An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework

Jean-Philippe Doyon[1,2], Sylvie Hamel[2], and Cedric Chauve[3]

[1] LIRMM, Université Montpellier2 and CNRS, UMR 5506 - CC 477, 161 rue Ada 34392 Montpellier Cedex 5, France. Jean-philippe.Doyon@lirmm.fr

[2] DIRO, Université de Montréal, CP6128, succ. Centre-Ville, H3C 3J7, Montréal (QC),Canada, hamelsyl@iro.umontreal.ca

[3] Department of Mathematics, Simon Fraser University, 8888 University Drive, V5A 1S6, Burnaby (BC), Canada, cedric.chauve@sfu.ca

Abstract. Background. Inferring an evolutionary scenario for a gene family is a fundamental problem both in functional and evolutionary genomics. The gene tree/species tree reconciliation approach has been widely used to address this problem, but mostly in a parsimony framework, that considers only the reconciliation that minimizes the number of duplication and/or loss events. Recently a probabilistic approach has been developed, based on the classical birth-death process, including efficient algorithms for computing posterior probabilities of reconciliations and orthology prediction.

Results. We recently proposed an efficient algorithm for exploring the whole space of gene tree/species tree reconciliations, that we adapt here to compute efficiently, either exactly or approximately depending on the space size, the posterior probability of the visited reconciliations. We use this algorithm to analyze the probabilistic landscape of the space of reconciliations for, both, a real dataset of fungal gene families and simulated data.

Conclusion. Our results suggest that with realistic gene duplication and loss rates, a very small subset of all reconciliations needs to be explored in order to approximate very closely the posterior probability of the most likely reconciliations. For cases where the posterior probability mass is more evenly dispersed, our method allows to explore efficiently the required subspace of reconciliations.

Keywords. Comparative genomics, species tree, gene tree, probability, reconciliation, parsimony.

1 Introduction

Genomes of contemporary species, especially eukaryotes, are the result of an evolutionary history that started with a common ancestor from which new species evolved through evolutionary events called speciations. One of the main objectives in molecular biology is the reconstruction of this evolutionary history, that can be depicted with a rooted binary tree, called a *species tree*, where the root represents the common ancestor, the internal nodes the ancestral species and speciation events, and the leaves the extant species. Other events than speciation can happen, that do not result immediately in the creation of new species but are essential in eukaryotic genes evolution, such as gene duplication and loss [13]. Duplication is the genomic process where one or more genes of a single genome are copied, resulting in two copies of each duplicated gene. Gene duplication allows one copy to possibly develop a new biological function through point mutation, while the other copy often preserves its original role; another outcome can be subfonctionalization, where each copy develops a specific subfunction of the function of the ancestral gene. A gene is considered to be lost when the corresponding sequence has been deleted by a genomic rearrangement or has completely lost any functional role (i.e. has become a pseudogene). (See [13] for example). Other genomic events such as lateral gene transfer, that occurs mostly in bacterial genomes, will not be considered here.

Genes of contemporary species that evolved from a common ancestor, through speciations and duplications, are said to be homologs [9] and are grouped into a gene family. Such gene families are in general inferred using protein sequence comparison and clustering methods. The evolution of a gene family can be depicted with a rooted binary tree, called a *gene tree*, where the leaves represent the homologous contemporary genes, the root their common ancestral gene and the internal nodes represent ancestral genes. Given a gene tree G and the species tree S of the corresponding genomes, a fundamental question is to infer the evolutionary history that led to G, which amounts to locate in G the evolutionary events of speciations and duplications and the branch or nodes of S where they occurred. Inferring the evolutionary scenario for a gene family has applications in phylogenomics [20], functional genomics, especially for the identification of orthologous genes [21], or comparative genomics and paleogenomics [18].

A *reconciliation* between G and S is a mapping of the genes (extant and ancestral) of G onto the nodes of S that induces such an evolutionary scenario, in terms of speciations, duplications and losses, for the gene family described by G. The notion of reconciliation was first introduced in the pioneering work of Goodman *et al.* [10] and a first formal definition was given in [19] to explain the discrepancies between gene and species trees. The Minimum Parsimony Reconciliation (MPR from now) is defined by the mapping of each gene u of G onto the most recent species of S that is ancestor of all genomes that contain a gene descendant of u (called the LCA-mapping, see Figure 1). It has been shown to be the reconciliation that induces the unique evolutionary scenario that minimizes both the number of gene duplication events and gene loss events [8]. It is generally accepted that parsimony is a pertinent criterion in evolutionary biology, but that it does not always reflect the true evolutionary history. This leads to the definition of more general notions of reconciliations between G and S [6, 11, 2, 8] and the natural problem of exploring all, or many, evolutionary scenarios for a given gene family [8].

In the context of probabilistic orthology analysis, an important breakthrough is due to Sennblad *et al.* [21], who developed a method for computing the posterior probability that a given pair of genes of a gene family tree G are orthologous, according to the definition of Fitch [9] (two extant genes are orthologs if their most recent ancestor (LCA) in G is a speciation vertex). For fixed gene duplication and gene loss rates along the branches of S, the method computes in polynomial time the exact posterior probability that a vertex of G is a speciation, and a Markov Chain Monte Carlo, based on a Bayesian framework, is used to integrate over the rates (that is given the trees G and S). With these two steps, their method estimates the posterior probability of orthology relationships between extant genes of G, with prior on the duplication and loss rates. They focus their study on orthology relationships between pairs of genes, or equivalently on the probability that a given vertex of G represents a speciation event, but they also provide algorithms to compute the posterior probability of a given reconciliation. With regard to this, their experimental results suggest that, for low gene duplication and loss rates, the posterior probability mass is dominated by the MPR, but using simulated data, they show that it is not rare that the MPR implies wrong orthology relationships and that their probabilistic framework improves on the traditional parsimony approach. They also argue against the explicit exploration of the whole space of all reconciliations, due to its possible huge size, and they develop in [2] efficient, but sophisticated, dynamic programming algorithms to both compute the posterior probability of a given reconciliation and sample reconciliations according to the posterior distribution. Both algorithms have a time complexity that is quadratic in the size of G and linear in the size of S.

The goal of the present study is to complement the recent work of Sennblad *et al.* by (1) providing more efficient algorithms to explore a subspace of the space of all reconciliations and (2) analyzing the shape of the space of all reconciliations between a gene tree and a species tree according to a probabilistic criterion, when gene duplication and loss rates are known. Our contribution is twofold. First, we extend the algorithm described in [8], to explore the whole space of the reconciliations between a gene tree and a species tree, in order to compute or estimate efficiently the posterior probability of the set of all visited reconciliations, whether this set represents the whole space or only a subspace if the latter is too large. This algorithm improves on the algorithms described in [2], as computing the probability for the visited reconciliations can be done in average linear time, both in the size of G and in the size of S, for each visited reconciliations. This makes possible to explore larger sets of reconciliations, a property we use in our experiments to study the probabilistic landscape of real and simulated datasets. We study a real dataset of fungal gene families and several simulated datasets, obtained with moderate but realistic gene duplication and loss rates, and we show that, in general, a small subset of reconciliations located close to the MPR cover the whole probability mass. Hence, the posterior probability of the plausible evolutionary scenarios of a gene tree G according to S can be estimated efficiently with very good precision.

2 Material and Methods

2.1 Trees and reconciliations

Let T be a binary tree with vertices $V(T)$ and edges $E(T)$, with labeled leaves. Let $r(T)$, $L(T)$, and $\Lambda(T)$ respectively denote its root, the set of its leaves, and the set of the labels of its leaves. For a vertex u of T, we denote by u_1 and u_2 its children and by T_u the subtree of T rooted at u. For a vertex $u \in V(T) \setminus \{r(T)\}$, we denote by $p(u)$ its parent and by $(p(u), u)$ the edge of T with $p(u)$ as the departure vertex.

Let G be a gene tree and S be a species tree, with $\Lambda(G) \subseteq \Lambda(S)$. Following [8], a reconciliation between G and S maps each vertex u of G onto either a vertex or an edge of S and is denoted $\alpha : V(G) \rightarrow V(S) \cup E(S)$. (see Appendix A for a complete formal definition and Figure 1 for an illustration). If $\alpha(u) = x \in V(S)$, u represents a gene that will be present in single copy in the two genomes x_1 and x_2 following a speciation event that happened to x. Otherwise, $\alpha(u) = (p(x), x) \in E(S)$ and u represents a gene of the ancestral species $p(x)$ that has been duplicated in the descendant species x. For simplicity, we assume that $r(G)$ has a singe copy in the ancestral genome $r(S)$, as the case where $r(G)$ is duplicated in the ancestor species $r(S)$ can be handled easily by adding a branch prior to $r(S)$.

