The Path-Label Reconciliation (PLR) Dissimilarity Measure for Gene Trees

Alitzel López Sánchez
Computer Science Department, Université de Sherbrooke, Canada

José Antonio Ramírez-Rafael
Center for Research and Advanced Studies of the National Polytechnic Institute, Irapuato, Gto., Mexico
Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

Alejandro Flores-Lamas
Center for Research and Advanced Studies of the National Polytechnic Institute, Irapuato, Gto., Mexico

Maribel Hernández-Rosales[1]
Center for Research and Advanced Studies of the National Polytechnic Institute, Irapuato, Gto., Mexico

Manuel Lafond[1]
Computer Science Department, Université de Sherbrooke, Canada

Abstract

In this study, we investigate the problem of comparing gene trees reconciled with the same species tree using a novel semi-metric, called the Path-Label Reconciliation (PLR) dissimilarity measure. This approach not only quantifies differences in the topology of reconciled gene trees, but also considers discrepancies in predicted ancestral gene-species maps and speciation/duplication events, offering a refinement of existing metrics such as Robinson-Foulds (RF) and their labeled extensions LRF and ELRF. A tunable parameter α also allows users to adjust the balance between its species map and event labeling components. We show that PLR can be computed in linear time and that it is a semi-metric. We also discuss the diameters of reconciled gene tree measures, which are important in practice for normalization, and provide initial bounds on PLR, LRF, and ELRF.

To validate PLR, we simulate reconciliations and perform comparisons with LRF and ELRF. The results show that PLR provides a more evenly distributed range of distances, making it less susceptible to overestimating differences in the presence of small topological changes, while at the same time being computationally efficient. Our findings suggest that the theoretical diameter is rarely reached in practice. The PLR measure advances phylogenetic reconciliation by combining theoretical rigor with practical applicability. Future research will refine its mathematical properties, explore its performance on different tree types, and integrate it with existing bioinformatics tools for large-scale evolutionary analyses. The open source code is available at: `https://pypi.org/project/parle/`.

2012 ACM Subject Classification Applied computing → Molecular evolution

Keywords and phrases Reconciliation, gene trees, species trees, evolutionary scenarios

Digital Object Identifier 10.4230/LIPIcs.WABI.2024.20

Related Version *Full Version*: `https://arxiv.org/abs/2407.06367`

Supplementary Material *Software (Source Code)*: `https://pypi.org/project/parle/` [46]

[1] Corresponding author

24th International Workshop on Algorithms in Bioinformatics (WABI 2024).
Editors: Solon P. Pissis and Wing-Kin Sung; Article No. 20; pp. 20:1–20:21
LIPICS Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Funding *Alitzel López Sánchez*: Bourse d'excellence de la Faculté de Sciences, Université de Sherbrooke.
José Antonio Ramírez-Rafael: CONAHCYT scholarship, Mexico.
Manuel Lafond: NSERC Discovery program

1 Introduction

During evolution, it is well-known that genes can be duplicated, lost, and transferred, resulting in evolutionary scenarios that differ from the history of the species that contain them. Gene trees can therefore be discordant with their species trees, and *reconciliation* aims to infer the macro-evolutionary events that explain the discrepancies. Several models have been proposed to achieve this task, allowing duplications and losses [24, 68, 27, 9, 20, 64, 39, 30, 21], horizontal gene transfer [25, 19, 4, 35, 33, 62, 50, 36, 66, 61], incomplete lineage sorting [69, 63, 56, 67, 14, 43], and others (see e.g. [17, 42, 28, 10, 2, 57, 45]). In addition, some of these models support segmental events that affect multiple genes at once [52, 13, 5, 53, 18], and some approaches infer histories based on parsimony whereas others are probabilistic [3, 1, 41].

