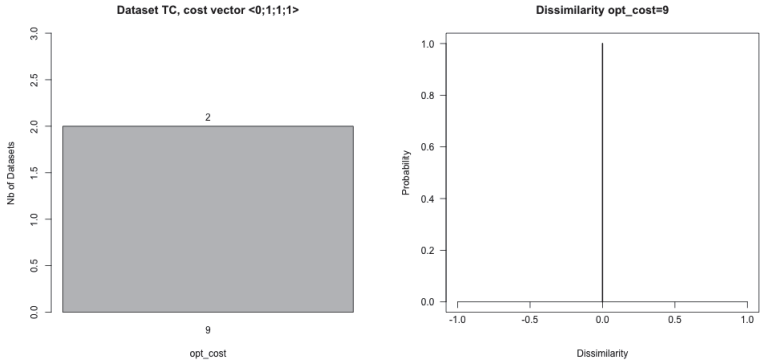


Figure B.70: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 9 (right) obtained on the datasets derived from the TC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

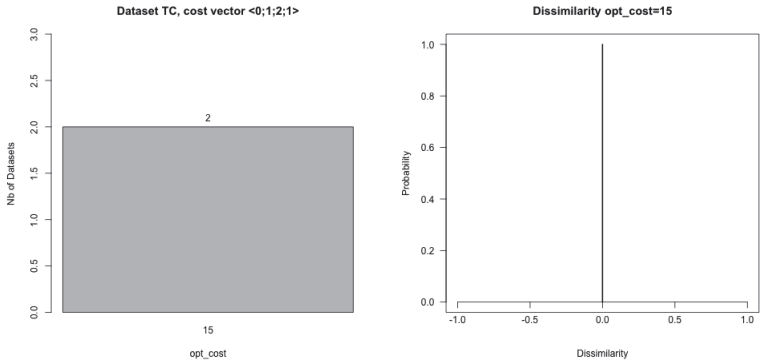


Figure B.71: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 15 (right) obtained on the datasets derived from the TC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

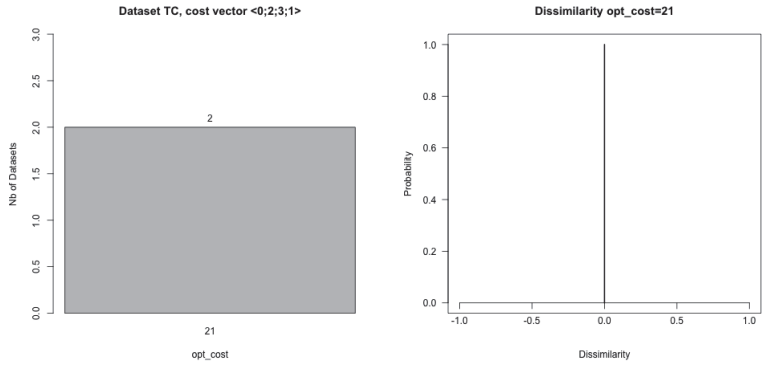


Figure B.72: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 21 (right) obtained on the datasets derived from the TC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

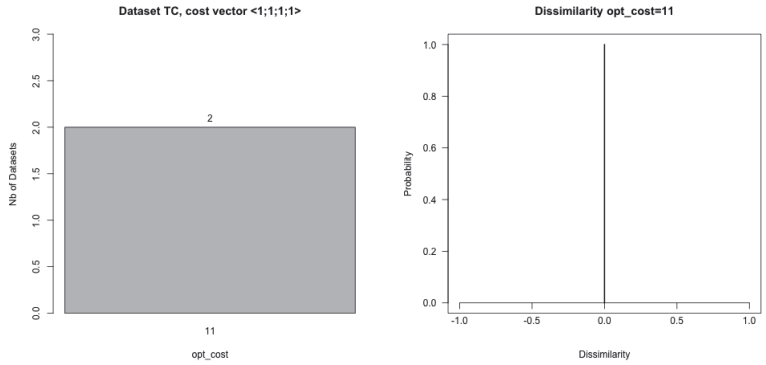


Figure B.73: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 11 (right) obtained on the datasets derived from the TC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$.

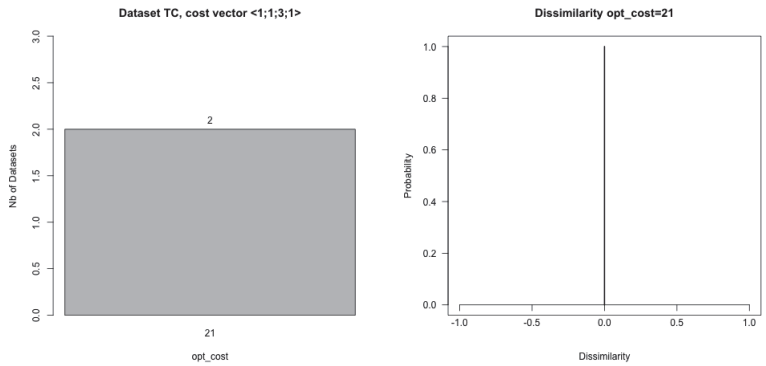


Figure B.74: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 21 (right) obtained on the datasets derived from the TC dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

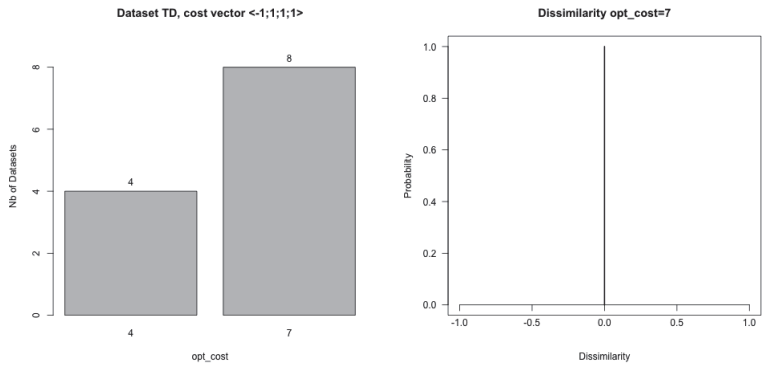


Figure B.75: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 7 (right) obtained on the datasets derived from the TD dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$.

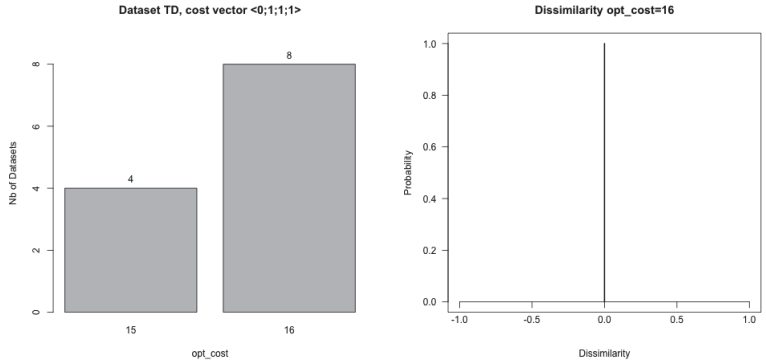


Figure B.76: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 16 (right) obtained on the datasets derived from the TD dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1 \rangle$.

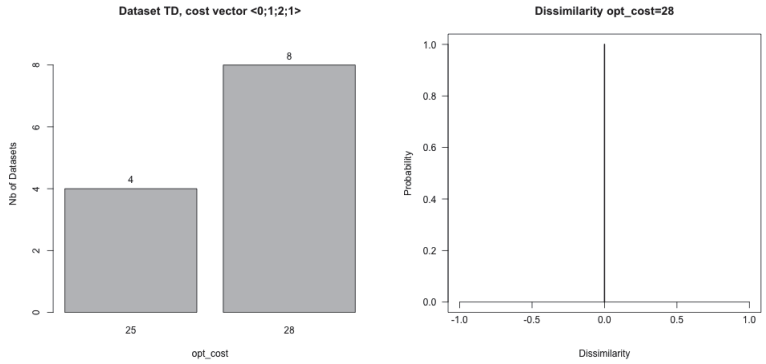


Figure B.77: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 28 (right) obtained on the datasets derived from the TD dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$.

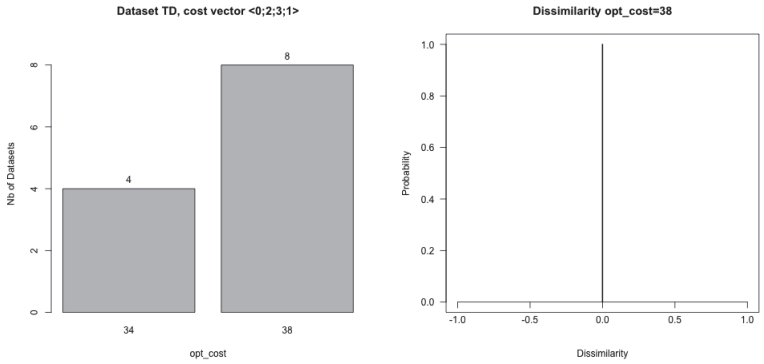


Figure B.78: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 38 (right) obtained on the datasets derived from the TD dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$.

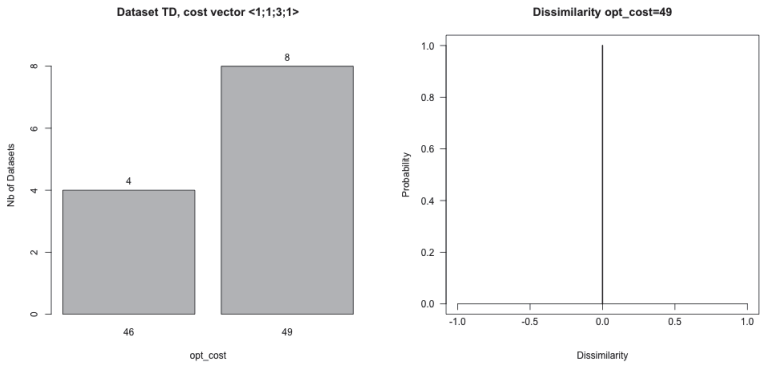


Figure B.79: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with optimal cost 49 (right) obtained on the datasets derived from the TD dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$.

B.2.2 Empirical distribution of the dissimilarity for real datasets

This section presents the results for the study of the empirical distribution of the dissimilarity for real datasets. Figures B.80 to B.90 show the results for 15 datasets that present multiple associations with the 6 cost vectors in $\mathcal{C}$. Let us recall that for each biological dataset, we generated 1000 datasets by permuting the labels of host and symbiont trees and keeping the associations between them fixed.

There are some cases where the number of resolutions into simple associations is too large (namely there are too many ways of resolving the multiple associations into simple ones). For this reason, the results for the datasets AP, CA and RM could not be computed. In the case of SHA, for all cost vectors c, there is only one pattern for the most frequent optimal cost and thus there are no results to show.

Observe that for the dataset CP and for all cost vectors different from $\langle 1,1,1,1\rangle$ and $\langle 1,1,3,1\rangle$, there are only cyclic reconciliations. As only time-feasible reconciliations are retained, there are no results to show in this case (Figure B.81).

Figure B.90 shows the histograms for the dissimilarity derived from TD. In this case we did not obtain any results for the cost vector $c = \langle 1,1,1,1\rangle$, because the number of reconciliations was too large and our program could not handle all these solutions (due to memory capacities)

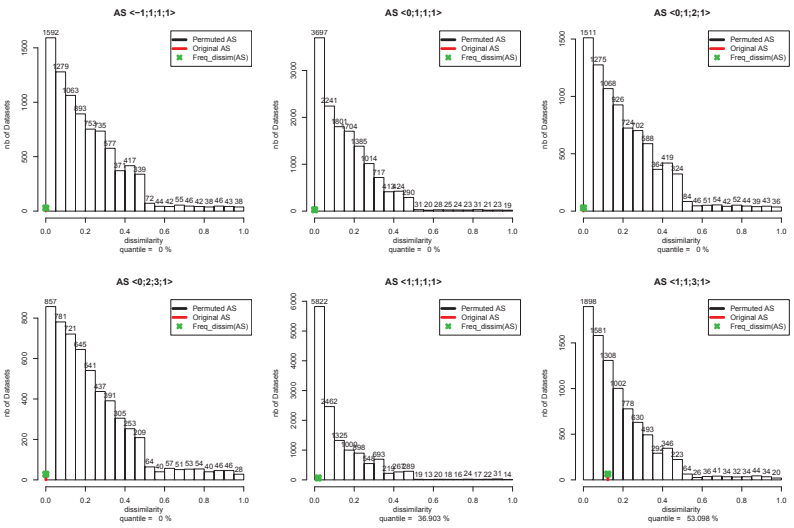


Figure B.80: Histograms of the dissimilarity derived from the AS dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset AS. The green crosses are the $freq_{dissim}(AS)$.

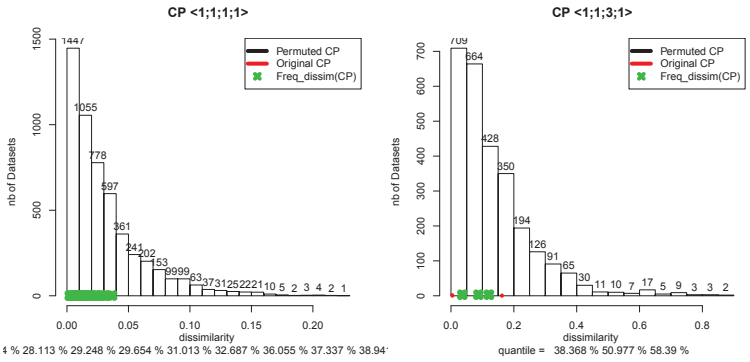


Figure B.81: Histograms of the dissimilarity derived from the CP dataset with the cost vectors $\langle 1,1,1,1\rangle$ and $\langle 1,1,3,1\rangle$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset CP. The green crosses are the $freq_{dissim}(CP)$.

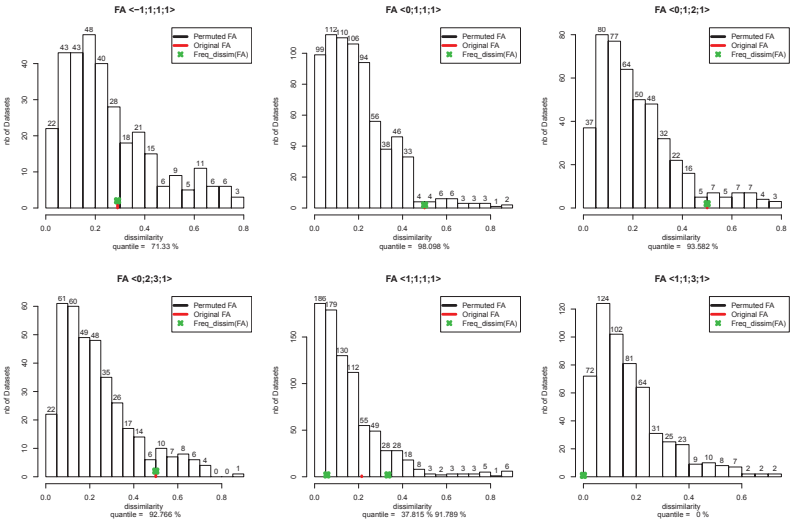


Figure B.82: Histograms of the dissimilarity derived from the FA dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset FA. The green crosses are the $freq_{dissim}(FA)$.

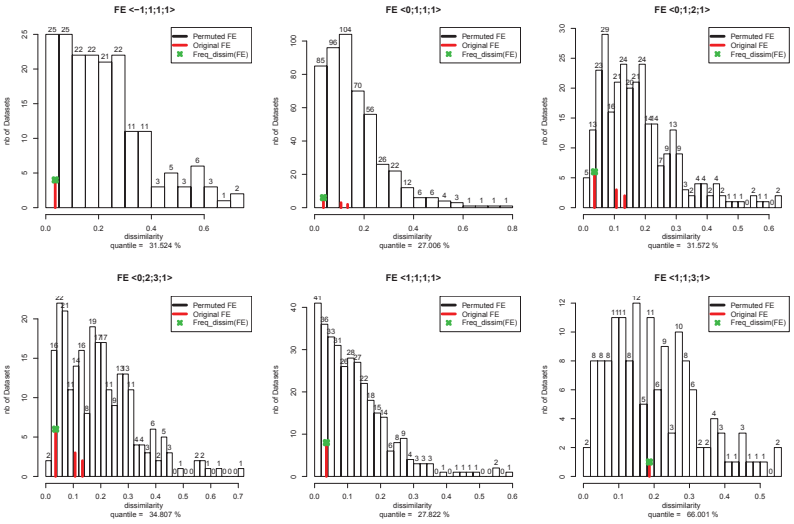


Figure B.83: Histograms of the dissimilarity derived from the FE dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset FE. The green crosses are the $freq_{dissim}(FE)$.

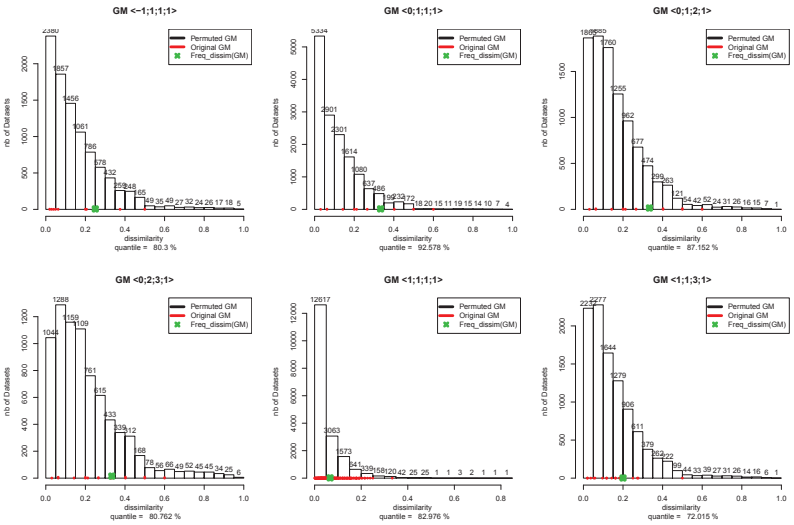


Figure B.84: Histograms of the dissimilarity derived from the GM dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset GM. The green crosses are the $freq_{dissim}(GM)$.

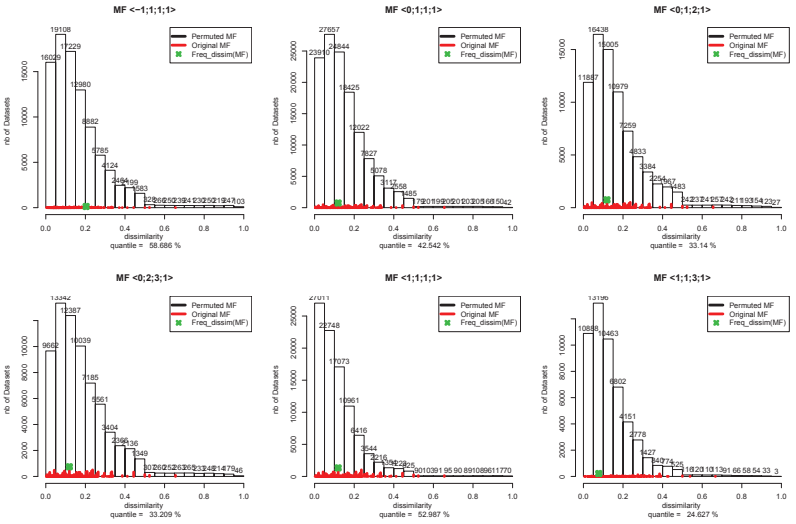


Figure B.85: Histograms of the dissimilarity derived from the MF dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset MF. The green crosses are the $freq_{dissim}(MF)$.

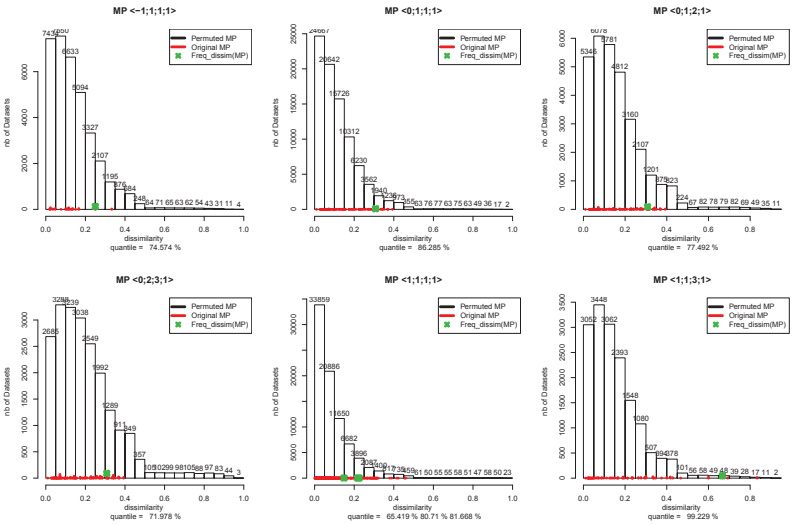


Figure B.86: Histograms of the dissimilarity derived from the MP dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset MP. The green crosses are the $freq_{dissim}(MP)$.

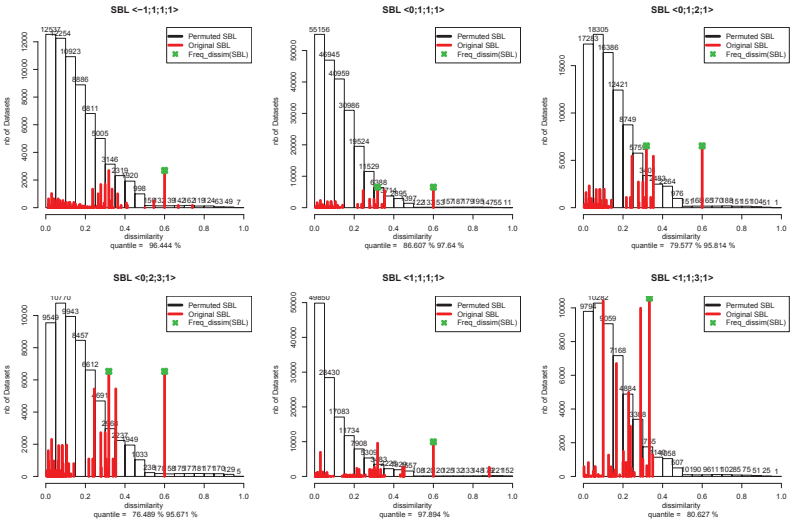


Figure B.87: Histograms of the dissimilarity derived from the SBL dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset SBL. The green crosses are the $freq_{dissim}(SBL)$.

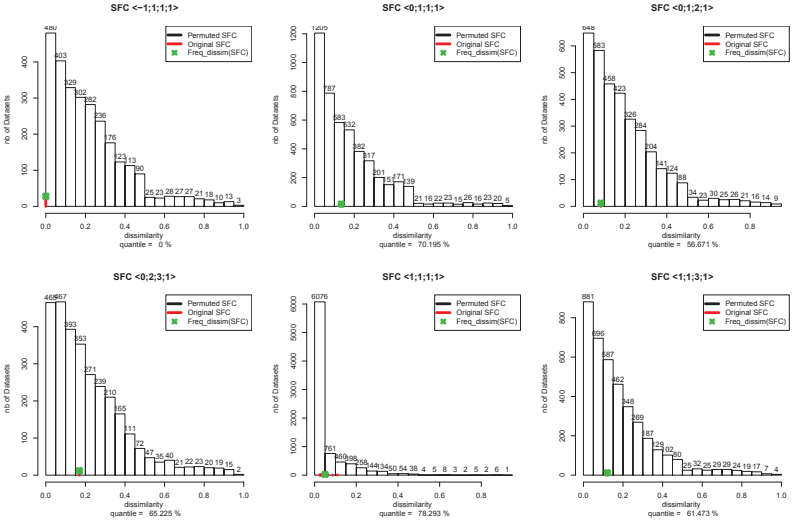


Figure B.88: Histograms of the dissimilarity derived from the SFC dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset SFC. The green crosses are the $freq_{dissim}(SFC)$.

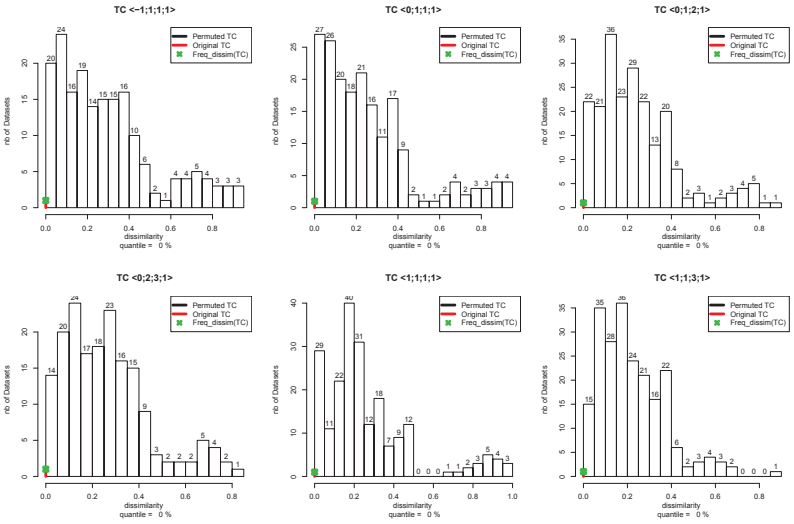


Figure B.89: Histograms of the dissimilarity derived from the TC dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset TC. The green crosses are the $freq_{dissim}(TC)$.

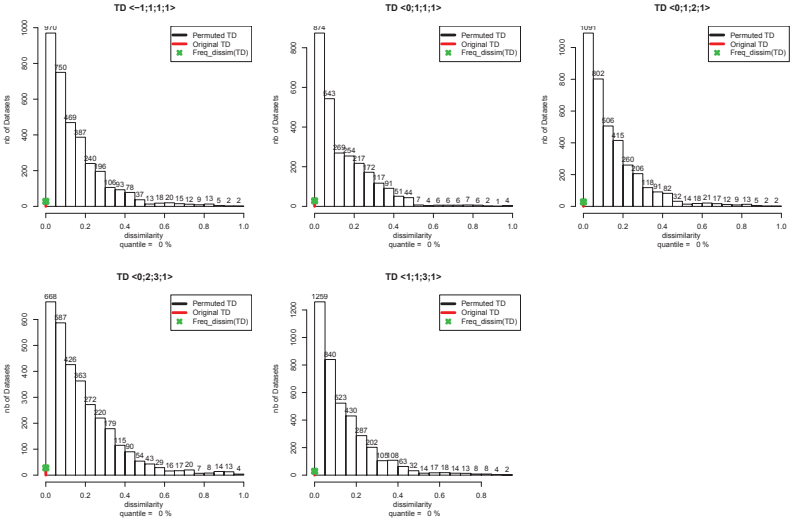


Figure B.90: Histograms of the dissimilarity derived from the TD dataset with the cost vectors in $\mathcal{C}$. For each of them, the black histogram is obtained by resolving the multiple associations in all the possible ways for the permuted datasets. The red lines are obtained by resolving the multiple associations in all possible ways for the original dataset TD. The green crosses are the $freq_{dissim}(TD)$.

B.2.3 Changing associations for simulated datasets

Figures B.91 to B.155 show the results for the simulated datasets $D_{x\%}$ with all cost vectors in $\mathcal{C}$. Only time-feasible reconciliations are retained.

Observe that sometimes some of the results are missing. This happens in the following three cases: (a) when the experiments did not finish due to the large computation time (this happened for $CA_{x\%}$ with x=15, 20, 25, 30, 40, 45, 50 for all the cost vectors in $\mathcal{C}$; $RH_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50 for all the cost vectors in $\mathcal{C}$; $TD_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50 for the cost vector $\langle 1, 1, 1, 1\rangle$), (b) when the reconciliations for the dataset $D_{x\%}$ are all time-unfeasible (this happened for $CP_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50 for the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$ and $\langle 0, 2, 3, 1\rangle$; $AP_{x\%}$ with x=10 for the cost vector $\langle -1, 1, 1, 1\rangle$), (c) when there is only one multiset of patterns, either because it is the only one generated, or because the other reconciliations are time-unfeasible.

In general, the number of optimal reconciliations and the dissimilarity increase with the value of x. Observe that in $MP_{x\%}$, when we consider a low cost for the host switch (Figures B.127, B.128 and B.131), namely for the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$ and $\langle 1, 1, 1, 1\rangle$, the number of optimal reconciliations is less than what is obtained for the other cost vectors (Figures B.129, B.130 and B.132).

Observe also that for these simulated datasets, we did not explore the significance of the dissimilarity observed.

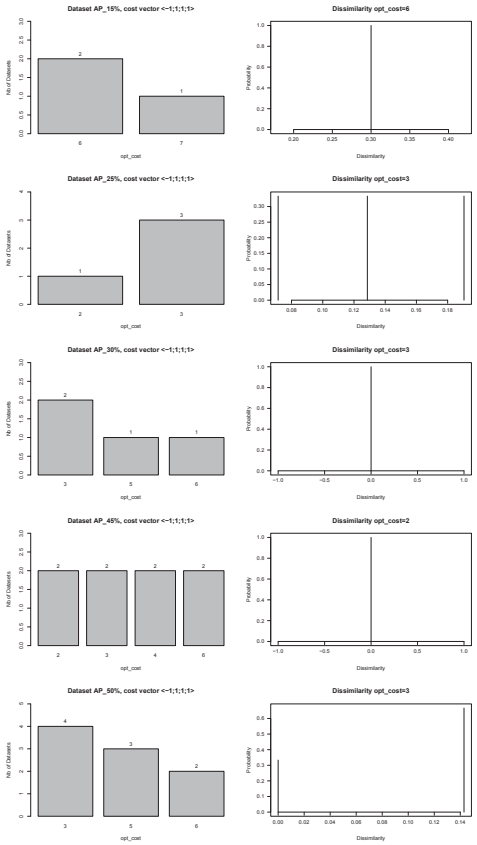


Figure B.91: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $AP_{x\%}$ with x=15, 25, 30, 45, 50.

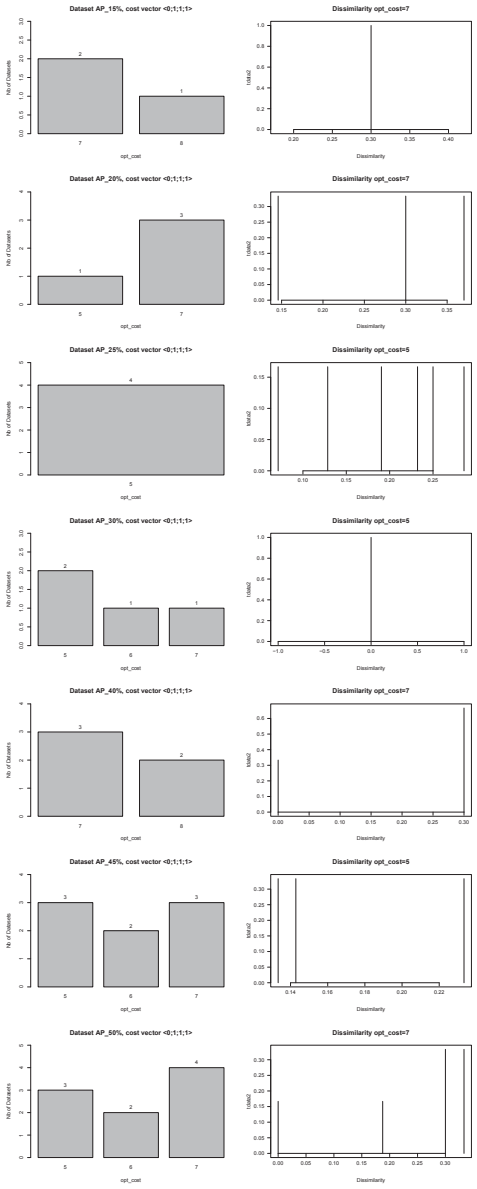


Figure B.92: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $AP_{x\%}$ with x=15, 20, 25, 30, 40, 45, 50.

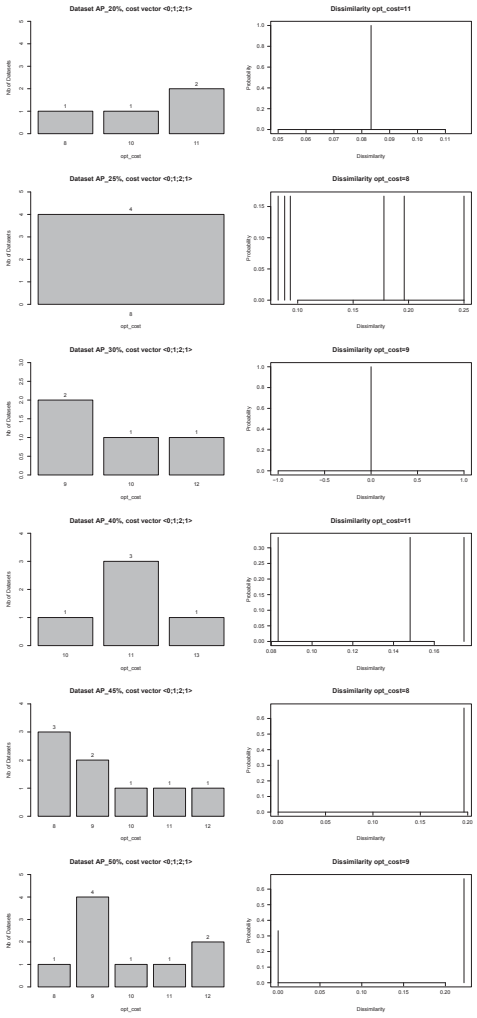


Figure B.93: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $AP_{x\%}$ with x=20, 25, 30, 40, 45, 50.

