Accounting for Gene Tree Uncertainties Improves Gene Trees and Reconciliation Inference.

Thi-Hau Nguyen, Jean-Philippe Doyon, Stéphanie Pointet, Anne-Muriel Arigon Chifolleau, Vincent Ranwez, Vincent Berry

Accounting for gene tree uncertainties improves gene trees and reconciliation inference

Thi Hau Nguyen*+, Jean-Philippe Doyon*, Stéphanie Pointet*, Anne-Muriel Arigon Chifolleau*, Vincent Ranwez+, and Vincent Berry*

*LIRMM, Université Montpellier 2 - CNRS, France
+Montpellier SupAgro (UMR AGAP)

Abstract. We propose a reconciliation heuristic accounting for gene duplications, losses and horizontal transfers that specifically takes into account the uncertainties in the gene tree. Rearrangements are tried for gene tree edges that are weakly supported, and are accepted whenever they improve the reconciliation cost. We prove useful properties on the dynamic programming matrix used to compute reconciliations, which allows to speed-up the tree space exploration when rearrangements are generated by Nearest Neighbor Interchanges (NNI) edit operations. Experimental results on simulated and real data confirm that running times are greatly reduced when considering the above-mentioned optimization in comparison to the naïve rearrangement procedure. Results also show that gene trees modified by such NNI rearrangements are closer to the correct (simulated) trees and lead to more correct event predictions on average. The program is available at http://www.atgc-montpellier.fr/Mowgli/

1 Introduction

A phylogenetic tree or *phylogeny* is a tree depicting evolutionary relationships among biological entities that are believed to have a common ancestor. A gene family is a group of genes descended from a common ancestor that retains similar sequences and often similar functions. A species tree depicts the evolutionary history of a group of species, whereas a gene tree depicts the evolutionary history of a gene family. Gene trees and species tree are often inconsistent due to family-specific evolutionary events such as gene duplications, gene losses, horizontal gene transfers. By comparing gene trees with a species tree, reconciliation methods try to recover those major evolutionary events. Reconciliation is indeed the process of constructing a mapping between a gene tree and a species tree to explain their differences and similitudes with evolutionary events such as speciation ($\mathbb{S}$), duplication ($\mathbb{D}$), loss ($\mathbb{L}$), and horizontal gene transfer ($\mathbb{T}$) events. Reconciliations are most often inferred on the basis of a parsimony criterion: a cost is given to each event type, the cost of a reconciliation is the sum of the costs of the individual events it uses, and a reconciliation of minimum total cost is sought for. This computational problem is often called *Most Parsimonious Reconciliation*, or MPR in short, and many works have been devoted to it recently [1,2,3,4,5,6,7].

The first proposed models focused on reconciliations involving only duplications and losses (the DL model) [8,9,10] or only horizontal transfers and losses [11].Probabilistic methods have also been developed for the DL model, such as that of Arvestad et al. [12] (see Doyon et al. [13] for a review). Most recent works using a parsimony approach have been devoted to models incorporating duplications, losses and transfers all together (the DTL model) [1,2,4,6], which is necessary to handle prokaryotes. When accounting for transfer events, the history proposed by a reconciliation is consistent if, for any transfer, the donor and receiver species co-exist. Ensuring such a time consistency is difficult and leads to an NP-hard problem in the general case [14,5]. However, in the case divergence dates are available for nodes of the species tree, the problem becomes amenable [15,4]. The difficulty to handle transfers has led to a split within proposed DTL methods, namely those that ensure time-consistency [15,4] and those that don't [1,5,7]. The fastest algorithm for the later category runs in $O(mn \log n)$ where m and n are the sizes of the gene and species trees respectively [7], while the fastest time-consistent algorithm runs in $O(mn^2)$ [4].

A major problem, when applying reconciliation methods, is that parts of the gene trees can be incorrect. This leads reconciliation methods to overestimate ($\mathbb{S}$), ($\mathbb{D}$), ($\mathbb{L}$) and ($\mathbb{T}$) events [16,17]. Errors within the gene tree can be due to sequence alignment problems or phylogenetic reconstruction artifacts such as long branch attraction. Such errors are well-known in phylogenetics and several support measures, such as bootstrap values or bayesian posterior probabilities, have been proposed to detect unreliable edges in a gene tree. Up to now, very few works have solved the reconciliation problem in the presence of unsupported edges, and most of them consider only the DL model [18,19,16,20,21]. Durand et al. proposed an exponential exact algorithm to find the best rearrangement of a gene tree while preserving its strongly supported edges[16]. Some approaches collapse unsupported edges, leading to the creation of nodes with more than two children, called *polytomies* [18,19,20]. They then rely on a generalization of the least common ancestor mapping (LCA) to avoid the need for examining all possible binary rearrangements of the polytomies. In this way, Chang et al.[19], resp. Vernot et al.[20], proposed polynomial time algorithms when considering non-binary gene trees, resp. species trees. Berglund et al. proved that when dealing with species and gene tree that are both non-binary, the problem becomes NP-complete[18]. They thus proposed a heuristic approach tackling a variant of the MPR problem where duplications and losses are optimized separately. Durand et al. used Nearest Neighbor Interchange (NNI) edit operations to rearrange the local topology of the gene tree in the regions of low supports but, unlike Berglund et al., they optimized simultaneously duplications and losses. Chaudhary et al. investigated Subtree Prune and Regraft (SPR) and Tree Bisection and reconnection (TBR) edit operations to search for the gene tree rearrangement that minimizes the number of duplications, regardless of losses and transferts[21].