This variety of reconciliation models and algorithms is accompanied by a large diversity of software and tools to reconcile gene trees with species trees (examples include NOTUNG [20], DLCoal [56], RANGER-DTL [6], ecceTERA [32], Jane [15]). Most of these tools infer, for each ancestral gene tree node, the ancestral species to which the gene belonged to, as well as the event that affected the gene. It is, however, difficult to assess the quality of the reconciliations produced by these approaches, even with the availability of high quality software to simulate gene tree evolution (e.g. SimPhy [48], Asymmetry [59], aevol [7], ZOMBI [16]). A standard benchmarking idea would be to simulate reconciled gene trees and to compare the inferred scenarios with the true simulated ones. However, it is not straightforward to perform this comparison. Indeed, reconciled gene trees exhibit three types of valuable information: the tree topology, the gene-species map, and the event labeling. While there exist metrics to measure discrepancies for each of those three criteria individually, we are not aware of any established method to measure disagreements in all three simultaneously. There is a large body of literature on measuring topological differences between trees (e.g. [54], [23], [58],[47],[49], [65]). In terms of gene-species mapping discordance, the *path distance* metric [31] applies to gene trees with identical topologies but possibly different species maps, and quantifies how far the species of corresponding nodes are in the gene trees. The metric was mainly introduced to obtain medians in the reconciliation spaces of gene trees. If the gene trees differ, though, the metric cannot be used.

Perhaps the most relevant metric to compare reconciled gene trees is the recent *labeled Robinson-Foulds (RF) distance*, now called ELRF, which accounts for differences in topology and event labeling. Given two gene trees, the distance is the minimum number of edge contractions, edge expansions, and node label substitutions required to transform one gene tree into the other [11]. It is unknown whether this distance can be computed in polynomial time, the main difficulty being that edge operations must have the same label on both endpoints. The authors then proposed a variant of this metric, called LRF, in which edge contractions/expansions are replaced with node insertions/deletions, which can be computed in linear time [12]. Although these are perhaps the only approaches specifically tailored for gene tree comparison, their usage has some disadvantages. First, these distances do not take gene-species maps into consideration. Second, the metric suffers from the same well-known shortcomings as the RF distance, see [44] for a discussion on this (for instance, a single misplaced leaf can increase the distance dramatically). Another subtle but yet important aspect is the topological uncertainty that can be present in gene trees. In particular, when ancestral species undergo gene duplication episodes (see e.g. [26, 53]), the corresponding gene trees may contain large duplication subtrees. In this case, there is too little phylogenetic signal to infer the topology of such duplication subtrees accurately. However, most approaches penalize discrepancies in those local parts of the gene trees as in any other part, even though predicting different speciation patterns should be more heavily penalized than in duplication clusters.

In this work, we introduce a novel approach for comparing gene trees that considers all the aforementioned components that play a role in reconciliations: the species tree, the gene tree, the labeling of their internal nodes by species and events, as well as duplication clusters. This method effectively circumvents the shortcomings of the RF distance. Given two reconciled gene trees on the same set of genes, our dissimilarity measure establishes a correspondence between the gene tree nodes from both trees and applies a penalty if the matched nodes differ in species or event label. As we demonstrate, due to the constraints inherent in reconciliation models, this approach implicitly penalizes topological disagreements between the gene trees, except when the discordance is solely due to consecutive duplication rounds within the same species.

Our measure also has the advantage of being computable in linear time. We first explore some theoretical properties of our approach and show that it functions as a semi-metric in the space of reconciled gene trees. We demonstrate that if non-binary gene trees are considered, the measure does not necessarily satisfy the triangle inequality, although this remains an open question for binary trees. We also provide initial results on the diameters of the PLR, LRF, and ELRF measures, which are important in practice for normalization.

We then validate our approach through experiments involving simulated reconciliations on the same set of leaves and calculation of various measures. We show that, as can be expected from previous knowledge, RF, LRF, and ELRF tend to produce large distances overestimating tree differences, which can result from a rapid increase in the distance values when, for example, a single leaf is misplaced. In contrast, our measure effectively captures small, average, and large distances between reconciliations. Therefore, PLR is established as the first reconciliation measure with greater variability than RF variants, and sensitivity to differences in every component of evolutionary scenarios.

Note that due to space constraints, some of the proofs were replaced by a sketch of the main idea, and the full detailed arguments can be found in the arxiv version (`https://arxiv.org/abs/2407.06367`).