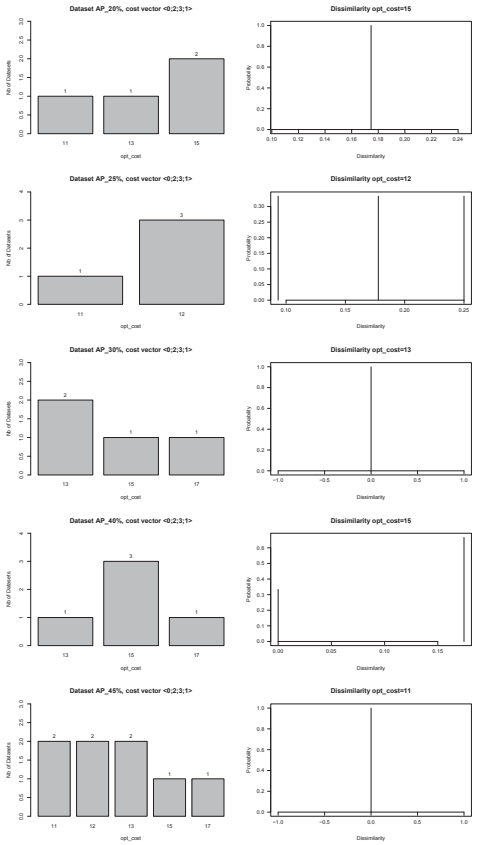


Figure B.94: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $AP_{x\%}$ with x=20, 25, 30, 40, 45.

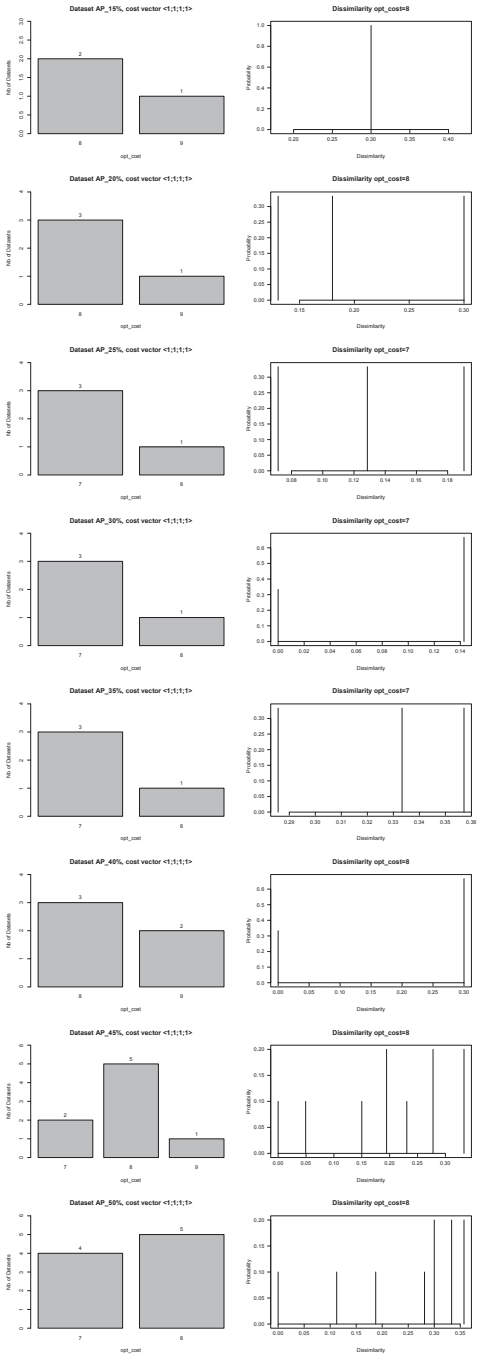


Figure B.95: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $AP_{x\%}$ with x=15, 20, 25, 30, 35, 40, 45, 50.

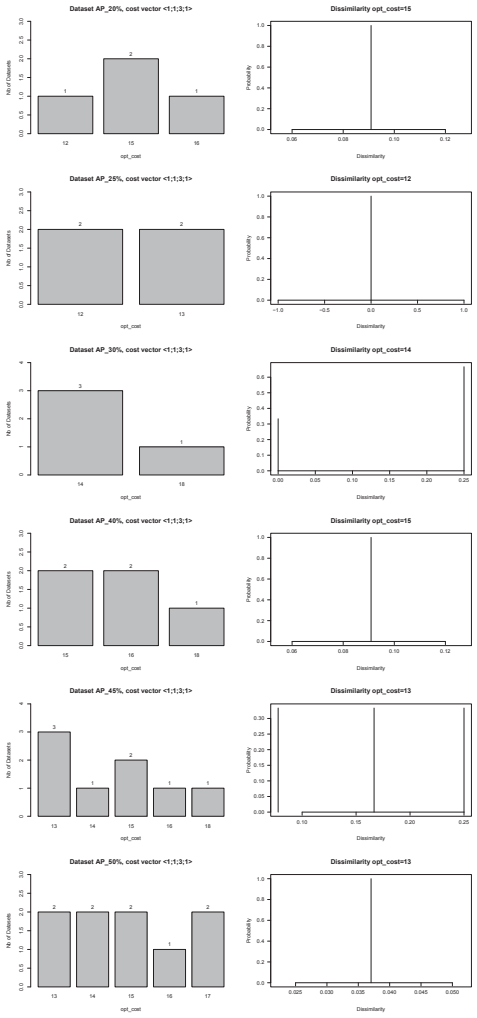


Figure B.96: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $AP_{x\%}$ with x=20, 25, 30, 40, 45, 50.

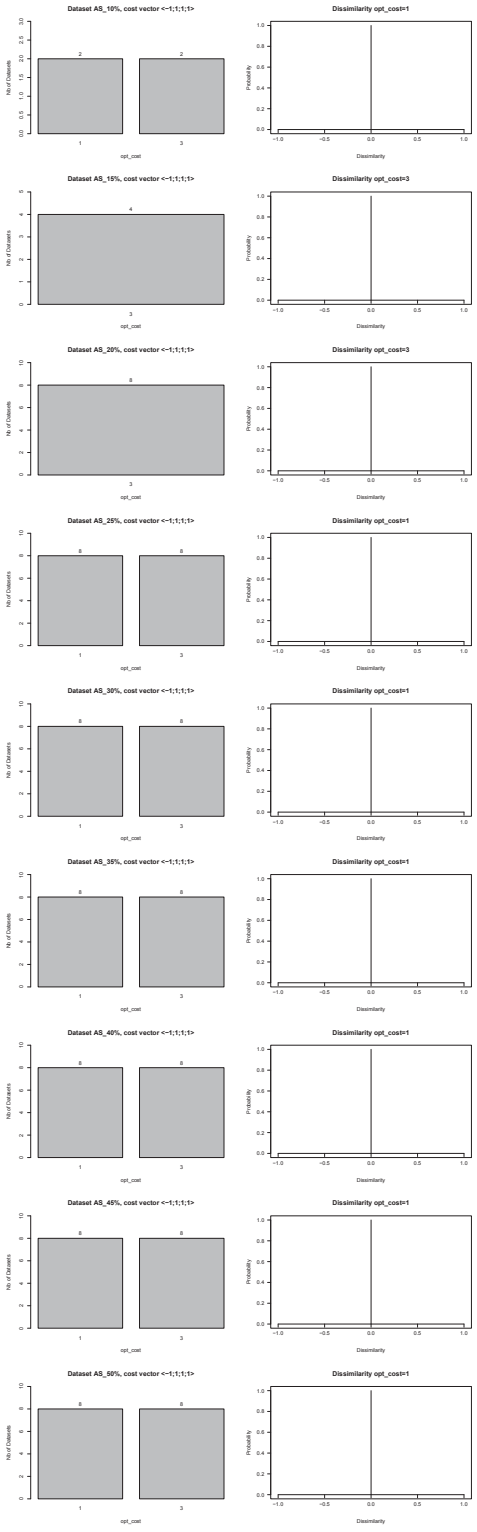


Figure B.97: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AS_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $AS_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

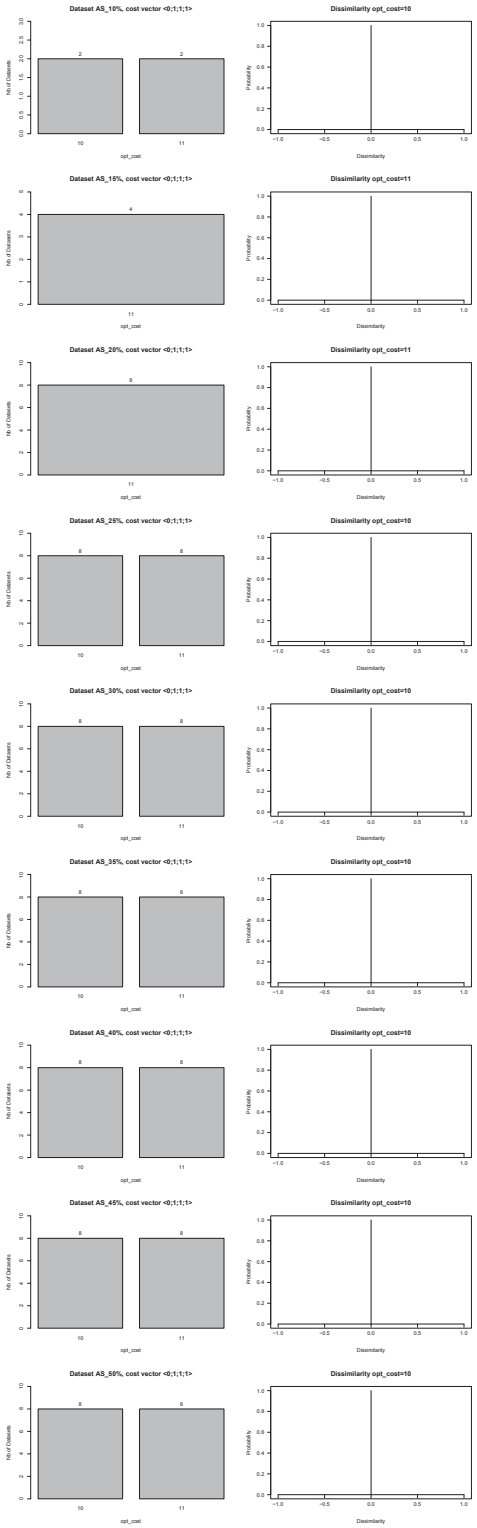


Figure B.98: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AS_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $AS_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

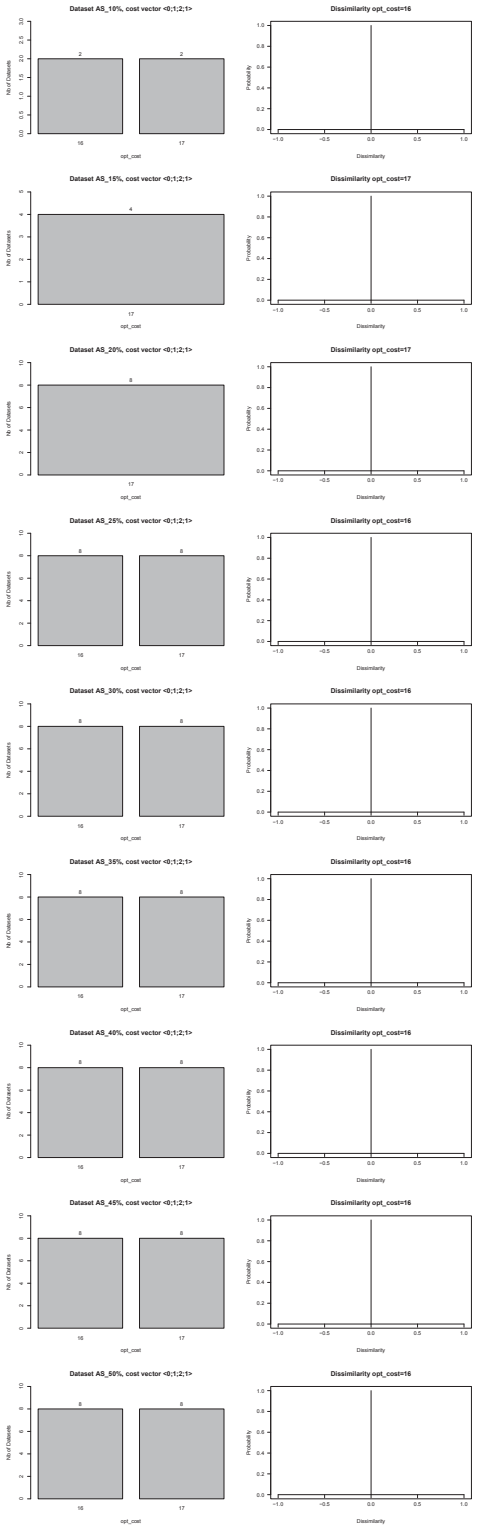


Figure B.99: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AS_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$. Each line is a different $AS_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

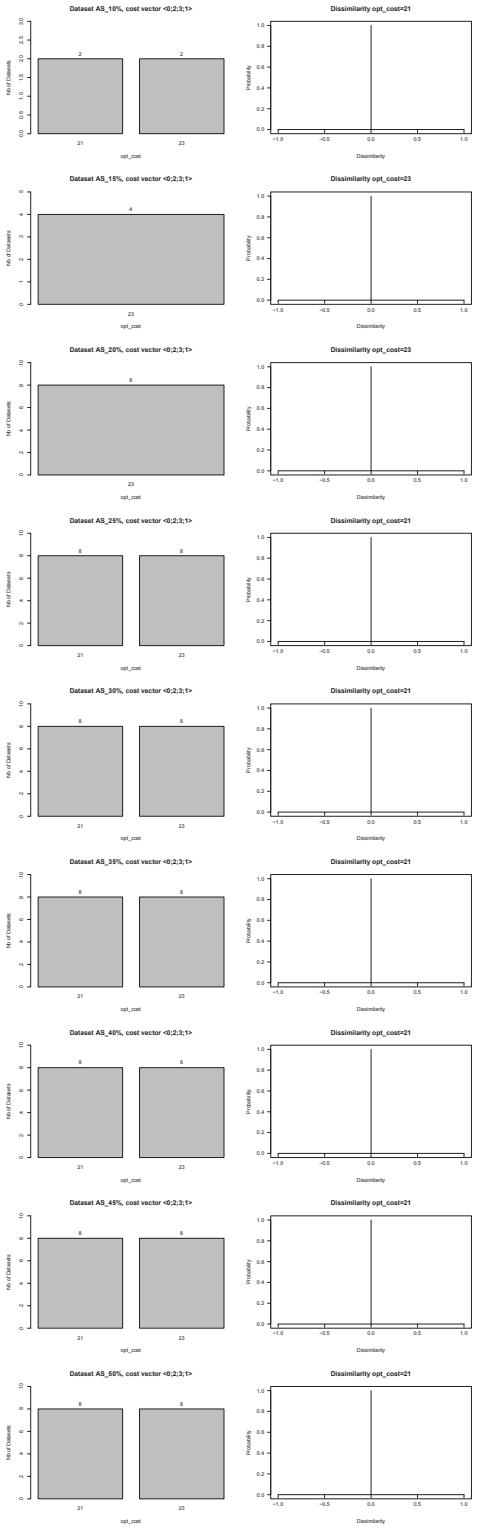


Figure B.100: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AS_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 3, 1 \rangle$. Each line is a different $AS_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

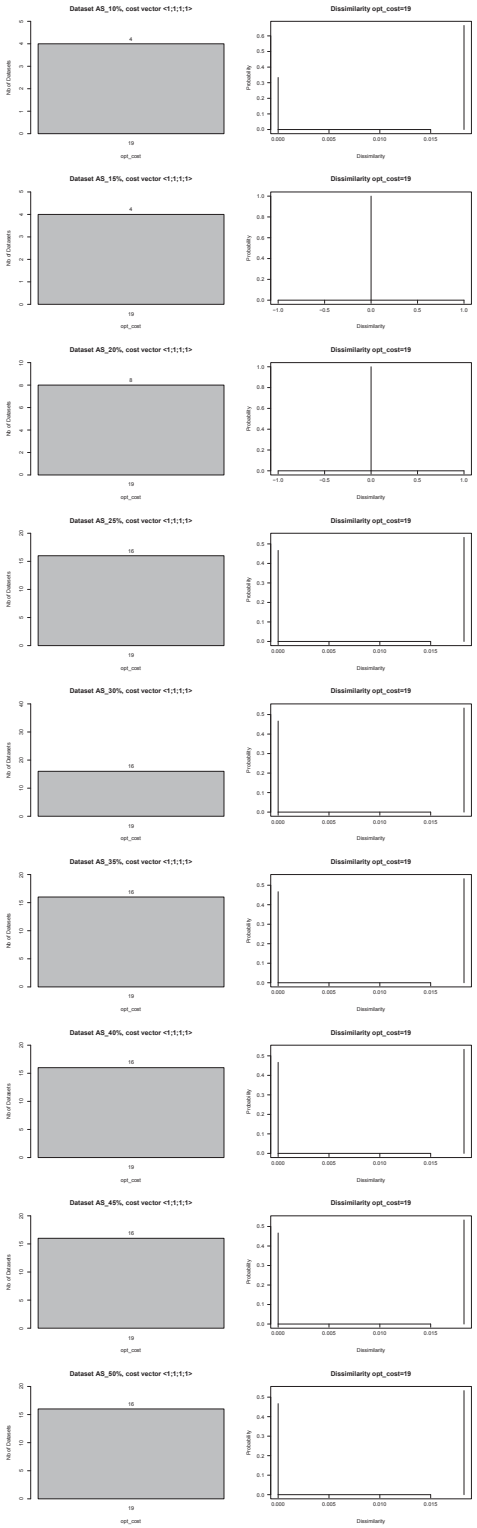


Figure B.101: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AS_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $AS_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

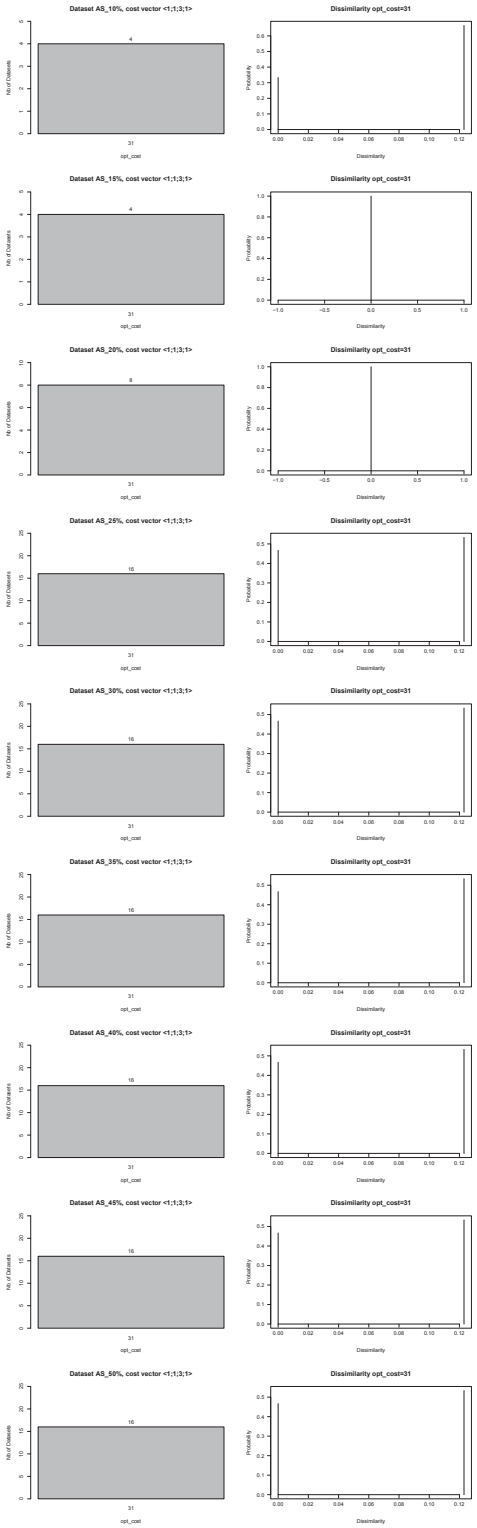


Figure B.102: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $AS_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $AS_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

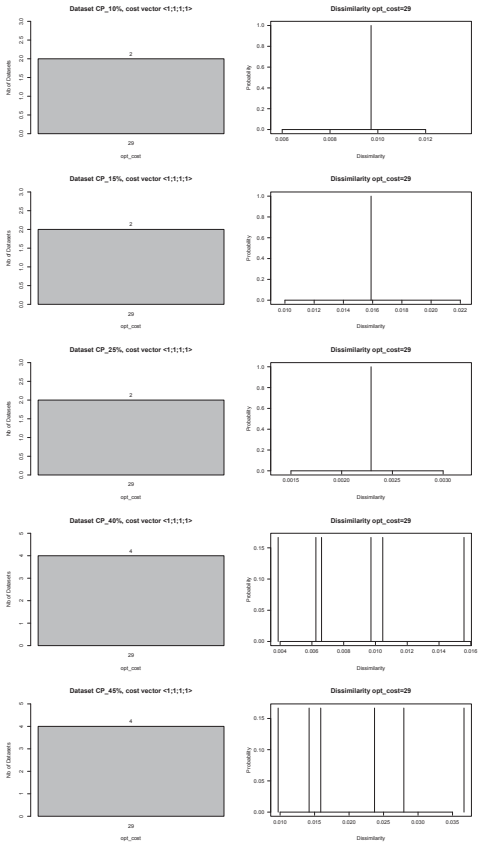


Figure B.103: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $CP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $CP_{x\%}$ with x=10, 15, 25, 40, 45.

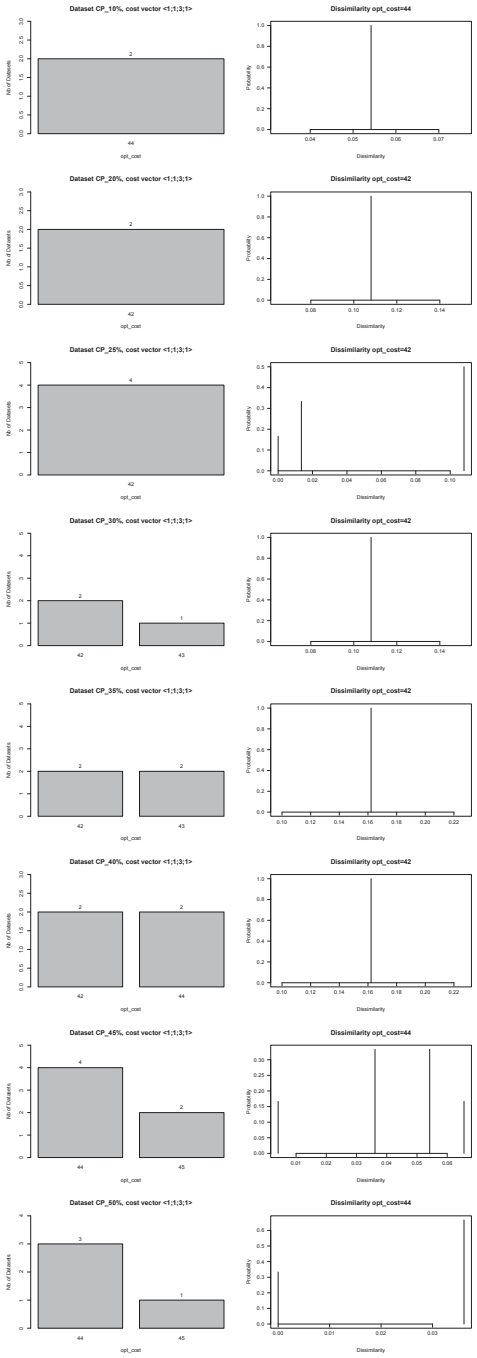


Figure B.104: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $CP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $CP_{x\%}$ with x=10, 20, 25, 30, 35, 40, 45, 50.

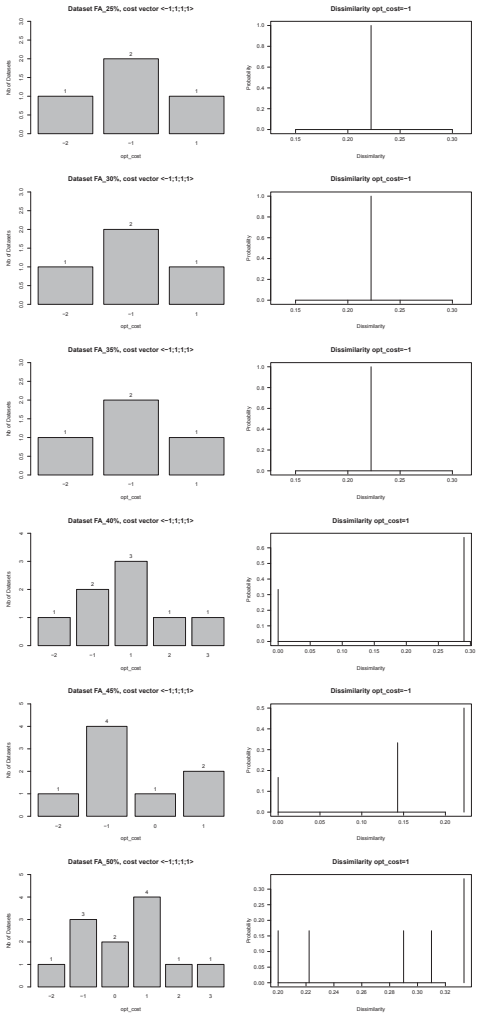


Figure B.105: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $FA_{x\%}$ with x= 25, 30, 35, 40, 45, 50.

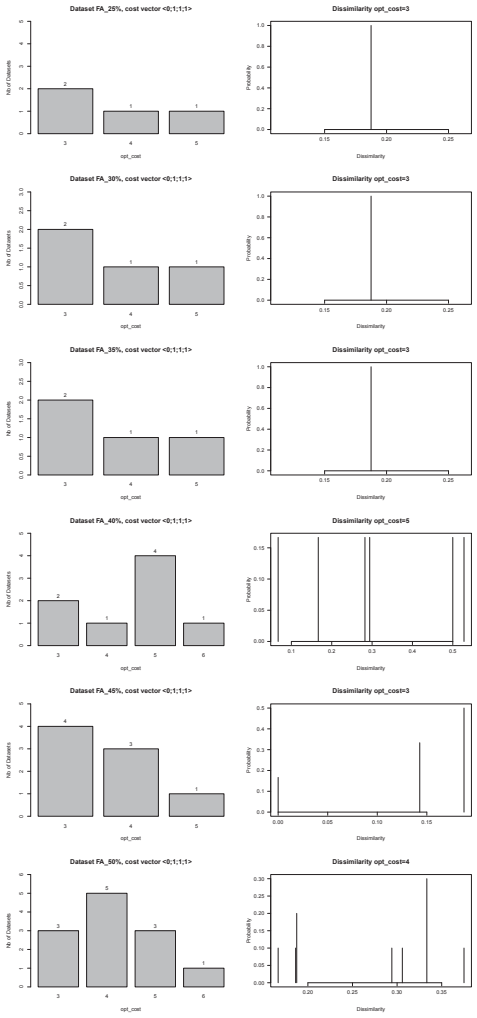


Figure B.106: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $FA_{x\%}$ with x= 25, 30, 35, 40, 45, 50.

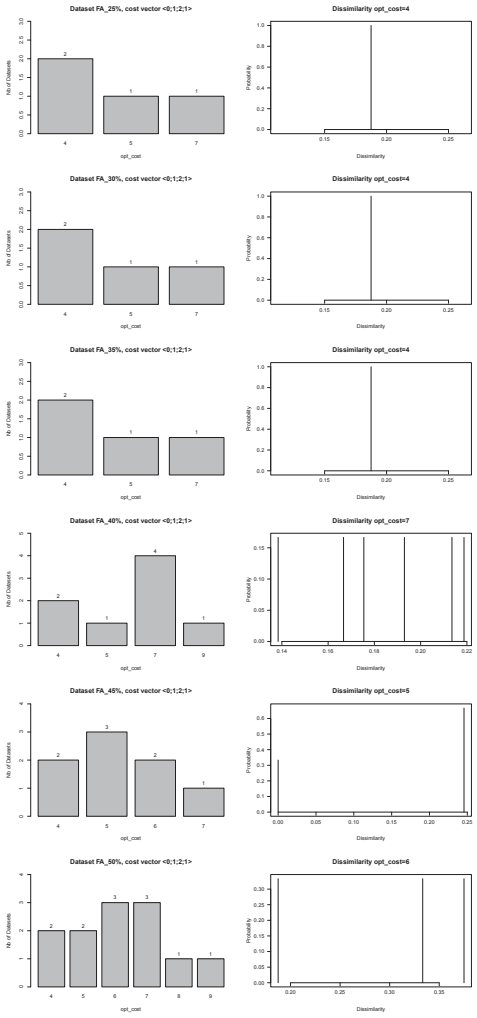


Figure B.107: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $FA_{x\%}$ with x=25, 30, 35, 40, 45, 50.

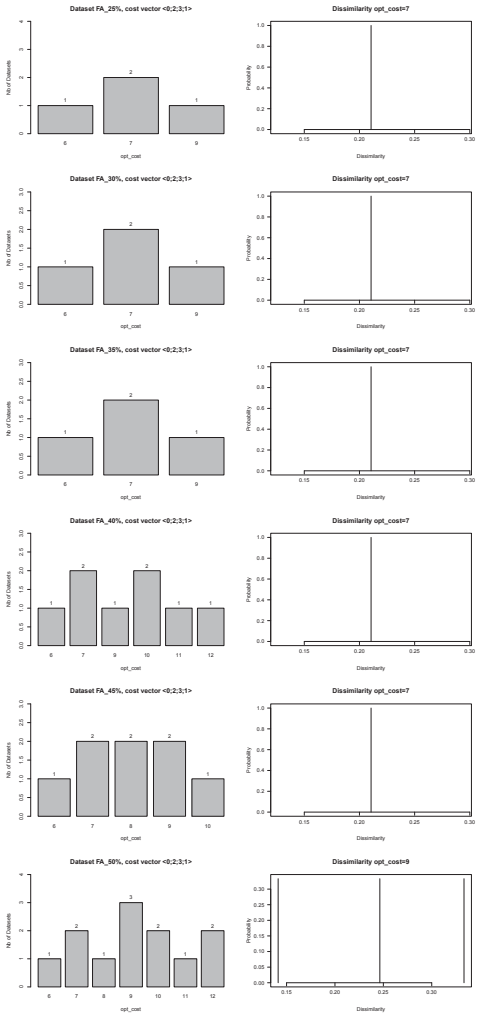


Figure B.108: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $FA_{x\%}$ with x=25, 30, 35, 40, 45, 50.

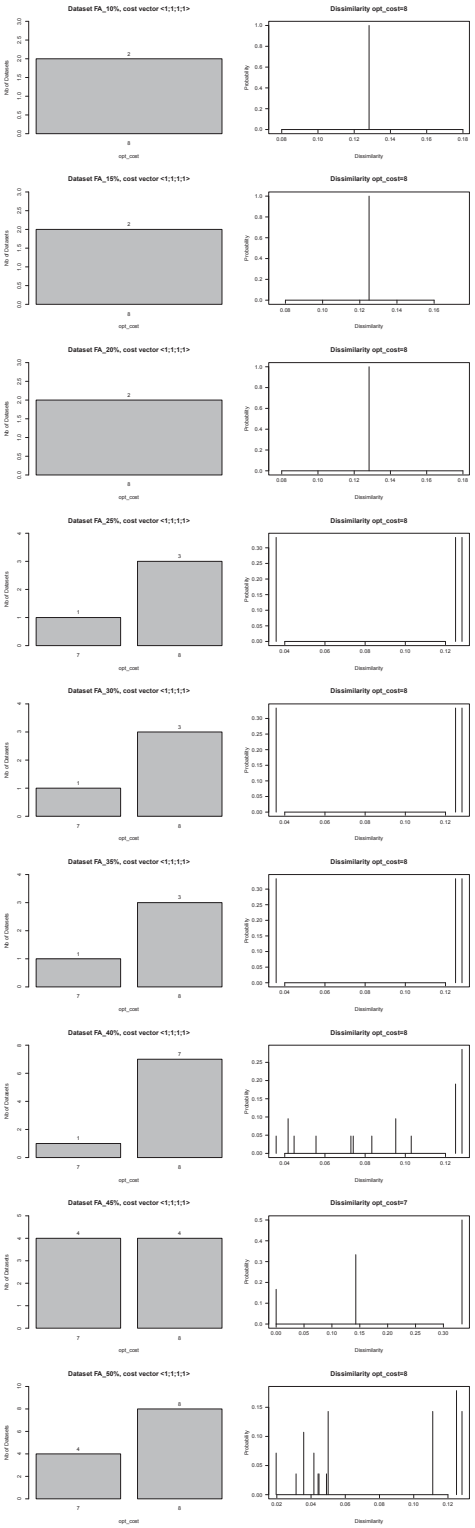


Figure B.109: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $FA_{x\%}$ with x=10, 15, 20, 25, 30, 35, 40, 45, 50.

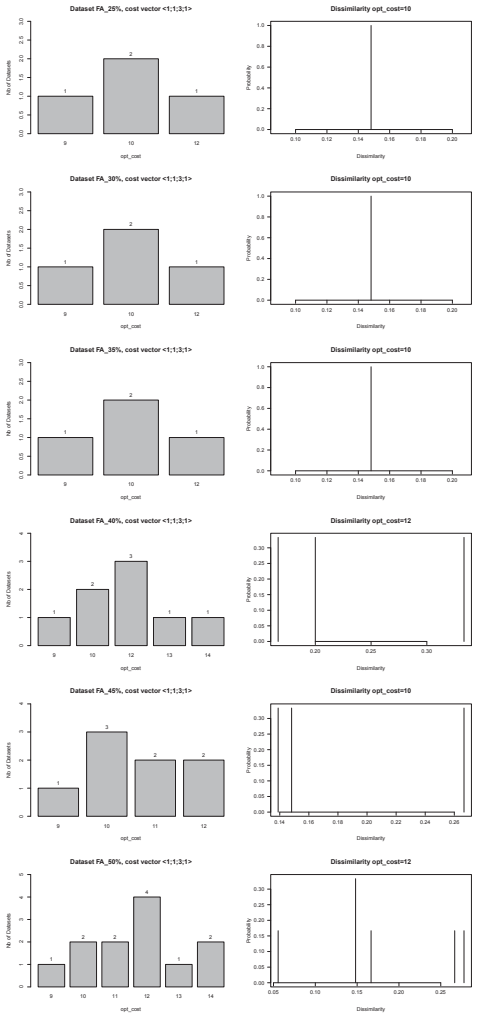


Figure B.110: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $FA_{x\%}$ with x=25, 30, 35, 40, 45, 50.