Due to transfer events, the LCA mapping can not be transposed to the DTL model and it seems hard to have an exact polynomial time algorithm for the MPR problem under this model even when the polytomies are present only

in the gene tree or in the species tree. Following the work of Berglund et al., Durand et al. for the DL model, we propose a heuristic method relying on NNI edit operations to search for a gene tree rearrangement that minimizes the cost of reconciliation to a fixed binary species tree, but in the context of the more complex DTL model.The resulting dynamic program, called MowgliNNI, is a generalization of Mowgli [4] a program initially developed for fixed binary gene trees. Experiments on simulated data show that MowgliNNI provides better $\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$, $\mathbb{S}$ prediction while improving gene tree inference, i.e. the modified gene tree is closer to the true evolutionary history of the gene family. Experiments on real data show a significant decrease in number of events and in the number of most parsimonious reconciliations.

2 Preliminaries

Trees considered in this paper are rooted and only labeled at their leaves, each leaf being labeled with the name of a studied species. Given a tree T, its nodes, edges, leaves and root are resp. denoted $V(T)$, $E(T)$, $L(T)$ and $r(T)$. The label of a leaf u of T is denoted by $\mathcal{L}(u)$ and the set of labels of leaves of T is denoted by $\mathcal{L}(T)$. When a node u has two children, they are denoted u_1 and u_2. Given two nodes u and v of T, $u \leq_T v$ (resp. $u <_T v$) if and only if v is on the unique path from u to $r(T)$ (resp. and $u \neq v$); if neither $u <_T v$ nor $v <_T u$ then u and v are said to be *incomparable*. As we consider rooted trees T only, we adopt the convention than an edge denoted (u, v) means that $v <_T u$. For a node u of T, T_u denotes the subtree of T rooted at u, u_p the parent node of u, while (u_p, u) is the *parent edge* of u. A tree T' is a *refinement* of a tree T if T can be obtained from T' by collapsing some edges in T', *i.e.* by merging the two extremities of these edges [22].

A *species tree* is a rooted binary tree depicting the evolutionary relationships of ancestral (internal nodes) species leading to a set of extant (leaf) species. A species tree S is considered here to be *dated*, that is associated to a time function $\theta : V(S) \to \mathbb{R}^+$ such that $y <_S x$ implies that $\theta(y) < \theta(x)$. Such times are usually estimated on the basis of molecular sequences [23] and fossil records. Note that to ensure the time consistency of inferred transfers, absolute dates are not required, the important information being the ordering of the nodes of S induced by the dating. Given a dated binary species tree S, the reconciliation model we rely on considers a variant of S called a *subdivision* and denoted S' (as done also in [24,3,4]). The subdivision S' is constructed from S as follows: for each node $x \in V(S) \setminus L(S)$ and each edge $(y_p, y) \in E(S)$ s.t. $\theta_S(y_p) > \theta_S(x) > \theta_S(y)$, an *artificial* node w is inserted along the edge (y_p, y) in S', with $\theta'_{S'}(w) = \theta_S(x)$.

A *gene tree* is a rooted binary tree explaining the evolutionary history of a gene family, that lead to a set of homologous sequences observed in current organisms. Each leaf of a species tree has a unique label, corresponding to a specific extant sequence of the gene. Though, several leaves of a gene tree can be associated to a same species due to duplication and transfert events. We denote by $s(u)$ the species associated to leaf $u \in V(G)$. Each edge (u, v) of $E(G)$ can be

uniquely identified by the subset $\mathcal{L}(T_v) \subseteq \mathcal{L}(G)$. A gene tree G *with supports* is a gene tree whose *internal* edges have a support value. Let $wk_t(G) \subseteq E(G)$ be the set of edges having a support value weaker than threshold t and let $str_t(G)$ be $E(G) - wk_t(G)$, that is the edges having a support equal or stronger than t.

Finding the most parsimonious reconciliation. Reconciling a (binary) gene tree G with a species tree S means building a mapping α that associates each gene $u \in V(G)$ to a sequence $\alpha(u)$ of nodes in the subdivision S'. The $\alpha(u)$ sequence models the evolution of gene u along S' with the following atomic events: ($\mathbb{C}$) contemporary gene, ($\mathbb{S}$) speciation, ($\mathbb{D}$) duplication, ($\mathbb{T}$) transfer, ($\mathbb{SL}$) speciation followed by a loss, ($\mathbb{TL}$) transfer followed by a loss for the donor, and ($\varnothing$) going from an artificial node of S' to its only child. Observe that each loss is coupled with either a speciation ($\mathbb{SL}$) or a transfer ($\mathbb{TL}$). Indeed, any most parsimonious reconciliation only needs to use a loss when it meets a speciation node of S' where G goes into only one descending edge, or when leaving an edge due to a transfer, with no part of G remaining in the donor edge. For more details, we refer the reader to Definition 3 of Doyon et al. [4] that we follow for reconciliation, except that the mapping α considered here concerns nodes instead of branches.