2 Preliminary notions

A *tree* is a connected acyclic graph. Unless stated otherwise, all trees in this paper are rooted. For a tree T, we denote by $r(T)$ the root of T, by $V(T)$ and $E(T)$ its set of nodes and edges, respectively, and by $L(T)$ its set of leaves. A non-leaf node is called *internal*. For $u, v \in V(T)$, we write $u \preceq_T v$ if u is a *descendant* of v, i.e., if v is on the path between $r(T)$ and u (we write $u \prec_T v$ if $u \neq v$). Then v is an *ancestor* of u. If $u \neq r(T)$, then the *parent* $p_T(u)$ of u if the ancestor v of u such that $uv \in E(T)$, and u is a *child* of v. A tree T is *binary* if each internal node has two children, and T is a *caterpillar* if all internal nodes have at most one child that is an internal node (that is, T is a path with leaves attached to its nodes).

For $X \subseteq V(T)$, we denote by $\text{lca}_T(X)$ the *lowest common ancestor* of all the nodes in X. When $|X| = 2$, we may write $\text{lca}_T(u, v)$ instead of $\text{lca}_T(\{u, v\})$. For $v \in V(T)$, we write $T(v)$ for the subtree of T rooted at v. Note that $L(T(v))$ is the set of leaves that descend from v, which we call the *clade* of v. As a shorthand, we may write $L_T(v)$ to denote the clade of v, or $L(v)$ if T is understood. The *distance* between two nodes u, v in T is denoted $dist_T(u, v)$, i.e., the length of the undirected path in T between u and v.

2.1 Species trees and reconciled gene trees

A *species tree* S is a tree which we assume to be binary. A *reconciled gene tree* (with S) is a tuple $\mathcal{G} = (G, S, \mu, l)$ where G is a tree in which each internal node has at least two children (possibly more), S is a species tree, $\mu : V(G) \to V(S)$ maps nodes of G to species in S, and $l : V(G) \to \{dup, spec, extant\}$ is an event labeling. We also have the following requirements:

1. *Leaves are from extant species:* for every leaf $v \in L(G)$, $\mu(v) \in L(S)$ and $l(v) = extant$. Moreover, every internal node $w \in V(G) \setminus L(G)$ satisfies $l(w) \in \{dup, spec\}$;
2. *Time-consistency:* for any two nodes $u, v \in V(G)$, $u \preceq_G v$ implies $\mu(u) \preceq_S \mu(v)$;
3. *Speciations separate species:* for any node $v \in V(G)$ such that $l(v) = spec$, we have $\mu(v) \in V(S) \setminus L(S)$ and v has exactly two children v_1, v_2.
   Moreover, denoting by s_1, s_2 the two children of $\mu(v)$ in S, we have that $\mu(v_1) \preceq_S s_1$ and $\mu(v_2) \preceq_S s_2$, or $\mu(v_2) \preceq_S s_1$ and $\mu(v_1) \preceq_S s_2$.

If μ satisfies $\mu(v) = lca_S(\{\mu(x) : x \in L(v)\})$ for every node $v \in V(G)$, then μ is called the *lca-mapping* [27, 9]. In this map, all genes map to the lowest possible species according to the rules of reconciliation. These concepts are illustrated in Figure 1, which presents two reconciled gene trees that use the lca-mapping (see caption). Note that our reconciled gene trees are not restricted to the *lca-mapping.* However, it is known that if $l(v) = spec$, then $\mu(v)$ must indeed be the lowest common ancestor of all the species that appear in the genes below v. However, the converse is not required to hold, that is, a duplication could be mapped to the lowest common ancestral species (or above).

Isomorphism between reconciled gene trees. Two reconciled gene trees $\mathcal{G}_1 = (G_1, S, \mu_1, l_1)$ and $\mathcal{G}_2 = (G_2, S, \mu_2, l_2)$ are *isomorphic* if they have the same sets of leaves, use the same species tree, have the same topology (i.e., they branch in identical ways), and their corresponding nodes map to the same species and have the same label. If this holds, we write $\mathcal{G}_1 \simeq \mathcal{G}_2$. Formally, $\mathcal{G}_1 \simeq \mathcal{G}_2$ if there exists a bijection $\phi : V(G_1) \to V(G_2)$ such that the following holds:

- $L(G_1) = L(G_2)$ and, for each leaf $x \in L(G_1)$, $\phi(x) = x$. In other words, each leaf of G_1 is mapped to the same leaf in G_2;
- $uv \in E(G_1)$ if and only $\phi(u)\phi(v) \in E(G_2)$;
- for every node $v \in V(G_1)$, $\mu_1(v) = \mu_2(\phi(v))$ and $l_1(v) = l_2(\phi(v))$.