2.2 Exploring the space of reconciliations

Following [8], we explore reconciliations between G and S according to an exploration tree denoted $\mathcal{T}(G, S)$ (see Figure 15 in Appendix A), whose root is the MPR, also denoted α_{min} from now, and such that two reconciliations that are incident in $\mathcal{T}(G, S)$ differ only by a single Nearest Mapping Change (NMC). An NMC is an operator introduced in [8] which transforms a first reconciliation into a second one by moving the mapping of an internal vertex of G by one vertex/edge of S either downward or upward. (See Figure 2 for an illustration).

$D_{NMC}(\alpha_{min}, \alpha)$ corresponds to the depth in $\mathcal{T}(G, S)$ of a given reconciliation α, i.e. the minimum number of NMC needed to transform α into α_{min}. $D(\mathcal{T}(G, S))$ denotes the maximum depth of a reconciliation of $\mathcal{T}(G, S)$; for a given depth d, $\mathcal{T}_d(G, S)$ is the subtree of $\mathcal{T}(G, S)$ rooted at α_{min} and formed of each reconciliation $\alpha \in \mathcal{T}(G, S)$ such that $D_{NMC}(\alpha_{min}, \alpha) \leq d$.

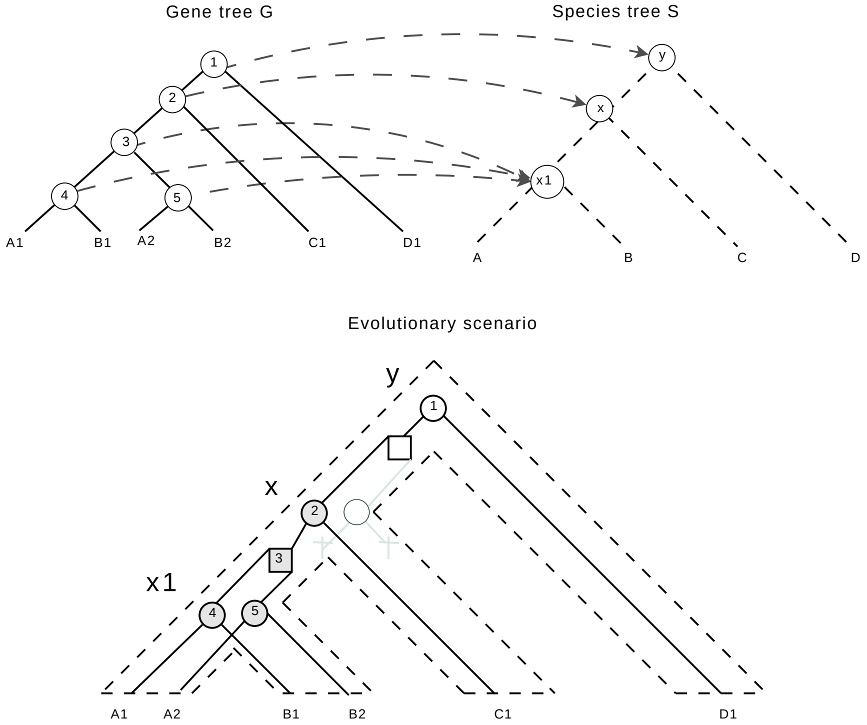

Fig. 1. Top: the species tree S has four (extant) species (A, B, C, and D). The gene tree G has six (extant) genes, where each gene belongs to one of the four species (i.e. gene A1 belongs to species A). The arrows represent the LCA-mapping between G and S. **Bottom:** Minimum Parsimony Reconciliation between G and S induced by the LCA mapping. A circle (square) represents an internal vertex of G that is mapped on an internal vertex (resp. edge) of S, that is a speciation (resp. duplication) event. A cross represents a gene loss. The right lineage of the first duplication has no extant gene that descents from it, as opposite to its left lineage. We then say that this duplication is hypothetical, because it is not a useful information for the evolutionary scenario of the extant genes of G along S. Hence, such duplication is not depicted by the reconciliation.

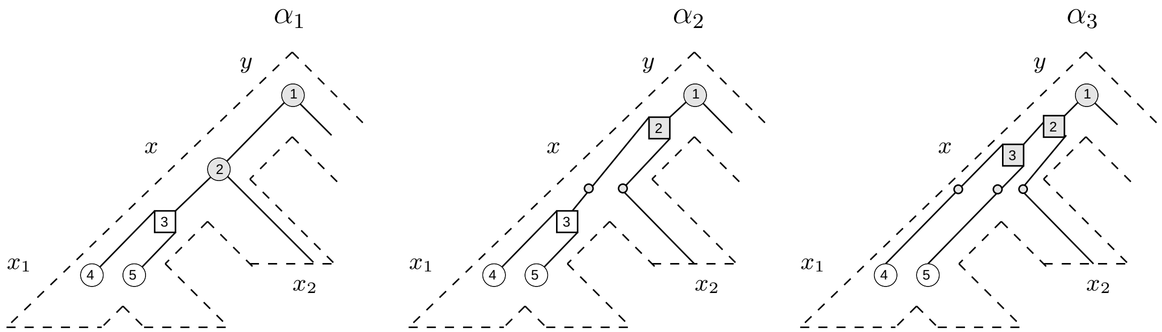

Fig. 2. Left: a section of the reconciliation depicted in Figure 1. Here, the mapping of vertex 2 forbids to move up vertex 3. **Center:** the vertex 2 changes from a speciation to a duplication by moving it up. **Right:** then, vertex 3 can be moved up and still is a duplication.

2.3 The probabilistic framework

We now assume that for each branch of S, the length of the branch, as well as a gene duplication rate and a gene loss rate are known.

For a reconciliation α between G and S, $P(G, \alpha)$ denotes the probability that a single gene of the ancestral species $r(S)$ evolves along S and generates a gene tree that is isomorphic to G following the evolutionary scenario induced by α. Hence, the probability $P(G)$ of generating G along S is the sum of $P(G, \alpha)$ over all reconciliations α between G and S, and the posterior probability of a reconciliation α, given G, is $P(\alpha|G) = \frac{P(G,\alpha)}{P(G)}$.

Given a set $\mathcal{T}$ of K reconciliations $\{\alpha_1, \ldots, \alpha_K\}$, the probability of G based on these reconciliations is defined as follows: $P_{\mathcal{T}}(G) = \sum_{i=1,\ldots,K} P(G, \alpha_i)$. If $\mathcal{T} = V(\mathcal{T}(G, S))$, then $P_{\mathcal{T}}(G)$ is the exact probability of G. Otherwise, it is a $\mathcal{T}$-approximation.

Given a subtree $\mathcal{T}_d(G, S)$ of $\mathcal{T}(G, S)$, the sum of the posterior probability of each reconciliation α located in $\mathcal{T}_d(G, S)$ is called its probability mass and is defined as follows $P(\mathcal{T}_d(G, S)|G) = \sum_{\alpha \in \mathcal{T}_d(G,S)} P(\alpha|G)$.

We now present our main algorithmic results.

Theorem 1. *Let G be a gene tree and S a species tree, with $|V(S)| = m$ and $|V(G)| = n$, and $\mathcal{T}$ a connected subtree of $\mathcal{T}(G, S)$ containing K reconciliations between G and S. Then,*

1. *computing the exact posterior probability $P(\alpha|G)$ for all K reconciliations α of $\mathcal{T}$ can be done in time and space $O(mn^2 + K(m + n))$,*
2. *computing the $\mathcal{T}$-approximation $P_{\mathcal{T}}(\alpha|G)$ of the posterior probability for all K reconciliations α of $\mathcal{T}$ can be done in time and space $O(mn + K(m + n))$.*

Proof. The computation scheme is based on

- an $O(mn^2)$ algorithm to compute $P(G)$ described in [2, Theorem 6.9], if one is interested in computing the exact posterior probabilities of the visited reconciliations,
- the exploration of $\mathcal{T}$ starting at α_{min}, using the general scheme described in [8], that requires time $O(K)$,
- the computation of the probability $P(G, \alpha_{min})$ in $O(mn)$ time [2, Theorem 5.21],
- Lemma 1 below that states that the probability of a newly visited reconciliation can be obtained from the previous one in $O(m + n)$ time .

All together, this gives the stated complexities

If one is interested in the posterior probability distribution of a subset of reconciliations, this result improves, from an efficiency point of view, on the algorithm presented in [2], that has an $O(mn)$ time complexity to compute $P(G, \alpha)$ for a given reconciliation α, while Lemma 1 below shows it can be updated in $O(m + n)$ time after a single NMC. Moreover, provided the $\mathcal{T}$-approximation of the posterior probability $P(\alpha|G)$ is a good approximation, the computational cost gain of our method makes it valuable in the context where duplication and loss rates are not known but are computed through an MCMC approach, where a large number of posterior probability computations is required (see [21]). The precise experimental results we present in the Results section, especially for very large sets of reconciliations, were obtained thanks to this computational complexity improvement.