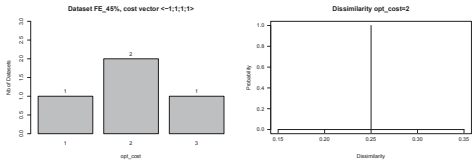


Figure B.111: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FE_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $FE_{x\%}$ with x=45.

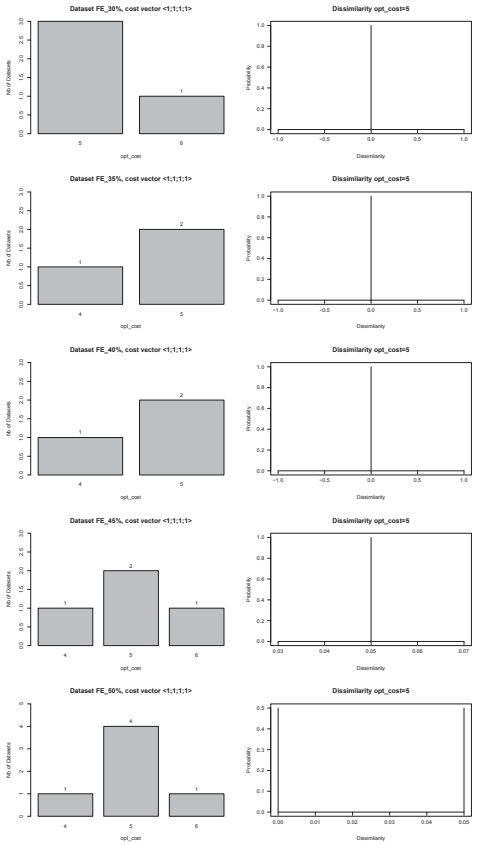


Figure B.112: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FE_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $FE_{x\%}$ with x=30, 35, 40, 45, 50.

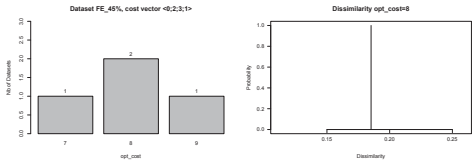


Figure B.113: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FE_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $FE_{x\%}$ with x=45.

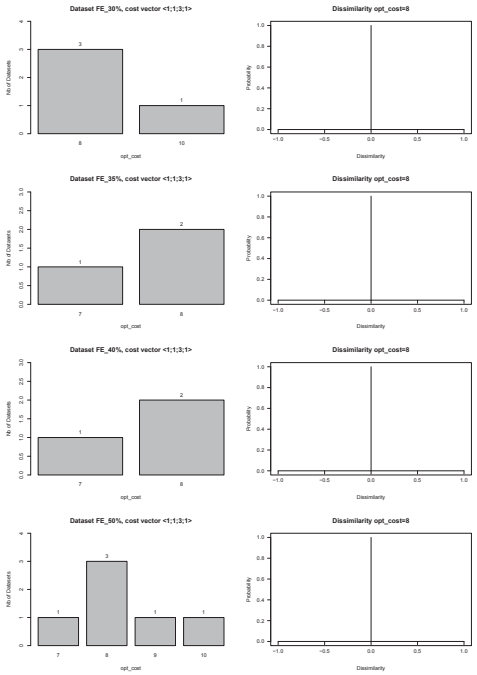


Figure B.114: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $FE_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $FE_{x\%}$ with x=30, 35, 40, 50.

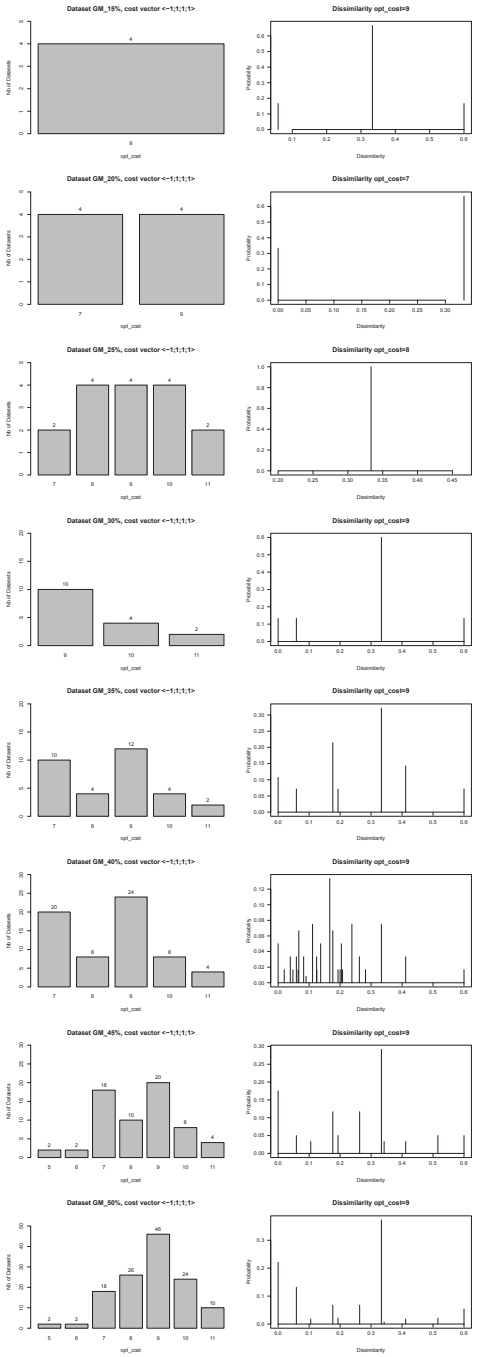


Figure B.115: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $GM_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $GM_{x\%}$ with x=15, 20, 25, 30, 35, 40, 45, 50.

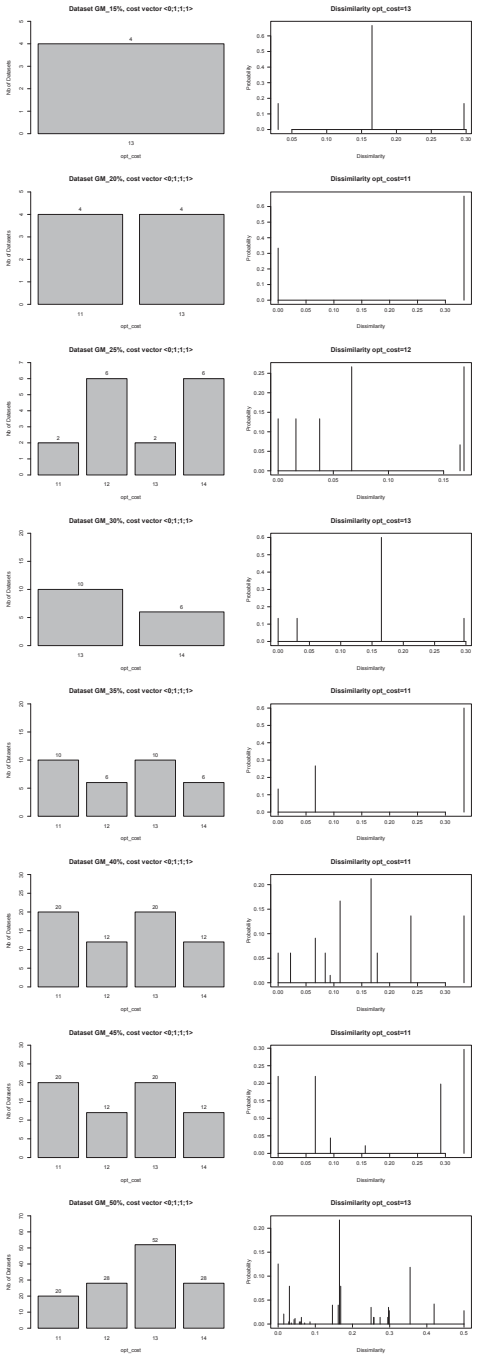


Figure B.116: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $GM_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $GM_{x\%}$ with x=15, 20, 25, 30, 35, 40, 45, 50.

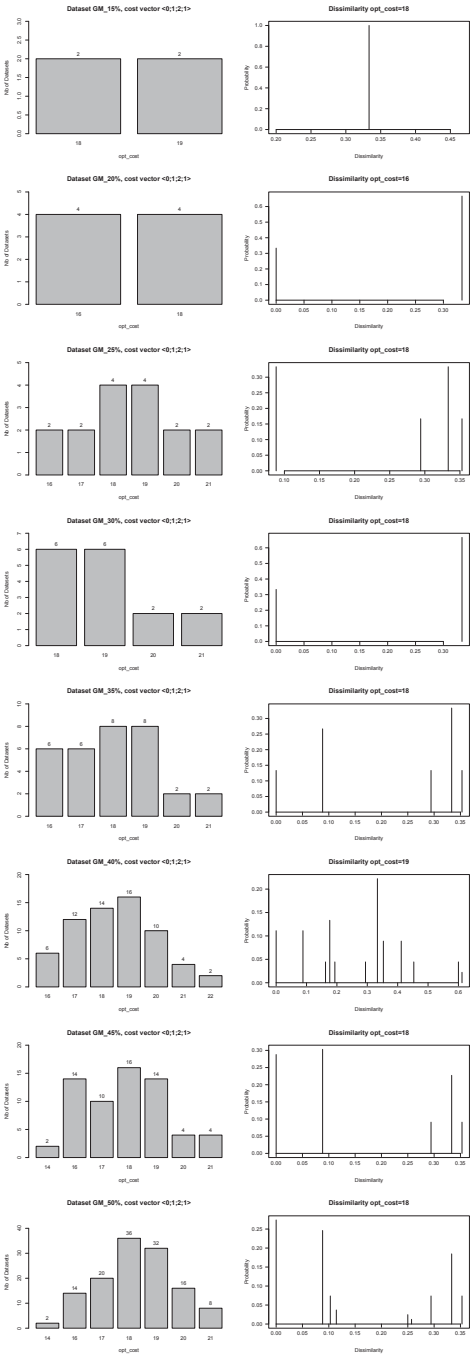


Figure B.117: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $GM_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $GM_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

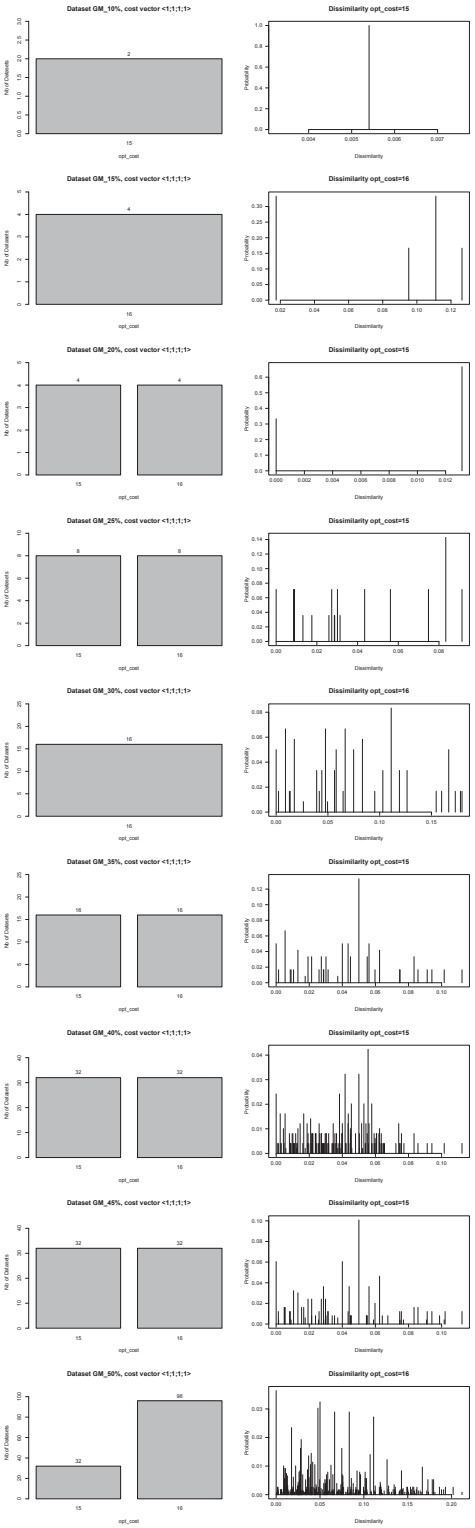


Figure B.118: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $GM_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $GM_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

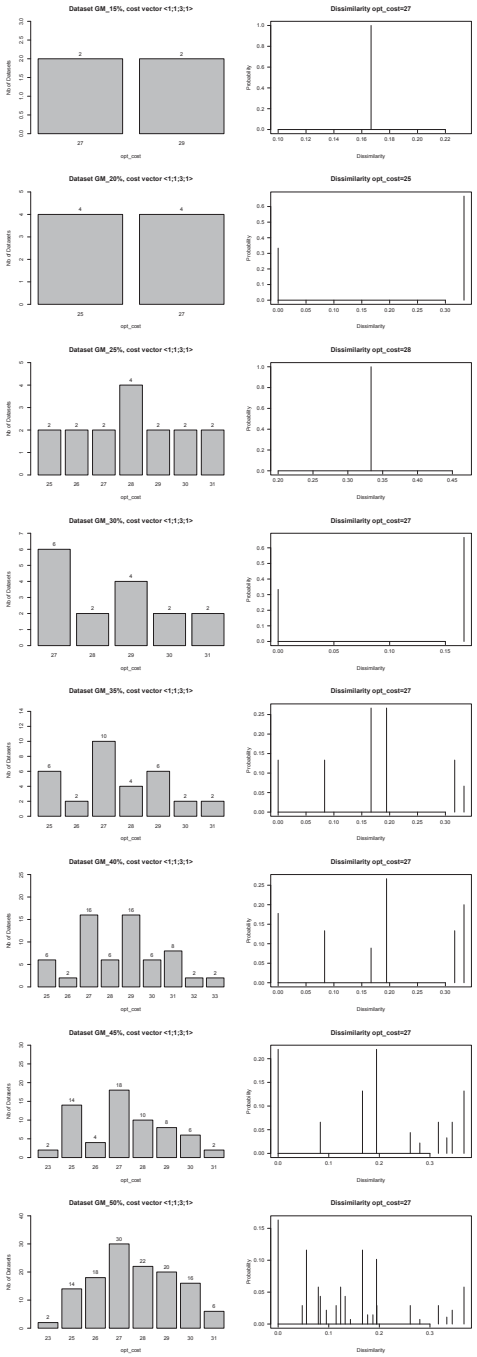


Figure B.119: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $GM_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $GM_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

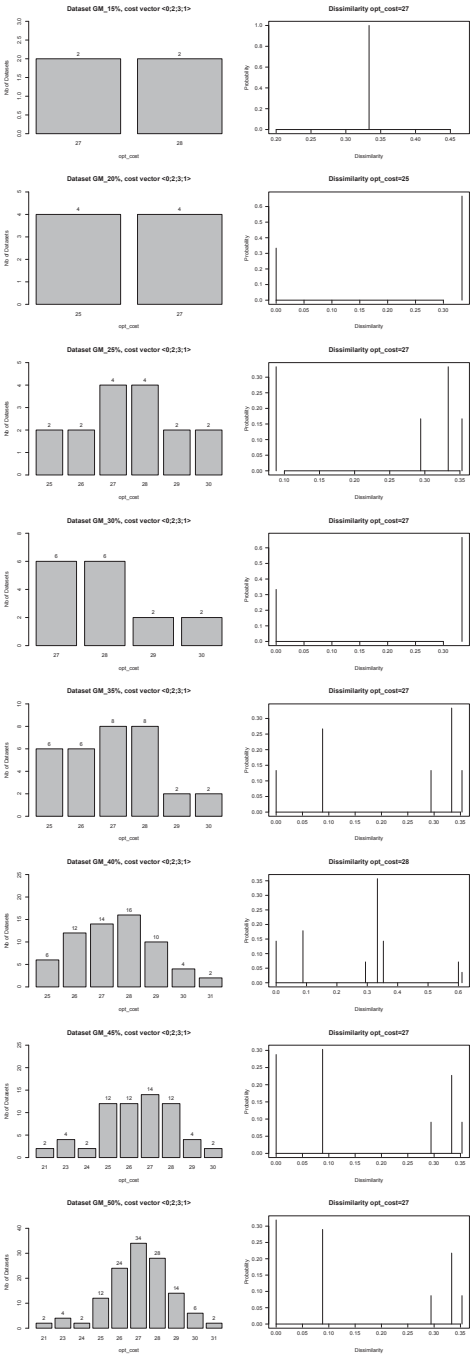


Figure B.120: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $GM_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $GM_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

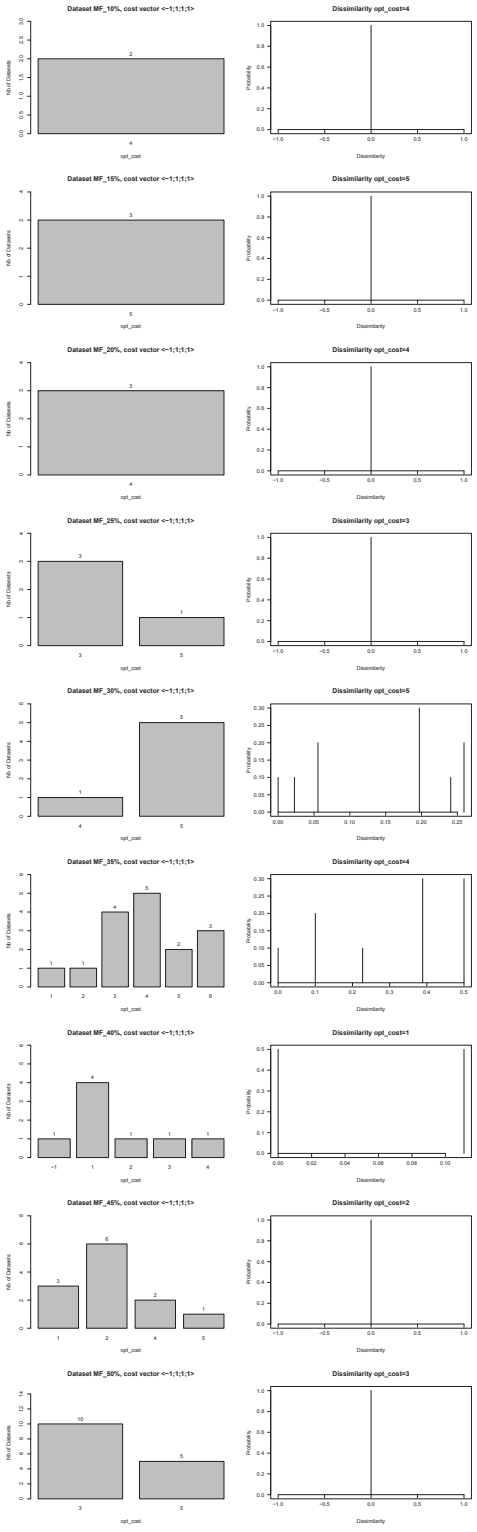


Figure B.121: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MF_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $MF_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

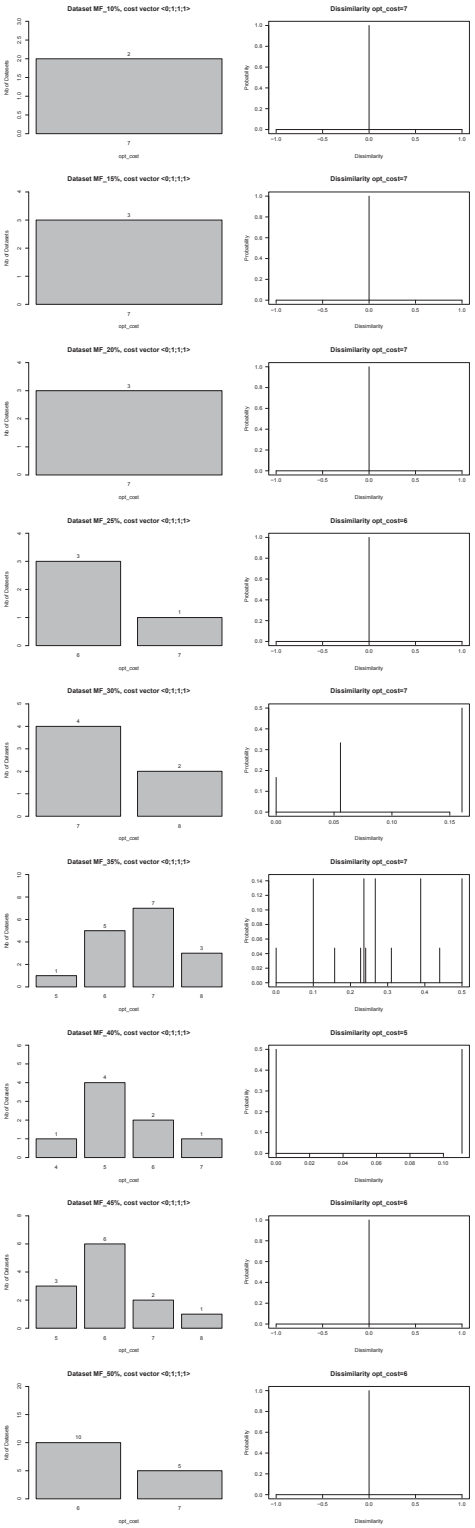


Figure B.122: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MF_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $MF_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

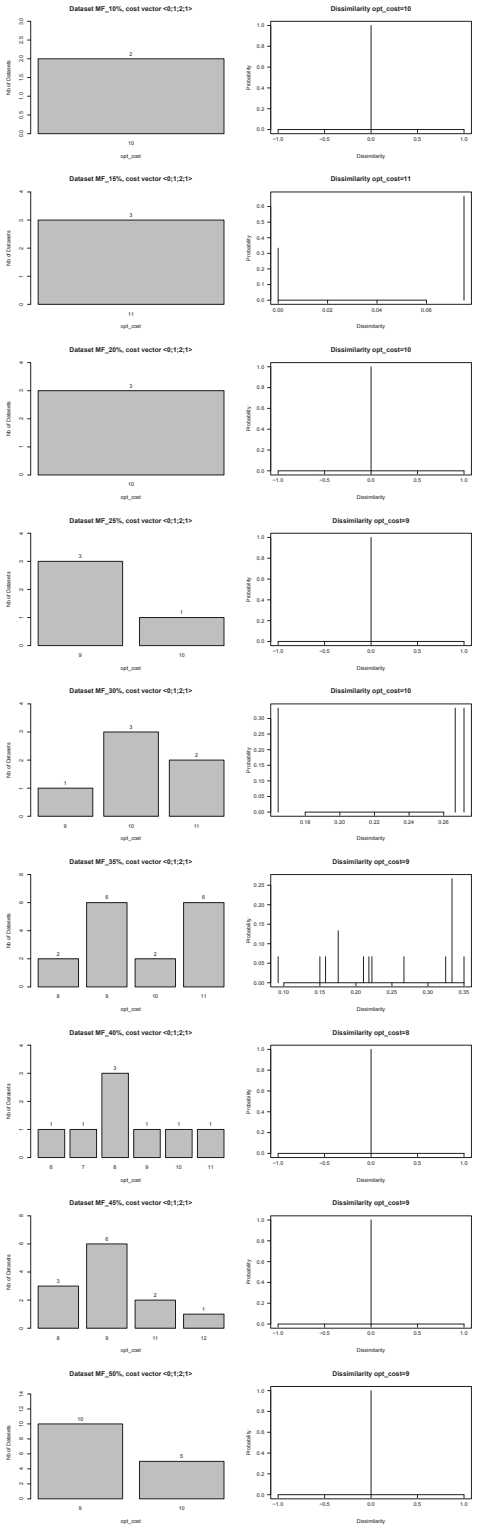


Figure B.123: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MF_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $MF_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

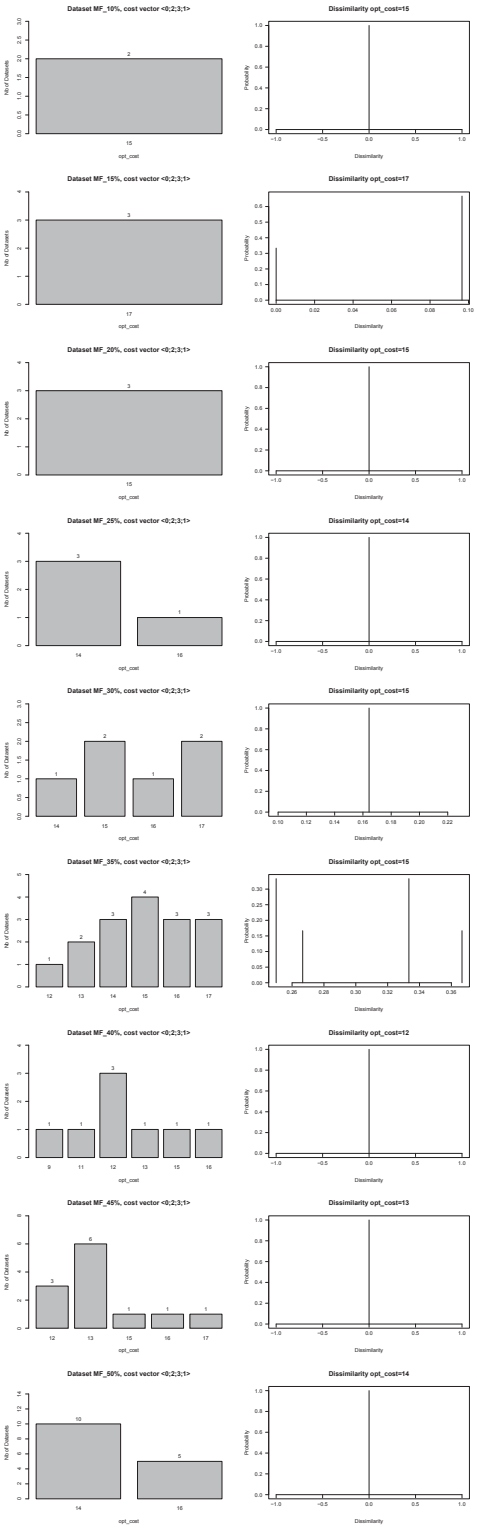


Figure B.124: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MF_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1\rangle$. Each line is a different $MF_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

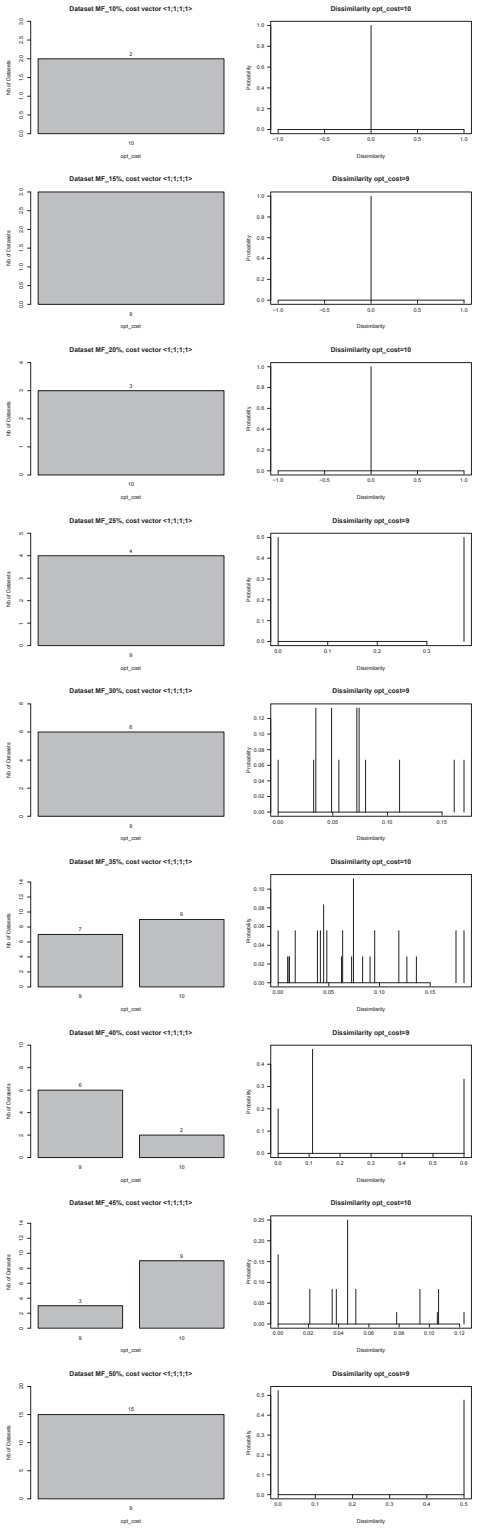


Figure B.125: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MF_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $MF_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

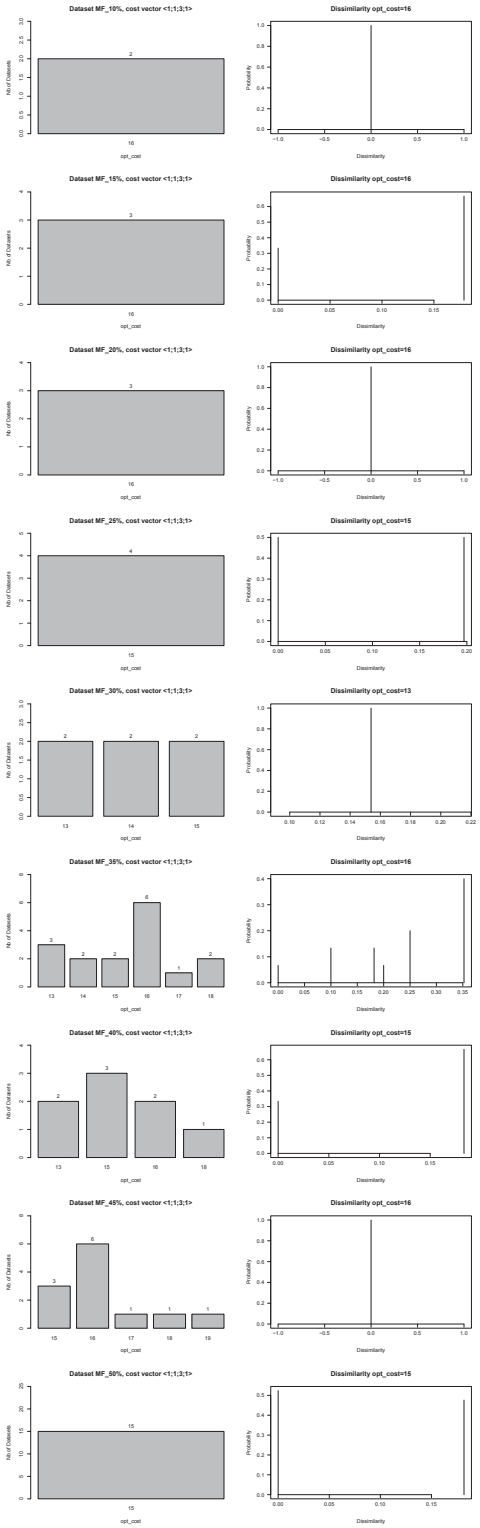


Figure B.126: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MF_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $MF_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

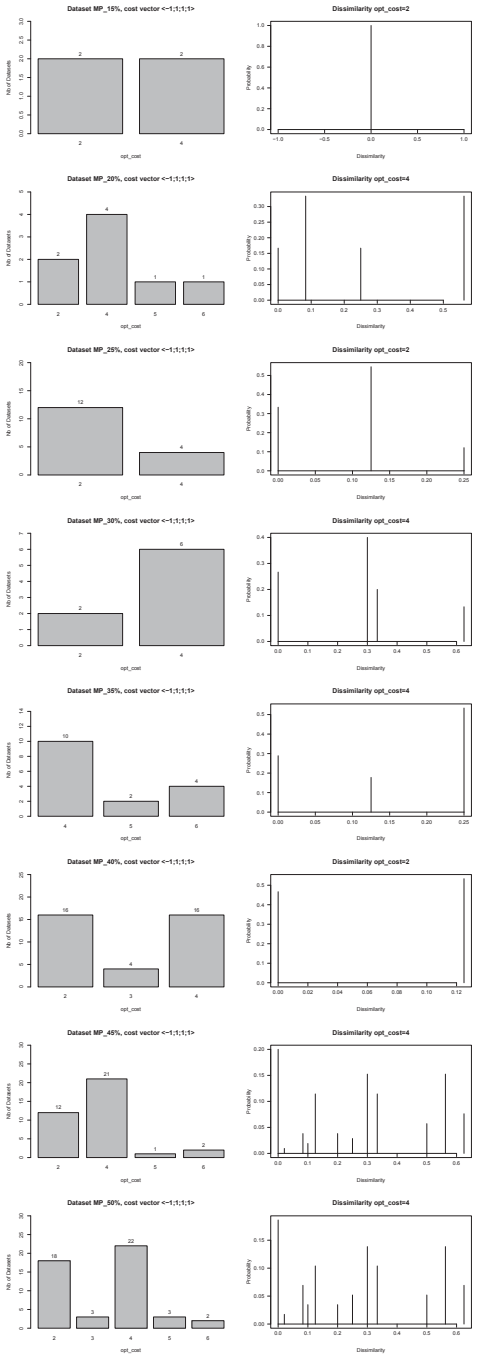


Figure B.127: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $MP_{x\%}$ with $x = 15, 20, 25, 30, 35, 40, 45, 50$.