The *cost* of a reconciliation α is denoted $cost(\alpha) = d\delta + t\tau + l\lambda$, where δ, τ, and λ respectively denote the cost of $\mathbb{D}$, $\mathbb{T}$, and $\mathbb{L}$ events, and d, t, and l denote the number of the corresponding events in α. Moreover, a $\mathbb{TL}$ event is atomic and costs $(\tau + \lambda)$ and a $\mathbb{SL}$ costs λ. The *optimal reconciliation cost* is denoted $C(G, S') = \min\{cost(\alpha) : \alpha \text{ is a reconciliation between } G \text{ and } S'\}$.

The optimal cost for mapping a node u of G on a vertex x of S' is defined according to the minimal cost among the events $\mathbb{C}$, $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, $\varnothing$, and $\mathbb{SL}$, together with the cost of a $\mathbb{TL}$ event, which are denoted $c_{\overline{\mathbb{TL}}}(u, x)$ and $c_{\mathbb{TL}}(u, x)$, respectively (see Definition 1 below). This directly follows from the dynamic programming algorithm that computes the optimal cost $C(G, S')$, where the computation of the cost for a $\mathbb{TL}$ event follows that of the other six atomic events, since a $\mathbb{TL}$ event is followed by a $\mathbb{C}$, $\mathbb{S}$, $\mathbb{D}$, $\mathbb{T}$, $\varnothing$, or $\mathbb{SL}$ event [4].

To ensure time consistency of $\mathbb{T}$ and $\mathbb{TL}$ events, the donor $x \in V(S')$ and the receiver $y \in V(S')$ have to be located at the same *time slice* $h(x) = h(y)$ of S' (the term time slice of a vertex refers to its height in S').

These intricated notions are formally detailled in definitions 1 and 2. Though, these definitions depend on one another, there is no circularity as either we progress in the gene tree or we switch from a $\mathbb{TL}$ event to a non-$\mathbb{TL}$ event.

Definition 1 (Reconciliation cost matrix). *Consider a gene tree G and the subdivision S' of a species tree S. Let $c : V(G) \times V(S') \to \mathbb{R}^+$ denote the cost matrix recursively defined as follows for a node u of G and a vertex x of S': $c_{\overline{\mathbb{TL}}}(u, x) = \min\{c_{\mathbb{E}}(u, x) : \mathbb{E} \in \{\mathbb{C}, \mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}\}\}$ and $c(u, x) = \min\{c_{\mathbb{TL}}(u, x), c_{\overline{\mathbb{TL}}}(u, x)\}$, where the costs $c_{\mathbb{E}}(u, x)$ for $\mathbb{E} \in \{\mathbb{C}, \mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}, \mathbb{TL}\}$ are defined below.*

- $c_{\mathbb{C}}(u, x) = 0$, *if* $u \in L(G)$, $x \in L(S')$ *and* $\mathcal{L}(x) = s(u)$.

- $c_{\mathbb{S}}(u,x) = \min\{c(u_1,x_1) + c(u_2,x_2), c(u_1,x_2) + c(u_2,x_1)\}$
  if $u \notin L(G)$ *and* $x \notin L(S')$.
- $c_{\mathbb{D}}(u,x) = c(u_1,x) + c(u_2,x) + \delta$, *if* $u \notin L(G)$.
- $c_{\mathbb{T}}(u,x) = \min\{c(u_1,x) + c(u_2,z), c(u_1,y) + c(u_2,x)\} + \tau$
  if u *has two children and where* z *(resp.* y*) denotes* $BR_{\mathbb{T}}(u_2,x)$ *(resp.* $BR_{\mathbb{T}}(u_1,x)$*).*
- $c_{\varnothing}(u,x) = c(u,x_1)$, *if* x *has a single child. event)*
- $c_{\mathbb{SL}}(u,x) = \min\{c(u,x_1), c(u,x_2)\} + \lambda$, *if* x *has two children.*
- $c_{\mathbb{TL}}(u,x) = c_{\overline{\mathbb{TL}}}(u,y) + \tau + \lambda$, *where* y *denotes* $BR_{\mathbb{TL}}(u,x)$.

If the above constraints for an event $\mathbb{E} \in \{\mathbb{C}, \mathbb{S}, \mathbb{D}, \mathbb{T}, \varnothing, \mathbb{SL}, \mathbb{TL}\}$ *on node* u *and vertex* x *are not respected, the corresponding cost* $c_{\mathbb{E}}(u,x)$ *is set to* ∞.

Definition 2 (Best receiver). *Consider a node* u *of* G *and a vertex* x *of* S'. *Let* $BR_{\mathbb{T}}(u,x)$ *denote a vertex* y *of* S' *that minimizes* $c(u,y) = \min\{c(u,z) : z \in V_{h(x)}(S')$ *and* $z \neq x\}$. *Similarly, let* $BR_{\mathbb{TL}}(u,x)$ *denote a vertex* y *of* S' *that minimizes* $c_{\overline{\mathbb{TL}}}(u,y) = \min\{c_{\overline{\mathbb{TL}}}(u,z) : z \in V_{h(x)}(S')$ *and* $z \neq x\}$

The value $c(u,x)$ is the optimal cost when mapping gene node u to node x in S' or on the edge above it. The optimal cost for reconciling G with S', denoted $C(G,S')$, is then $min_{x \in V(S')}(c(r(G),x)$. The algorithm of Doyon et al.[4], called *Mowgli*, fills the dynamic programming cost matrix $c : V(G) \times V(S') \to \mathbb{R}^+$ by two embedded loops that visit all slices of S' in backward order and nodes of G in postorder. Due to an optimization in precomputing the best receiver edge for transfer events of nodes u in a given time slice, this algorithm runs in $O(|S|^2.|G|)$ time and space.