Lemma 1. *Given two reconciliations α and α' of $\mathcal{T}(G, S)$ that are separated by a single NMC, the time complexity to compute $P(G, \alpha')$ given $P(G, \alpha)$ is $O(m + n)$.*

Proof. The proof relies on the recursion described in [2] to compute the probability of generating (G, α) (called a reconciled tree from now) along S, and we outline here its main properties.

Following the notation of [2] and given an internal vertex x of S and a vertex u of G such that $\alpha(u) = x$, $r_V(x, u)$ denotes the probability that the evolution of the gene u along the species tree S_x results in the reconciled tree (G_u, α_u), where α_u is the reconciliation between G_u and S_x that is induced by α. The computation of this probability is based on Recursion 1 below and on the four components:

- $r_A(x_1, u)$ is the component of $r_V(x, u)$ for the edge (x, x_1) and the subtree S_{x_1} ([4]).
- $Q_{x_1}(l)$ is the probability that the evolution of u along this edge generates l non-ghost genes[5] that belong to x_1.
- If the subtree of G induced by this evolution and rooted at u is denoted $G_{u||x_1}$, which is called a sliced subtree (see Figure 1: the colored vertices of G correspond to $G_{u||x_1}$, where u denotes the speciation vertex 2), $h(G_{u||x_1})$ is the probability that the sliced subtree of G generated by this evolution is isomorphic to $G_{u||x_1}$, assuming that all such subtrees are equiprobable.
- Finally, $W(x_1, u)$ corresponds to the number of ways the reconciled trees rooted at the leaves of $G_{u||x_1}$ can be exchanged to produce a reconciled tree that is isomorphic to (G_u, α_u).

Recursion 1. ([2, Recursion 5.20]) The probability of a reconciled tree (G, α). Let $x \in V(S)$,

$$r_V(x, u) = \begin{cases} 1 & \text{if } x \in L(S), u \in L(G), u \in \alpha^{-1}(x) \\ r_A(x_1, u)\ r_A(x_2, u) & \text{otherwise} \end{cases}$$

$$r_A(x, u) = \begin{cases} Q_x(0) & \text{if } L(G_{u||x}) = \emptyset \\ Q_x(|L(G_{u||x})|)\ W(x, u)\ h(G_{u||x}) \displaystyle\prod_{v \in L(G_{u||x})} r_V(x, v) & \text{otherwise.} \end{cases}$$

According to [2, Theorem 5.21], if the root of G is mapped on the root of S (that is $\alpha(r(G)) = r(S)$), the probability that an evolutionary scenario produces the reconciled tree (G, α) is $P(G, \alpha) = r_V(r(S), r(G))$ and can be computed in time $O(mn)$, where $|V(S)| = m$ and $|V(G)| = n$ ([6]).

We now prove the lemma. Consider the three reconciliations of Figure 2, and let v and u, and a_1 and a_2 respectively denotes the vertices labeled 1 and 2, and the left and right artificial vertices that are mapped on species x in α_2. Given that the probability $P(G, \alpha_1)$ is computed and the NMC that moves u upward is applied on α_1 and results in α_2, the new probability $P'(G, \alpha_2)$ is updated as follows: $r'_A(x_1, a_1) = r_A(x_1, u)$, $r'_A(x_2, a_1) = Q_{x_2}(0)$, $r'_A(x_1, a_2) = Q_{x_1}(0)$, and $r'_A(x_2, a_2) = r_A(x_2, u)$. As the sliced subtree $G_{v||x}$ changes following the NMC applied on u, both $h(G_{v||x})$ and $W(x, v)$ have to be recomputed. According to [2], the time complexities to compute $Q_{x_1}(0)$ and $Q_{x_2}(0)$ and to compute $h(G_{v||x})$ and $W(x, v)$ respectively are $O(m)$ and $O(n)$. It is then immediate that the time complexity to compute $P'(G, \alpha_2)$ given $P(G, \alpha_1)$ is $O(n + m)$. For the three others NMCs, the proof is similar.

[4] In $r_V(x, u)$ (resp. $r_A(x_1, u)$), the r stands for reconciliation and the subscript V (resp. A) indicates that it starts at the considered vertex (resp. arc) of S.

[5] Given an evolutionary scenario for a gene family G, an ancestral gene that belongs to an ancestral species x of S that goes extinct before reaching the leaves of S_x is called a ghost gene. Hence, a non-ghost gene in such an evolutionary scenario is a gene that has at least one descendant among the extinct genes of G.

[6] The case where $r(G)$ is not mapped on $r(S)$, but is mapped on an edge or another vertex of S, is similar to the one described above and we omit it for simplicity.

2.4 A dataset of fungal gene families

We considered 12 fungal genomes, whose species tree is given in Figure 16 of Appendix B. Althought there is debate (see [16]) over the true phylogeny, we argue that the signals of our experiments are sufficiently clear so that a phylogenetic error would have a small impact. For these 12 species, we considered 1278 gene trees from [24], which originally contains a total of 20598 gene families from 12 fungal genomes. After keeping only a single copy for sets of isomorphic gene trees, 1543 gene trees remain. From these 1543 gene trees, we conserved the 1278 ones whose reconciliation space contains between 10 and 10^7 reconciliations, called from now the A-trees. The distribution of all the A-trees G according to the number of genes and species present in G and to the size and depth of $\mathcal{T}(G, S)$ are depicted in Figure 17 of Appendix B.

The branch length (in Million of Years) of the species tree S were computed by a Bayesian framework that takes as input homologous DNA sequences and assumes that the rates of nucleotide substitutions is constant over any branch, but it can differ among the branches (relaxed molecular clock) [22]. The duplication and loss rates along the branches of S were estimated by CAFE [5], which takes as input the branch lengths and the profiles (i.e. the number of genes per species) of the 20598 fungal gene families, and performs an Expectation-Maximization algorithm to find the rates that maximizes the probability of the observed profiles. We performed several runs of the algorithm, without regrouping the branches of S into rate classes, and observed that it converges to the same rate for each branch of S.

2.5 Datasets of synthetic gene trees

We computed synthetic sets of gene trees obtained from the fungi species tree S, with three different rates of gene duplication and loss. We defined three classes of rates using 1, 1.4, and 1.8 as the Increasing Factor (I.F. for short) on the previous ones [7] and, for each I.F., we generated synthetic gene trees along S using a birth-and-death process starting from a single ancestral gene. Given length and duplication and loss rates along each branch of S, the generation of a gene tree starts with a single gene u at $r(S) = x$, simulates its evolution along the branch (x, x_1) (resp. (x, x_2)) using the birth-and-death [17] process, which computes the probability that u has n descendants that belong to x_1 (resp. x_2), and recursively repeats this process throw the extant species of S. Afterward, the synthetic gene tree G is obtained by the removal of each ghost gene, and the resulted evolutionary scenario corresponds to the real reconciliation α_{real} between G and S. The I.F. 1.8 is the highest one considered because it induces 14 gene trees that cover the 12 genomes, as opposite with an I.F. of 2 for which none such tree were generated. For each considered I.F., the considered synthetic gene trees were obtained by 2000 such simulations and the conservation of all the unique A-trees, after the removal of multiple copies. We then obtained 1051 A-trees with I.F. 1, 1025 with I.F. 1.4 and 924 with I.F. 1.8. The characteristics of these genes trees, in terms of number of genes, species, and reconciliation spaces are depicted in Figure 18 (see Appendix B).

It is important to point out that the birth-and-death process does not take into consideration some evolutionary properties: genes that belong to the same protein complexes tend to have similar evolutionary scenarios [12]; whole genome or large segmental duplications (such events happened in yeasts evolution [23, 25]); after duplication of a gene, the evolution of one copy tends to affect the other [7]; the duplication and loss rates may not be constant along the branches of S. These discrepancies may be the reason why the distributions of the number of genes and species present in a gene tree differ between the trees generated by the rates estimated by CAFE and the 20598

[7] That is each (duplication and loss) rate along any branch of S is multiplied by the considered factor.

real ones mentioned before, which are used by CAFE to estimates these rates (see Figure 17 and 18). However, these synthetic gene trees provide valuable and reasonable datasets to study the probabilistic landscape of the space of reconciliations, which is our goal in the present work.

3 Results

The main concerns of our experiments is to sustain the two following observations.

1. The probability mass of the whole tree of reconciliations is technically covered (i.e. approximated with very high precision) by a small set $\mathcal{T}$ of reconciliations located close to α_{min}.
2. For a given reconciliation α that belongs to $\mathcal{T}$, the approximation $P_{\mathcal{T}}(\alpha|G)$ is a very precise approximation of the exact posterior probability $P(\alpha|G)$.

We first illustrate these facts on a real dataset of fungal gene families, then we show they remain true with synthetic gene trees obtained with higher duplication and loss rates.