2.2 The Path-Label Reconciliation (PLR) dissimilarity measure

Let $\mathcal{G}_1 = (G_1, S, \mu_1, l_1)$ and $\mathcal{G}_2 = (G_2, S, \mu_2, l_2)$ be two reconciled gene trees. We say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are *comparable* if: (1) they are reconciled with the same species tree S; (2) $L(G_1) = L(G_2)$; and (3) for each leaf $x \in L(G_1)$, $\mu_1(x) = \mu_2(x)$ (that is, extant genes map to the same species in both trees). Unless stated otherwise, we assume that all pairs of reconciled trees mentioned are comparable, although (3) could be dropped, see remark below.

For a node $v \in V(G_1)$, we need a corresponding node for v in G_2. This can be done in multiple ways, and here we assign this corresponding node as the lowest possible node of G_2 that is an ancestor of all the descendants of v. To put it more formally, define

$$m_{\mathcal{G}_1,\mathcal{G}_2}(v) = lca_{G_2}(L(G_1(v)))$$

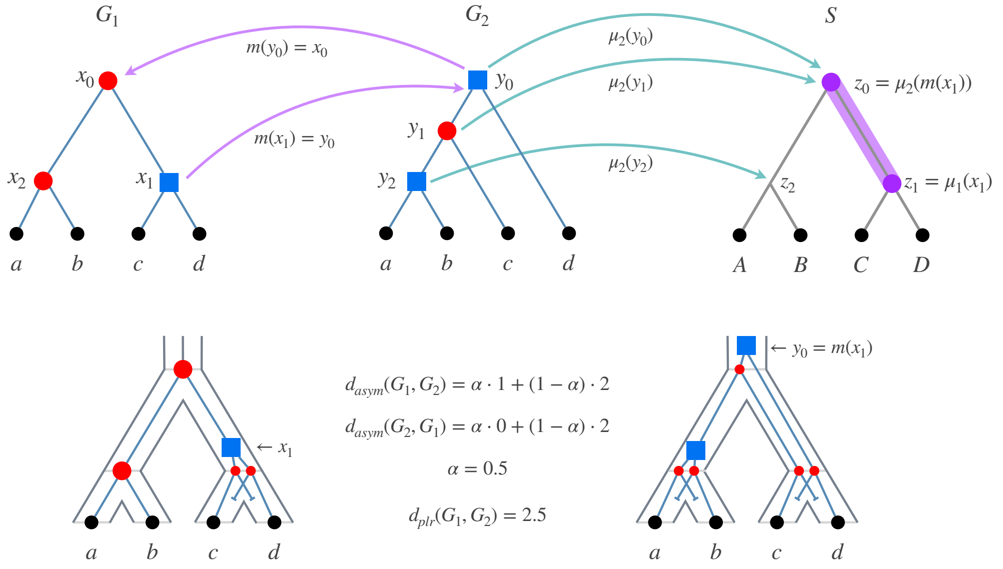

Figure 1 In the upper row, there are two reconciled gene trees G_1 and G_2 as well as a species tree S. The event labelings are shown as red circles and blue squares, which represent speciations and duplications, respectively. Lowercase letters a, b, c, d depict extant genes, while the corresponding uppercase letters are the species where genes reside. The maps μ_1, μ_2 use the lca-mapping, that is, $\mu_1(x_0) = z_0, \mu_1(x_1) = z_1, \mu_1(x_2) = z_2$, and $\mu_2(y_0) = \mu_2(y_1) = z_0, \mu_2(y_2) = z_2$. The gene trees have the same set of leaves but different topology and event labeling. Purple arrows exemplify the maps $m_{\mathcal{G}_1,\mathcal{G}_2}(x_1)$, which is the lca of genes c and d, and $m_{\mathcal{G}_2,\mathcal{G}_1}(y_0)$, while green arrows illustrate the species map μ_2. The shaded edge in S displays the path distance between $\mu_1(x_1) = z_1$ and $\mu_2(m(x_1)) = \mu_2(y_0) = z_0$. The lower row shows the explicit evolution of the gene trees within the species tree. The contribution of x_1 to the d_{path} component is 1, because $dist_S(\mu_1(x_1), \mu_2(m(x_1))) = 1$, whereas its contribution to d_{lbl} is 0 because $l(x_1) = l(m(x_1)) = dup$. On the other hand, the node y_0 from G_2 contributes 0 to d_{path} since its correspondent x_0 is mapped to the same species, but contributes 1 to d_{lbl} since $l(y_0) = dup$ and $l(x_0) = spec$.