The problem considered in this paper is the following:

Most Parsimonious Reconciliation Gene Tree (MPR-GT)
INPUT: a dated species tree S with a time function θ_S, a gene tree G with supports on the same set of species, costs δ, τ, resp. λ for $\mathbb{D}$, $\mathbb{T}$, resp. $\mathbb{L}$, and a threshold t.
OUTPUT: a tree G' s.t. $str_t(G) \subseteq E(G')$ and $C(G',S')$ is minimum among all such trees.

3 Methods

We describe here a heuristic for the MPR-GT problem that relies on a hill-climbing strategy to seek a (rooted) gene tree G of minimum reconciliation cost (see Def. 1) using NNI edit operations [25].

Performing an NNI operation around an *internal* edge (w,v) means swapping the position of one of the two subtrees connected to v with that of the subtree connected to the sibling of v. Given an initial gene tree G and an edge (w,v) of G, two "alternative" trees can be obtained from G by performing an NNI operation around (w,v). (see Fig. 1 for an example). The hill-climbing proceeds as follows: (1) select a weak edge of G; (2) compute the reconciliation cost for the two alternative gene trees obtained by NNI on that edge; (3) if none of these trees decreases the reconciliation cost, then try another weak edge; if none of the weak edges allows to progress, then G is a local minimum and the hill climbing stops; (4) otherwise one of the alternative gene trees leads to a decrease

in reconciliation cost, and the above process continues with the alternative tree of minimum reconciliation cost. MowgliNNI outputs the final binary rearrangement along with its most parsimonious reconciliation. In the worst cases, MowgliNNI examines all unreliable edges and does not find any better binary rearrangement of the given gene tree G since the topology G is already (locally) optimal.

Consider now the time complexity of *MowgliNNI*. Identifying the weak edges is done in $O(|G|)$ and generating the two alternative gene trees for an NNI operation is done in constant time. Hence, the complexity bottleneck of *MowgliNNI* is the number of times (denoted N) the $\Theta(|S|^2 \cdot |G|)$ *Mowgli* algorithm is called. Overall, the time complexity of *MowgliNNI* is $\Theta(|S|^2 \cdot |G| \cdot N)$. The next section describes how we can avoid recomputing large parts of the cost matrix, and hence greatly reduce the running time of *MowgliNNI*.

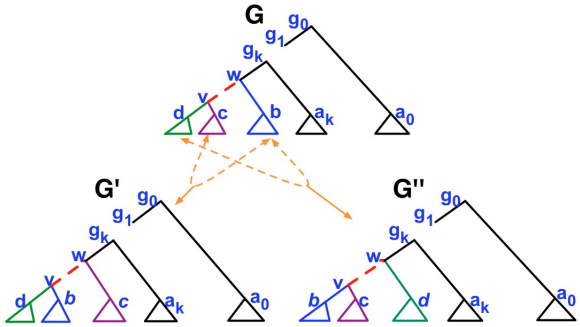

Fig. 1. A gene tree G with a weak edge (w, v) selected for an NNI. v is connected to two subtrees G_c and G_d, while w is connected to v and to the subtree G_b. Performing an NNI operation around (w, v) means exchanging subtree G_b with either G_c or G_d, leading to trees G' and G" respectively.

Combinatorial optimization. We now present results that take advantage of the way the dynamic programming matrix is computed (Def. 1) to avoid recomputing from scratch the cost matrix associated to a gene tree G' obtained by an NNI edit operation from a gene tree G. Consider the gene tree G of Figure 1, the NNI operation applied on edge (w, v) that swaps the two subtrees G_b and G_c, and the resulting gene tree denoted G'. We can observe that despite the global architecture of G and G' differs, the local architectures of subtrees $G_b, G_c, G_d, G_{a_0}, \ldots G_{a_k}$ remain unchanged. Hence, any cost that differs between the matrices $c : V(G) \times V(S') \to \mathbb{R}^+$ and $c' : V(G') \times V(S') \to \mathbb{R}^+$ (see Definition 1) is located in a column (i.e. node of the gene tree) associated to an ancestor of v (including v itself). For each of those nodes, there is two cases: (i) the node belongs to the NNI edge and its two children have subtree that have been modified (e.g. nodes w and v); (ii) the node is a strict ancestor of the NNI edge (w, v) and has exactly one child with a subtree that has been modified (e.g. $g_k, \ldots, g_0$). Lemma 1 below indicates which columns of the cost matrix don't need to be recomputed.

Algorithm 1 $MowgliNNI(G, c)$: seek a gene tree G' of minimum reconciliation cost, starting from a gene tree G and the precomputed matrix reconciliation cost $c : V(G) \times V(S') \to \mathbb{R}^+$, where S' is the subdivided species tree.