3.1 Fungal gene trees

To begin, let us describe two observations regarding the MPR α_{min} over the 1278 A-trees.

1. In 1276 cases, the MPR is the most probable reconciliation, and in the two remaining cases, the most probable reconciliation α^* is one NMC away from α_{min}.
2. The average probability $P(\alpha_{min}|G)$ is 0.94672 with a standard deviation of 0.03906, and varies from 0.98 when the depth of $\mathcal{T}(G,S)$ is 5 or less to 0.88 when the depth of $\mathcal{T}(G,S)$ is between 55 and 60 (See Figure 3). Note however that it happens that α_{min} has a significantly lower posterior probability, pointing at gene families where the probability mass is more evenly distributed.

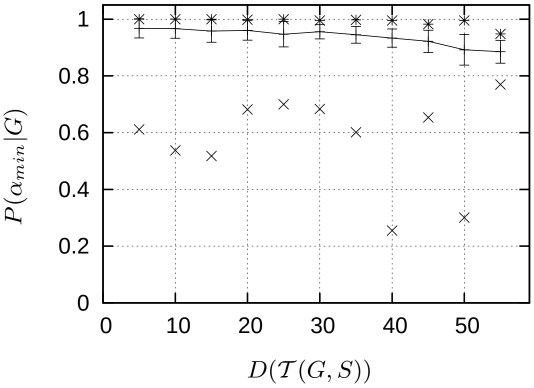

Fig. 3. Over all 1278 A-trees G, average posterior probability of α_{min} (y axis) according to the depth of $\mathcal{T}(G,S)$ (x axis; gene trees are grouped by classes for which the depth divided by 5 is equal), together with the standard deviation, the maximum and minimum probabilities.

Next, we recorded the variation of the exact posterior probability $P(\alpha|G)$ with respect to the depth of α. Our results, described in Figure 4 below show that, on the average, this probability decreases very quickly with the depth.

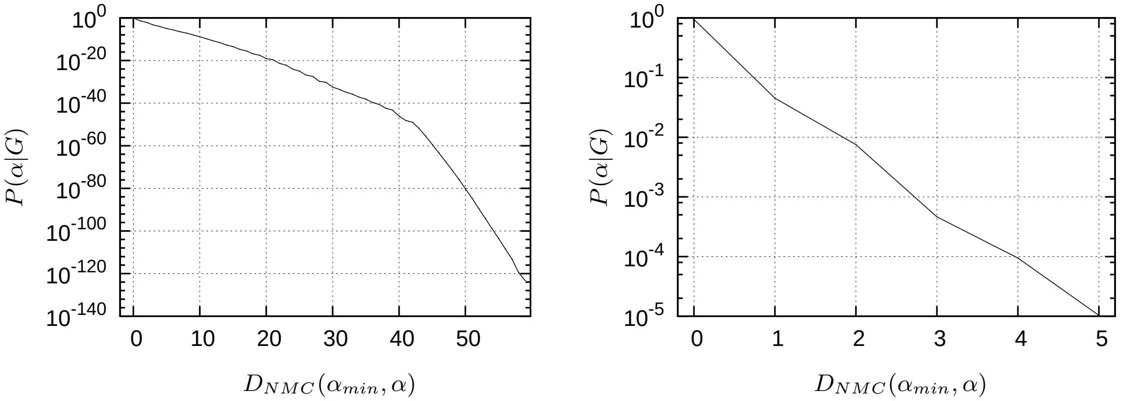

Fig. 4. Over all 1278 A-trees G, (**left**) average probability (y axis) of each reconciliation $\alpha \in \mathcal{T}(G, S)$ located at the same depth in $\mathcal{T}(G, S)$ (x axis), and (**right**) zoom on the reconciliations of depth at most 5.

These observations agree with previous ones for datasets with low gene duplication and loss rates [21] and give a clear insight on the fact that the probability mass of the whole space tree is concentrated around α_{min}, and the depth of this tree has a relative small impact (although not negligible) on these probabilities. Note also that, although the standard variation of the average probability of the MPR is relatively low, this does not prevent some cases where the MPR has a relatively low probability a posteriori (down to less than 0.3 in some cases). This suggests that, for some gene families, even with low rates, exploring a relatively large subspace of the space of reconciliations is worth it.

Next, we analyzed the depth required to explore sufficiently many reconciliations to capture most of the probability mass. Let $d^\circ(\mathcal{T}(G, S))$ be the smallest depth for which the probability mass of $\mathcal{T}_{d^\circ(\mathcal{T}(G,S))}(G, S)$ is technically equal to one (that is according to the usual C++ floating point precision [15], the sum of the probabilities of the reconciliations in this subspace is 1). Figure 5(left) below plots $d^\circ(\mathcal{T}(G, S))$ against $D(\mathcal{T}(G, S))$. It is clear that, even for very deep reconciliation spaces, a small depth is enough to capture the probability mass: for $D(\mathcal{T}(G, S)) = 55$, the average value of $d^\circ(\mathcal{T}(G, S))$ is only 4.5, and the maximum required depth is only 9, even when the posterior probability of α_{min} is low, as shown by Figure 5(right).

To complement these observations, we show in Figure 6 the respective sizes, on average, of the whole reconciliation spaces and of the subspace that covers the probability mass. It shows that as few as a thousand reconciliations are necessary to cover the probability mass even with spaces of up to ten millions of reconciliations.

Another argument for the concentration of the whole probability mass around the MPR is how fast it increases for a subtree $\mathcal{T}_d(G, S)$ according to a given depth d. We performed such analysis on the 24 A-trees G for which $D(\mathcal{T}(G, S))$ belongs to the highest depth range (that is between 55 and 59), which are the ones with the largest required depth and the smallest probability $P(\alpha_{min}|G)$ (see Figures 5(left) and 3), and as we can see in Figure 7 below, with a depth d as small as 2, the probability mass covered by $\mathcal{T}_d(G, S)$ is almost equal to one.

Although the immediate neighborhood $\mathcal{T}$ of α_{min} of depth $d^\circ(\mathcal{T}(G, S))$ technically covers the probability mass of the whole space of reconciliations between G and S, the probability of each reconciliation (that is reconciled tree) located beyond this minimal depth may have a non negligible contribution in the computation of the (exact) probability $P(G)$. This question is important due to the difference in terms of computational complexity between the exact posterior probability $P(\alpha|G)$

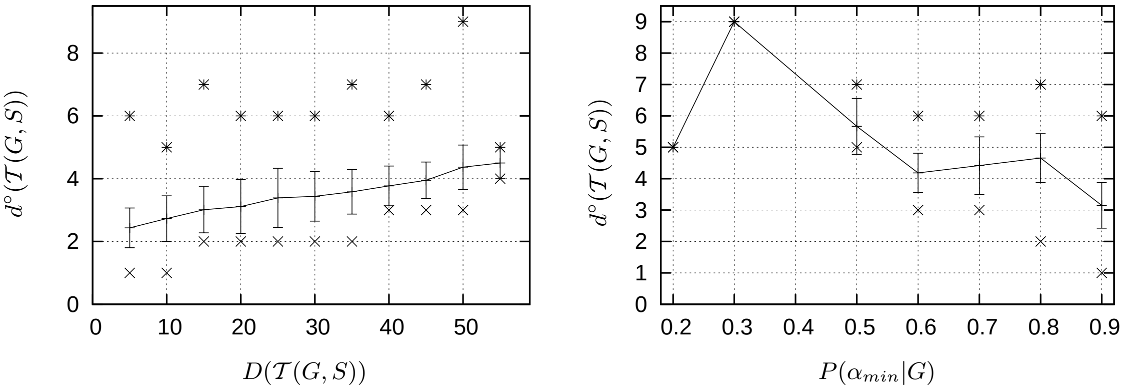

Fig. 5. Over all 1278 A-trees G, **(left)** average depth $d^\circ(\mathcal{T}(G,S))$ (y axis) for each space tree $\mathcal{T}(G,S)$ for which $D(\mathcal{T}(G,S))$ belongs to the same depth range (x axis), together with standard deviation, minimum and maximum depth, and **(right)** average depth $d^\circ(\mathcal{T}(G,S))$ (y axis) according to the probability of α_{min} (x axis; gene trees are grouped by classes for which $P(\alpha_{min}|G)$ divided by 0.1 is equal), together with the standard deviation, minimum and maximum depth.