which is the lowest common ancestor in G_2 of the clade of v. Note that this is well-defined since $L(G_1) = L(G_2)$. For instance in Figure 1, $m_{\mathcal{G}_1,\mathcal{G}_2}(x_1) = y_0$. When $\mathcal{G}_1, \mathcal{G}_2$ are clear from the context, we may write $m(v)$ instead of $m_{\mathcal{G}_1,\mathcal{G}_2}(v)$. In essence, this is the lca-mapping, but applied between two gene trees. Note that such mappings are usually applied between gene and species trees, but [37] also introduced the ancestral gene-gene map idea (or more specifically, ancestral RNA-gene maps).

Our measure has two components: one for the discrepancies in the species mappings, and one for the labelings. These components are defined as:

$$d_{path}(\mathcal{G}_1, \mathcal{G}_2) = \sum_{v \in V(G_1)} dist_S(\mu_1(v), \mu_2(m(v)))$$

$$d_{lbl}(\mathcal{G}_1, \mathcal{G}_2) = |\{v \in V(G_1) : l_1(v) \neq l_2(m(v))\}|$$

In words, in d_{path}, each term $dist_S(\mu_1(v), \mu_2(m(v)))$ penalizes v by how far its species is from the species of its correspondent $m(v)$, and d_{lbl} is simply the number of nodes of G_1 whose label differ from their correspondent in G_2.

We assume the existence of a given parameter $\alpha \in [0,1]$ to weigh these components, and define the *asymmetric dissimilarity* between $\mathcal{G}_1$ and $\mathcal{G}_2$ as:

$$d_{asym}(\mathcal{G}_1, \mathcal{G}_2) = \alpha \cdot d_{path}(\mathcal{G}_1, \mathcal{G}_2) + (1-\alpha) \cdot d_{lbl}(\mathcal{G}_1, \mathcal{G}_2).$$

Note that when $\alpha = 1$ and G_1, G_2 have the same topology, then d_{asym} is exactly the *path distance metric* studied in [31]. Our dissimilarity measure generalizes this by allowing trees with different topologies and by considering node labels. One could ignore the α parameter by weighing d_{path} and d_{lbl} equally, which can be achieved with $\alpha = 0.5$. Also notice that d_{path} may be adapted to species trees with branch lengths.

It is easy to see that d_{asym} is not symmetric. For instance, suppose that $\mathcal{G}_1$ consists of a binary gene tree with several internal nodes mapping to different species, and $\mathcal{G}_2$ consists of a star tree with a single internal node, such that both roots are duplications that map to the same species. Then $d_{asym}(\mathcal{G}_1, \mathcal{G}_2)$ can be proportional to the number of internal nodes of G_1, whereas $d_{asym}(\mathcal{G}_2, \mathcal{G}_1) = 0$.

The Path-Label Reconciliation (PLR) dissimilarity is therefore defined as

$$d_{plr}(\mathcal{G}_1, \mathcal{G}_2) = d_{asym}(\mathcal{G}_1, \mathcal{G}_2) + d_{asym}(\mathcal{G}_2, \mathcal{G}_1)$$

If $\mathcal{G}_1$ and $\mathcal{G}_2$ are not comparable, then we define $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) = \infty$.

In Figure 1 we exemplify all the components of the dissimilarity measure. In the example, following the μ_1, μ_2 maps given in the caption, if we count the respective costs of x_0, x_1, x_2, we have $d_{path}(\mathcal{G}_1, \mathcal{G}_2) = 0 + 1 + 0 = 1$ and $d_{lbl}(\mathcal{G}_1, \mathcal{G}_2) = 1 + 0 + 1 = 2$. If we put $\alpha = 0.5$, we get $d_{asym}(\mathcal{G}_1, \mathcal{G}_2) = 0.5 \cdot 1 + 0.5 \cdot 2 = 1.5$. As for the costs of y_0, y_1, y_2, we get $d_{path}(\mathcal{G}_2, \mathcal{G}_1) = 0 + 0 + 0$ and $d_{lbl}(\mathcal{G}_2, \mathcal{G}_1) = 1 + 0 + 1 = 2$, and thus $d_{asym}(\mathcal{G}_2, \mathcal{G}_1) = 1$. Therefore, $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) = 2.5$.