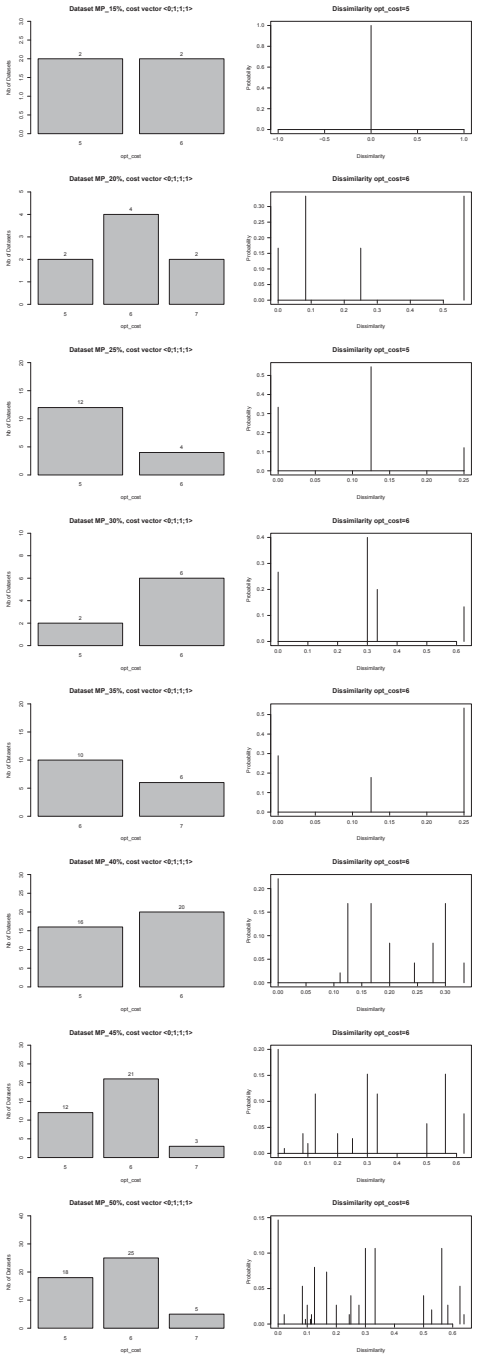


Figure B.128: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $MP_{x\%}$ with $x = 15, 20, 25, 30, 35, 40, 45, 50$.

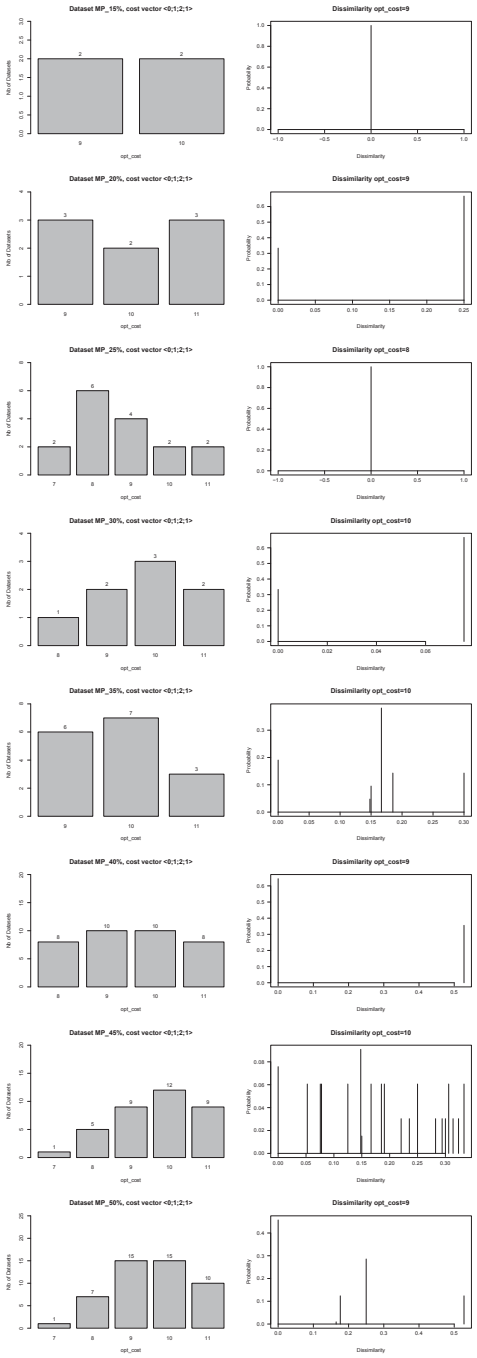


Figure B.129: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $MP_{x\%}$ with $x = 15, 20, 25, 30, 35, 40, 45, 50$.

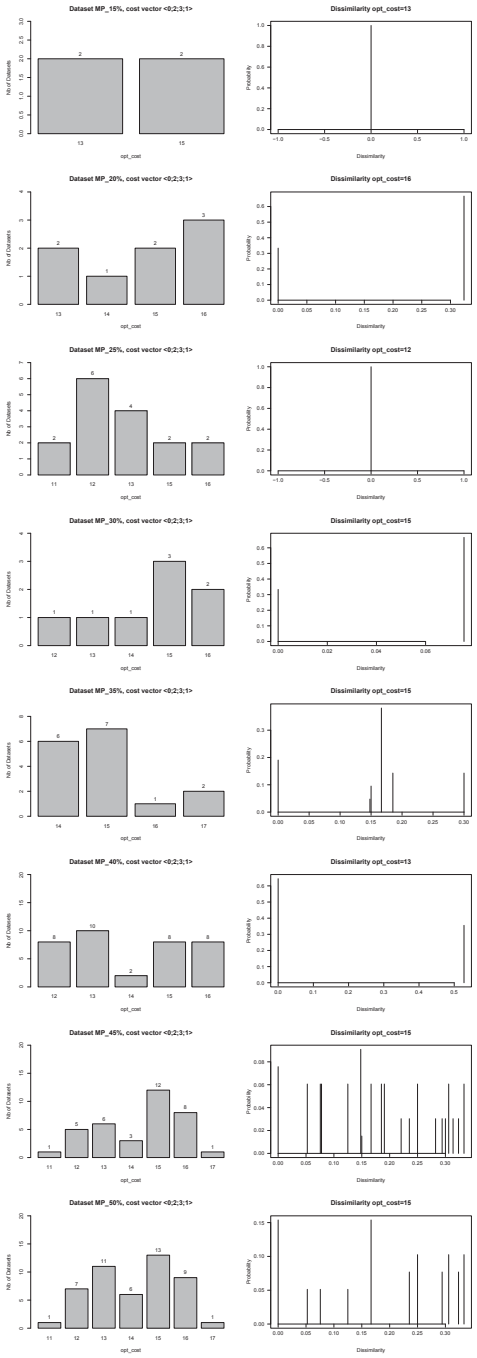


Figure B.130: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $MP_{x\%}$ with $x = 15, 20, 25, 30, 35, 40, 45, 50$.

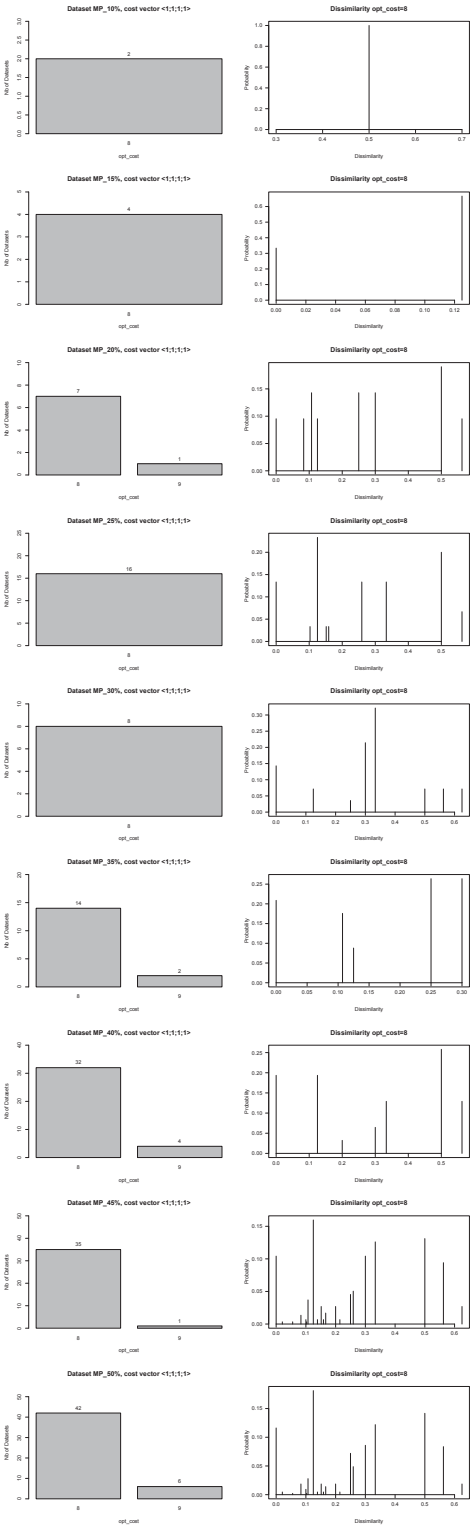


Figure B.131: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $MP_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

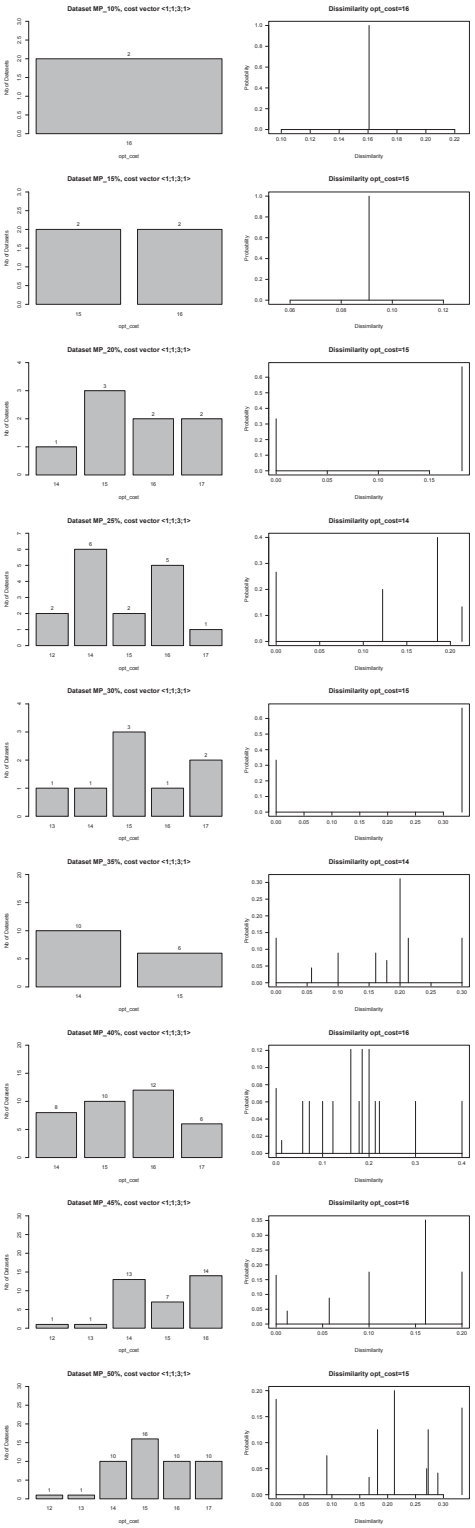


Figure B.132: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $MP_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $MP_{x\%}$ with $x = 10, 15, 20, 25, 30, 35, 40, 45, 50$.

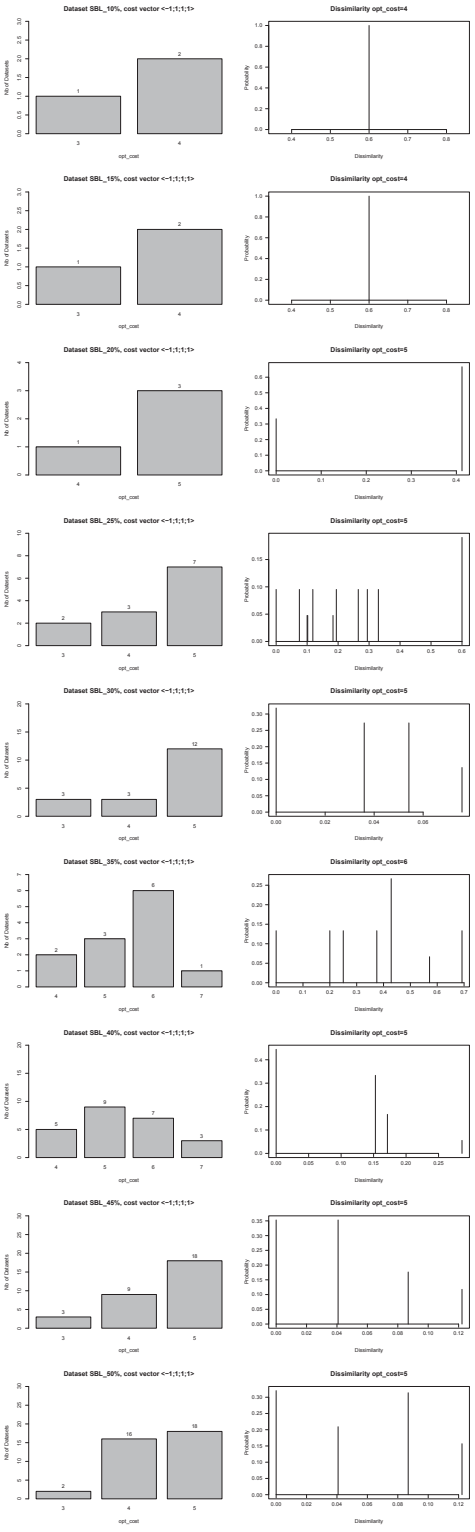


Figure B.133: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $SBL_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

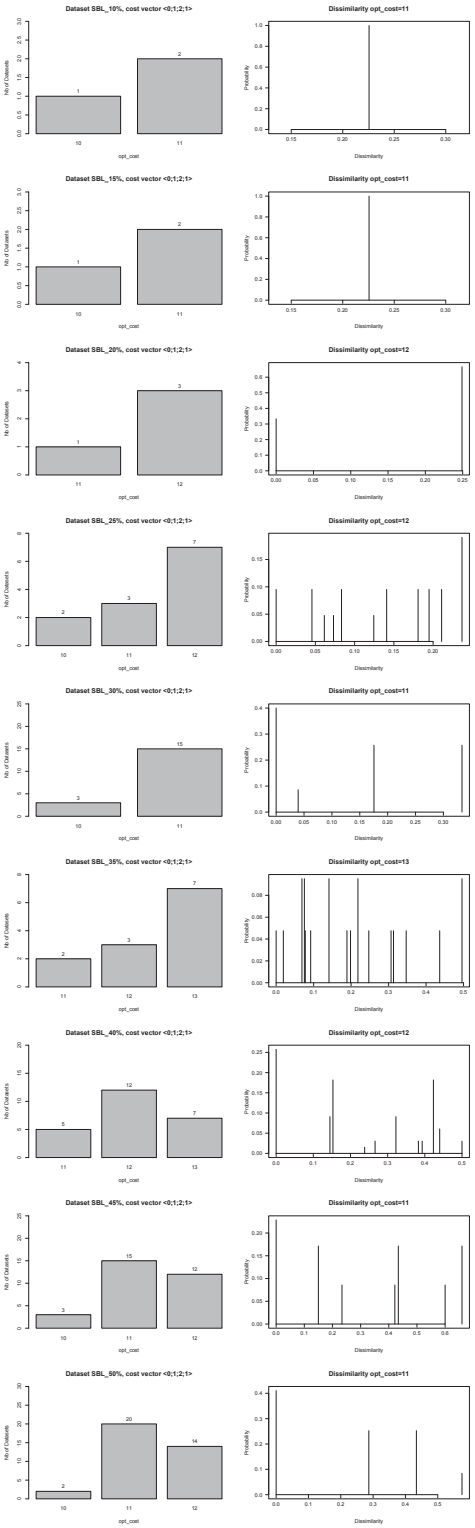


Figure B.134: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$. Each line is a different $SBL_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

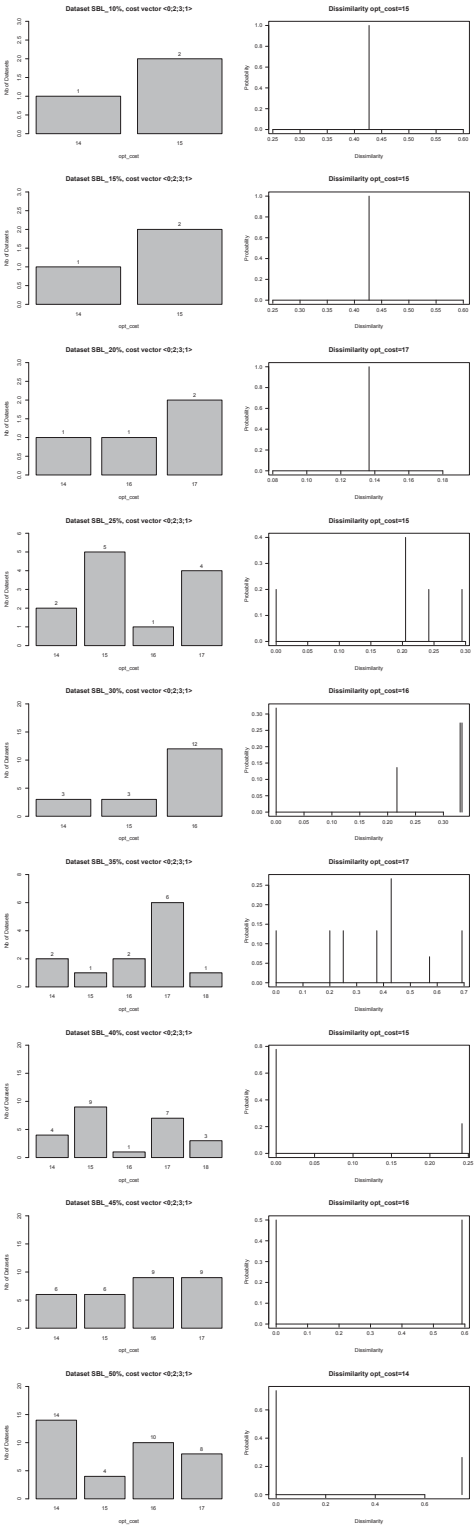


Figure B.135: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $SBL_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

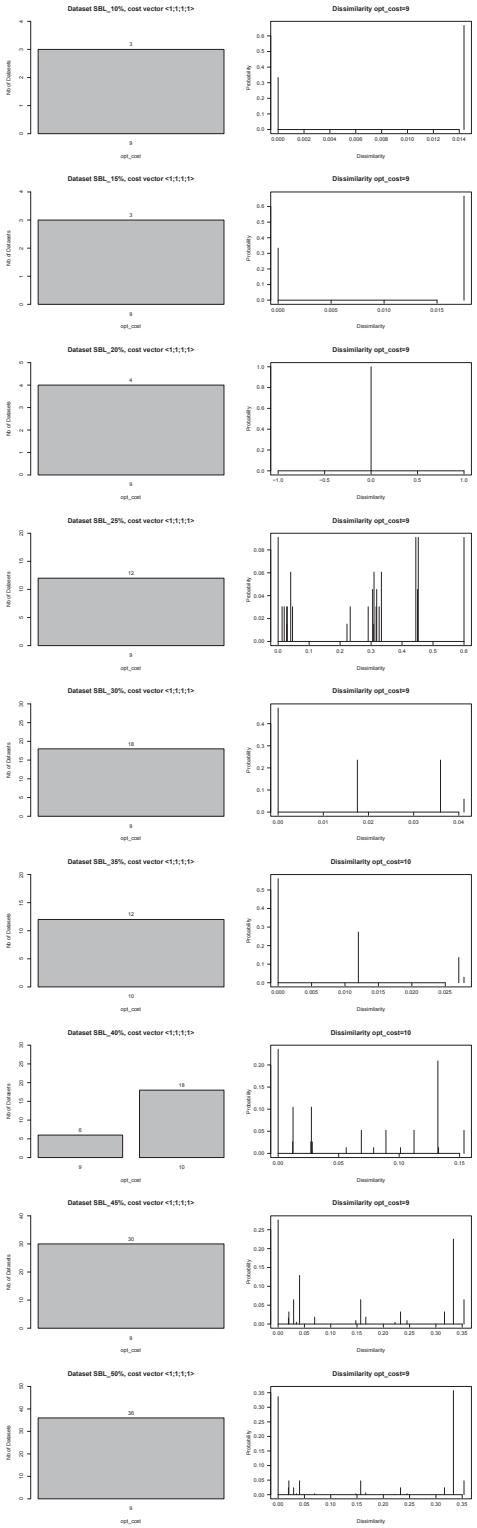


Figure B.136: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $SBL_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

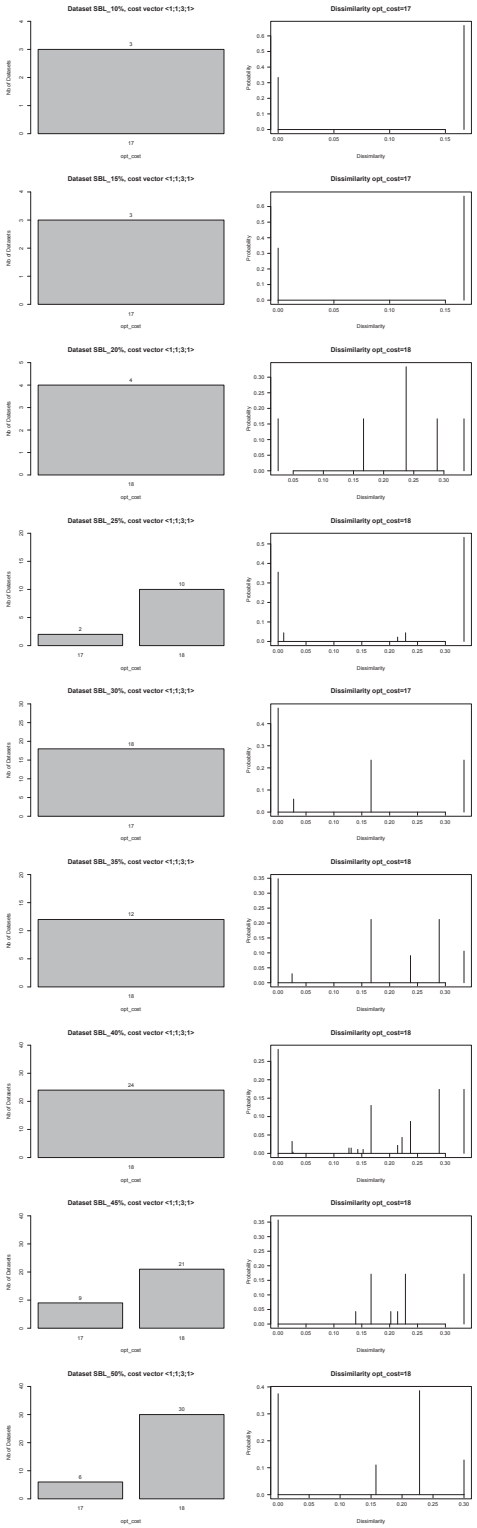


Figure B.137: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SBL_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $SBL_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

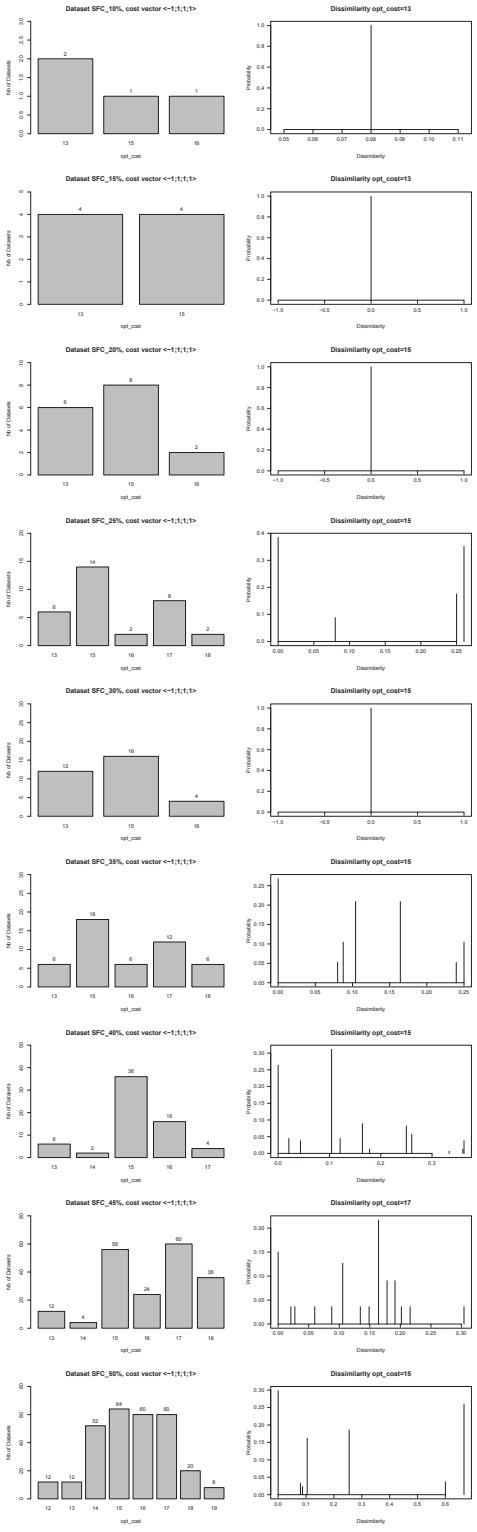


Figure B.138: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SFC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $SFC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

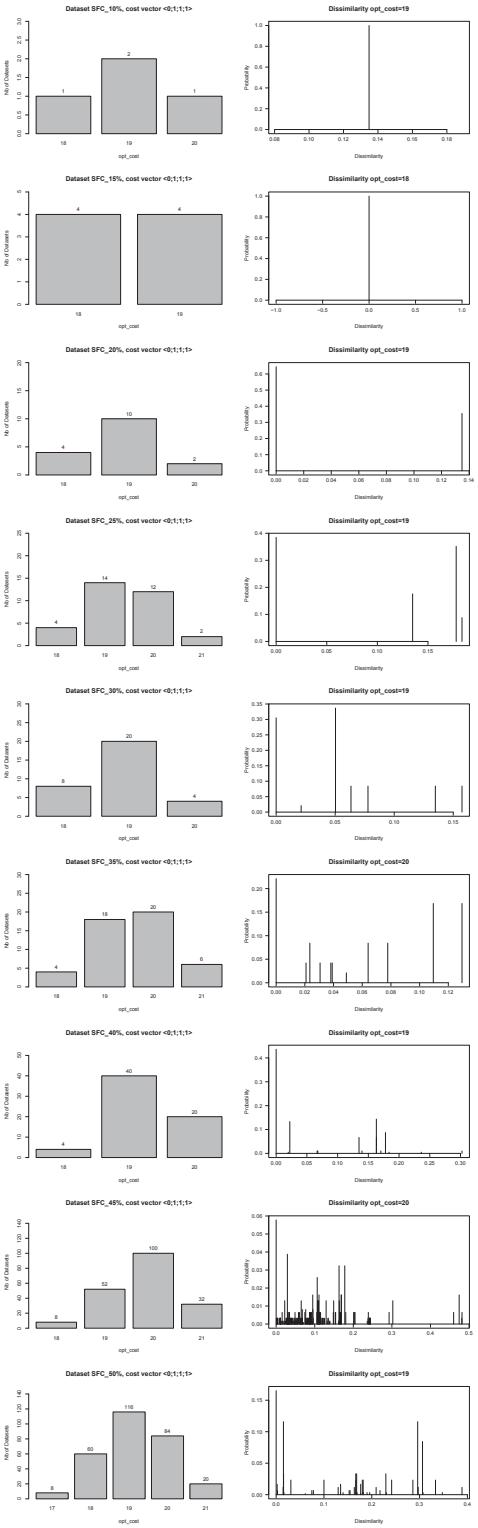


Figure B.139: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SFC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $SFC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

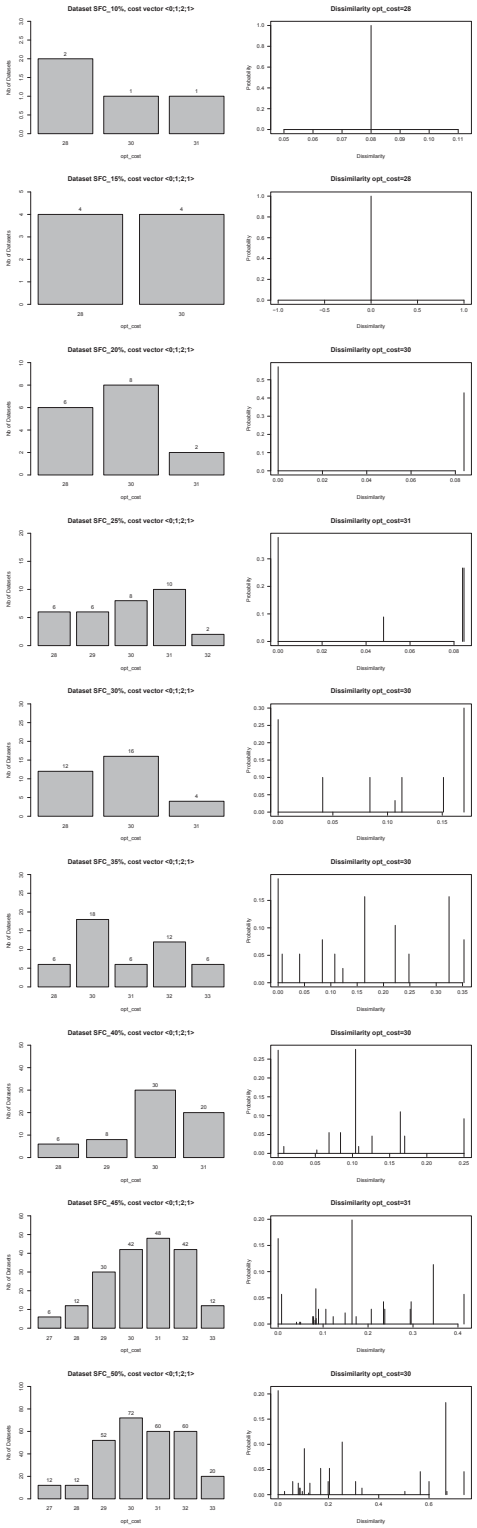


Figure B.140: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SFC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $SFC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

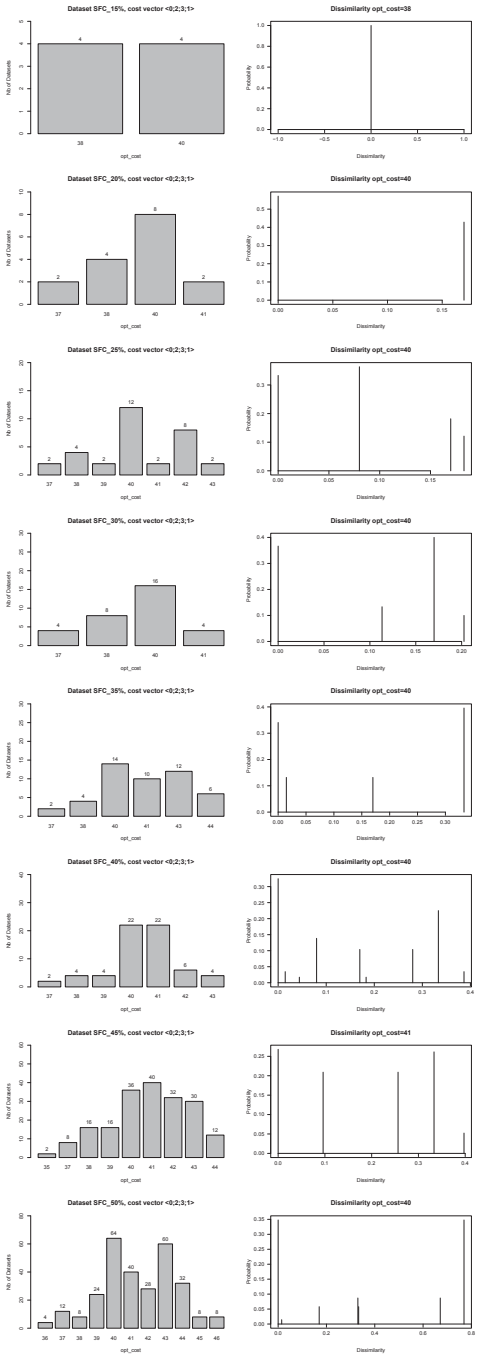


Figure B.141: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SFC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $SFC_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