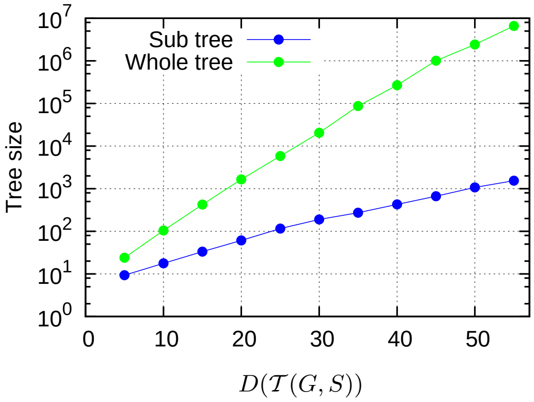

Fig. 6. Over all 1278 A-trees G, average size of the subtree $\mathcal{T}_{d^\circ}(G,S)$ and of the whole tree $\mathcal{T}(G,S)$ (y axis) for each space tree $\mathcal{T}(G,S)$ for which $D(\mathcal{T}(G,S))$ belongs to the same depth range (x axis).

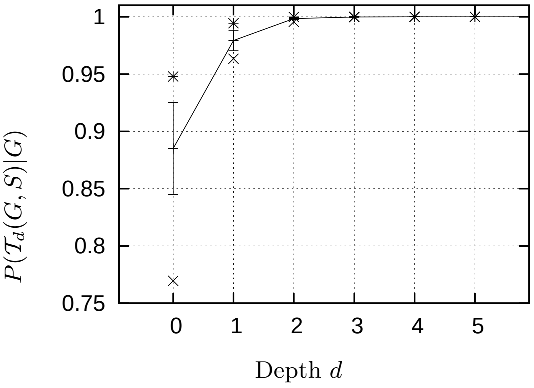

Fig. 7. Over all 24 A-trees such that $55 \leq D(\mathcal{T}(G,S)) \leq 59$, the average probability mass covered by $\mathcal{T}_d(G,S)$ (y axis) for the considered depth d (x axis).

for each visited reconciliation α (Theorem 1, point 1) and its $\mathcal{T}$-approximation $P_{\mathcal{T}}(\alpha|G)$ (Theorem 1, point 2). To assess this point, we compared the exact probability $P(G)$ and its $\mathcal{T}$-approximation $P_{\mathcal{T}}(G)$, where the error ratio of the latter according to the former is $1 - P_{\mathcal{T}}(G)/P(G)$. The results are depicted in Figures 8, 9 and 10 below: (1) the error ratio is inversely proportional to the depth d of the considered subspace $\mathcal{T}$; (2) with a depth d as small as 1, the average ratio is 0.02 and the average number of visited reconciliations is 10, and (3) for each gene tree, the approximation $P_{\mathcal{T}}(G)$ computed with a subtree $\mathcal{T}$ of depth $d \geq 8$ is equal to $P(G)$.

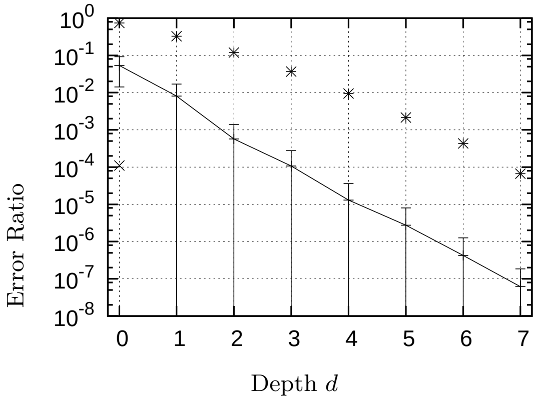

Fig. 8. Over all 1278 A-trees G, the error ratio of the approximated probability $P_{\mathcal{T}}(G)$ (y axis) for the subtree $\mathcal{T}$ (of $\mathcal{T}(G,S)$) of depth $d \in \{0, 1, \ldots, 7\}$ (x axis).

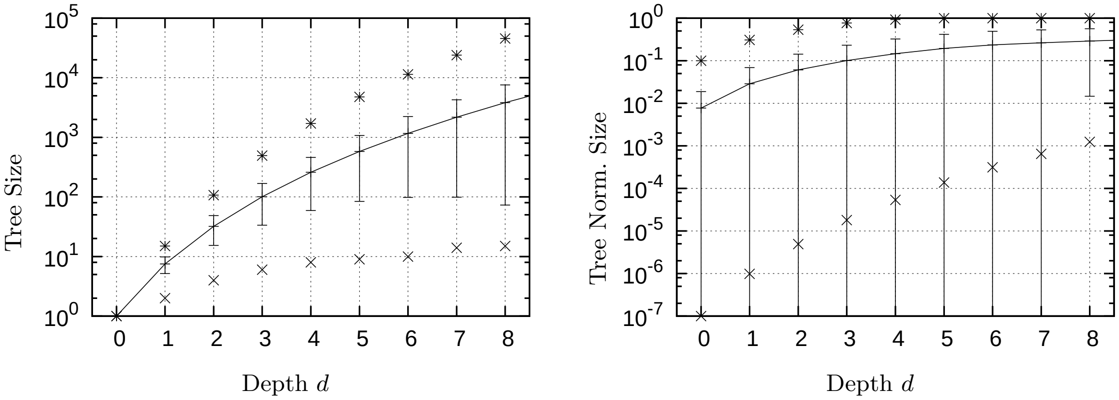

Fig. 9. Over all 1278 A-trees G and for each depth $d \in \{0, 1, \ldots, 8\}$ (x axis), the average size of the corresponding subtree of $\mathcal{T}(G,S)$ (y axis) both in absolute value (**left**) and normalized by the number of reconciliations (**right**).

3.2 Synthetic gene trees

With the duplication and loss rates used above for the 1278 real gene trees, the conclusion is that the immediate neighborhood of the MPR mostly covers the probability mass of the whole tree of

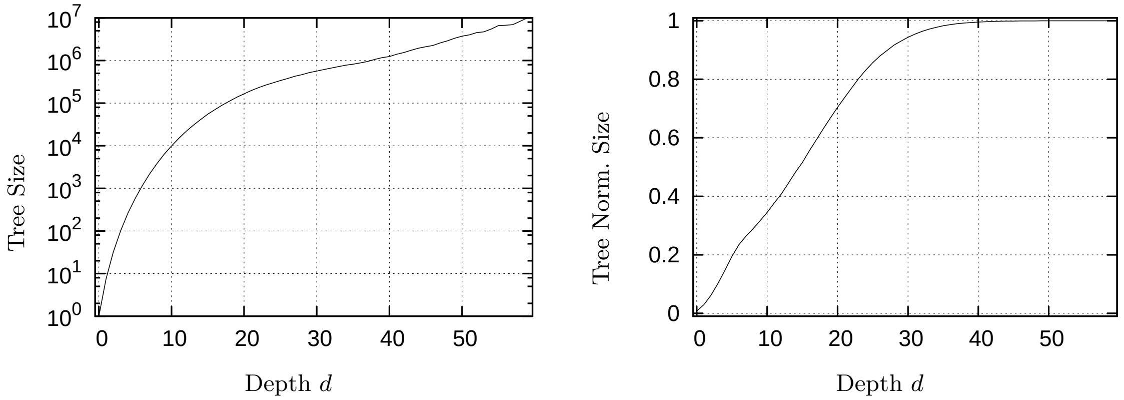

Fig. 10. Same than Figure 9, but for all possible depths.

reconciliations. The question that we address now is whether or not this is true for gene trees that would be generated with higher rates? As we can see in Table 1 below, increasing the duplication and loss rates suggests that the average probability of the MPR α_{min} decreases and the frequency where it is not the most likely one (denoted by α^*) increases. Figures 11 to 14 below show that with higher duplication and loss rates, the probability mass is more evenly dispersed among the reconciliations, but the highest concentration is still located among the most parsimonious evolutionary scenarios and a small subset of reconciliations need to be explored to cover the probability mass.

I.F.	Nb. of gene trees	Average $P(\alpha_{min}\|G)$	% of the Nb. of G s.t. $\alpha_{min} \neq \alpha^*$
1	1051	0.97876	0.09% (1)
1.4	1025	0.95234	0.78% (8)
1.8	924	0.91781	1.51% (14)

Table 1. For each I.F.: the number of gene trees (A-trees) generated, the average probability $P(\alpha_{min}|G)$, and the number of gene trees G such that $\alpha_{min} \neq \alpha^*$. The maximal NMC distance between α_{min} and α^* is 2.

4 Discussion and Conclusion

In this work, we presented an efficient algorithm to compute, either exactly or approximately, the posterior probabilities of a subspace of the space of all reconciliations between a given gene tree and a given species tree, provided the gene duplication and loss rates are known. Based on this algorithm, we were able to explore large reconciliations spaces both for real and simulated datasets and we showed that, for realistic species tree and duplication and loss rates, only a very small subset of reconciliations need to be explored to obtain in a very short time very precise approximations of the posterior probabilities of the most likely reconciliations. Such computational speed-up allows to analyze gene families with potentially a very large number of reconciliations, as we demonstrated in our experiments. It can also have applications in a Bayesian framework where duplications and losses rates are estimated using an MCMC approach [21], by reducing the time required in each state of the Markov Chain, for given duplication and loss rates. Our second contribution is experimental.