A remark on leaves belonging to the same species. Recall that condition (3) of comparability requires $\mu_1(x) = \mu_2(x)$ for every leaf $x \in L(G_1)$. Although this assumption usually follows from the knowledge of the species of a gene, it may not hold in some contexts. Indeed, in metagenomics even the species of extant genes is unknown and needs to be inferred (see for example [26]). Therefore, for an extant gene x, two different reconciliation algorithms may predict that x belongs to a different species, leading to $\mu_1(x) \neq \mu_2(x)$. Although condition (3) is useful in the proofs that follow, we note that it is not required in the definition of d_{plr}, and the latter remains well-defined even if we drop this condition. Therefore, d_{plr} could be used to also compare gene trees with predicted gene-species maps that differ even at the level of leaves (although the theory developed hereafter may need revision for this case).

A remark on setting α. The reader may notice that if α is ignored in d_{plr}, or set to a constant, the d_{path} component can easily outweigh the d_{lbl} component. This is because in the worst case, $d_{path}(\mathcal{G}_1, \mathcal{G}_2)$ can be in $\Theta(nm)$, where n is the number of species leaves and m is the number of gene tree leaves, which occurs if most nodes of $\mathcal{G}_1$ are mapped to nodes of $\mathcal{G}_2$ with $\Theta(n)$ path distance in S (see the diameter section for a detailed analysis). On the other hand, the $d_{lbl}(\mathcal{G}_1, \mathcal{G}_2)$ component is always $O(m)$, as it only depends on the number of nodes in the gene tree. This quadratic-versus-linear effect can be prevented by making α depend on n. For instance, one may put $\alpha = 1/n$, or more generally $\alpha = c/n$ for some constant c.

A remark on scenarios with horizontal transfer events. In the presence of horizontal gene transfers, gene tree nodes can also undergo a *transfer* event, and a different notion of time-consistency than ours is typically used (see e.g. [51]). Nonetheless, such reconciliations also include a gene-species map μ and a labeling function l, and d_{plr} is also well-defined in this context. On the other hand, it is unclear whether path distances are appropriate to compare transferred genes, and again, the theory that follows may need to be adapted to allow transfers.

Least duplication-resolved gene trees

Consider a reconciled gene tree $\mathcal{G} = (G, S, \mu, l)$. If, in G, there is a connected subtree consisting only of duplication nodes, all mapped to the same species, then it is difficult to postulate on the exact topology of the duplication subtree due to the lack of clear phylogenetic signals. One solution is to contract the subtree into a single node to model the uncertainty. Contracting weakly supported branches in gene trees can be useful to detect and correct errors in dubious duplication nodes [40]. Moreover, special cases of least-duplication resolved trees such as discriminating co-trees arise in the context of orthology detection [29, 22]. To this end, we say that an edge $uv \in E(G)$ is *redundant* if $\mu(u) = \mu(v)$ and $l(u) = l(v) = dup$. We then say that $\mathcal{G}$ is *least duplication-resolved* if no edge uv of G is redundant.

Suppose that $\mathcal{G}$ is *not* least duplication-resolved, and let $uv \in E(G)$ be a redundant edge, with $u = p_G(v)$. We denote by $\mathcal{G}/uv$ the reconciled gene tree obtained by contracting uv in G and updating μ and l accordingly. More specifically, $\mathcal{G}/uv = (G', S, \mu', l')$, where: G' is obtained from G by deleting v and its incident edges and, for each child v' of v in G, adding the edge uv'; and then putting $\mu'(w) = \mu(w)$ and $l'(w) = l(w)$ for every $w \in V(G')$. If $R \subseteq E(G)$ is a set of redundant edges of $\mathcal{G}$, then $\mathcal{G}/R$ is the reconciled gene tree obtained after contracting every edge in R, in any order. If R is the set of all redundant edges of $\mathcal{G}$, then we define $LR(\mathcal{G}) = \mathcal{G}/R$, called the *least duplication-resolved subtree* of $\mathcal{G}$. It is not difficult to see that such a subtree is unique, least duplication-resolved, and satisfies all conditions of a reconciled gene tree. Figure 2 shows two gene trees and their least duplication-resolved version (note that two consecutive duplications in distinct species remain).