1: **for all** edges $(w, v) \in wk_t(G)$ **do**
2: For each node s of G that is not an ancestor of v, set the column $c'(s, \cdot)$ to $c(s, \cdot)$.
3: For each vertex x of S', recompute the cost $c'(v, x)$ according to Def. 1.
4: **for all** strict ancestors s of v according to a bottom-up traversal of G **do**
5: For each vertex x of S', recompute the cost $c'(s, x)$ according to Def. 1.
6: If $c(s, x) \leq c'(s, x)$ holds for each vertex x of S', then examine the next edge of loop at line 1 {the NNI rearrangement tree G' is refused}.
7: **end for**
8: Return $MowgliNNI(G', c')$ {The rearranged tree G' is accepted}.
9: **end for**
10: Return G {No successful rearrangement of G}

Lemma 1. *Consider a gene tree G, the subdivision S' of a species tree S, an edge (w, v) of G, and G' obtained from G by an NNI operation on (w, v). For each node z of G that is incomparable to v and for each vertex x of S', $c(z, x) = c'(z, x)$ holds.*

Unfortunately, there is no extension of Lemma 1 to ensure that when an edge has already been unsuccessful tried for an NNI it is useless to reconsider it later, even if it is a descendant in G of the edge leading to the last successful NNI.

Theorem 1. *Consider a gene tree G, the subdivision S' of a species tree S, an edge (w, v) of G, a gene tree G' obtained by an NNI operation on (w, v), and any strict ancestor u of w in G where the unique child of u that is an ancestor of w is u_1 w.l.o.g. (i.e. $w \leq u_1$ in both G and G'). If $c(u_1, x) \leq c'(u_1, x)$ holds for all $x \in V(S')$, then (1) $c(u, x) \leq c'(u, x)$ holds for all $x \in V(S')$; and (2) $C(G, S') \leq C(G', S')$.*

Computing the cost matrix $c' : V(G') \to V(S')$ given $c : V(G) \to V(S')$ is then achieved in worst-case time $O(|S'| \cdot h(G))$, where $h(G)$ is the height of G.

Theorem 2. MowgliNNI *has worst case running time $O(|S|^2 \cdot |G| + |S|^2 \cdot h(G) \cdot N)$.*

Indeed the steps of Algorithm 1 can be described as follows: initializing the reconciliation matrix for the initial gene tree is done in $O(|S|^2 \cdot |G|)$ time; updating the matrix for each NNI now only costs $O(|S'| \cdot h(G)) = O(|S|^2 \cdot h(G))$.

In *MowgliNNI* 's naïve implementation each rearrangement requires to recompute the cost associated to each and every node of the gene tree. In contrast, in the optimized version, an NNI around edge (w, v) is examined after updating only those costs associated to ancestral nodes of w. This has no impact on the worst case complexity (when the gene tree is a caterpillar $h(G)$ is in $O(|G|)$) but significantly reduces the running times in practice since in most cases the number of nodes in G is much larger than their average height. For some random tree models the average height of a node in an n-leaf tree is indeed proportional to $log(n)$ [26].

4 Experimental Evaluation

4.1 Experiments on simulated datasets

The phylogeny of 37 proteobacteria proposed by David and Alm [6] was used as a reference species tree. Along this tree, we simulated the evolutionary histories (denoted R_{True}) of 985 gene families (G_{True}), containing 10 to 100 genes, according to the birth and death process [27]. The *initial* gene trees (G_{ML}) were inferred from the simulated molecular sequences of length 1500 - 3000 bp by RAxML under GTR model [28]. *Mowgli* [4] and Ranger-dtl-D [7] were used to infer the most parsimonious evolutionary history (R_{ML}) between the initial gene tree and the reference species tree. Then, *MowgliNNI* was used to search for an alternative gene tree topology (G_{NNI}) of lower reconciliation cost, along with its most parsimonious evolutionary history (R_{NNI}). The cost of each $\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$ event considered in reconciliations was computed as follows:

$$Cost_{\mathbb{E}} = \begin{cases} \log(\frac{|\mathbb{D}_{R_{True}}|+|\mathbb{T}_{R_{True}}|+|\mathbb{L}_{R_{True}}|}{|\mathbb{E}_{R_{True}}|}) & \text{if } |\mathbb{E}_{R_{True}}| \neq 0 \\ \log(\frac{|\mathbb{D}_{R_{True}}|+|\mathbb{T}_{R_{True}}|+|\mathbb{L}_{R_{True}}|}{0.1}) & \text{otherwise} \end{cases} \quad (1)$$

where $\mathbb{E}_{R_{True}}$ (with $\mathbb{E}$ beeing $\mathbb{D}$, $\mathbb{T}$ or $\mathbb{L}$) stands for the true events of this type.

We explored the ability of *MowgliNNI* to improve the set of G_{ML} trees using six different bootstrap values as threshold for defining weak edges, i.e. 20, 40, 60, 80, 90, and 95. The G_{ML} trees were inferred from relatively long sequences, they thus contained a large proportion of high bootstrap values, *e.g.* more than 65% edges had a bootstrap value ≥ 80. Though this left only a moderate number of edges in each gene tree to be considered by *MowgliNNI* for rearrangement, the method was still able to improve their quality (see below).