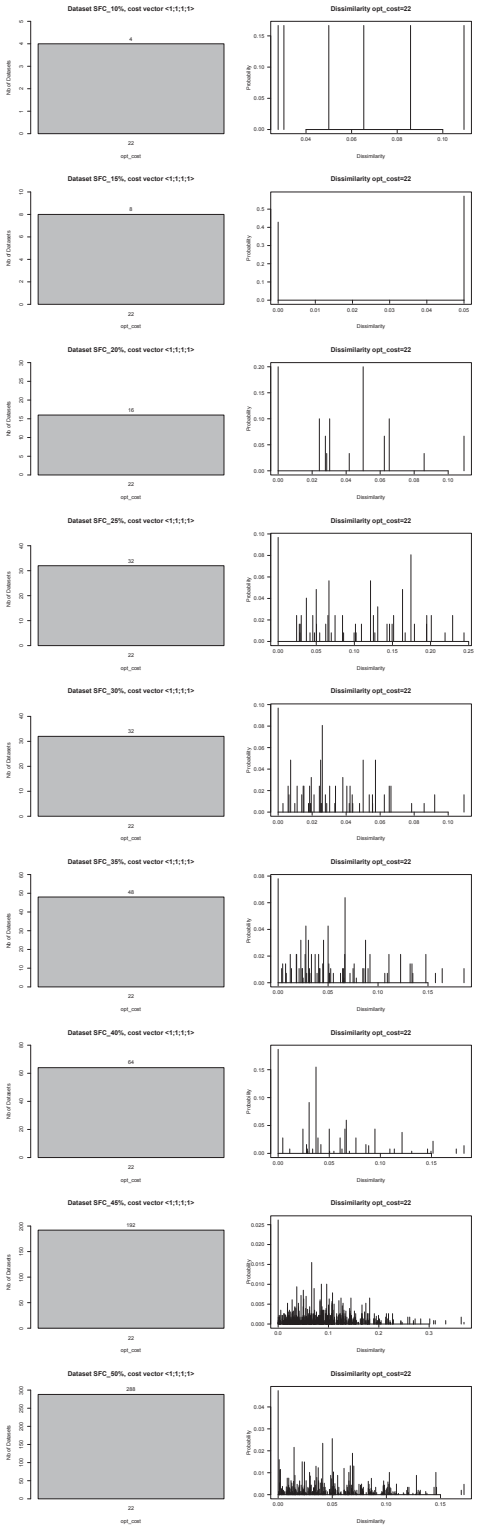


Figure B.142: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SFC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $SFC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

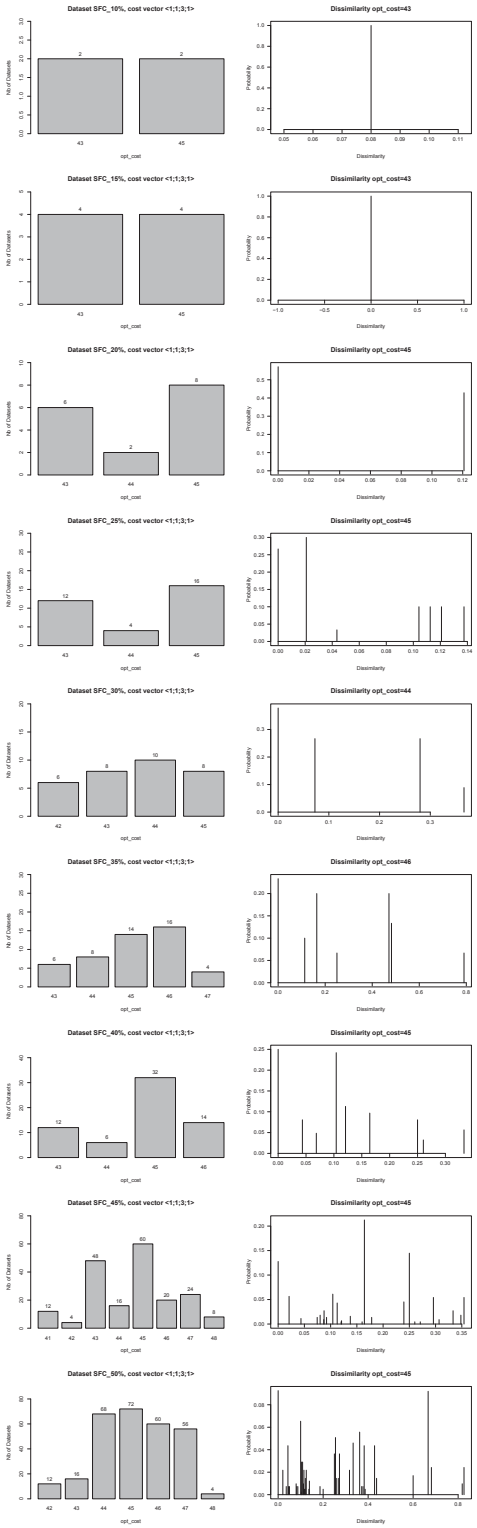


Figure B.143: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SFC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $SFC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

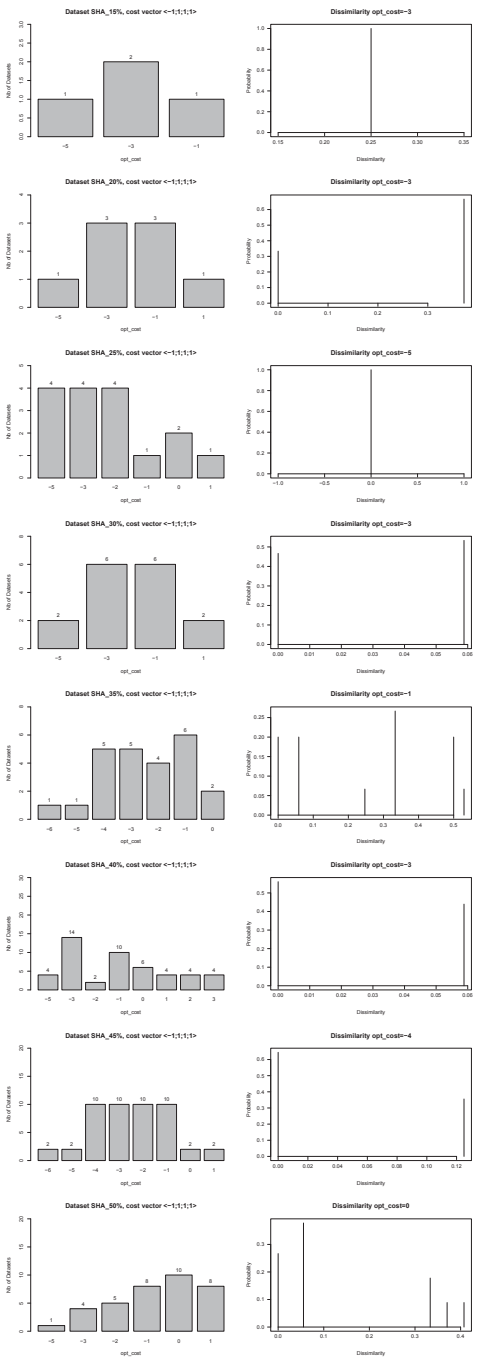


Figure B.144: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SHA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $SHA_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

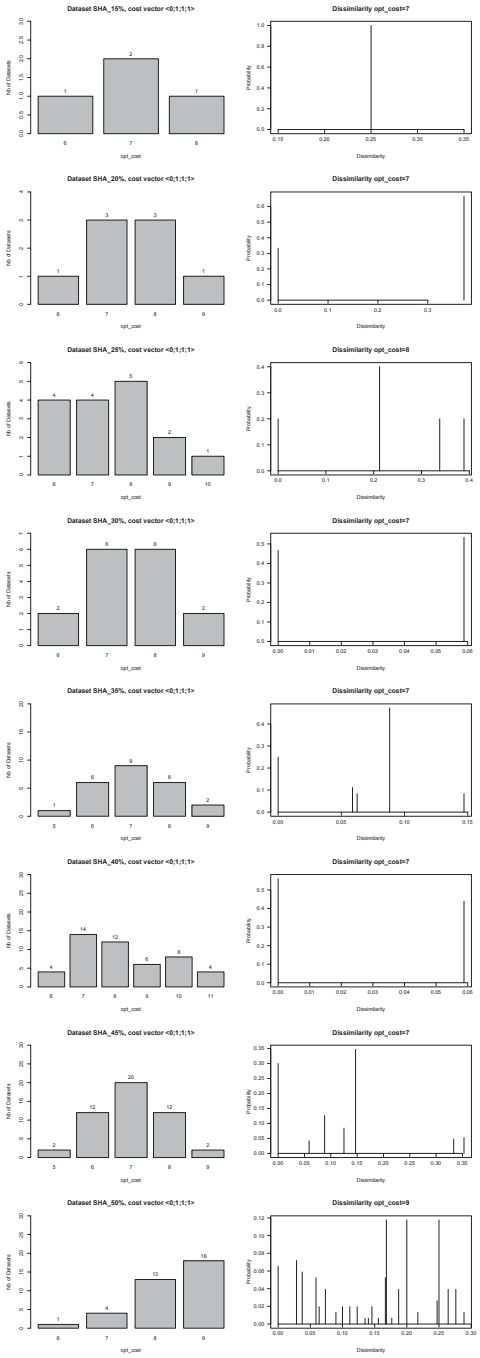


Figure B.145: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SHA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $SHA_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

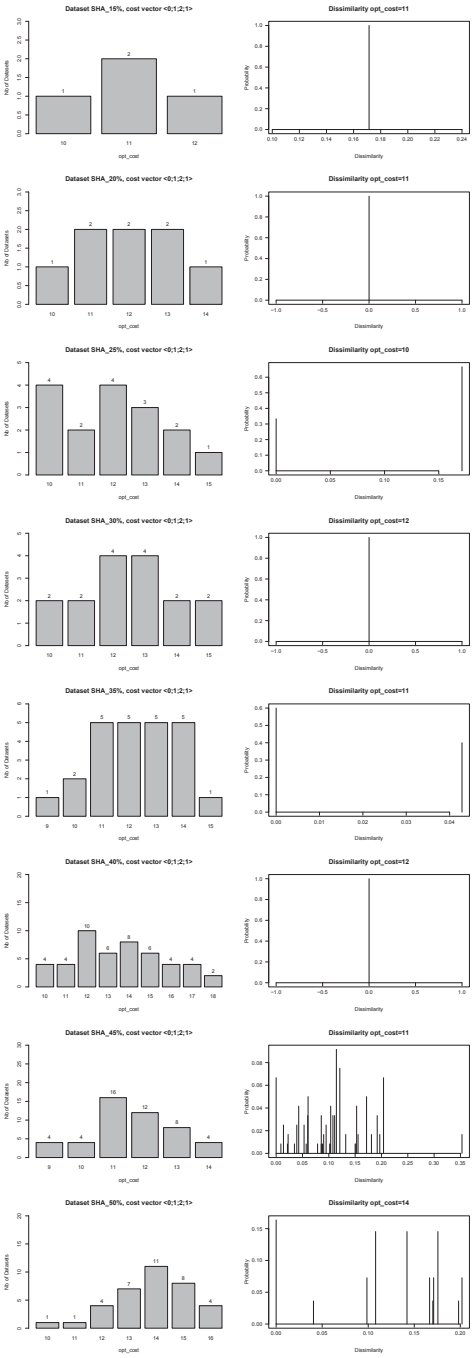


Figure B.146: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SHA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1 \rangle$. Each line is a different $SHA_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

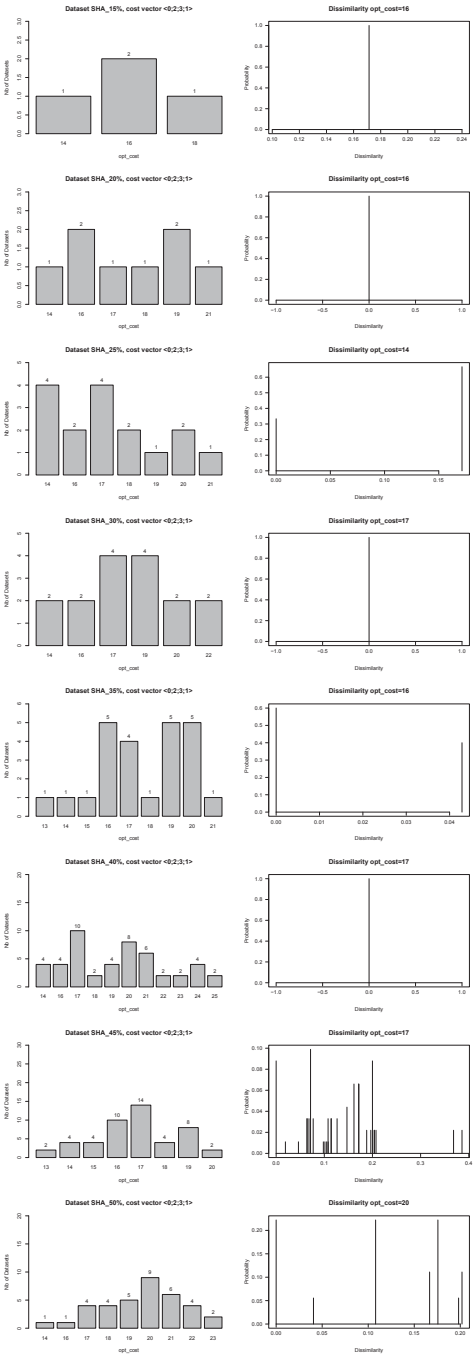


Figure B.147: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SHA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $SHA_{x\%}$ with x= 15, 20, 25, 30, 35, 40, 45, 50.

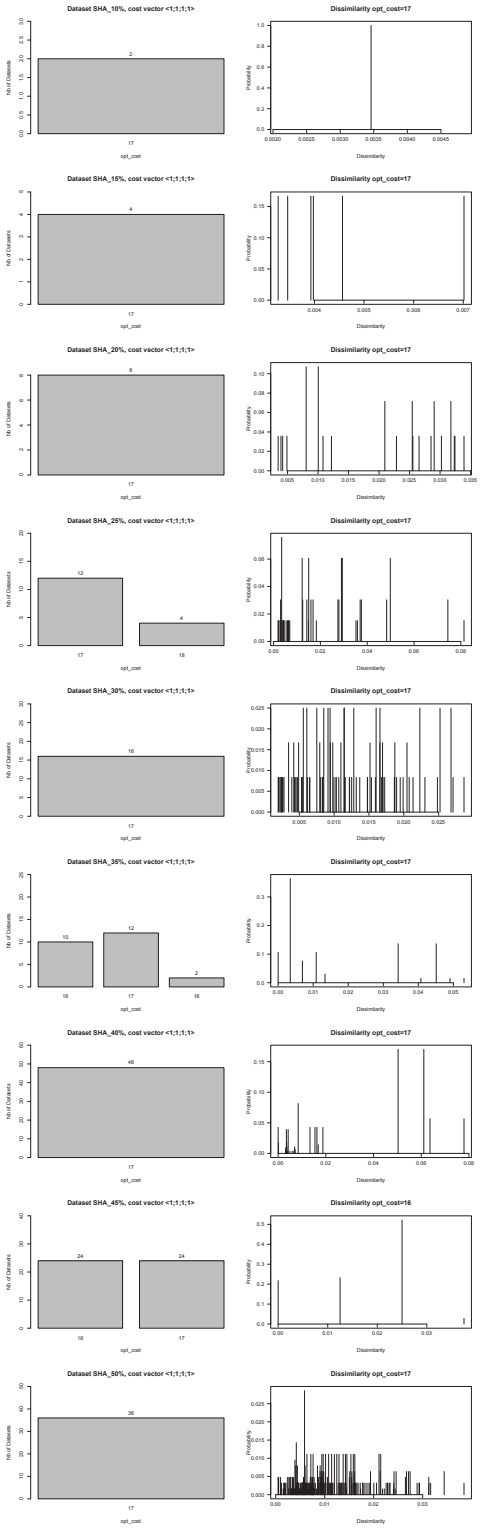


Figure B.148: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SHA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $SHA_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

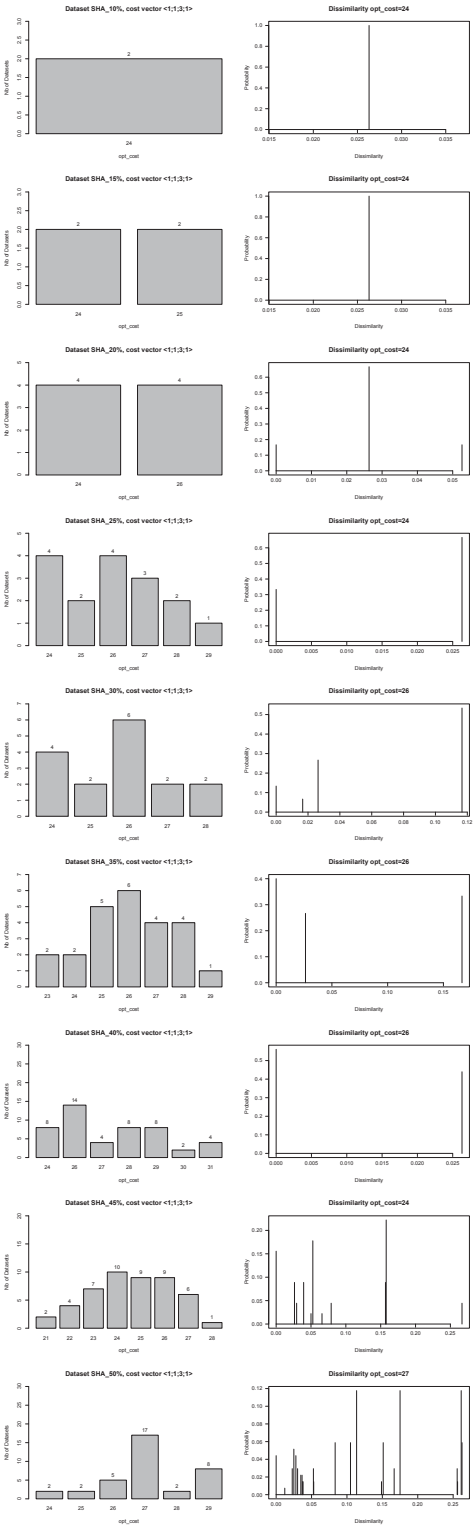


Figure B.149: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $SHA_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $SHA_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

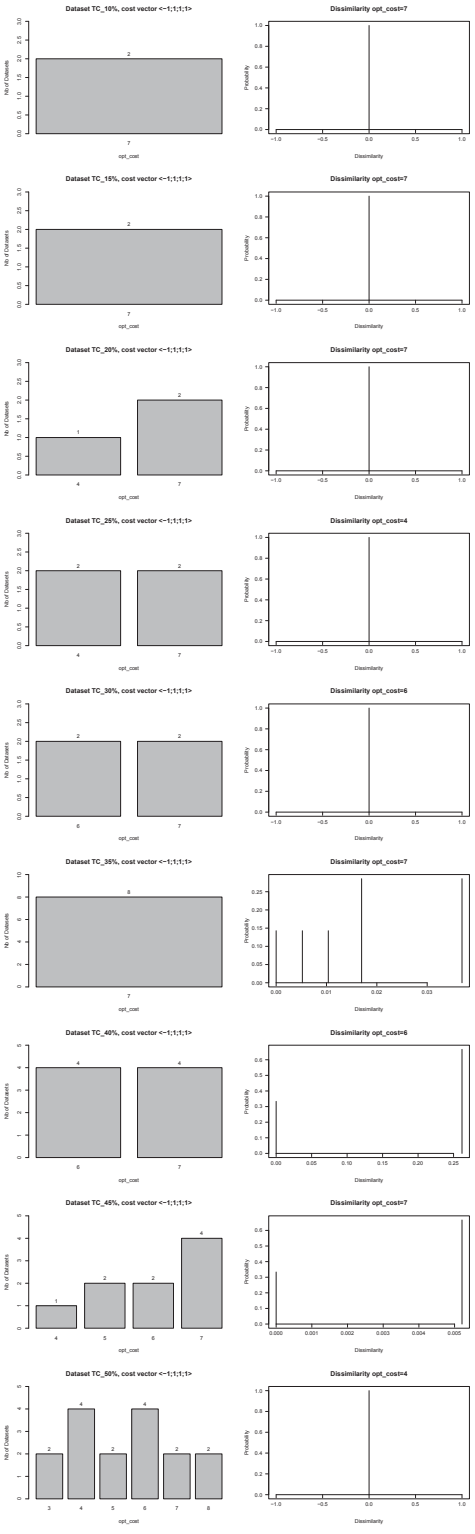


Figure B.150: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $TC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle -1, 1, 1, 1 \rangle$. Each line is a different $TC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

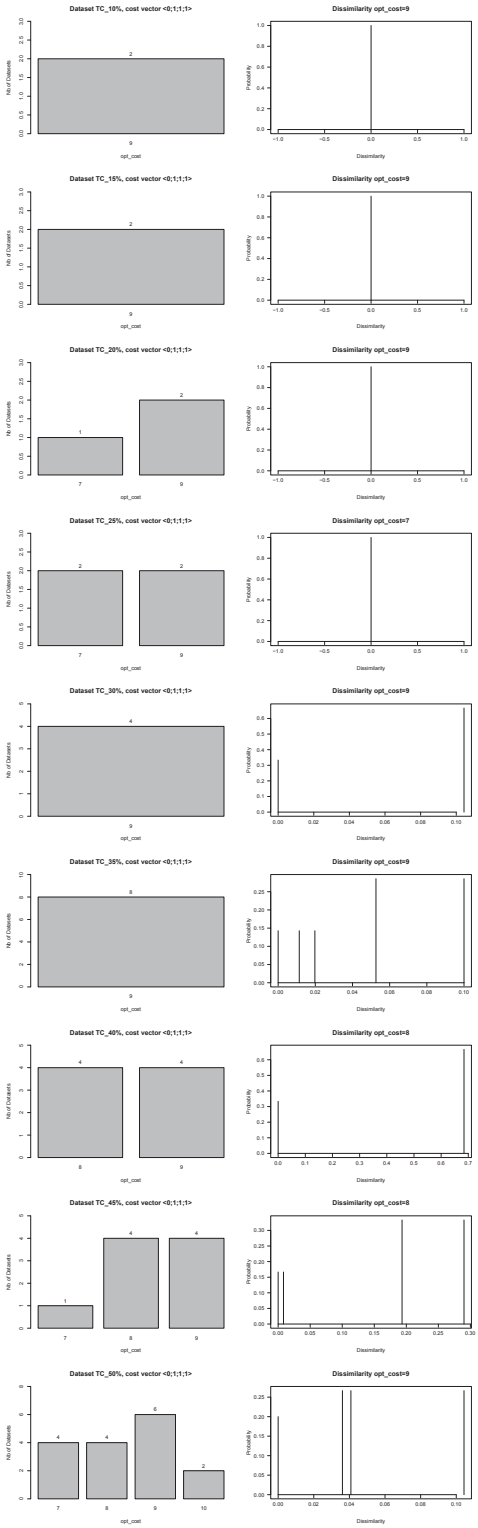


Figure B.151: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $TC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 1, 1\rangle$. Each line is a different $TC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

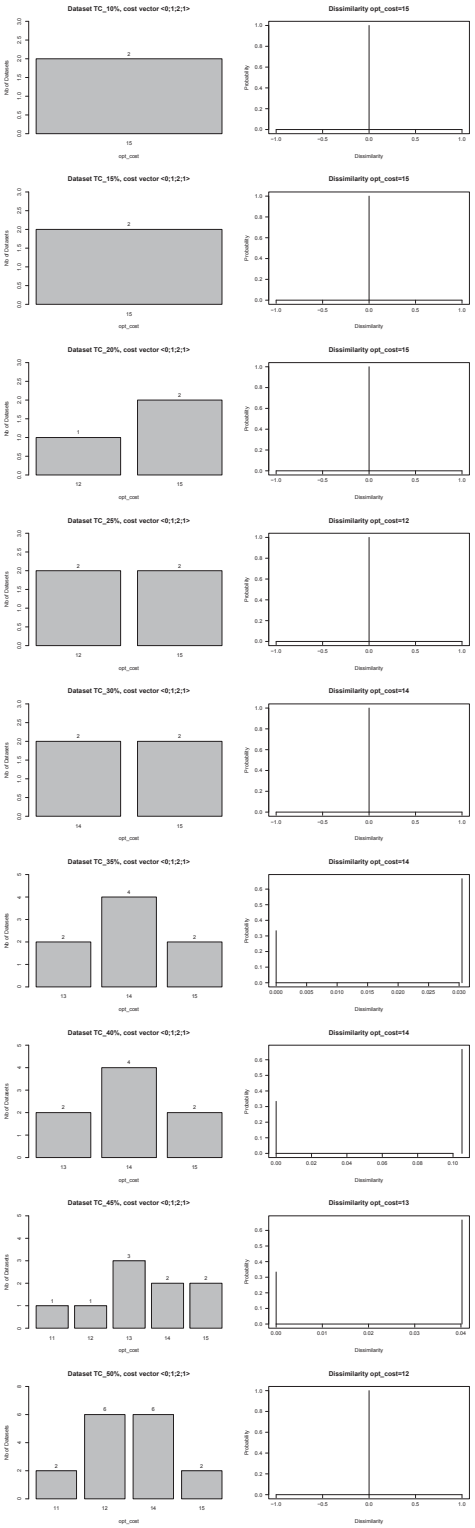


Figure B.152: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $TC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 1, 2, 1\rangle$. Each line is a different $TC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

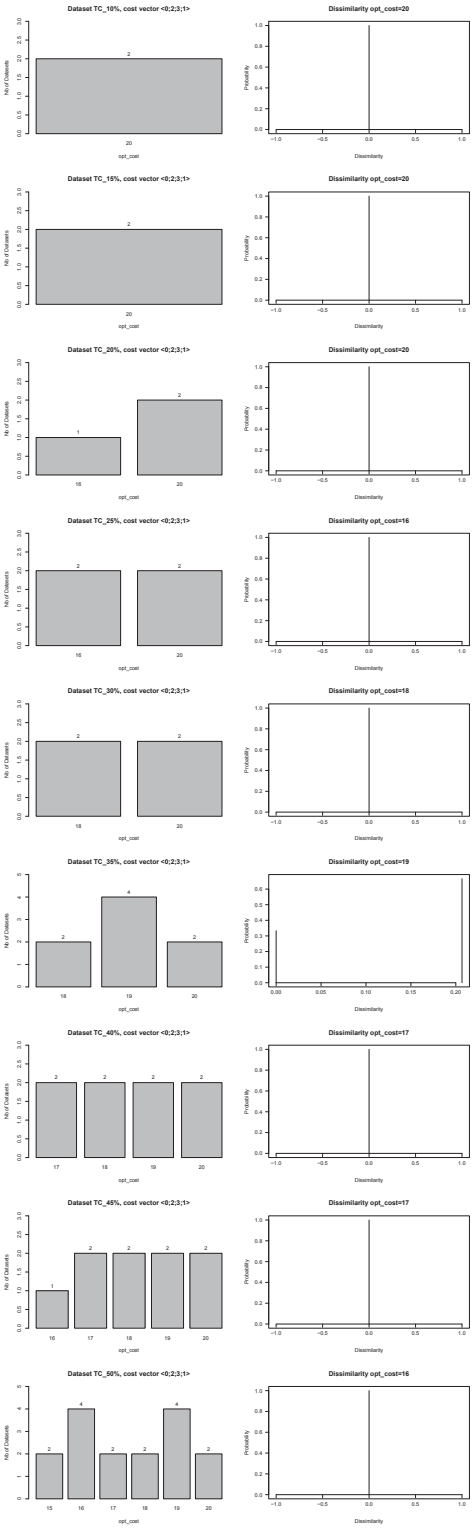


Figure B.153: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $TC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 0, 2, 3, 1 \rangle$. Each line is a different $TC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

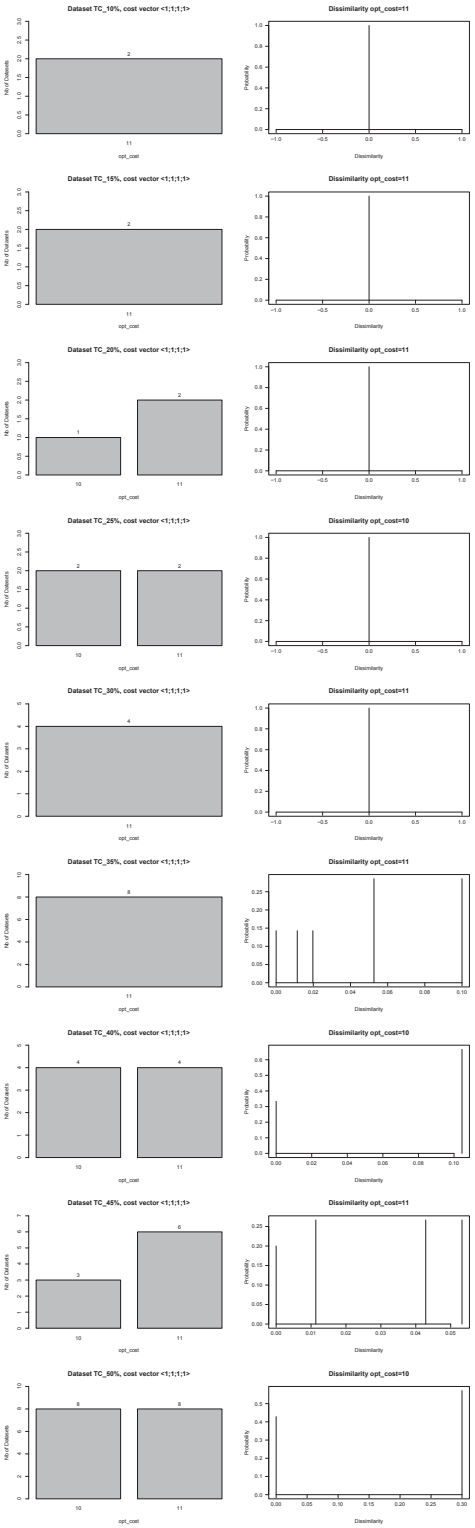


Figure B.154: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $TC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 1, 1 \rangle$. Each line is a different $TC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

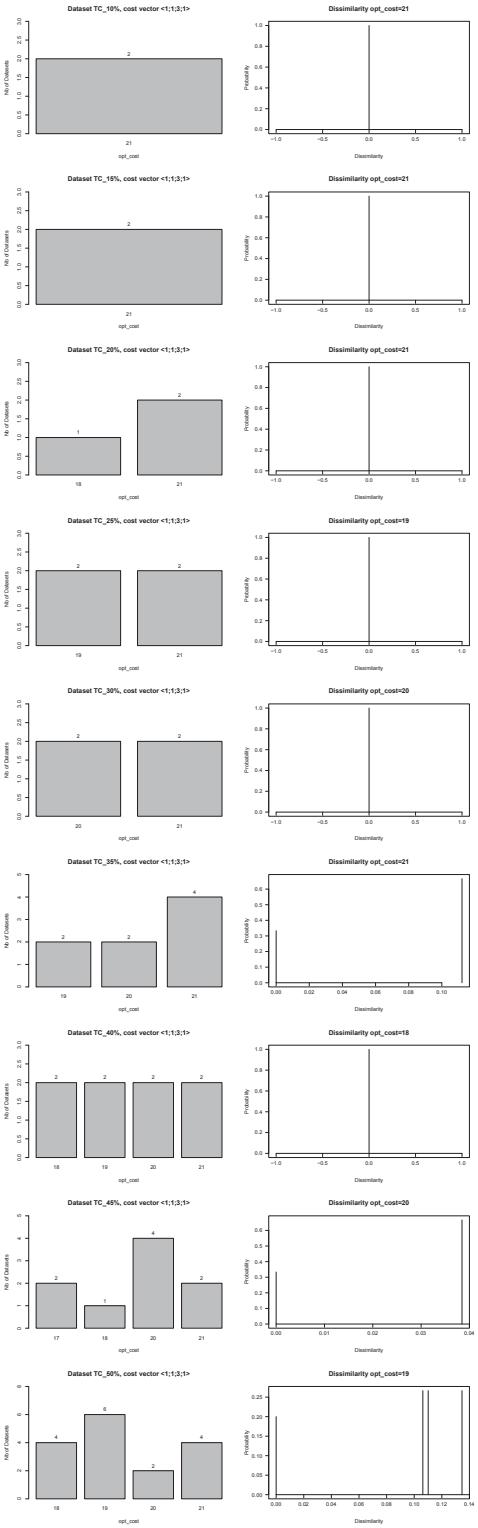


Figure B.155: Barplots of optimal cost (left) and dissimilarity between pairs of reconciliations with the most frequent optimal cost (right) obtained on the datasets derived from the $TC_{x\%}$ dataset by resolving the multiple associations in all the possible ways and computed with the cost vector $\langle 1, 1, 3, 1 \rangle$. Each line is a different $TC_{x\%}$ with x= 10, 15, 20, 25, 30, 35, 40, 45, 50.

B.2.4 Re-rooting: the plateau property

This section presents the results for the re-rooting experiment focusing on the plateau property. Tables B.1 and B.2 contain the results for the 28 biological datasets evaluated with the 6 cost vectors in $\mathcal{C}$.

Notice that for 2 of the 28 datasets, namely PP and RH, we did not obtain any result for the cost vector $\langle 1, 1, 1, 1\rangle$, due to the long computational time.

Tables B.3 and B.4 are dedicated to the simulated datasets. Let us recall that for each biological dataset D, we simulated 50 datasets, the whole set of which is being called D-sim, relying on parameter values estimated on the corresponding biological dataset. Due to the large number of simulated datasets, in these tables we present the results summarised. Notice that for the simulated datasets, we did not obtain any results for the cost vector $\langle 1, 1, 1, 1\rangle$, because of the very long computational time.

In general, we observed between 1 and 5 plateaux in the simulated datasets and Tables B.3 and B.4 report the number of datasets exhibiting between 1 and 5 plateaux among each set of 50 simulated datasets. In three cases, that were not reported in these tables due to space limits, we observe exactly 6 plateaux (these are CA-sim for the cost vector $\langle 1, 1, 2, 1\rangle$, FE-sim for the cost vectors $\langle -1, 1, 1, 1\rangle$ and $\langle 0, 1, 1, 1\rangle$).

Table B.1: Tables showing the results of the re-rooting for the first 16 biological datasets. Only time-feasible reconciliations are retained. Each line shows a summary of the time-feasible reconciliations obtained with all possible rootings for one dataset analysed with one cost vector. Column A indicates the cost vector, column B shows the smallest optimal cost obtained among all possible reconciliations, column C shows the number of plateaux, column D shows the optimal reconciliation cost for the original root (with no value when there is no time-feasible solution), column E indicates whether the root belongs to a plateau.