With gene families from 12 fungal genomes and realistic duplication and loss rates along the corresponding species phylogeny, our analysis on the probabilistic landscape of the space of rec-

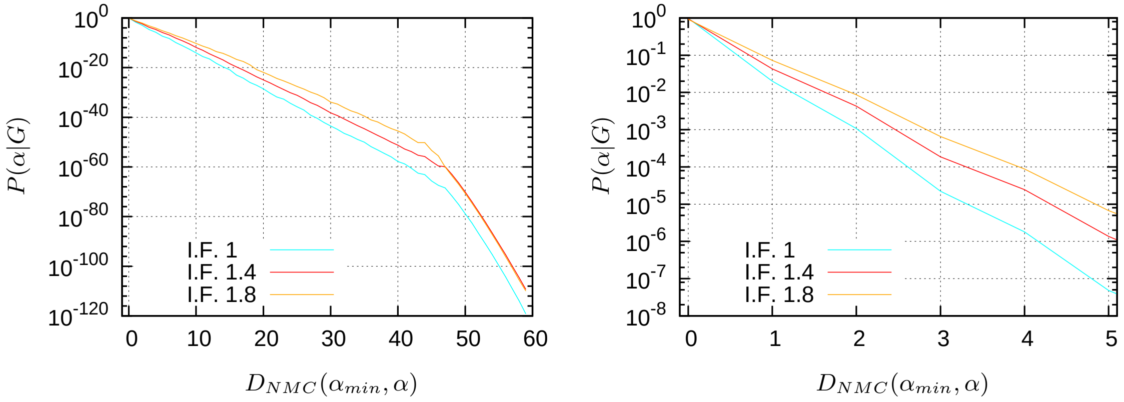

Fig. 11. For each I.F and the considered gene trees (A-trees), (**left**) average probability $P(\alpha|G)$ (y axis) of each reconciliation $\alpha \in \mathcal{T}(G,S)$ at the distance $D_{NMC}(\alpha_{min}, \alpha)$ to α_{min} (x axis), (**right**) zoom on depth less than or equal to 5.

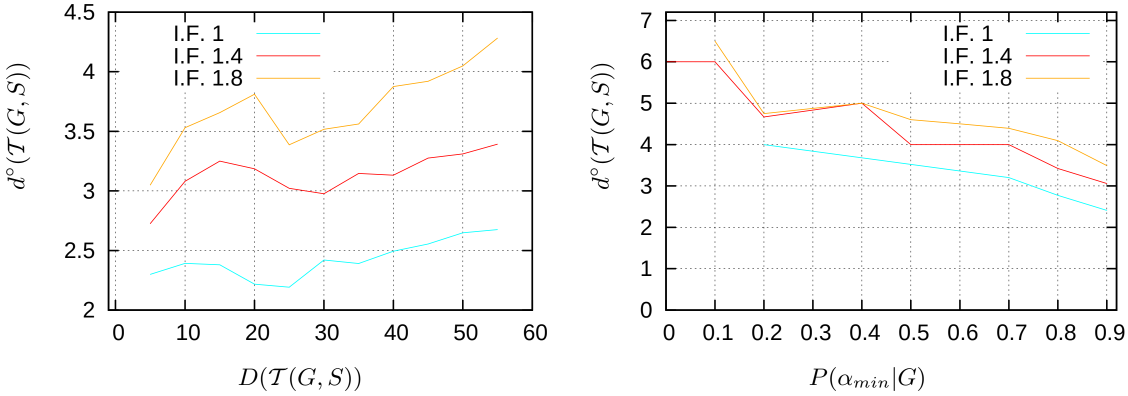

Fig. 12. For each I.F and the considered gene trees (A-trees), **(left)** average depth $d^\circ(\mathcal{T}(G,S))$ (y axis) for each space tree $\mathcal{T}(G,S)$ for which $D(\mathcal{T}(G,S))$ belongs to the same depth range (x axis), and **(right)** average depth $d^\circ(\mathcal{T}(G,S))$ (y axis) for each space tree $\mathcal{T}(G,S)$ for which $P(\alpha_{min}|G)$ belongs to the same probability range (x axis).

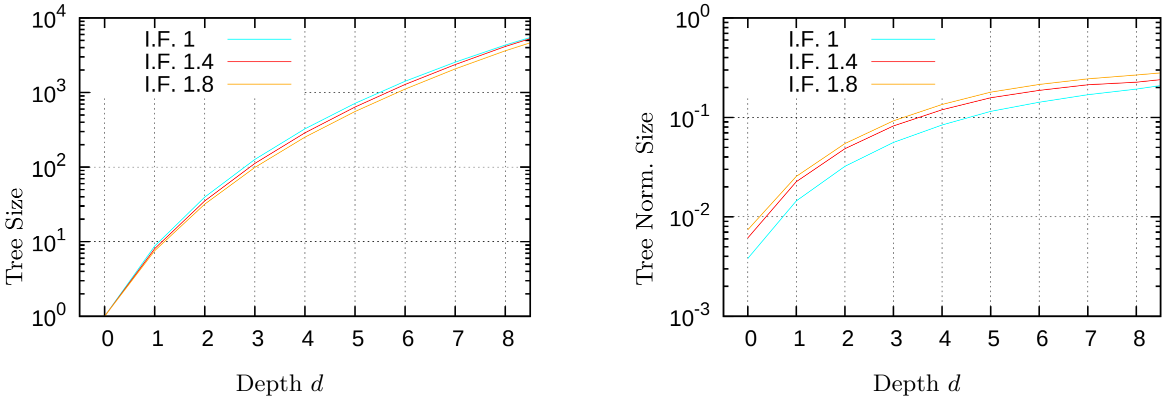

Fig. 13. For each I.F and the considered gene trees (A-trees), and for each depth $d \in \{0, 1, \ldots, 8\}$ (x axis), the average size of the corresponding subtree of $\mathcal{T}(G,S)$ (y axis) both in absolute value (**left**) and normalized by the number of reconciliations (**right**).

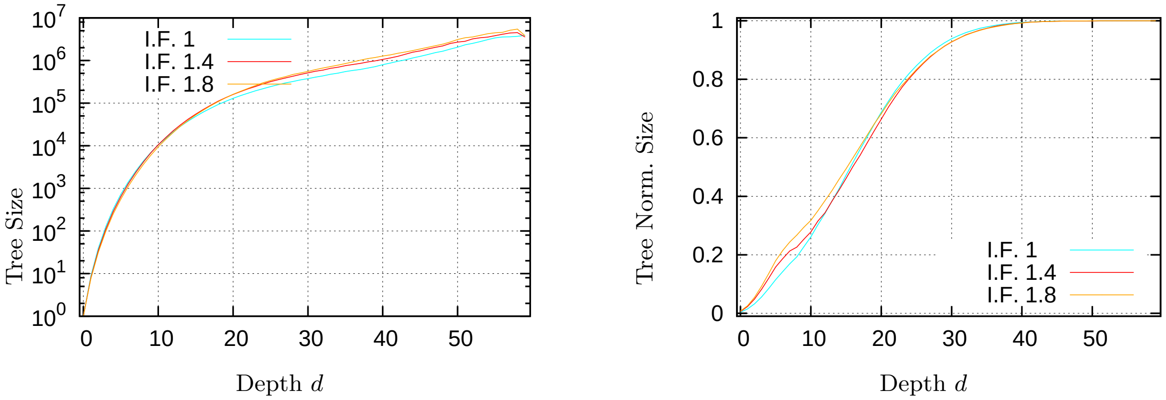

Fig. 14. Same than Figure 10, but for all possible depths.

onciliations show that the more probable is a reconciliation, the more it is located close (in term of NMC operators) to the MPR and the immediate neighborhood of the latter covers most of the whole space probability mass. For synthetic gene trees generated with higher rates, although its probability mass is more evenly dispersed, the same property holds. We believe these results offer the first detailed probabilistic analysis of the space of reconciliations and, together with the recent works of Sennblad *et al.* [2, 21, 1], clearly indicate that the probabilistic analysis of gene family evolution is applicable to large datasets, even if more experiments have to be done with different data (higher rates along the considered phylogeny, different duplication and loss rates along a given branch, larger species and gene trees). Recent works on ancestral genome reconstruction for example [18] could benefit from such algorithms.