Mowgli and Ranger-dtl-D showed a similar accuracy in inferring duplications and transfers (Fig. 2(a)), though Ranger-dtl-D proposed reconciliations with higher costs in 13% of the cases. As moreover the Ranger-dtl-D software does not provide the mapping of loss events, below we mainly compare the events inferred by *MowgliNNI* with those inferred by *Mowgli* and those of R_{True}.

Table 1 reports the accuracy of the G_{ML} and G_{NNI} trees. The number of families considered for improvement logically increases as the threshold for identifying an edge as weak increases (row 2 of Table 1). Even though bootstrap values in the initial gene trees are high on average, a large number of the 985 processed families are still subject to possible improvements, as for threshold 20, already 43% of the families are concerned, and this goes up to 99% of the families at threshold 95. The first conclusion is that there is a large number of cases where *MowgliNNI* can propose a modified gene tree. The percentage of cases where it actually did is provided in row 3 of Table 1, showing *e.g.* that at threshold 80, *MowgliNNI* proposed a new gene tree in 88% of the cases (853 cases over 965). Even for the lowest considered threshold of 20, a new gene tree is obtained for 79% of the families having weak edges, representing more than a third of the initial 985 families. These modified gene trees (G_{NNI}) represent an improvement over the initial trees (G_{ML}) since they are in most cases closer to the true gene trees (rows 4, 5, 6) and allow to obtain better reconciliations (rows 7, 8, 9). For instance at threshold 80, G_{NNI} is better in 71% of the cases, and worse in only 8%. Similarly, the reconciliation is better in 79% of the cases, and worse in only 6%. The higher the threshold value, the more edges are considered for NNI moves by *MowgliNNI*. Up to a certain point, broadening the search space of *MowgliNNI* allows to improve both the gene trees and the reconciliations. Yet, for threshold greater than 80, *MowgliNNI*'s

Threshold	20	40	60	80	90	95
Number of gene families containing weak edges	422	708	911	965	979	981
% **of cases where** $Cost(S, G_{NNI}) < Cost(S, G_{ML})$	79	81	85	88	88	89
% **of cases where** $RF(G_{True}, G_{NNI}) < RF(G_{True}, G_{ML})$	38	55	66	71	70	69
% **of cases where** $RF(G_{True}, G_{NNI}) = RF(G_{True}, G_{ML})$	59	40	28	21	21	21
% **of cases where** $RF(G_{True}, G_{NNI}) > RF(G_{True}, G_{ML})$	3	5	6	8	9	10
% **of cases where** $ED(R_{True}, R_{NNI}) < ED(R_{True}, R_{ML})$	59	70	75	79	79	78
% **of cases where** $ED(R_{True}, R_{NNI}) = ED(R_{True}, R_{ML})$	27	23	19	15	16	15
% **of cases where** $ED(R_{True}, R_{NNI}) > ED(R_{True}, R_{ML})$	14	8	6	6	6	6

Table 1. Quality of the gene trees and reconciliations inferred by *MowgliNNI*. For each tested threshold value, the second row indicates the number of gene families containing some weak edges (among the 985 simulated gene families). Third row indicates the percentage of these families where *MowgliNNI* proposes a modified tree of lower reconciliation cost. The last six rows provide the percentage of the former families where *MowgliNNI* provides modified gene trees (resp. reconciliations) that are closer, equally far or farther from the true gene trees (resp. the true evolutions). $RF(G_{True}, G_X)$ denotes the Robinson Fould distance between G_{True} and G_X, $ED(R_{True}, R_X) = |R_{True} - R_X| + |R_X - R_{True}|$, where X stands for NNI or ML.

performance starts to decrease due to the fact that the sequence signal is no longer sufficiently taken into account.

MowgliNNI progressively reduced the number of predicted duplications, transfers and losses as the threshold increased. At threshold 0 (where *MowgliNNI* = *Mowgli*), 5403 duplications, 2460 transfers and 12007 losses were predicted on the whole dataset; going to threshold 80, these numbers drop to 4510 duplications, 1160 transfers and 8016 losses, *i.e.* values that are much closer to the 4443 duplications and 8142 losses contained in the true reconciliations.

The average number of false positive events (FP) of the R_{NNI} reconciliations decreases as the threshold increases (Fig. 2(b)). However, as in Doyon et al. [4], the average number of FP transfers is quite high compared to that of duplications and losses. This can be explained by several reasons. First, a transfer is judged incorrect as soon as i) it does not depart or end in the same edges of the species tree as the corresponding true transfer, or ii) it does not concern the same edge in the gene tree. Overall, there is an additional constraint w.r.t. duplications and loss events, leading on average to more incorrect events. This point is all the more sensitive that several most parsimonious reconciliations (MPR) are obtained in a number of cases, while we just accounted for one of them for each gene family. Hence, event error rates we report are pessimistic (note that this does not affect RF distance results). Last, incorrect gene trees lead to incorrect event inferences, but the latter are very sensitive to only small errors in gene trees. The event FP error grows exponentially when the RF distance between the initial and the true tree increases from 0 to 10% (data not shown).