Dataset	*A*	*B*	*C*	*D*	*E*
AP	$\langle -1;1;1;1\rangle$	1	1	5	no
AP	$\langle 0;1;1;1\rangle$	4	1	6	no
AP	$\langle 0;1;2;1\rangle$	6	1	11	no
AP	$\langle 0;2;3;1\rangle$	9	1	15	no
AP	$\langle 1;1;1;1\rangle$	7	1	7	yes
AP	$\langle 1;1;3;1\rangle$	11	1	16	no
AS	$\langle -1;1;1;1\rangle$	1	1	3	no
AS	$\langle 0;1;1;1\rangle$	10	1	11	no
AS	$\langle 0;1;2;1\rangle$	15	1	17	no
AS	$\langle 0;2;3;1\rangle$	20	1	23	yes
AS	$\langle 1;1;1;1\rangle$	19	1	19	no
AS	$\langle 1;1;3;1\rangle$	29	1	31	no
AW	$\langle -1;1;1;1\rangle$	3	1	6	no
AW	$\langle 0;1;1;1\rangle$	8	1	10	no
AW	$\langle 0;1;2;1\rangle$	14	1	17	no
AW	$\langle 0;2;3;1\rangle$	19	1	23	no
AW	$\langle 1;1;1;1\rangle$	13	1	14	no
AW	$\langle 1;1;3;1\rangle$	25	1	28	no
CA	$\langle -1;1;1;1\rangle$	16	2	6	yes
CA	$\langle 0;1;1;1\rangle$	21	1	10	yes
CA	$\langle 0;1;2;1\rangle$	36	2	17	yes
CA	$\langle 0;2;3;1\rangle$	49	2	23	no
CA	$\langle 1;1;1;1\rangle$	25	1	14	yes
CA	$\langle 1;1;3;1\rangle$	56	2	28	yes
CP	$\langle -1;1;1;1\rangle$	10	1	-	no
CP	$\langle 0;1;1;1\rangle$	19	1	-	no
CP	$\langle 0;1;2;1\rangle$	29	1	-	no
CP	$\langle 0;2;3;1\rangle$	49	1	-	no
CP	$\langle 1;1;1;1\rangle$	28	1	29	no
CP	$\langle 1;1;3;1\rangle$	41	1	42	no
CT	$\langle -1;1;1;1\rangle$	9	2	-	no
CT	$\langle 0;1;1;1\rangle$	15	1	-	no
CT	$\langle 0;1;2;1\rangle$	27	2	-	no
CT	$\langle 0;2;3;1\rangle$	33	1	-	no
CT	$\langle 1;1;1;1\rangle$	20	1	20	yes
CT	$\langle 1;1;3;1\rangle$	43	2	-	no
EC	$\langle -1;1;1;1\rangle$	1	1	1	yes
EC	$\langle 0;1;1;1\rangle$	6	2	6	yes
EC	$\langle 0;1;2;1\rangle$	10	1	10	yes
EC	$\langle 0;2;3;1\rangle$	14	1	14	yes
EC	$\langle 1;1;1;1\rangle$	10	2	10	yes
EC	$\langle 1;1;3;1\rangle$	16	1	16	yes
FA	$\langle -1;1;1;1\rangle$	0	1	1	no
FA	$\langle 0;1;1;1\rangle$	5	1	5	yes
FA	$\langle 0;1;2;1\rangle$	6	1	7	no
FA	$\langle 0;2;3;1\rangle$	8	1	9	no
FA	$\langle 1;1;1;1\rangle$	8	1	8	yes
FA	$\langle 1;1;3;1\rangle$	12	1	12	yes

Dataset	*A*	*B*	*C*	*D*	*E*
FD	$\langle -1;1;1;1\rangle$	31	1	34	no
FD	$\langle 0;1;1;1\rangle$	42	1	44	no
FD	$\langle 0;1;2;1\rangle$	63	1	66	no
FD	$\langle 0;2;3;1\rangle$	106	1	-	no
FD	$\langle 1;1;1;1\rangle$	51	1	52	no
FD	$\langle 1;1;3;1\rangle$	91	1	94	no
FE	$\langle -1;1;1;1\rangle$	-1	1	4	no
FE	$\langle 0;1;1;1\rangle$	2	1	6	no
FE	$\langle 0;1;2;1\rangle$	3	1	8	no
FE	$\langle 0;2;3;1\rangle$	4	1	10	no
FE	$\langle 1;1;1;1\rangle$	5	1	7	no
FE	$\langle 1;1;3;1\rangle$	7	1	12	no
GL	$\langle -1;1;1;1\rangle$	-2	1	-2	yes
GL	$\langle 0;1;1;1\rangle$	4	1	4	yes
GL	$\langle 0;1;2;1\rangle$	7	1	7	yes
GL	$\langle 0;2;3;1\rangle$	10	1	10	yes
GL	$\langle 1;1;1;1\rangle$	10	1	10	yes
GL	$\langle 1;1;3;1\rangle$	15	1	16	no
GM	$\langle -1;1;1;1\rangle$	7	1	10	no
GM	$\langle 0;1;1;1\rangle$	11	1	13	no
GM	$\langle 0;1;2;1\rangle$	17	1	20	no
GM	$\langle 0;2;3;1\rangle$	26	1	29	no
GM	$\langle 1;1;1;1\rangle$	15	1	16	no
GM	$\langle 1;1;3;1\rangle$	27	1	30	no
IFL	$\langle -1;1;1;1\rangle$	-3	1	-1	no
IFL	$\langle 0;1;1;1\rangle$	8	1	8	yes
IFL	$\langle 0;1;2;1\rangle$	13	1	15	no
IFL	$\langle 0;2;3;1\rangle$	18	1	21	no
IFL	$\langle 1;1;1;1\rangle$	17	1	17	yes
IFL	$\langle 1;1;3;1\rangle$	29	1	31	no
MF	$\langle -1;1;1;1\rangle$	5	1	6	no
MF	$\langle 0;1;1;1\rangle$	8	1	10	no
MF	$\langle 0;1;2;1\rangle$	11	1	12	no
MF	$\langle 0;2;3;1\rangle$	15	1	16	no
MF	$\langle 1;1;1;1\rangle$	10	1	11	no
MF	$\langle 1;1;3;1\rangle$	16	1	18	no
MP	$\langle -1;1;1;1\rangle$	9	1	9	yes
MP	$\langle 0;1;1;1\rangle$	11	1	11	yes
MP	$\langle 0;1;2;1\rangle$	13	1	13	yes
MP	$\langle 0;2;3;1\rangle$	18	1	18	yes
MP	$\langle 1;1;1;1\rangle$	13	1	13	yes
MP	$\langle 1;1;3;1\rangle$	17	1	17	yes
PML	$\langle -1;1;1;1\rangle$	2	1	2	yes
PML	$\langle 0;1;1;1\rangle$	11	1	11	yes
PML	$\langle 0;1;2;1\rangle$	19	1	19	yes
PML	$\langle 0;2;3;1\rangle$	27	1	27	yes
PML	$\langle 1;1;1;1\rangle$	18	1	18	yes
PML	$\langle 1;1;3;1\rangle$	36	1	36	yes

Table B.2: Tables showing the results of the re-rootings for the last 12 biological datasets. Only time-feasible reconciliations are retained. Each line shows a summary of the time-feasible reconciliations obtained with all possible rootings for one dataset analysed with one cost vector. Column A indicates the cost vector, column B shows the smallest optimal cost obtained among all possible reconciliations, column C shows the number of plateaux, column D shows the optimal reconciliation cost for the original root (with no value when there is no time-feasible solution), column E indicates whether the root belongs to a plateau.

Dataset	*A*	*B*	*C*	*D*	*E*
PMP	$\langle -1;1;1;1\rangle$	-3	1	-3	yes
PMP	$\langle 0;1;1;1\rangle$	8	1	8	yes
PMP	$\langle 0;1;2;1\rangle$	14	1	14	yes
PMP	$\langle 0;2;3;1\rangle$	20	1	20	yes
PMP	$\langle 1;1;1;1\rangle$	18	1	18	yes
PMP	$\langle 1;1;3;1\rangle$	31	1	31	yes
PP	$\langle -1;1;1;1\rangle$	1	1	1	yes
PP	$\langle 0;1;1;1\rangle$	25	1	25	yes
PP	$\langle 0;1;2;1\rangle$	37	1	37	yes
PP	$\langle 0;2;3;1\rangle$	53	1	53	yes
PP	$\langle 1;1;1;1\rangle$	??	?	??	??
PP	$\langle 1;1;3;1\rangle$	72	1	72	yes
RH	$\langle -1;1;1;1\rangle$	10	1	-	no
RH	$\langle 0;1;1;1\rangle$	27	1	28	no
RH	$\langle 0;1;2;1\rangle$	45	1	-	no
RH	$\langle 0;2;3;1\rangle$	66	1	-	no
RH	$\langle 1;1;1;1\rangle$	??	?	??	??
RH	$\langle 1;1;3;1\rangle$	77	1	-	no
RM	$\langle -1;1;1;1\rangle$	-1	1	-1	yes
RM	$\langle 0;1;1;1\rangle$	2	1	2	yes
RM	$\langle 0;1;2;1\rangle$	3	1	3	yes
RM	$\langle 0;2;3;1\rangle$	4	1	4	yes
RM	$\langle 1;1;1;1\rangle$	5	1	5	yes
RM	$\langle 1;1;3;1\rangle$	7	1	7	yes
RP	$\langle -1;1;1;1\rangle$	0	1	0	yes
RP	$\langle 0;1;1;1\rangle$	8	1	8	yes
RP	$\langle 0;1;2;1\rangle$	12	1	12	yes
RP	$\langle 0;2;3;1\rangle$	16	1	16	yes
RP	$\langle 1;1;1;1\rangle$	13	1	13	yes
RP	$\langle 1;1;3;1\rangle$	24	1	24	yes
SBL	$\langle -1;1;1;1\rangle$	4	1	4	yes
SBL	$\langle 0;1;1;1\rangle$	7	1	7	yes
SBL	$\langle 0;1;2;1\rangle$	11	1	11	yes
SBL	$\langle 0;2;3;1\rangle$	14	1	14	yes
SBL	$\langle 1;1;1;1\rangle$	9	1	9	yes
SBL	$\langle 1;1;3;1\rangle$	18	1	18	yes

Dataset	*A*	*B*	*C*	*D*	*E*
SC	$\langle -1;1;1;1\rangle$	-3	1	-3	yes
SC	$\langle 0;1;1;1\rangle$	6	1	6	yes
SC	$\langle 0;1;2;1\rangle$	10	1	10	yes
SC	$\langle 0;2;3;1\rangle$	14	1	14	yes
SC	$\langle 1;1;1;1\rangle$	14	1	14	yes
SC	$\langle 1;1;3;1\rangle$	23	1	23	yes
SFC	$\langle -1;1;1;1\rangle$	4	1	12	no
SFC	$\langle 0;1;1;1\rangle$	11	1	17	no
SFC	$\langle 0;1;2;1\rangle$	19	1	27	no
SFC	$\langle 0;2;3;1\rangle$	26	1	37	no
SFC	$\langle 1;1;1;1\rangle$	17	1	21	no
SFC	$\langle 1;1;3;1\rangle$	34	1	42	no
SHA	$\langle -1;1;1;1\rangle$	-5	1	-2	no
SHA	$\langle 0;1;1;1\rangle$	6	1	8	no
SHA	$\langle 0;1;2;1\rangle$	9	1	12	no
SHA	$\langle 0;2;3;1\rangle$	13	1	17	no
SHA	$\langle 1;1;1;1\rangle$	17	1	18	no
SHA	$\langle 1;1;3;1\rangle$	22	1	25	no
SSA	$\langle -1;1;1;1\rangle$	-6	1	-6	yes
SSA	$\langle 0;1;1;1\rangle$	2	1	2	yes
SSA	$\langle 0;1;2;1\rangle$	3	1	3	yes
SSA	$\langle 0;2;3;1\rangle$	4	1	4	yes
SSA	$\langle 1;1;1;1\rangle$	10	1	10	yes
SSA	$\langle 1;1;3;1\rangle$	12	1	12	yes
TC	$\langle -1;1;1;1\rangle$	5	3	7	no
TC	$\langle 0;1;1;1\rangle$	8	1	9	no
TC	$\langle 0;1;2;1\rangle$	12	1	15	no
TC	$\langle 0;2;3;1\rangle$	17	1	21	no
TC	$\langle 1;1;1;1\rangle$	10	1	11	no
TC	$\langle 1;1;3;1\rangle$	19	2	21	no
TD	$\langle -1;1;1;1\rangle$	1	1	4	no
TD	$\langle 0;1;1;1\rangle$	13	1	15	no
TD	$\langle 0;1;2;1\rangle$	22	1	25	no
TD	$\langle 0;2;3;1\rangle$	30	1	34	no
TD	$\langle 1;1;1;1\rangle$	22	1	23	no
TD	$\langle 1;1;3;1\rangle$	43	1	46	no

Table B.3: Results of the re-rooting for the simulated datasets. Only time-feasible reconciliations are retained. The columns N_i represent the number of datasets that have exactly i plateaux (for $1 \leq i \leq 5$).

Dataset	*CostVector*	$N1$	$N2$	$N3$	$N4$	$N5$
AP-sim	$\langle -1, 1, 1, 1\rangle$	37	11	0	2	0
AP-sim	$\langle 0, 1, 1, 1\rangle$	35	12	1	1	0
AP-sim	$\langle 0, 1, 2, 1\rangle$	40	7	1	1	0
AP-sim	$\langle 0, 2, 3, 1\rangle$	41	7	0	1	0
AP-sim	$\langle 1, 1, 3, 1\rangle$	45	4	0	0	0
AW-sim	$\langle -1, 1, 1, 1\rangle$	41	8	1	0	0
AW-sim	$\langle 0, 1, 1, 1\rangle$	34	14	1	0	0
AW-sim	$\langle 0, 1, 2, 1\rangle$	38	10	1	0	0
AW-sim	$\langle 0, 2, 3, 1\rangle$	39	8	2	0	0
AW-sim	$\langle 1, 1, 3, 1\rangle$	35	12	1	0	1
CA-sim	$\langle -1, 1, 1, 1\rangle$	50	0	0	2	0
CA-sim	$\langle 0, 1, 1, 1\rangle$	46	2	1	0	0
CA-sim	$\langle 0, 1, 2, 1\rangle$	49	0	0	0	0
CA-sim	$\langle 0, 2, 3, 1\rangle$	47	3	0	0	0
CA-sim	$\langle 1, 1, 3, 1\rangle$	47	2	1	0	0
CP-sim	$\langle -1, 1, 1, 1\rangle$	43	6	0	1	0
CP-sim	$\langle 0, 1, 1, 1\rangle$	34	9	5	1	1
CP-sim	$\langle 0, 1, 2, 1\rangle$	44	5	1	0	0
CP-sim	$\langle 0, 2, 3, 1\rangle$	46	2	2	0	0
CP-sim	$\langle 1, 1, 3, 1\rangle$	42	6	2	0	0
CT-sim	$\langle -1, 1, 1, 1\rangle$	49	0	1	0	0
CT-sim	$\langle 0, 1, 1, 1\rangle$	47	1	2	0	0
CT-sim	$\langle 0, 1, 2, 1\rangle$	50	0	0	0	0
CT-sim	$\langle 0, 2, 3, 1\rangle$	49	1	0	0	0
CT-sim	$\langle 1, 1, 3, 1\rangle$	49	1	0	0	0
EC-sim	$\langle -1, 1, 1, 1\rangle$	45	3	1	0	0
EC-sim	$\langle 0, 1, 1, 1\rangle$	41	6	1	0	0
EC-sim	$\langle 0, 1, 2, 1\rangle$	49	1	0	0	0
EC-sim	$\langle 0, 2, 3, 1\rangle$	47	2	0	0	0
EC-sim	$\langle 1, 1, 3, 1\rangle$	48	2	0	0	0
FA-sim	$\langle -1, 1, 1, 1\rangle$	47	3	0	0	0
FA-sim	$\langle 0, 1, 1, 1\rangle$	47	3	0	0	0
FA-sim	$\langle 0, 1, 2, 1\rangle$	47	3	0	0	0
FA-sim	$\langle 0, 2, 3, 1\rangle$	47	3	0	0	0
FA-sim	$\langle 1, 1, 3, 1\rangle$	50	0	0	0	0

Dataset	*CostVector*	$N1$	$N2$	$N3$	$N4$	$N5$
FE-sim	$\langle -1, 1, 1, 1\rangle$	33	8	3	4	1
FE-sim	$\langle 0, 1, 1, 1\rangle$	34	8	3	3	1
FE-sim	$\langle 0, 1, 2, 1\rangle$	41	8	1	0	0
FE-sim	$\langle 0, 2, 3, 1\rangle$	38	7	2	3	0
FE-sim	$\langle 1, 1, 3, 1\rangle$	44	6	0	0	0
GL-sim	$\langle -1, 1, 1, 1\rangle$	49	1	0	0	0
GL-sim	$\langle 0, 1, 1, 1\rangle$	48	2	0	0	0
GL-sim	$\langle 0, 1, 2, 1\rangle$	48	1	0	0	1
GL-sim	$\langle 0, 2, 3, 1\rangle$	49	0	1	0	0
GL-sim	$\langle 1, 1, 3, 1\rangle$	48	1	1	0	0
GM-sim	$\langle -1, 1, 1, 1\rangle$	47	1	1	0	0
GM-sim	$\langle 0, 1, 1, 1\rangle$	45	4	1	0	0
GM-sim	$\langle 0, 1, 2, 1\rangle$	47	1	2	0	0
GM-sim	$\langle 0, 2, 3, 1\rangle$	48	1	0	1	0
GM-sim	$\langle 1, 1, 3, 1\rangle$	47	3	0	0	0
IFL-sim	$\langle -1, 1, 1, 1\rangle$	46	4	0	0	0
IFL-sim	$\langle 0, 1, 1, 1\rangle$	44	6	0	0	0
IFL-sim	$\langle 0, 1, 2, 1\rangle$	47	3	0	0	0
IFL-sim	$\langle 0, 2, 3, 1\rangle$	47	3	0	0	0
IFL-sim	$\langle 1, 1, 3, 1\rangle$	48	2	0	0	0
MF-sim	$\langle -1, 1, 1, 1\rangle$	42	8	0	0	0
MF-sim	$\langle 0, 1, 1, 1\rangle$	41	8	1	0	0
MF-sim	$\langle 0, 1, 2, 1\rangle$	43	6	1	0	0
MF-sim	$\langle 0, 2, 3, 1\rangle$	45	5	0	0	0
MF-sim	$\langle 1, 1, 3, 1\rangle$	43	6	1	0	0
MP-sim	$\langle -1, 1, 1, 1\rangle$	44	3	1	1	1
MP-sim	$\langle 0, 1, 1, 1\rangle$	46	1	2	1	0
MP-sim	$\langle 0, 1, 2, 1\rangle$	46	3	1	0	0
MP-sim	$\langle 0, 2, 3, 1\rangle$	45	2	2	1	0
MP-sim	$\langle 1, 1, 3, 1\rangle$	47	2	1	0	0
PML-sim	$\langle -1, 1, 1, 1\rangle$	50	0	0	0	0
PML-sim	$\langle 0, 1, 1, 1\rangle$	50	0	0	0	0
PML-sim	$\langle 0, 1, 2, 1\rangle$	48	2	0	0	0
PML-sim	$\langle 0, 2, 3, 1\rangle$	49	1	0	0	0
PML-sim	$\langle 1, 1, 3, 1\rangle$	45	4	0	0	0

Table B.4: Results of the re-rooting for the simulated datasets. Only time-feasible reconciliations are retained. The columns N_i represent the number of datasets that have exactly i plateaux (for $1 \leq i \leq 5$).

Dataset	*CostVector*	$N1$	$N2$	$N3$	$N4$	$N5$
PMP-sim	$\langle -1, 1, 1, 1\rangle$	48	2	0	0	0
PMP-sim	$\langle 0, 1, 1, 1\rangle$	49	1	0	0	0
PMP-sim	$\langle 0, 1, 2, 1\rangle$	47	3	0	0	0
PMP-sim	$\langle 0, 2, 3, 1\rangle$	47	3	0	0	0
PMP-sim	$\langle 1, 1, 3, 1\rangle$	47	2	0	0	0
PP-sim	$\langle -1, 1, 1, 1\rangle$	49	1	0	0	0
PP-sim	$\langle 0, 1, 1, 1\rangle$	47	3	0	0	0
PP-sim	$\langle 0, 1, 2, 1\rangle$	49	1	0	0	0
PP-sim	$\langle 0, 2, 3, 1\rangle$	49	1	0	0	0
PP-sim	$\langle 1, 1, 3, 1\rangle$	50	0	0	0	0
RM-sim	$\langle -1, 1, 1, 1\rangle$	48	1	0	0	1
RM-sim	$\langle 0, 1, 1, 1\rangle$	47	1	1	0	1
RM-sim	$\langle 0, 1, 2, 1\rangle$	46	4	0	0	0
RM-sim	$\langle 0, 2, 3, 1\rangle$	48	2	0	0	0
RM-sim	$\langle 1, 1, 3, 1\rangle$	47	3	0	0	0
RP-sim	$\langle -1, 1, 1, 1\rangle$	47	3	0	0	0
RP-sim	$\langle 0, 1, 1, 1\rangle$	47	3	0	0	0
RP-sim	$\langle 0, 1, 2, 1\rangle$	47	3	0	0	0
RP-sim	$\langle 0, 2, 3, 1\rangle$	47	3	0	0	0
RP-sim	$\langle 1, 1, 3, 1\rangle$	45	5	0	0	0
RH-sim	$\langle -1, 1, 1, 1\rangle$	45	3	1	1	0
RH-sim	$\langle 0, 1, 1, 1\rangle$	44	5	0	1	0
RH-sim	$\langle 0, 1, 2, 1\rangle$	46	3	0	1	0
RH-sim	$\langle 0, 2, 3, 1\rangle$	46	4	0	0	0
RH-sim	$\langle 1, 1, 3, 1\rangle$	46	2	1	1	0
SBL-sim	$\langle -1, 1, 1, 1\rangle$	45	4	2	0	0
SBL-sim	$\langle 0, 1, 1, 1\rangle$	44	4	1	0	0
SBL-sim	$\langle 0, 1, 2, 1\rangle$	47	1	1	0	0
SBL-sim	$\langle 0, 2, 3, 1\rangle$	47	2	0	0	0
SBL-sim	$\langle 1, 1, 3, 1\rangle$	44	5	0	0	0
SC-sim	$\langle -1, 1, 1, 1\rangle$	45	4	1	0	0
SC-sim	$\langle 0, 1, 1, 1\rangle$	45	4	0	1	0
SC-sim	$\langle 0, 1, 2, 1\rangle$	46	4	0	0	0
SC-sim	$\langle 0, 2, 3, 1\rangle$	46	4	0	0	0
SC-sim	$\langle 1, 1, 3, 1\rangle$	48	2	0	0	0

Dataset	*CostVector*	$N1$	$N2$	$N3$	$N4$	$N5$
SFC-sim	$\langle -1, 1, 1, 1\rangle$	40	7	2	0	1
SFC-sim	$\langle 0, 1, 1, 1\rangle$	40	8	2	0	0
SFC-sim	$\langle 0, 1, 2, 1\rangle$	41	6	3	0	0
SFC-sim	$\langle 0, 2, 3, 1\rangle$	43	7	0	0	0
SFC-sim	$\langle 1, 1, 3, 1\rangle$	47	3	0	0	0
SHA-sim	$\langle -1, 1, 1, 1\rangle$	46	4	0	0	0
SHA-sim	$\langle 0, 1, 1, 1\rangle$	48	2	0	0	0
SHA-sim	$\langle 0, 1, 2, 1\rangle$	44	4	2	0	0
SHA-sim	$\langle 0, 2, 3, 1\rangle$	46	3	1	0	0
SHA-sim	$\langle 1, 1, 3, 1\rangle$	47	1	2	0	0
SSA-sim	$\langle -1, 1, 1, 1\rangle$	49	0	1	0	0
SSA-sim	$\langle 0, 1, 1, 1\rangle$	49	0	1	0	0
SSA-sim	$\langle 0, 1, 2, 1\rangle$	48	1	1	0	0
SSA-sim	$\langle 0, 2, 3, 1\rangle$	48	1	1	0	0
SSA-sim	$\langle 1, 1, 3, 1\rangle$	49	0	1	0	0
TC-sim	$\langle -1, 1, 1, 1\rangle$	47	2	1	0	0
TC-sim	$\langle 0, 1, 1, 1\rangle$	48	1	1	0	0
TC-sim	$\langle 0, 1, 2, 1\rangle$	47	2	1	0	0
TC-sim	$\langle 0, 2, 3, 1\rangle$	48	2	0	0	0
TC-sim	$\langle 1, 1, 3, 1\rangle$	44	4	0	0	0
TD-sim	$\langle -1, 1, 1, 1\rangle$	48	1	1	0	0
TD-sim	$\langle 0, 1, 1, 1\rangle$	48	1	1	0	0
TD-sim	$\langle 0, 1, 2, 1\rangle$	46	3	0	0	1
TD-sim	$\langle 0, 2, 3, 1\rangle$	48	1	1	0	0
TD-sim	$\langle 1, 1, 3, 1\rangle$	48	2	0	0	0

B.2.5 Re-rooting at distance k, biological datasets

Figures B.158 to B.182 present the analysis on the re-rooting at distance k for the remaining 27 biological datasets (dataset MP appearing in the paper). Only time-feasible reconciliations are retained. Figures are missing whenever the original positioning of the root did not produce any time-feasible reconciliation (which corresponds also to no value in column D of Tables B.1 and B.2). This is the case for datasets CP (all cost vectors except $\langle 1,1,1,1\rangle$ and $\langle 1,1,3,1\rangle$), CT (all cost vectors except $\langle 1,1,1,1\rangle$), FD (with cost vector $\langle 0,2,3,1\rangle$), PP (with cost vector $\langle 1,1,1,1\rangle$) and RH (all cost vectors except $\langle 0,1,1,1\rangle$). Notice also that for 2 of the 28 datasets, namely FD and PP, we did not obtain any results for the cost vector $\langle 1,1,1,1\rangle$. Indeed, this cost vector results in the longest computation time and those datasets contain large trees.

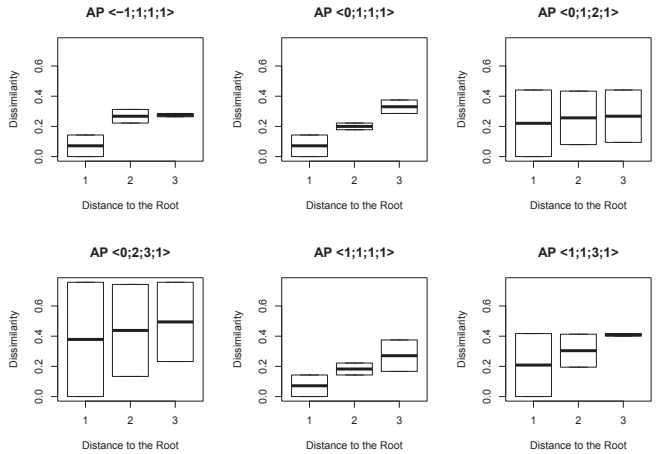


Figure B.156: Boxplots of the dissimilarities between reconciliations obtained for the original dataset AP and all datasets derived from AP by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

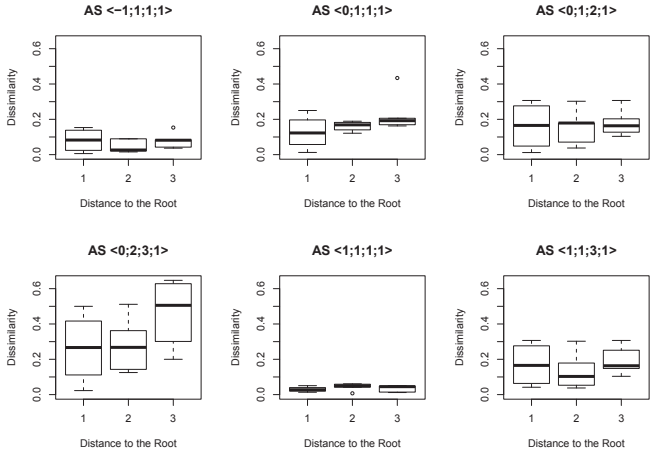


Figure B.157: Boxplots of the dissimilarities between reconciliations obtained for the original dataset AS and all datasets derived from AS by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

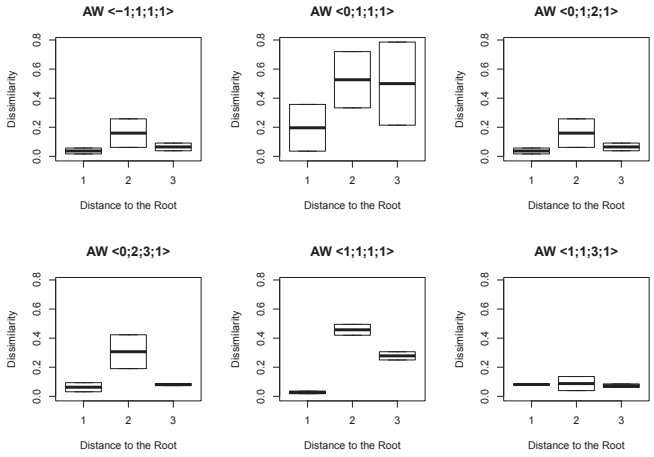


Figure B.158: Boxplots of the dissimilarities between reconciliations obtained for the original dataset AW and all datasets derived from AW by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

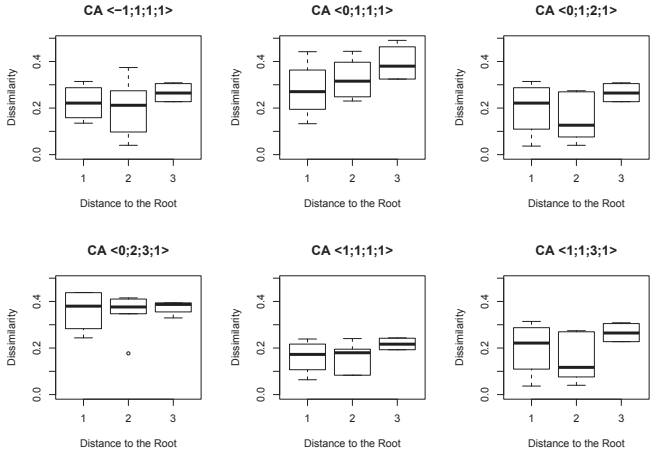


Figure B.159: Boxplots of the dissimilarities between reconciliations obtained for the original dataset CA and all datasets derived from CA by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

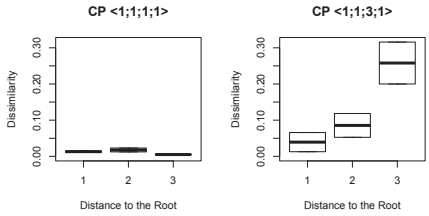


Figure B.160: Boxplots of the dissimilarities between reconciliations obtained for the original dataset CP and all datasets derived from CP by re-rooting the symbiont tree at distance k from the original root. The two plots correspond to the cost vectors $\langle 1;1;1;1\rangle$ and $\langle 1;1;1;3\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

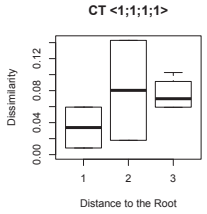


Figure B.161: Boxplots of the dissimilarities between reconciliations obtained for the original dataset CT and all datasets derived from CT by re-rooting the symbiont tree at distance k from the original root. The plot correspond to the cost vector $\langle 1; 1; 1; 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

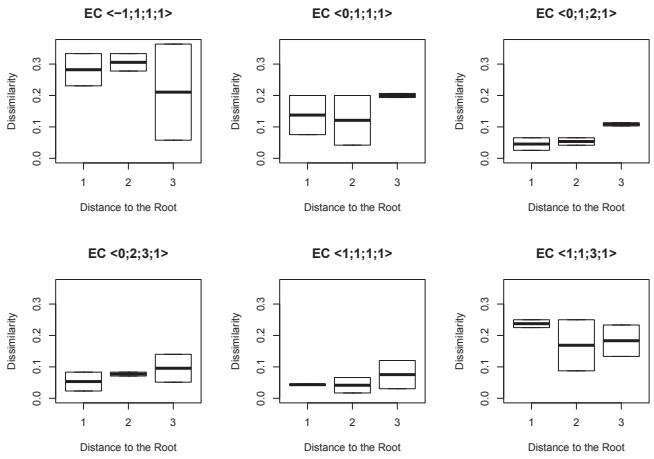


Figure B.162: Boxplots of the dissimilarities between reconciliations obtained for the original dataset EC and all datasets derived from EC by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

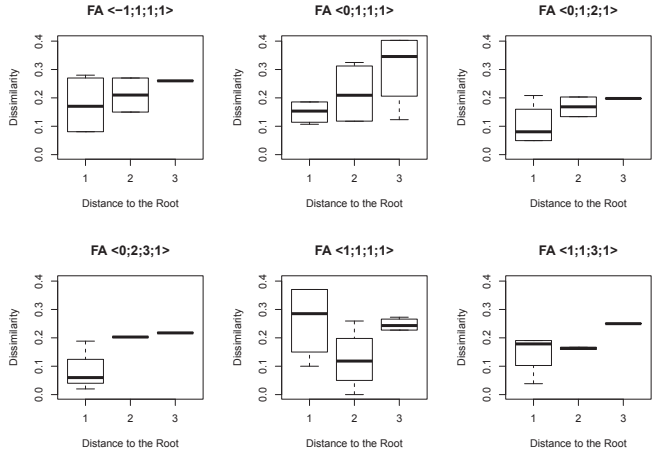


Figure B.163: Boxplots of the dissimilarities between reconciliations obtained for the original dataset FA and all datasets derived from FA by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