It would also be important to study different types of data (duplication and loss rates and species phylogeny) where the most probable reconciliation is located far from the MPR and the probability mass of the whole space is not concentrated around a single reconciliation. For such problematic data, it would also be of particular interest to take a look over the generated gene trees (gene family profiles and sizes) to point out gene tree characteristics for which our approach would not be relevant. As a major problem of a Markov Chain Monte Carlo approach is caused by the presence of several peaks in the probability distribution, which forces it to stay in a small region of the space for a long period of time, these informations would obviously be useful when such an approach is used to approximate the posterior probabilities mentioned above, with prior on the rates. With a similar Bayesian framework, our observations can also be useful to develop a MCMC approach to approximate the posterior probabilities of duplication and loss rates given one (or more) gene tree, and that will be an alternative to the Expectation-Maximization approach of [5].

One fundamental application of such probabilistic analysis could be to detect and correct wrong gene trees. One of the major problems in using gene families trees is related to uncertainties and errors in such trees. Hahn illustrated this problem, in a parsimony framework, in [14], while [1] illustrates with the same fungal gene families we considered in a probabilistic framework. One possible approach to detect possible erroneous gene trees could then be to search the neighborhood of a given gene tree G (in terms of operations such as Nearest Neighbor Interchange [3] or the Tree Pruning and Regrafting [4]) and to see if some of its neighbors has a higher probability. This would require to design efficient algorithms to update efficiently the probability of a gene tree after such an operation is performed.

A Definitions

Let $\sigma : L(G) \to L(S)$ be the function that maps each leaf of G to the unique leaf of S with the same label. The *LCA-mapping* $M : V(G) \to V(S)$ maps each vertex u of G to the unique vertex $M(u)$ of S such that $\Lambda(S_{M(u)})$ is the smallest cluster of S containing $\Lambda(G_u)$.

Given two cells (either a vertex or an edge) c and c' of a tree T, $c' \leq_T c$ (resp. $c' <_T c$) if and only if c is on the unique path from c' to $r(T)$ (resp. and $c \neq c'$), in that case c' is said to be a (resp. strict) descendant of c.

Definition 1. A reconciliation between a gene tree G and a species tree S is a mapping $\alpha : V(G) \to V(S) \cup E(S)$ such that

1. (Base constraint) $\forall u \in L(G), \alpha(u) = M(u) = \sigma(u)$.
2. (Tree Mapping Constraint) For any vertex $u \in V(G) \setminus L(G)$,
   (a) if $\alpha(u) \in V(S)$, then $\alpha(u) = M(u)$.
   (b) If $\alpha(u) \in E(S)$, then $M(u) <_S \alpha(u)$.
3. (Ancestor Consistency Constraint) For any two vertices $u, v \in V(G)$, such that $v <_G u$,
   (a) if $\alpha(u), \alpha(v) \in E(S)$, then $\alpha(v) \leq_S \alpha(u)$,
   (b) otherwise, $\alpha(v) <_S \alpha(u)$.

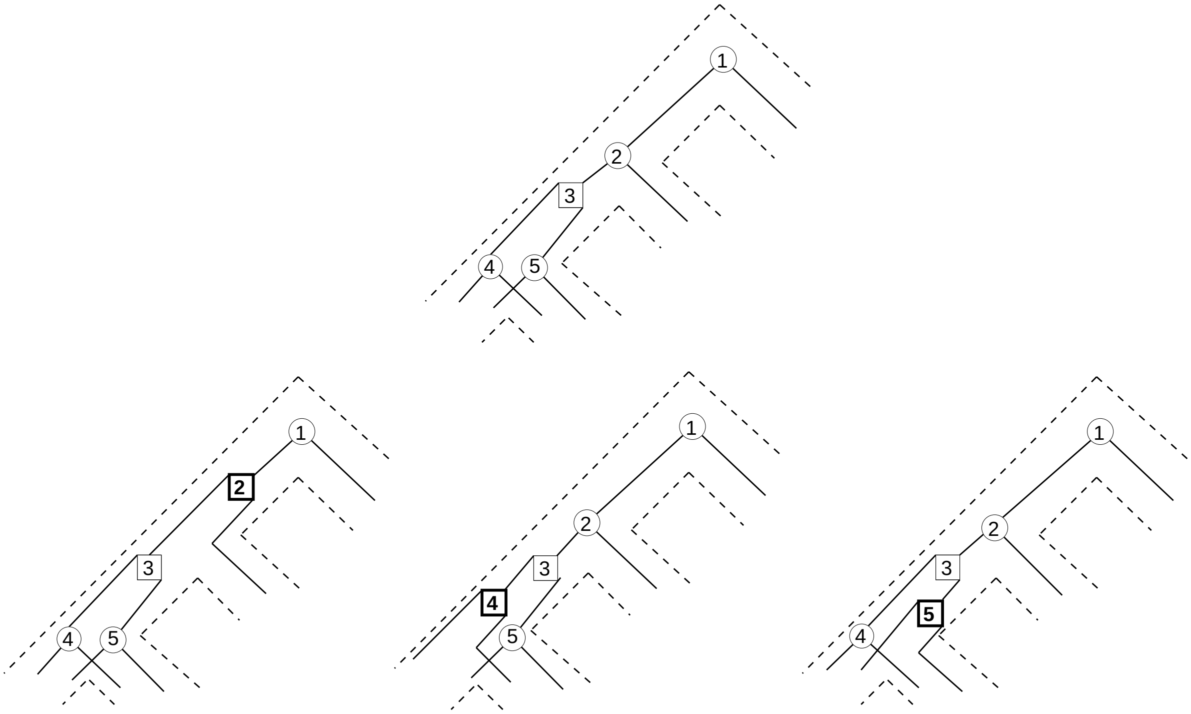

Fig. 15. The subtree of $\mathcal{T}(G, S)$ rooted at α_{min} for the trees G and S depicted in Figure 1. α_{min} and its children respectively are at the top and bottom of the figure. For each child, the vertex that has been moved upward is in boldface.

B Input Data

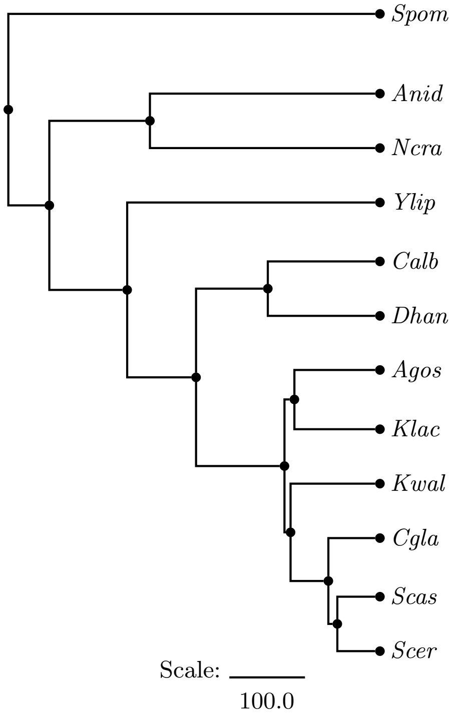

Fig. 16. The species tree S for the 12 fungal genomes, where divergence time is in Million Years.