Inferring G_{ML} trees from shorter sequences (400 bp), led to a decrease in their quality both in terms of RF distance to G_{True} and in event distance between inferred and true reconciliation. Starting from a less accurate gene tree, the accuracy of the NNI trees and of their evolutionary histories is also lowered, though the relative improvement provided by *MowgliNNI* over *Mowgli* is higher than with long sequences (data not shown).

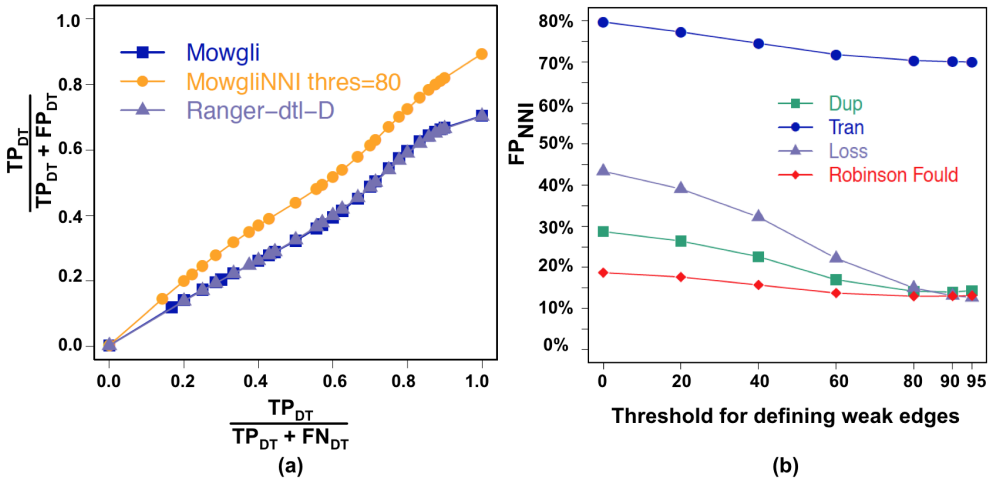

Fig. 2. (a) The accuracy of *Mowgli*, Ranger-dtl-D and *MowgliNNI* (threshold=80) in inferring duplications and transfers, where TP_{DT} (resp. FP_{DT}, FN_{DT}) denotes the true positive (resp. false positive, false negative) of duplications and transfers predicted. (b) Average false positive (FP) of the NNI trees – note that FP values at threshold 0 correspond to *Mowgli* results.

In order to measure the dependance of *MowgliNNI* on the precise costs used for each kind of event, we ran the method on G_{ML} trees with costs varying up to 10%, 20%, then 50% w.r.t. those computed from Formula (1). The paired t-test for RF distances shows that the G_{NNI} trees obtained with the new costs are not significantly different from those obtained with the former costs (p-value=0.296, 0.2723, 0.2028 respectively). The accuracy of inferred events also does not change much. Transfers have the highest variation with 3.6% (resp. 3%) increase in FP (resp. FN) when the event costs vary up to 50%. Thus, *MowgliNNI* is quite robust to changes in the event costs.

In summary, *MowgliNNI* successfully uses the reconciliation cost as additional information to resolve the uncertain parts of gene trees inferred from sequences only. Though the gene tree resolutions are partly guided by reconciliations with the species tree, they are not attracted away from the true gene trees, but are closer to them than the initial gene trees. As a result, *MowgliNNI* infers gene events more accurately, which is of prior importance to distinguish orthologs from paralogs and xenologs [13].

4.2 Experiments on real data

We constructed a dataset of $\approx$ 30000 homologous gene families (3 to 312 taxa) on Bacteria from the HOGENOM database (release 04) [30] and ran *Mowgli* and *MowgliNNI* on this dataset with fixed parameters ($\tau = 3, \delta = 3.5, \lambda = 1$).

MowgliNNI allows to change the gene tree, hence to lower the reconciliation cost, in 24% of the $\approx$ 30000 families. This gain is non-negligible and is uttermost important as changing the gene tree topology has an important impact on the inferred events (as shown on the simulated data sets and below) that are used in turn to predict the function of genes on the basis of ortholog and paralog relationships. When allowing

rearrangements on weak edges under the DL model, Berglund-Sonnhammer et al. reported that 10% of their families were improved [18], while Chaudhary el al. improved all their gene trees in a pure D model when rearranging gene trees with *Subtree Prune and Regraft* (SPR) operations [21]. However, it is hard to know whether the datasets are comparable.

For all gene families, we counted the number of events of each kind ($\mathbb{D}$, $\mathbb{T}$, $\mathbb{L}$) inferred by *Mowgli* and *MowgliNNI*. As a rule, *MowgliNNI* led to a decrease in the number of events in inferred evolutionary histories, the reduction being considerable for transfers and losses (88.3% and 59.9% resp.) but quite small for duplications (5.2%). These results obtained in the DTL model are consistent with those of Durand et al. reporting that in the DL model gene tree rearrangements substantially reduce the number of events needed to explain the data [16]. The differences in reductions we observed depending on the kind of events can be explained by the fact that given the costs we used for the events ($\tau = 3, \delta = 3.5, \lambda = 1$), it is usually more parsimonious to explain the conflicts between gene and species tree by a combination of $\mathbb{T}$ and $\mathbb{L}$ rather than a combination of $\mathbb{D}$, and $\mathbb{L}$. Thus, when *MowgliNNI* infers a gene tree closer to the species tree, it mostly removes the need for artificial transfers (and losses to a lesser extent), while not altering that much the number of duplications.