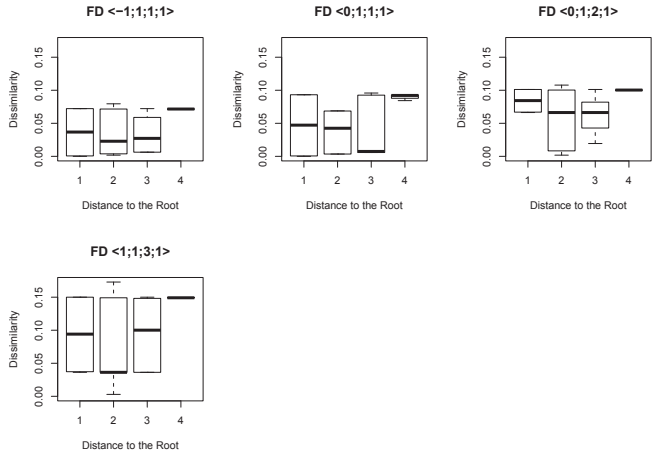


Figure B.164: Boxplots of the dissimilarities between reconciliations obtained for the original dataset FD and all datasets derived from FD by re-rooting the symbiont tree at distance k from the original root. The four plots correspond to the 4 cost vectors in $\langle -1;1;1;1\rangle$, $\langle 0;1;1;1\rangle$, $\langle 0;1;2;1\rangle$ and $\langle 1;1;1;3\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

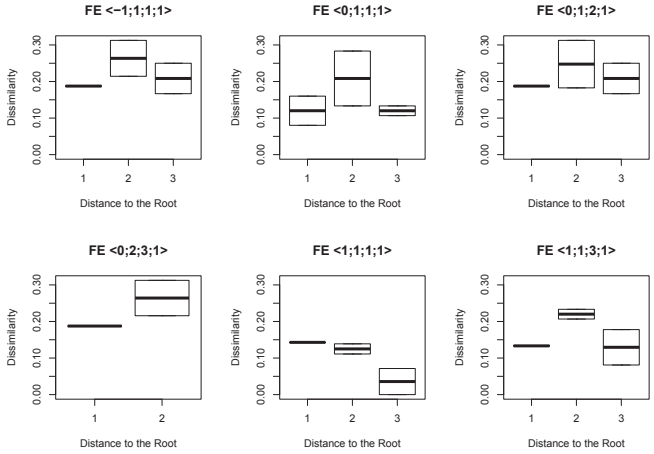


Figure B.165: Boxplots of the dissimilarities between reconciliations obtained for the original dataset FE and all datasets derived from FE by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

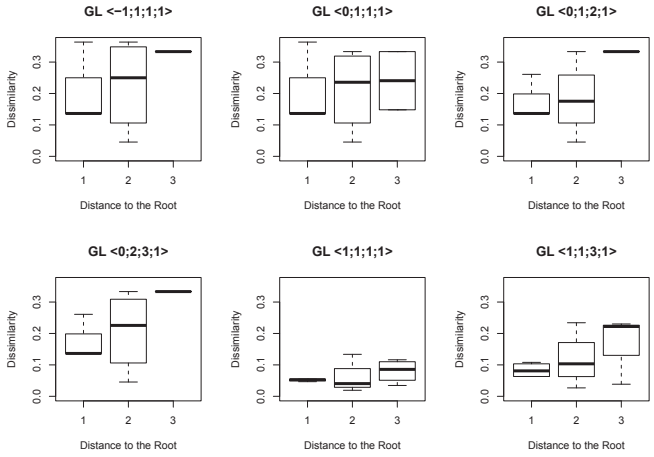


Figure B.166: Boxplots of the dissimilarities between reconciliations obtained for the original dataset GL and all datasets derived from GL by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

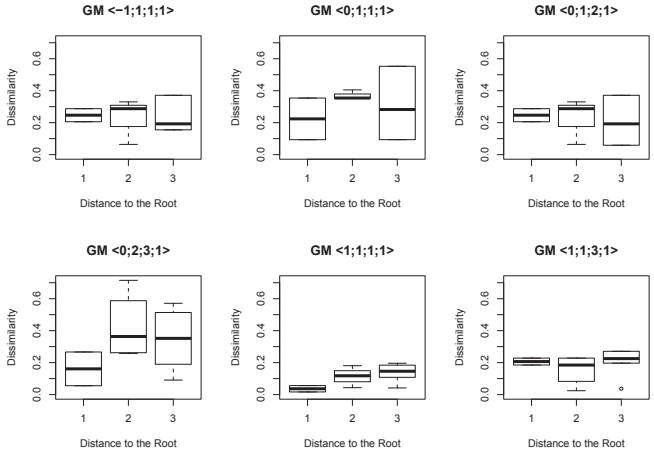


Figure B.167: Boxplots of the dissimilarities between reconciliations obtained for the original dataset GM and all datasets derived from GM by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

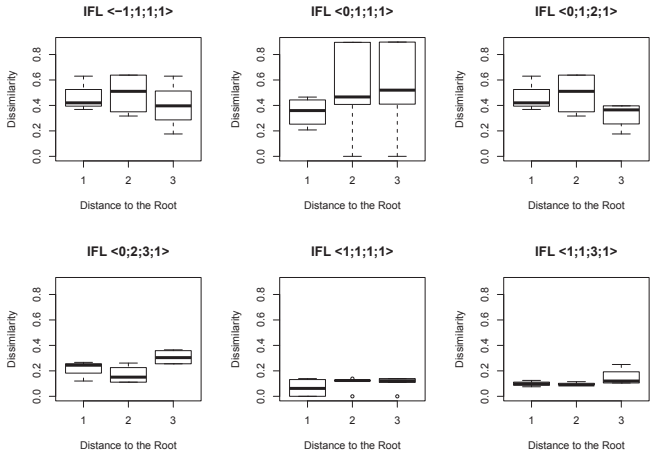


Figure B.168: Boxplots of the dissimilarities between reconciliations obtained for the original dataset IFL and all datasets derived from IFL by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

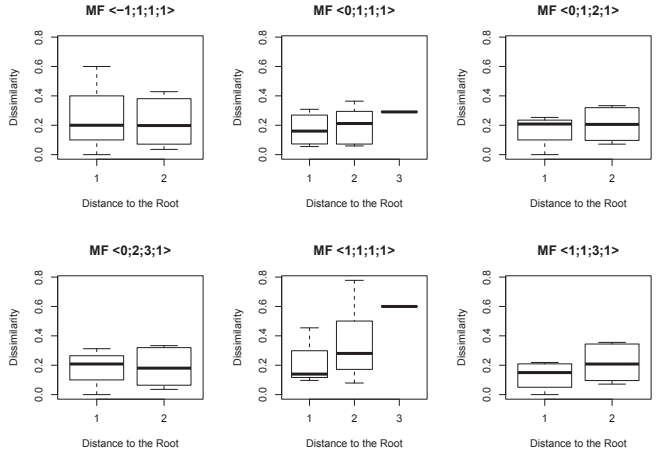


Figure B.169: Boxplots of the dissimilarities between reconciliations obtained for the original dataset MF and all datasets derived from MF by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

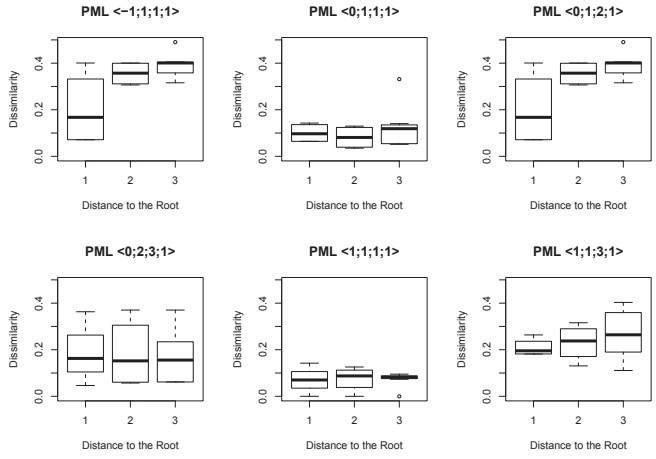


Figure B.170: Boxplots of the dissimilarities between reconciliations obtained for original dataset PML and all datasets derived from PML by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

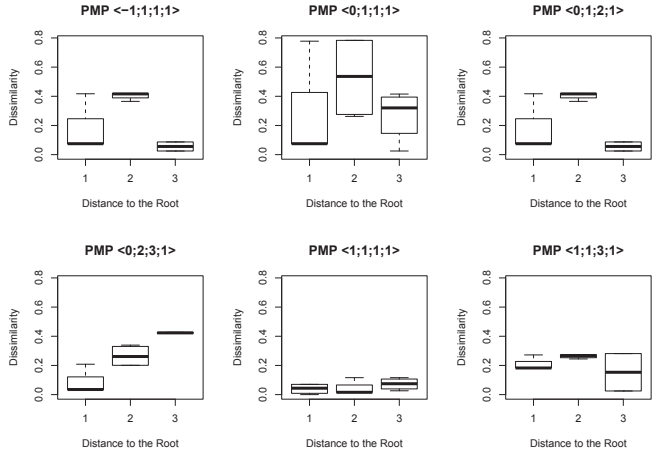


Figure B.171: Boxplots of the dissimilarities between reconciliations obtained for the original dataset PMP and all datasets derived from PMP by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

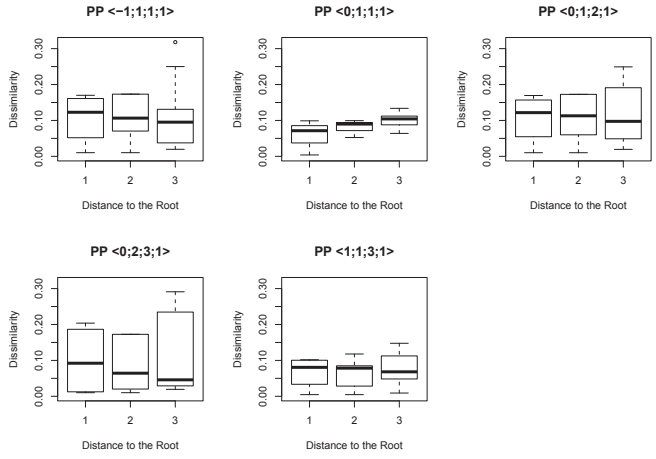


Figure B.172: Boxplots of the dissimilarities between reconciliations obtained for the original dataset PP and all datasets derived from PP by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the 5 cost vectors $\langle -1; 1; 1; 1\rangle$, $\langle 0; 1; 1; 1\rangle$, $\langle 0; 1; 2; 1\rangle$, $\langle 0; 2; 3; 1\rangle$ and $\langle 1; 1; 1; 3\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

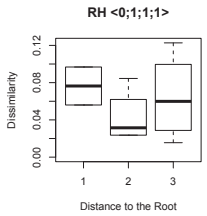


Figure B.173: Boxplots of the dissimilarities between reconciliations obtained for the original dataset RH and all datasets derived from RH by re-rooting the symbiont tree at distance k from the original root. The plot corresponds to the cost vector $\langle 0; 1; 1; 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

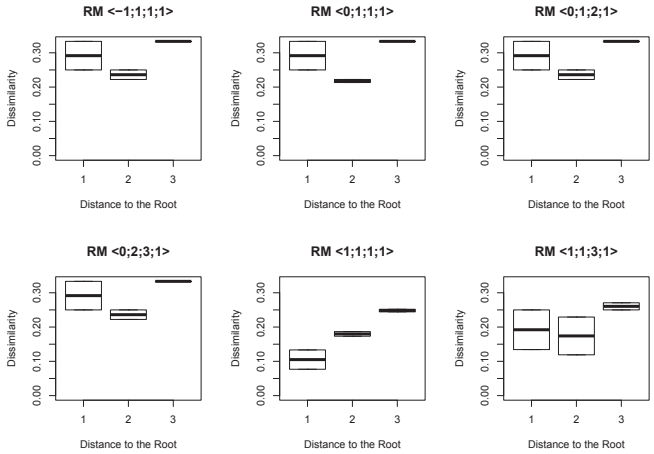


Figure B.174: Boxplots of the dissimilarities between reconciliations obtained for the original dataset RM and all datasets derived from RM by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

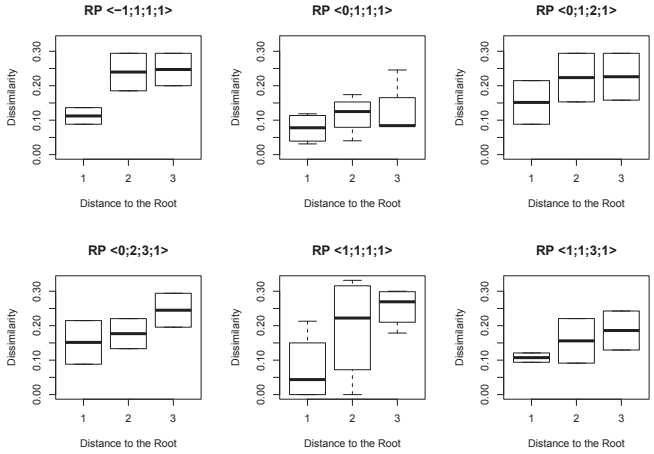


Figure B.175: Boxplots of the dissimilarities between reconciliations obtained for the original dataset RP and all datasets derived from RP by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

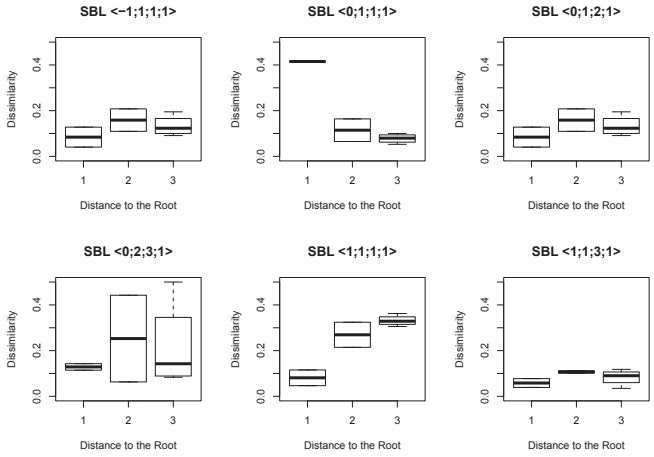


Figure B.176: Boxplots of the dissimilarities between reconciliations obtained for the original dataset SBL and all datasets derived from SBL by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

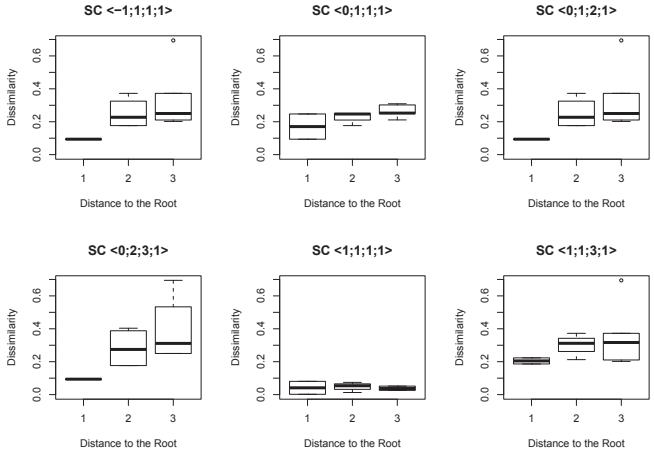


Figure B.177: Boxplots of the dissimilarities between reconciliations obtained for the original dataset SC and all datasets derived from SC by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

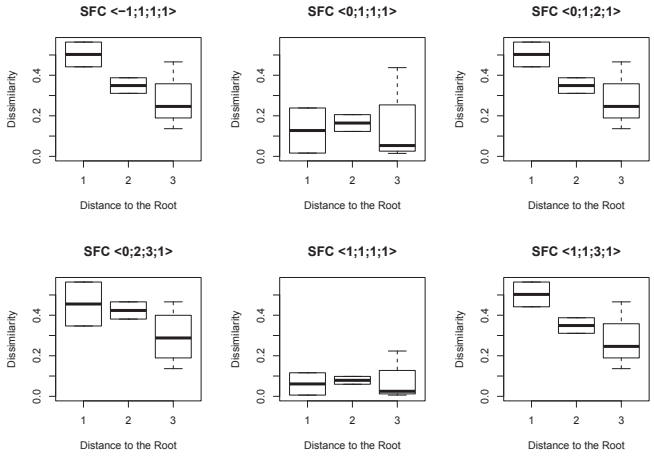


Figure B.178: Boxplots of the dissimilarities between reconciliations obtained for the original dataset SFC and all datasets derived from SFC by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

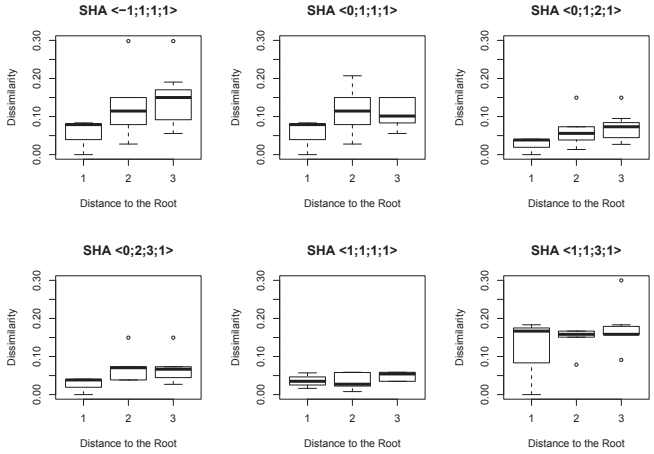


Figure B.179: Boxplots of the dissimilarities between reconciliations obtained for the original dataset SHA and all datasets derived from SHA by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

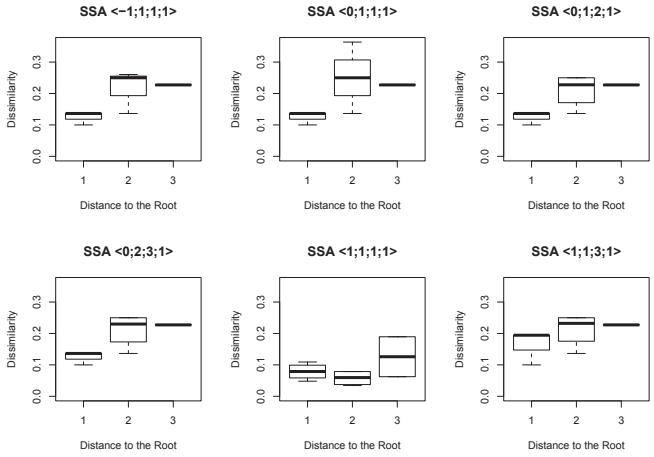


Figure B.180: Boxplots of the dissimilarities between reconciliations obtained for the original dataset SSA and all datasets derived from SSA by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

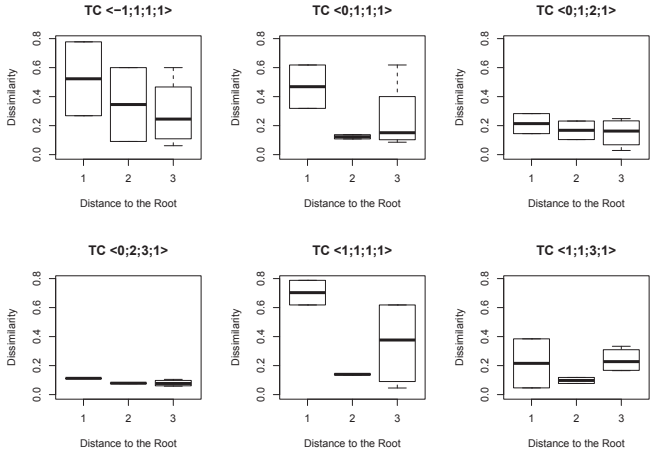


Figure B.181: Boxplots of the dissimilarities between reconciliations obtained for the original dataset TC and all datasets derived from TC by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

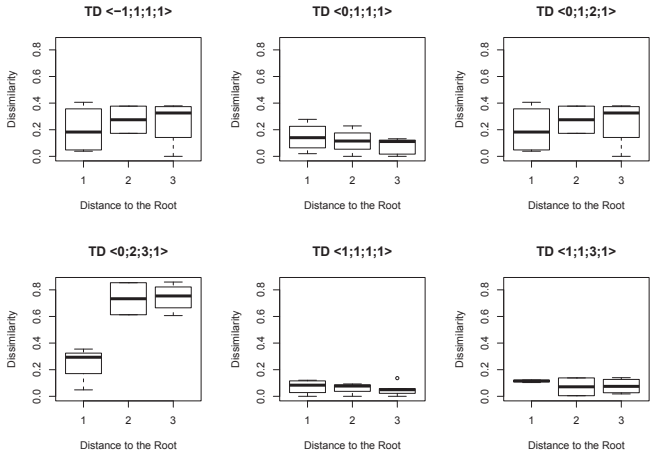


Figure B.182: Boxplots of the dissimilarities between reconciliations obtained for the original dataset TD and all datasets derived from TD by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the 6 cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

B.2.6 Re-rooting at distance k, simulated datasets

Figures B.185 to B.205 show the results for 14 sets of simulated datasets (each containing 50 datasets). Only time-feasible reconciliations are retained. Dataset FD contains the symbiont tree with the largest number of leaves and our procedure did not produce simulated datasets corresponding to this biological dataset. Moreover, notice that the cost vector $\langle 1, 1, 1, 1\rangle$ is always missing here as this cost vector induces the largest computation time.

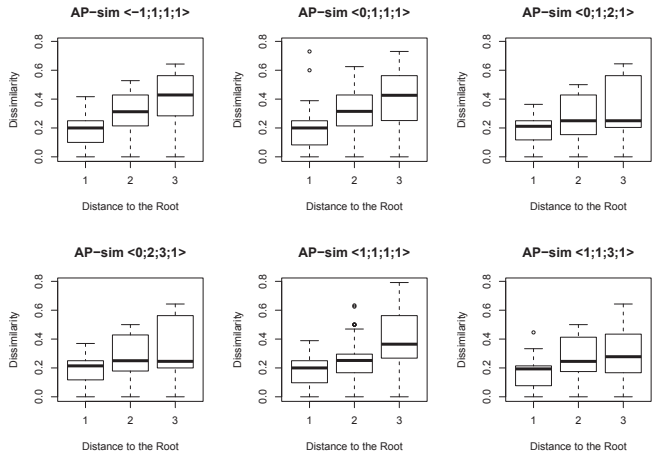


Figure B.183: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets AP-sim and all datasets derived from AP-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

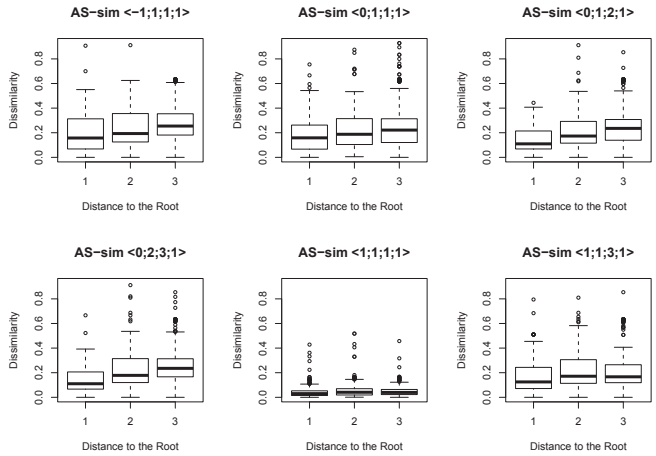


Figure B.184: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets AS-sim and all datasets derived from AS-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

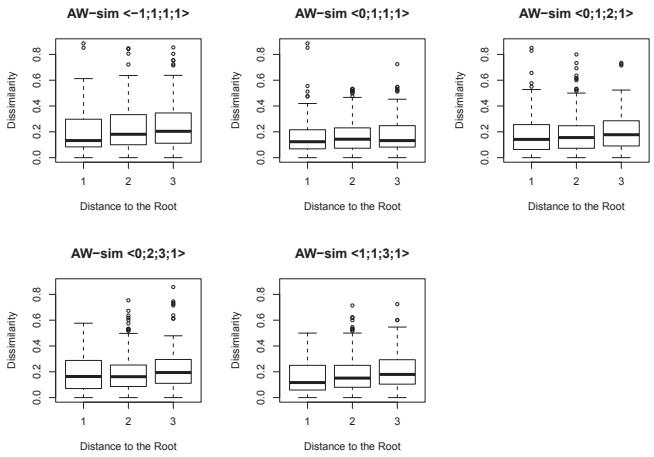


Figure B.185: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets AW-sim and all datasets derived from AW-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

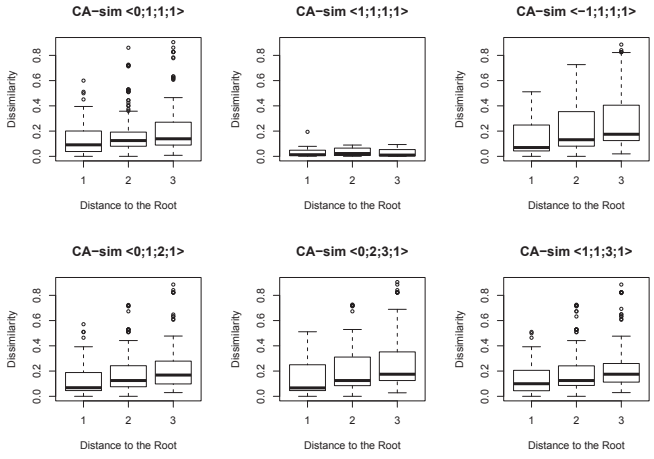


Figure B.186: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets CA-sim and all datasets derived from CA-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

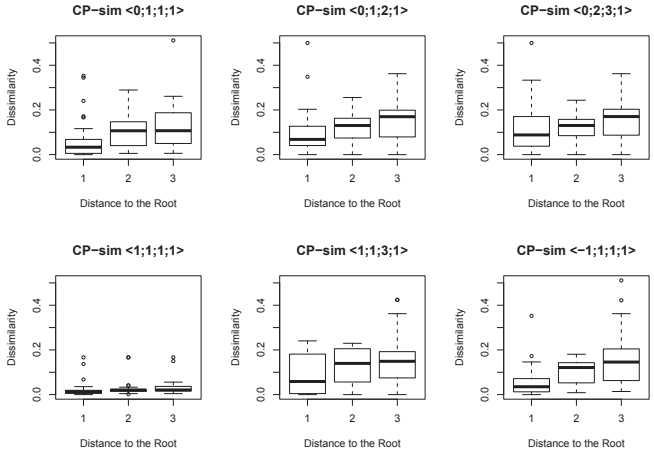


Figure B.187: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets CP-sim and all datasets derived from CP-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

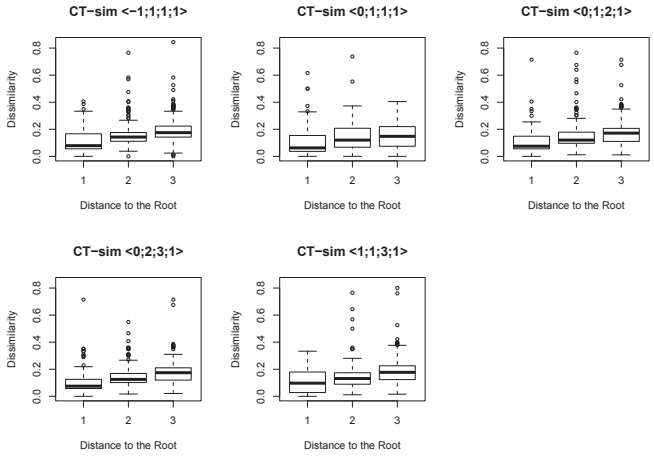


Figure B.188: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets CT-sim and all datasets derived from CT-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

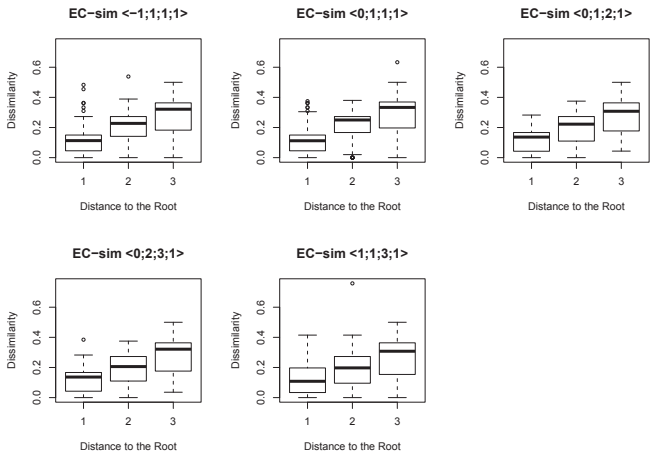


Figure B.189: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets EC-sim and all datasets derived from EC-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

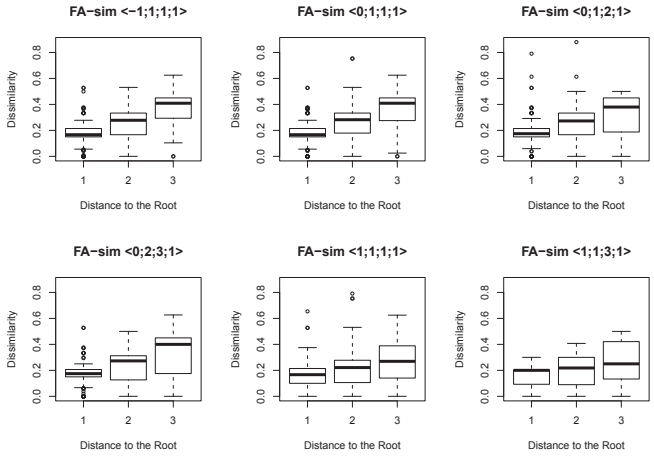


Figure B.190: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets FA-sim and all datasets derived from FA-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

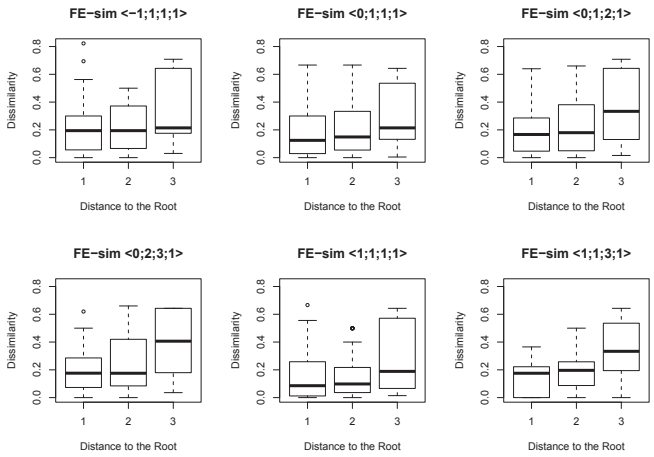


Figure B.191: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets FE-sim and all datasets derived from FE-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

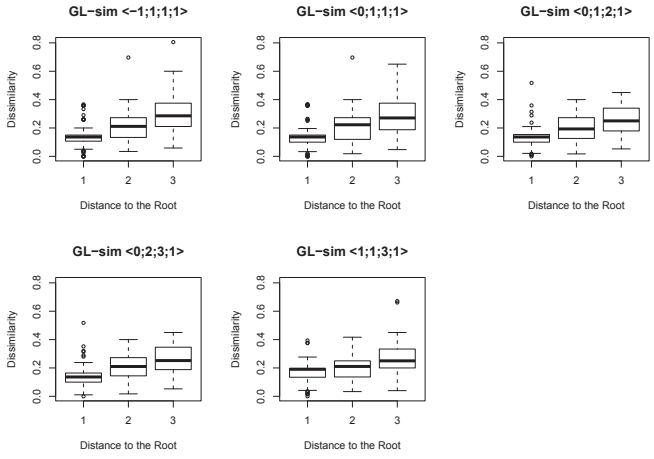


Figure B.192: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets GL-sim and all datasets derived from GL-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

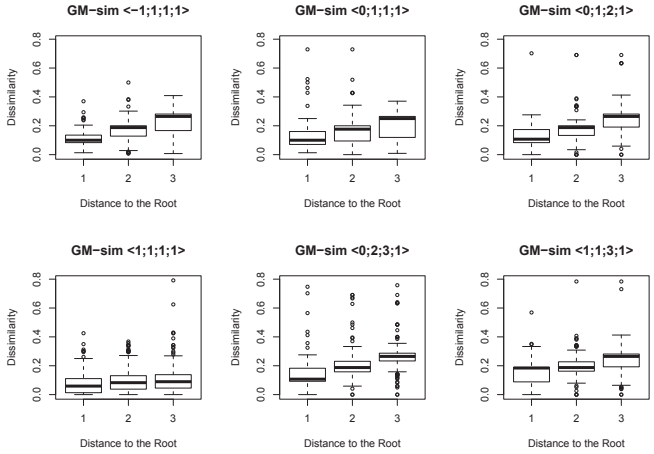


Figure B.193: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets GM-sim and all datasets derived from GM-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

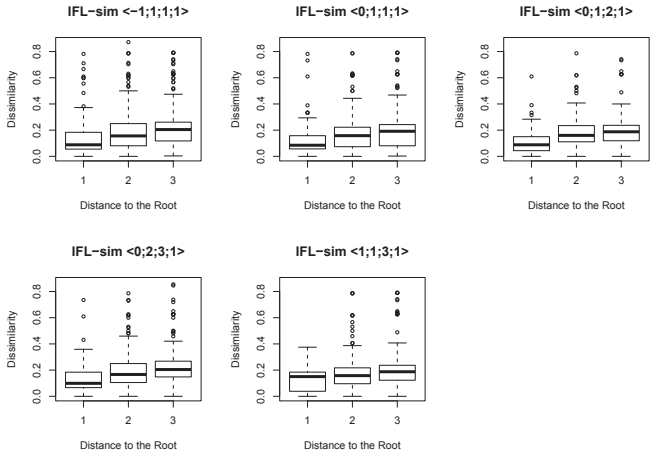


Figure B.194: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets IFL-sim and all datasets derived from IFL-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