For two reconciled gene trees $\mathcal{G}_1, \mathcal{G}_2$, we write $\mathcal{G}_1 \simeq_d \mathcal{G}_2$ if $LR(\mathcal{G}_1) \simeq LR(\mathcal{G}_2)$. This means that $\mathcal{G}_1$ and $\mathcal{G}_2$ may differ, but every form of disagreement is due to redundant edges, and they become identical in their least duplication-resolved form. The following will be useful.

▶ **Lemma 1.** *Let $\mathcal{G} = (G, S, \mu, l)$ be a reconciled gene tree that is least duplication-resolved. Let $u, v \in V(G)$ be such that $v \prec_G u$. Then either $\mu(u) \neq \mu(v)$ or $l(u) \neq l(v)$.*

Proof sketch. If there is a *spec* node on the path from u to v (excluding v), then the *speciation separates species* conditions implies that $\mu(v) \prec_S \mu(u)$. If all nodes on the path are *dup* nodes, and if v is a *spec*, then $l(u) \neq l(v)$. Finally, if v is also a *dup*, the least duplication-resolved properties imply that $\mu(v) \neq \mu(u)$. ◀

3 Properties of the Path-Label Reconciliation (PLR) dissimilarity

We first show that in terms of time complexity, $d_{plr}(\mathcal{G}_1, \mathcal{G}_2)$ can be computed in linear time, using appropriate data structures, in a very straightforward manner as shown in Algorithm 1. The details of a linear-time implementation can be found in the proof of Theorem 2.

▶ **Theorem 2.** *The value $d_{plr}(\mathcal{G}_1, \mathcal{G}_2)$ can be computed in time $O(|V(G_1)| + |V(G_2)| + |V(S)|)$.*

Algorithm 1 Computing d_{asym} in one direction.

```
function getAsymmetricDist($\mathcal{G}_1 = (G_1, S, \mu_1, l_1), \mathcal{G}_2 = (G_2, S, \mu_2, l_2)$, $\alpha$)
    $d_{path} \leftarrow 0, d_{lbl} \leftarrow 0$;
    $m \leftarrow lcamap(G_1, G_2)$;                // Computes all $m(v) = lca_{G_2}(L(G_1(v)))$
    foreach $v \in V(G_1)$ do
        $v' \leftarrow m(v)$;
        $d_{path} \leftarrow d_{path} + dist_S(\mu_1(v), \mu_2(v'))$;
        if $l_1(v) \neq l_2(v')$ then $d_{lbl} \leftarrow d_{lbl} + 1$;
    return $\alpha \cdot d_{path} + (1 - \alpha) \cdot d_{lbl}$;
```

Proof. We argue that Algorithm 1 can be implemented to run in time $O(|V(G_1)| + |V(G_2)| + |V(S)|)$, which clearly proves the statement since we only need to run it twice (once for $\mathcal{G}_1$ versus $\mathcal{G}_2$, and once for $\mathcal{G}_2$ versus $\mathcal{G}_1$). We assume that G_1, G_2, and S are pre-processed to answer lowest common ancestor queries between any two nodes in constant time. This pre-processing time is linear for each tree [8], and therefore this step takes time $O(|V(G_1)| + |V(G_2)| + |V(S)|)$. We also assume that we know the depth of each node x of S, denoted $depth(x)$, which is the distance between x and the root. This can easily be computed by a linear-time preorder traversal of S. It is not difficult to compute $m = lcamap(G_1, G_2)$ in time $O(|V(G_1)| + |V(G_2)|)$ using the *lca* pre-processing and dynamic programming. Indeed, for a gene tree node $v \in V(G_1)$ with children $v_1, \ldots, v_l$, we have $m(v) = lca_{G_2}(\{m(v_1), \ldots, m(v_l)\})$. The latter *lca* expression can be computed with $l - 1$ *lca* queries as follows. Define $w_{1,i} = lca_{G_2}(\{m(v_1), \ldots, m(v_i)\})$. First compute $w_{1,2} = lca_{G_2}(m(v_1), m(v_2))$, then $w_{1,3} = lca_{G_2}(w_{12}, m(v_3))$, and so on until $m(v) = w_{1,l} = lca_{G_2}(w_{1,l-1}, m(v_l))$, each in $O(1)$ time. Since l is the number of edges between v and its children, the number of *lca* queries required throughout the execution of the whole algorithm is less than the number of edges of G_1, which is $O(|V(G_1)|)$.