In addition to reductions in errors and number of events, the new gene tree proposed by *MowgliNNI* usually reduced the number of alternative MPRs, i.e. histories. On a random sample of two dozens new gene trees, the number of MPRs is reduced in 63% of the cases (by a factor of 18 in the best case), and increased in 21% (by a factor 3 at worst). This echoes similar findings of Durand et al. in a DL model [16].

We measured the running time of *Mowgli* and of both the non-optimized and optimized versions of *MowgliNNI* (see Methods) on a random sample of 100 families having from 10 to 80 taxa. The results show that the optimized version of *MowgliNNI* is 20 (resp. 50 and 80) times faster than the non-optimized one, when facing 1-20 (resp. 20-40 and 40-60) weak edges. The increase in accuracy due to *MowgliNNI* is obtained at the price of a small computation time overcost.

Acknowledgement

We thank Gergely J. Szöllősi for his help in determining the event costs of the real dataset and the referees that helped strengthen the experimental validation. This work was funded by the *Languedoc-Roussillon Chercheur d'Avenir* program and by the french *Agence Nationale de la Recherche Investissements d'avenir / Bioinformatique* (ANR-10-BINF-01-02, *Ancestrome*) and *Programme 6ème Extinction* (ANR-09-PEXT-000 *PhyloSpace*).

References

1. M. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. In *RECOMB '04*, pages 347–356, New York, NY, USA, 2004. ACM.
2. P. Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. In Philip E. Bourne and Dan Gusfield, editors, *RECOMB*, pages 316–325. ACM, 2004.

3. C. Conow, D. Fielder, Y. Ovadia, and R. Libeskind-Hadas. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol*, 5:16, 2010.
4. J-P. Doyon, C. Scornavacca, K. Y. Gorbunov, G. Szöllösi, V. Ranwez, and V. Berry. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *RECOMB-CG*, 2010.
5. A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM TCBB*, 8(2):517–535, 2011.
6. L. A. David and E. J. Alm. Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328):93–6, Jan 2011.
7. M. S. Bansal, E. J. Alm, and M. Kellis. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer, and loss. In *Proceeding ISBM'12*, 2012.
8. M. Goodman, J. Czelusniak, G. W. Moore, Romero 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.
9. R. D. Page. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Mol Phylogenet Evol*, 14:89–106, 2000.
10. Bin Ma, Ming Li, and Louxin Zhang. From gene trees to species trees. *SIAM Journal on Computing*, 30(3):729–752, 2001.
11. L. Nakhleh, T. Warnow, and C. R. Linder. Reconstructing reticulate evolution in species: theory and practice. In *Proceedings of the eighth annual international conference on Research in computational molecular biology*, RECOMB '04, pages 337–346, New York, NY, USA, 2004. ACM.
12. L. Arvestad, J. Lagergren, and B. Sennblad. The gene evolution model and computing its associated probabilities. *J. ACM*, 56(2), 2009.
13. J-P. Doyon, V. Ranwez, V. Daubin, and V. Berry. Models, algorithms and programs for phylogeny reconciliation. *Brief Bioinform*, 12(5):392–400, 2011.
14. Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas. The co phylogeny reconstruction problem is NP-complete. *J. Comput. Biol.*, 18(1):59–65, Jan 2011.
15. R. Libeskind-Hadas and M. A. Charleston. On the computational complexity of the reticulate cophylogeny reconstruction problem. *JCB*, 16(1):105–117, 2009.
16. D. Durand, B. V. Halldorsson, and B. Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.*, 13(2):320–335, Mar 2006.
17. M. W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology*, 8(R141), 2007.
18. AC Berglund-Sonnhammer, P Steffansson, MJ Betts, and DA Liberles. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol*, 63(2):240–50, Aug 2006.
19. W. Chang and Oliver Eulenstein. Reconciling gene tree with apparent polytomies. *COCOON, LNCS*, 4112:235–244, 2006.
20. B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *J. Comput. Biol.*, 15:981–1006, 2008.
21. R. Chaudhary, J. G. Burleigh, and O. Eulenstein. Algorithms for rapid error correction for the gene duplication problem. In *Proceedings of the 7th international conference on Bioinformatics research and applications*, ISBRA'11, pages 227–239, Berlin, Heidelberg, 2011. Springer-Verlag.
22. C. Semple and M. A. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, 2003.
23. M. J. Sanderson. inferring absolute rates of evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19:301–302, 2003.

24. A. Tofigh. *Using Trees to Capture Reticulate Evolution, Lateral Gene Transfers and Cancer Progression*. PhD thesis, KTH Royal Institute of Technology, Sweden, 2009.
25. J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Inc., 2004.
26. Donald Ervin Knuth. *The Art of Computer Programming*, volume 3. Addison-Wesley, 1973.
27. DG Kendall. On the generalized birth-and-death process. *Ann Math Stat*, 19:1–15, 1948.
28. A. Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.
29. D.F. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
30. S. Penel, A. M. Arigon, J. F. Dufayard, A. S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perriere. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6:S3, 2009.