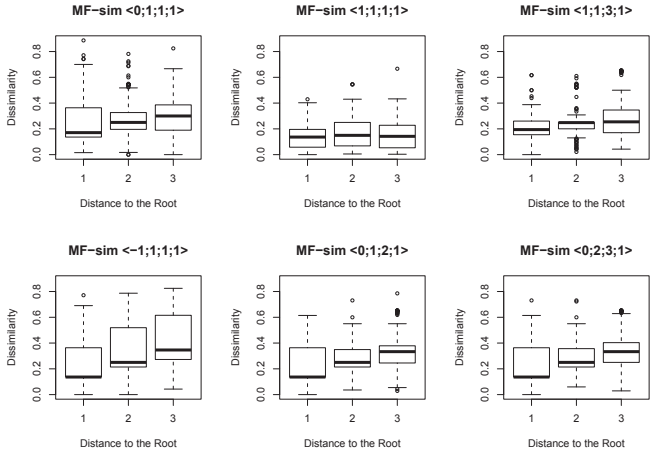


Figure B.195: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets MF-sim and all datasets derived from MF-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

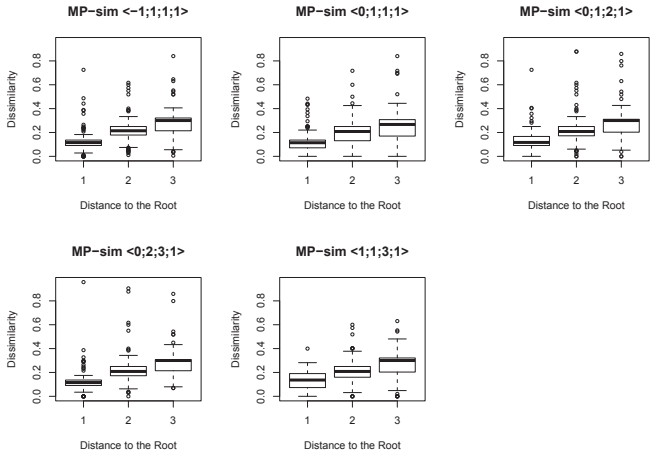


Figure B.196: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets MP-sim and all datasets derived from MP-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

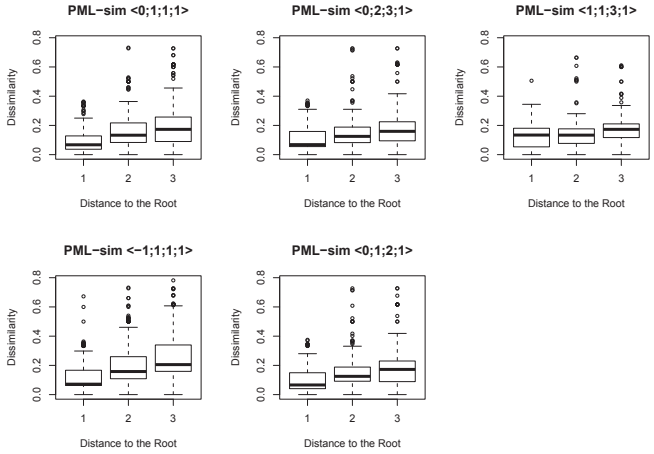


Figure B.197: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets PML-sim and all datasets derived from PML-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

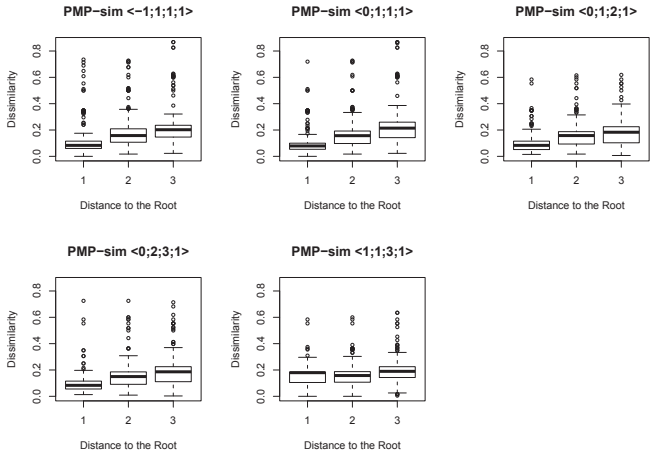


Figure B.198: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets PMP-sim and all datasets derived from PMP-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

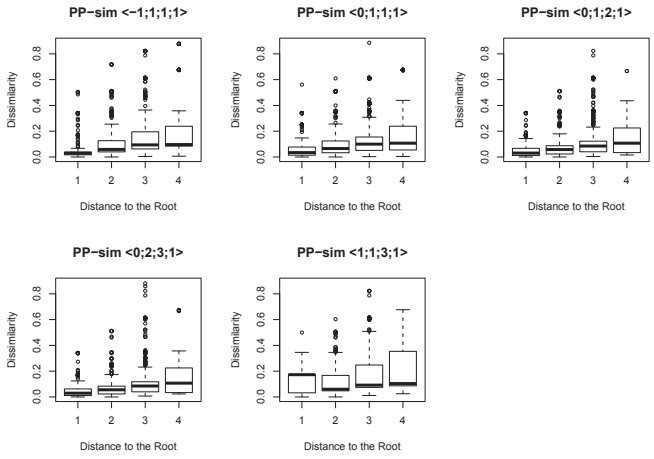


Figure B.199: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets PP-sim and all datasets derived from PP-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

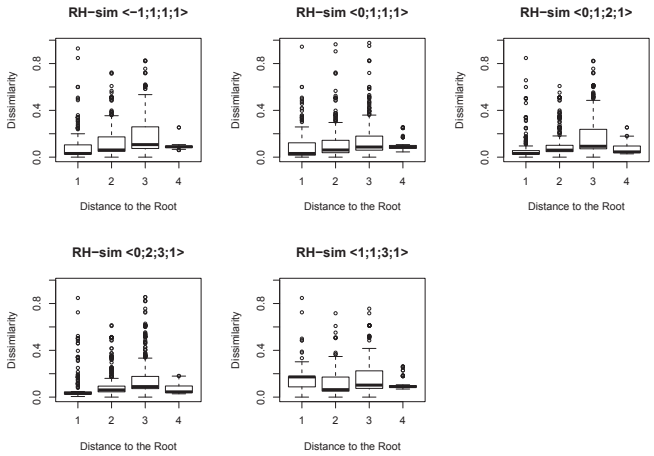


Figure B.200: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets RH-sim and all datasets derived from RH-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

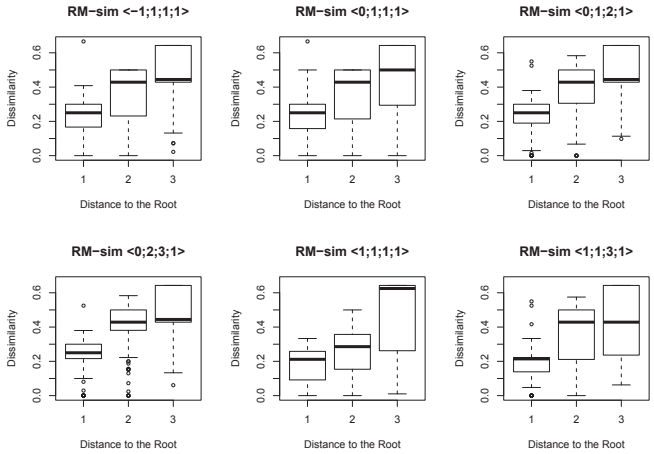


Figure B.201: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets RM-sim and all datasets derived from RM-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

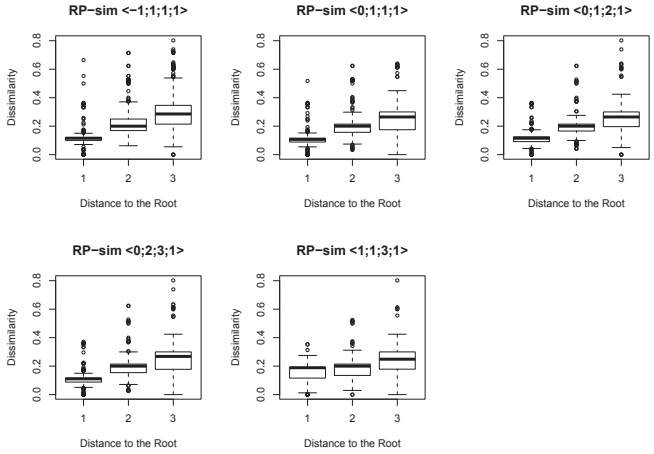


Figure B.202: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets RP-sim and all datasets derived from RP-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

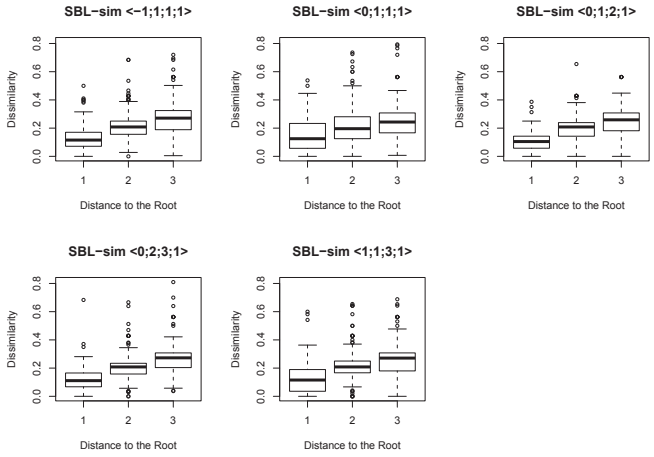


Figure B.203: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets SBL-sim and all datasets derived from SBL-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

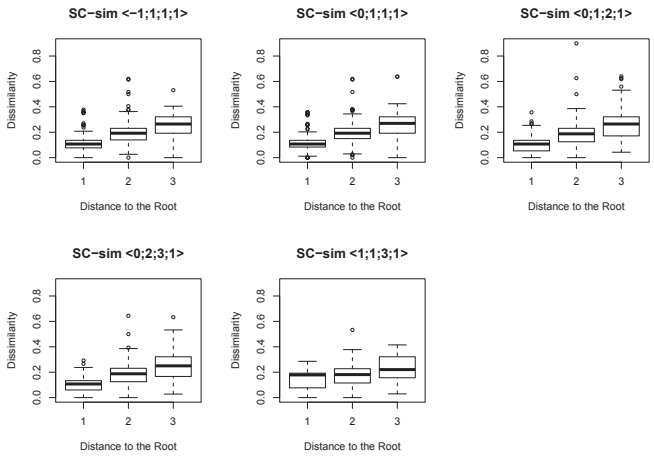


Figure B.204: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets SC-sim and all datasets derived from SC-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

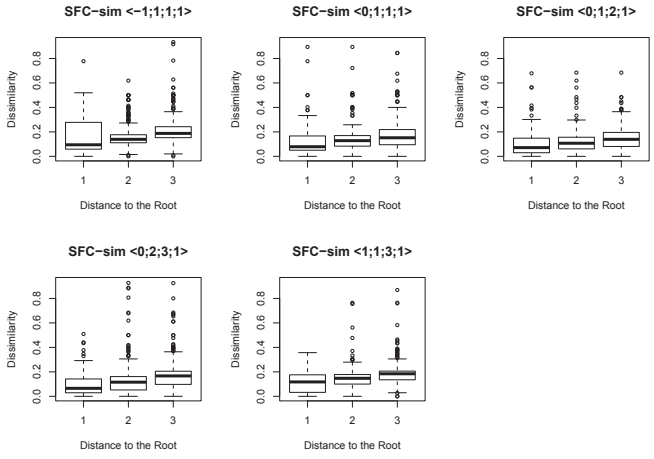


Figure B.205: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets SFC-sim and all datasets derived from SFC-sim by re-rooting the symbiont tree at distance k from the original root. The five plots correspond to the cost vectors $\langle -1, 1, 1, 1\rangle$, $\langle 0, 1, 1, 1\rangle$, $\langle 0, 1, 2, 1\rangle$, $\langle 0, 2, 3, 1\rangle$, $\langle 1, 1, 3, 1\rangle$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

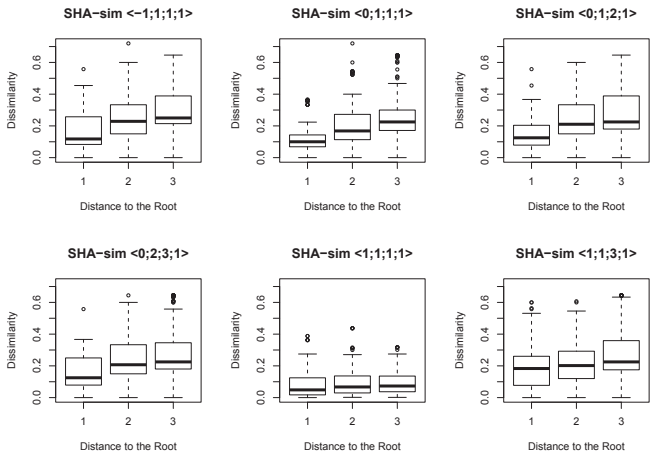


Figure B.206: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets SHA-sim and all datasets derived from SHA-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

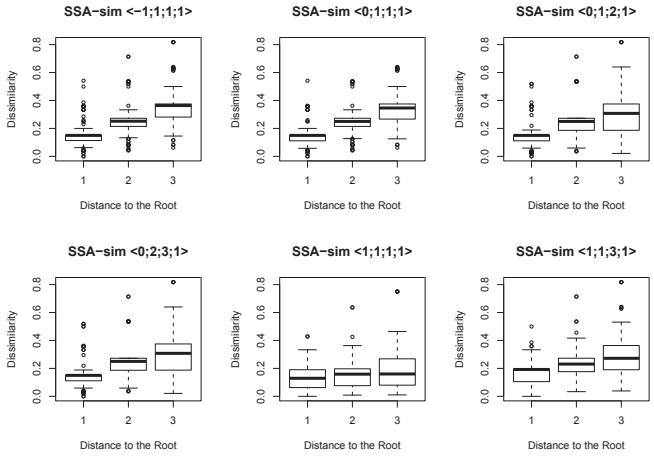


Figure B.207: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets SSA-sim and all datasets derived from SSA-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

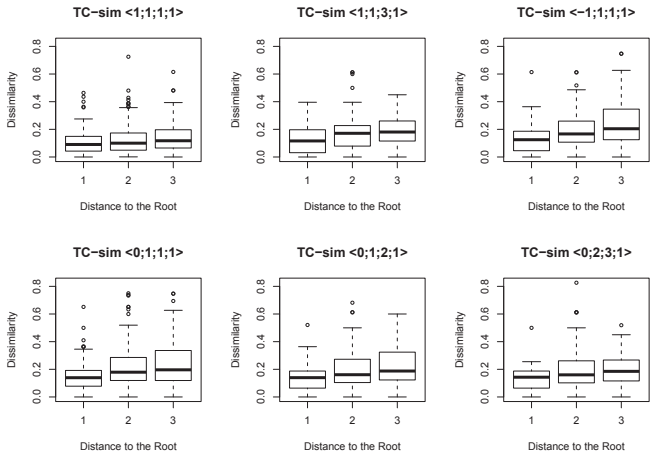


Figure B.208: Boxplots of the dissimilarities between reconciliations obtained for all simulated datasets TC-sim and all datasets derived from TC-sim by re-rooting the symbiont tree at distance k from the original root. The six plots correspond to the cost vectors in $\mathcal{C}$. The x-axis shows the distance k between the new and the original root. The y-axis shows the value d of dissimilarity of the reconciliation patterns.

Bibliography

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

[2] Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19(suppl_1):i7–i15.

[3] Balbuena, J. A., Míguez-Lozano, R., and Blasco-Costa, I. (2013). PACo: A novel procrustes application to cophylogenetic analysis. *PloS One*, 8(4):e61048.

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

[5] Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., and Sagot, M.-F. (2014). Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology*, 64(3):416–431.

[6] Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990.

[7] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

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

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

[10] Charleston, M. A. (2003). Recent results in cophylogeny mapping. *Advances in Parasitology*, 54:303–330.

[11] Charleston, M. A. (2012). TreeMap 3b. https://sites.google.com/site/cophylogeny/.

[12] Charleston, M. A. and Perkins, S. L. (2006). Traversing the tangle: Algorithms and applications for cophylogenetic studies. *Journal of Biomedical Informatics*, 39(1):62–71.

[13] Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: A new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(16):1–10.

[14] de Vienne, D. M., Giraud, T., and Martin, O. C. (2007). A congruence index for testing topological similarity between trees. *Bioinformatics*, 23(23):3119–3124.

[15] De Vienne, D. M., Giraud, T., and Shykoff, J. A. (2007). When can host shifts produce congruent host and parasite phylogenies? A simulation approach. *Journal of Evolutionary Biology*, 20(4):1428–1438.

[16] Deng, J., Yu, F., Li, H.-B., Gebiola, M., Desdevises, Y., Wu, S.-A., and Zhang, Y.-Z. (2013). Cophylogenetic relationships between Anicetus parasitoids (Hymenoptera: Encyrtidae) and their scale insect hosts (Hemiptera: Coccidae). *BMC Evolutionary Biology*, 13(1):275.

[17] Donati, B. (2014). *Graph models and algorithms in (co-)evolutionary contexts*. PhD thesis, Université Claude Bernard Lyon 1 and Università degli studi di Firenze.

[18] Donati, B., Baudet, C., Sinaimeri, B., Crescenzi, P., and Sagot, M.-F. (2015). Eucalypt: Efficient tree reconciliation enumerator. *Algorithms for Molecular Biology*, 10(1):3.

[19] Doyon, J.-P., Ranwez, V., Daubin, V., and Berry, V. (2011a). Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics*, 12(5):392–400.

[20] Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V., and Berry, V. (2011b). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In Tannier, E., editor, *RECOMB International Workshop on Comparative Genomics*, volume 6398 of *LNBI/LNCS*, pages 93–108. Springer-Verlag Berlin Heidelberg.

[21] Drinkwater, B., Qiao, A., and Charleston, M. A. (2016). WiSPA: A new approach for dealing with widespread parasitism. *arXiv preprint arXiv:1603.09415*.

[22] Escudero, M. (2015). Phylogenetic congruence of parasitic smut fungi (Anthracoidea, Anthracoideaceae) and their host plants (Carex, Cyperaceae): Cospeciation or host-shift speciation? *American Journal of Botany*, 102(7):1108–1114.

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

[24] Gómez-Acevedo, S., Rico-Arce, L., Delgado-Salinas, A., Magallón, S., and Eguiarte, L. E. (2010). Neotropical mutualism between Acacia and Pseudomyrmex: Phylogeny and divergence times. *Molecular Phylogenetics and Evolution*, 56(1):393–408.

[25] Górecki, P., Eulenstein, O., and Tiuryn, J. (2013). Unrooted tree reconciliation: A unified approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(2):522–536.

[26] Hafner, M. S. and Nadler, S. A. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332(6161):258–259.

[27] Hallett, M. T. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In Lengauer, T., editor, *Proceedings of the Fifth Annual International Conference on Computational Biology (RECOMB)*, pages 149–156, New York, USA. ACM.

[28] Hendricks, S. A., Flannery, M. E., and Spicer, G. S. (2013). Cophylogeny of quill mites from the genus *Syringophilopsis* (Acari: Syringophilidae) and their North American passerine hosts. *The Journal of Parasitology*, 99(5):827–834.

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

[30] Hommola, K., Smith, J. E., Qiu, Y., and Gilks, W. R. (2009). A permutation test of host–parasite cospeciation. *Molecular Biology and Evolution*, 26(7):1457–1468.

[31] Huelsenbeck, J. P., Rannala, B., and Larget, B. (2000). A Bayesian framework for the analysis of cospeciation. *Evolution*, 54:352–364.

[32] Huelsenbeck, J. P., Rannala, B., and Yang, Z. (1997). Statistical tests of host-parasite cospeciation. *Evolution*, 51(2):410–419.

[33] Hughes, A. L. and Friedman, R. (2000). Evolutionary diversification of protein-coding genes of hantaviruses. *Molecular Biology and Evolution*, 17(10):1558–1568.

[34] Hughes, J., Kennedy, M., Johnson, K. P., Palma, R. L., and Page, R. D. (2007). Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and pectinopygus lice. *Systematic Biology*, 56(2):232–251.

[35] Hugot, J. (1999). Primates and their pinworm parasites: The Cameron hypothesis revisited. *Systematic Biology*, 48(3):523–546.

[36] Hugot, J. (2003). New evidence for hystricognath rodent monophyly from the phylogeny of their pinworms. *Tangled trees: Phylogeny, cospeciation and coevolution*, pages 144–173.

[37] Jackson, A. P. and Andrew, P. (2005). The effect of paralogous lineages on the application of reconciliation analysis by cophylogeny mapping. *Systematic Biology*, 54(1):127–145.

[38] Jackson, A. P. and Charleston, M. A. (2004). A cophylogenetic perspective of RNA–virus evolution. *Molecular Biology and Evolution*, 21(1):45–57.

[39] Jacox, E., Chauve, C., Szöllosi, G. J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

[40] Jansen, G., Vepsäläinen, K., and Savolainen, R. (2011). A phylogenetic test of the parasite-host associations between *Maculinea* butterflies (Lepidoptera: Lycaenidae) and *Myrmica* ants (Hymenoptera: Formicidae). *European Journal of Entomology*, 108(1):53–62.

[41] Johnson, D. S., Yannakakis, M., and Papadimitriou, C. H. (1988). On generating all maximal independent sets. *Information Processing Letters*, 27(3):119–123.

[42] Kellner, K., Fernández-Marín, H., Ishak, H., Sen, R., Linksvaye, T., and Mueller, U. (2013). Co-evolutionary patterns and diversification of ant–fungus associations in the asexual fungus-farming ant Mycocepurus smithii in Panama. *Journal of Evolutionary Biology*, 26(6):1353–1362.

[43] Kosters, W. A. and Laros, J. F. J. (2008). Metrics for mining multisets. In *Research and Development in Intelligent systems XXIV*, pages 293–303. Springer.

[44] Legendre, P., Desdevises, Y., and Bazin, E. (2002). A statistical test for host–parasite coevolution. *Systematic Biology*, 51(2):217–234.

[45] Libeskind-Hadas, R., Wu, Y.-C., Bansal, M. S., and Kellis, M. (2014). Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics*, 30(12):i87–i95.

[46] Martínez-Aquino, A., Ceccarelli, F. S., Eguiarte, L. E., Vázquez-Domínguez, E., and de León, G. P.-P. (2014). 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*, 9(7):e101700.

[47] McLeish, M. J. and Van Noort, S. (2012). Codivergence and multiple host species use by fig wasp populations of the Ficus pollination mutualism. *BMC Evolutionary Biology*, 12(1):1.

[48] Mendlová, M., Desdevises, Y., Civáňová, K., Pariselle, A., and Šimková, A. (2012). Monogeneans of West African cichlid fish: evolution and cophylogenetic interactions. *PLoS One*, 7(5):e37268.

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

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

[51] Murray, E. A., Carmichael, A. E., and Heraty, J. M. (2013). Ancient host shifts followed by host conservatism in a group of ant parasitoids. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1759):20130495.

[52] Nei, M. and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford Univ. Press.

[53] Nemirov, K., Vaheri, A., and Plyusnin, A. (2004). Hantaviruses: co-evolution with natural hosts. *Recent Res. Devel. Virol.*, 6:201–228.

[54] Ovadia, Y., Fielder, D., Conow, C., and Libeskind-Hadas, R. (2011). The cophylogeny reconstruction problem is NP-complete. *Journal of Computational Biology*, 18(1):59–65.

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

[56] Page, R. D. and Charleston, M. A. (1998). Trees within trees: Phylogeny and historical associations. *Trends in Ecology & Evolution*, 13(9):356–359.

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

[58] Paterson, A. and Poulin, R. (1999). Have chondracanthid copepods co-speciated with their teleost hosts? *Systematic Parasitology*, 44(2):79–85.

[59] Paterson, A. M. and Banks, J. (2001). Analytical approaches to measuring cospeciation of host and parasites: Through a glass, darkly. *International Journal for Parasitology*, 31(9):1012–1022.

[60] Paterson, A. M., Palma, R. L., and Gray, R. D. (2003). Drowning on arrival, missing the boat, and x-events: How likely are sorting events. *Tangled trees: Phylogeny, Cospeciation, and Coevolution*, pages 287–309.

[61] Perkins, S. L. and Schall, J. (2002). A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *Journal of Parasitology*, 88(5):972–978.

[62] Plyusnin, A. and Morzunov, S. (2001). Virus evolution and genetic diversity of hantaviruses and their rodent hosts. In *Hantaviruses*, pages 47–75. Springer.

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

[64] Ramsden, C., Holmes, E. C., and Charleston, M. A. (2008). Hantavirus evolution in relation to its rodent and insectivore hosts: No evidence for codivergence. *Molecular Biology and Evolution*, 26(1):143–153.

[65] Refrégier, G., Le Gac, M., Jabbour, F., Widmer, A., Shykoff, J. A., Yockteng, R., Hood, M. E., and Giraud, T. (2008). 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.

[66] Ricci, I., Valzano, M., Ulissi, U., Epis, S., Cappelli, A., and Favia, G. (2012). Symbiotic control of mosquito borne disease. *Pathogens and Global Health*, 106(7):380–385.

[67] Ronquist, F. (2003). Parsimony analysis of coevolving species associations. In Page, R., editor, *Tangled trees: Phylogeny, Cospeciation, and Coevolution*, pages 22–64. University of Chicago Press.

[68] Ronquist, F. and Nylin, S. (1990). Process and pattern in the evolution of species associations. *Systematic Zoology*, 39(4):323–344.

[69] Rosenblueth, M., Sayavedra, L., Sámano-Sánchez, H., Roth, A., and Martínez-Romero, E. (2012). Evolutionary relationships of flavobacterial and enterobacterial endosymbionts with their scale insect hosts (Hemiptera: Coccoidea). *Journal of Evolutionary Biology*, 25(11):2357–2368.

[70] Sanderson, M. J. and Shaffer, H. B. (2002). Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics*, 33(1):49–72.

[71] Simões, P. (2012). *Diversity and dynamics of Wolbachia-host associations in arthropods from the Society archipelago, French Polynesia.* PhD thesis, Université Claude Bernard-Lyon I.

[72] Simões, P., Mialdea, G., Reiss, D., Sagot, M.-F., and Charlat, S. (2011). Wolbachia detection: An assessment of standard PCR Protocols. *Mol. Ecol. Resour.*, 11(3):567–572.

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

[74] Stolzer, M., Han, L., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012a). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415.

[75] Stolzer, M., Han, L., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012b). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415.

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

[77] Szöllosi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912.

[78] Szöllosi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. *Systematic Biology*, 64(1):e42–e62.

[79] Thompson, J. N. (1994). *The Coevolutionary Process.* University of Chicago Press.

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

[81] Urbini, L., Sinaimeri, B., Matias, C., and Sagot, M.-F. (2016). Robustness of the parsimonious reconciliation method in cophylogeny. In *International Conference on Algorithms for Computational Biology (AlCoB)*, volume 9702 of *Lecture Notes in BioInformatics/Lecture Notes in Computer Science*, pages 119–130. Springer.

[82] Urbini, L., Sinaimeri, B., Matias, C., and Sagot, M.-F. (2017). Exploring the robustness of the parsimonious reconciliation method in host-symbiont cophylogeny. *IEEE/ACM Trans. Comput. Biology Bioinform.* submitted.

[83] Vanhove, M. P. M., Pariselle, A., Van Steenberge, M., Raeymaekers, J. A. M., Hablützel, P. I., Gillardin, C., Hellemans, B., Breman, F. C., Koblmüller, S., Sturmbauer, C., Snoeks, J., Volckaert, F. A. M., and Huyse, T. (2015). Hidden biodiversity in an ancient lake: Phylogenetic congruence between Lake Tanganyika tropheine cichlids and their monogenean flatworm parasites. *Scientific Reports*, 5:13669.

[84] Viale, E., Martinez-Sanudo, I., Brown, J., Simonato, M., Girolami, V., Squartini, A., Bressan, A., Faccol, M., and Mazzon, L. (2015). 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*, 90:67–79.

[85] Weckstein, J. D. (2004). Biogeography explains cophylogenetic patterns in toucan chewing lice. *Systematic Biology*, 53(1):154–164.

[86] Wei, Z. and Jousset, A. (2017). Plant breeding goes microbial. *Trends Plant. Sci.*, 22(7):555–558.

[87] Wieseke, N., Hartmann, T., Bernt, M., and Middendorf, M. (2015). Cophylogenetic reconciliation with ILP. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(6):1227–1235.

Modèles et algorithmes pour étudier l'histoire évolutive commune des hôtes et des symbiotes.

Lors de cette thèse, je me suis intéressée aux modèles et aux algorithmes pour étudier l'histoire évolutive commune des hôtes et des symbiotes.

Le premier objectif était d'analyser la robustesse des méthodes de réconciliation des arbres phylogénétiques, qui sont très utilisées dans ce type d'étude. Celles-ci associent (ou lient) un arbre, d'habitude celui des symbiotes, à l'autre, en utilisant un modèle dit *basé sur des évènements*. Les évènements les plus utilisés sont la cospéciation, la duplication, le saut et la perte. Les phylogénies des hôtes et des symbiotes sont généralement considérés comme donnés, et sans aucune erreur. L'objectif était de comprendre les forces et les faiblesses du modèle parcimonieux utilisé et comprendre comment les résultats finaux peuvent être influencés en présence de petites perturbations ou d'erreurs dans les données en entrée. Ici deux cas sont considérés, le premier est le choix erroné d'une association entre les feuilles des hôtes et des symbiotes dans le cas où plusieurs existent, le deuxième est lié au mauvais choix de l'enracinement de l'arbre des symbiotes. Nos résultats montrent que le choix des associations entre feuilles et le choix de l'enracinement peuvent avoir un fort impact sur la variabilité de la réconciliation obtenue. Nous avons également remarqué que l'evènement appelé "saut" joue un rôle important dans l'étude de la robustesse, surtout pour le problème de l'enracinement.

Le deuxième objectif de cette thèse était d'introduire certains evènements peu ou pas formellement considérés dans la littérature. L'un d'entre eux est la "propagation", qui correspond à l'invasion de différents hôtes par un même symbiote. Dans ce cas, lorsque les propagations ne sont pas considérés, les réconciliations optimales sont obtenues en tenant compte seulement des coûts des évènements classiques (cospeciation, duplication, saut, perte). La nécessité de développer des méthodes statistiques pour assigner les coûts les plus appropriés est toujours d'actualité. Deux types de propagations sont introduites : verticaux et horizontaux. Le premier type correspond à ce qu'on pourrait appeler aussi un gel, à savoir que l'évolution du symbiote s'arrête et "gèle" alors que le symbiote continue d'être associé à un hôte et aux nouvelles espèces qui descendent de cet hôte. Le second comprend à la fois une invasion, du symbiote qui reste associé à l'hôte initial, mais qui en même temps s'associe ("envahit") un autre hôte incomparable avec le premier, et un gel par rapport à l'évolution des deux l'hôtes, celui auquel il était associé au début et celui qu'il a envahi. Nos résultats montrent que l'introduction de ces evènements rend le modèle plus réaliste, mais aussi que désormais il est possible d'utiliser directement des jeux de données avec un symbiote qui est associé plusieurs hôtes au même temps, ce qui n'était pas faisable auparavant.

MOTS-CLEFS en français : Cophilogenie; parsimonie; méthodes basées sur des evènements; robustesse; mesure pour la comparaison de reconciliations entre arbres; systèmes hôtes/symbiotes; calcul approximatif Bayésien, spread.

Models and algorithms to study the common evolutionary history of hosts and symbionts.

In this Ph.D. work, we proposed models and algorithms to study the common evolutionary history of hosts and symbionts.

The first goal was to analyse the robustness of the methods of phylogenetic tree reconciliations, which are a common way of performing such study. This involves mapping one tree, most often the symbiont's, to the other using a so-called event-based model. The events considered in general are cospeciation, duplication, host switch, and loss. The host and the symbiont phylogenies are usually considered as given and without any errors. The objective here was to understand the strengths and weaknesses of the parsimonious model used in such mappings of one tree to another, and how the final results may be influenced when small errors are present, or are introduced in the input datasets. This may correspond either to a wrong choice of present-day symbiont-host associations in the case where multiple ones exist, or to small errors related to a wrong rooting of the symbiont tree. Our results show that the choice of leaf associations and of root placement may have a strong impact on the variability of the reconciliation output. We also noticed that the host switch event has an important role in particular for the rooting problem.

The second goal of this Ph.D. was to introduce some events that are little or not formally considered in the literature. One of them is the spread, which corresponds to the invasion of different hosts by a same symbiont. In this case, as when spreads are not considered, the optimal reconciliations obtained will depend on the choice made for the costs of the events. The need to develop statistical methods to assign the most appropriate ones therefore remains of actuality. Two types of spread are introduced: vertical and horizontal. The first case corresponds to what could be called also 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. The second includes both an invasion, of the symbiont which remains with the initial host but at the same time gets associated with ("invades") another one incomparable with the first, and 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". Our results show that the introduction of these events makes the model more realistic, but also that it is now possible to directly use datasets with a symbiont that is associated with more than one host at the same time, which was not feasible before.

Keywords in english: Cophylogeny; parsimony; event-based methods; robusness; measure for tree reconciliation comparison; host/symbiont system; approximate Bayesian computation; spread.