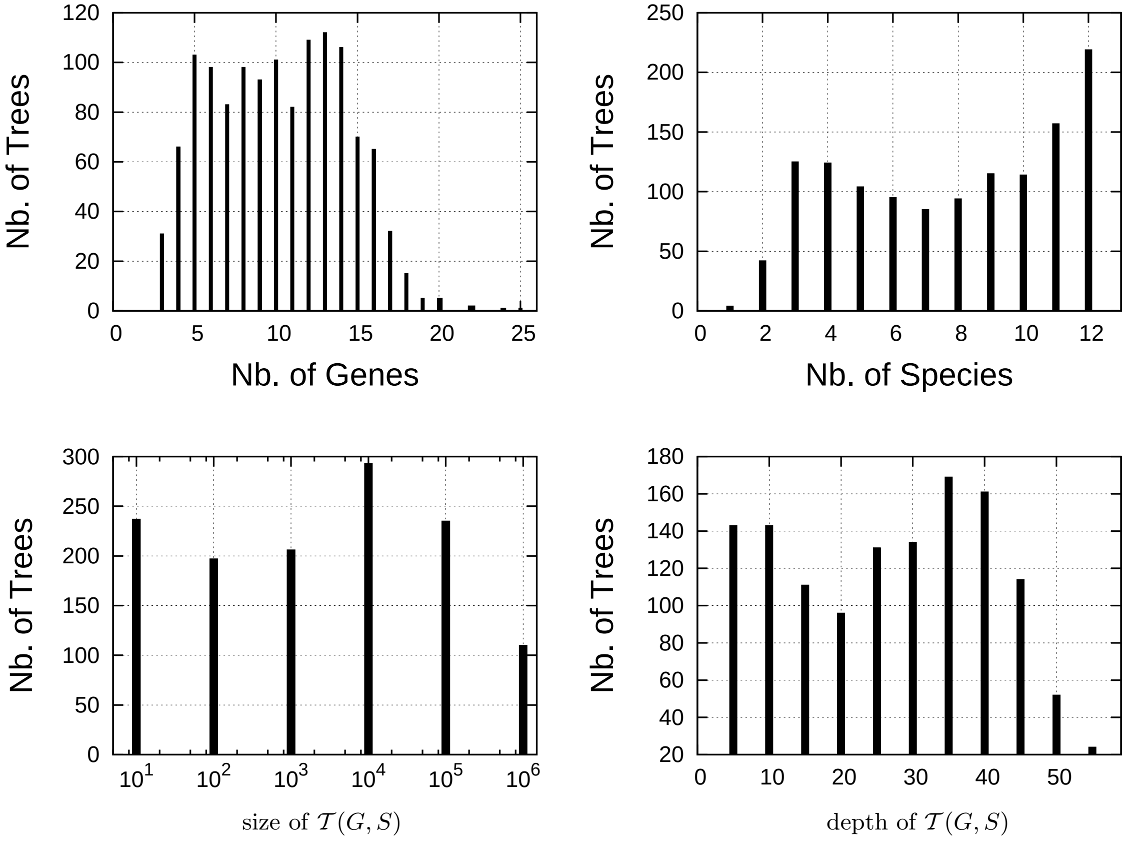

Fig. 17. Distribution of the 1278 real and A-trees G according to the number of genes and species present in G (**above**) and the size and depth of the space tree $\mathcal{T}(G,S)$ (**below**).

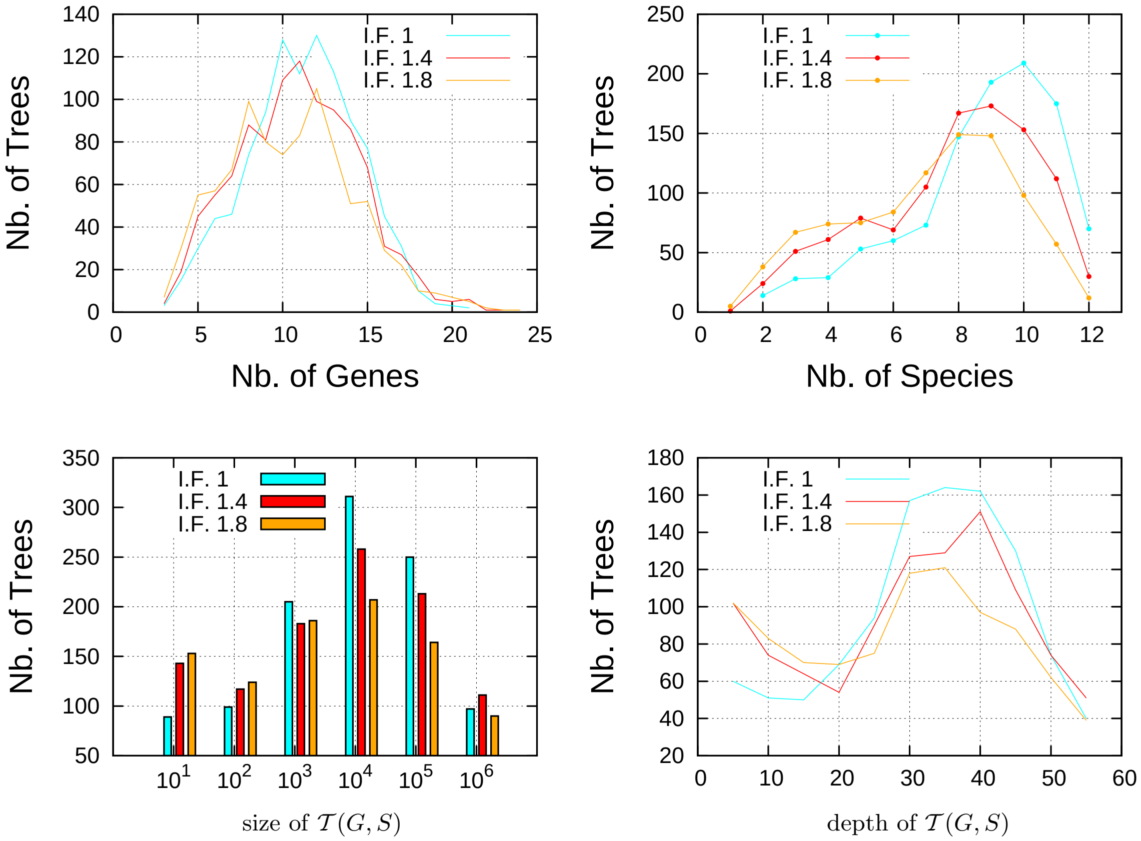

Fig. 18. For each I.F., distribution of the considered synthetic gene trees G according to the number of genes and species present in G (**above**) and the size and depth of the space tree $\mathcal{T}(G, S)$ (**below**).

Acknowledgements

C.C. and S.H. are supported, through the Discovery Grants program, by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank H. Philippe for providing the branch lengths of the fungi species tree.

References

1. Örjan Akerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc. National Academy Sci. U.S.A.*, 106(14):5714–5719, 2009.
2. Lars Arvestad, Jens Lagergren, and Bengt Sennblad. The gene evolution model and computing its associated probabilities. *J. ACM*, 56(2):1–44, 2009.
3. Mukul S. Bansa, Oliver Eulenstein, and Andre Wehe. The gene-duplication problem: Near-linear time algorithms for nni based local searches. *IEEE/ACM Trans. Comput. Biol. and Bioinform.*, 6(2):221–231, 2009.
4. Mukul S. Bansal and Oliver Eulenstein. An $\omega(n^2/\log n)$ speed-up of tbr heuristics for the gene-duplication problem. *IEEE/ACM Trans. Comput. Biol. and Bioinform.*, 5(4):514–524, 2008.
5. Tijl De Bie, Nello Cristianini, Jeffrey P. Demuth, and Matthew W. Hahn. Cafe: a computational tool for the study of gene family evolution. *Bioinformatics*, 22(10):1269–1271, 2006.
6. Paola Bonizzoni, Gianluca Della Vedova, and Riccardo Dondi. Reconciling a gene tree to a species tree under the duplication cost model. *Theoret. Comput. Sci.*, 347(1-2):36–53, 2005.
7. K.P. Byrne and K.H. Wolfe. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3):1341–1350, 2007.
8. J.P. Doyon, C. Chauve, and S. Hamel. Space of gene/species trees reconciliations and parsimonious models. *J Comput Biol*, 16(10):1399–1418, 2009.
9. Walter M. Fitch. Homology - a personal view on some of the problems. *Trends Genet.*, 16(5):227 – 231, 2000.
10. M. Goodman, J. Czelusniak, G.W. Moore, R.A. Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, 28:132–163, 1979.
11. PawełGórecki and Jerzy Tiuryn. Dls-trees: a model of evolutionary scenarios. *Theoret. Comput. Sci.*, 359(1):378–399, 2006.
12. A. De Grassi, C. Lanave, and C. Saccone. Genome duplication and gene-family evolution: the case of three oxphos gene families. *Gene*, 421(1-2):1–6, 2008.
13. Dan Graur and Wen-Hsiung Li. *Fundamentals of Molecular Evolution second edition*. Sinauer Associates, Sunderland, MA., 1999.
14. Matthew W. hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.*, 8:R141, 2007.
15. Harvey. *C++ How to Program (5th Edition) (How to Program)*. Prentice Hall, 2005.
16. Olivier Jeffroy, Henning Brinkmann, Frédéric Delsuc, and Hervé Philippe. Phylogenomics: the beginning of incongruence? *Trends Genet*, 22(4):225–31, Apr 2006.
17. David G. Kendall. On the generalized “birth-and-death” process. *Ann. Math. Statistics*, 19:1–15, 1948.
18. Jian Ma, Aakrosh Ratan, Brian J. Raney, Bernard B. Suh, Louxin Zhang, Webb Miller, and David Haussler. Dupcar: Reconstructing contiguous ancestral regions with duplications. *J. Comput. Biol.*, 15(8):1007–1027, 2008.
19. Roderic D. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, 43:58–77, 1994.
20. Michael Sanderson and Michelle McmMahon. Inferring angiosperm phylogeny from est data with widespread gene duplication. *BMC Evolutionary Biology*, 7(Suppl 1), 2007.
21. Bengt Sennblad and Jens Lagergren. Probabilistic orthology analysis. *Syst. Biol.*, 58(4):411–424, 2009.
22. J. Thorne, H. Kishino, and I. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, 15:1647–1657, 1998.
23. M.J. van Hoeck and P. Hogeweg. Metabolic adaptation after whole genome duplication. *Mol. Biol. Evol.*, 2009. To appear. DOI:10.1093/molbev/msp160.
24. Ilan Wapinski, Avi Pfeffer, Nir Friedman, and Aviv Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61, 2007.
25. K.H. Wolfe and D.C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997.