For each $v \in V(G_1)$, we can obtain $dist_S(\mu_1(v), \mu_2(v'))$ in constant time, since it is equal to $depth(\mu_1(v)) + depth(\mu_2(v')) - 2 \cdot depth(lca_S(\mu_1(v), \mu_2(v')))$. It follows that each $v \in V(G_1)$ can be dealt with in $O(1)$ time and the loop of the algorithm takes time $O(|V(G_1)|)$, which does not add to the complexity. ◀

A semi-metric under least duplication-resolved equivalence

Let us recall the mathematical notion of a metric, which can be defined as a triple $(X, d, \equiv)$ where X is a set, $d : X \times X \to \mathbb{R}$ is a dissimilarity function, and $\equiv$ is a binary equality operator on X, such that the following conditions are satisfied:

- (identity) for all $x \in X$, $d(x, x) = 0$;
- (positivity) for all $x, y \in X$, if $x \not\equiv y$, then $d(x, y) > 0$;
- (symmetry) for all $x, y \in X$, $d(x, y) = d(y, x)$;
- (triangle inequality) for all $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.

If all the above conditions are satisfied, except the triangle inequality, then $(X, d, \equiv)$ is a *semi-metric*. If X is clear from the context, we call d a *metric (or semi-metric) under* $\equiv$.

In our case, we consider the set of all reconciled gene trees, with d_{plr} as our dissimilarity function. As for the equality operator, we may consider $\simeq$ or $\simeq_d$. In general, d_{plr} does not always meet the *positivity* requirement under $\simeq$. That is, $\mathcal{G}_1 \not\simeq \mathcal{G}_2$ does not necessarily imply $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) > 0$. Consider for example two gene trees with different topologies, but whose internal nodes are all duplications in the same species (in which case all internal nodes incur a path and label penalty of 0). For a more elaborate example, see Figure 2.

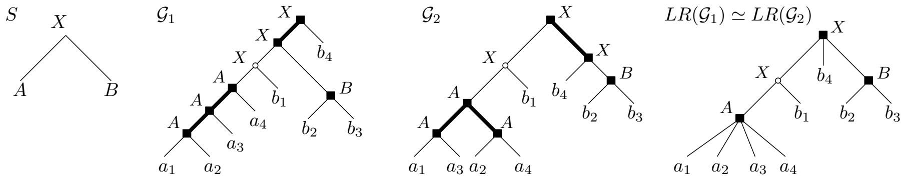

Figure 2 Two different reconciled gene trees $\mathcal{G}_1, \mathcal{G}_2$, where redundant edges are bold (again, lowercase letters indicate the species). Their d_{plr} value is 0 (one can check that all duplications in species $W \in \{A, X, B\}$ in either tree maps to a duplication in the same W in the other tree, and the X speciation to an X speciation. On the right, the least duplication-resolved version of the trees, showing that $\mathcal{G}_1 \simeq_d \mathcal{G}_2$.

However, we can show that d_{plr} is a semi-metric under $\simeq_d$. The most difficult part is to show that $\mathcal{G}_1 \not\simeq_d \mathcal{G}_2$ implies $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) > 0$. We first need to show that contracting the trees to their least duplication-resolved form cannot increase the dissimilarity.

▶ **Lemma 3.** *Let $\mathcal{G}_1 = (G_1, S, \mu_1, l_1), \mathcal{G}_2 = (G_2, S, \mu_2, l_2)$ be comparable reconciled gene trees, and let $uv \in E(G_1)$ be a redundant edge. Then $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) \geq d_{plr}(\mathcal{G}_1/uv, \mathcal{G}_2)$.*

Proof sketch. Denote $\mathcal{G}_1/uv = \mathcal{G}_1' = (G_1', S, \mu_1', l_1')$. For each node $w \in V(G_1) \setminus \{v\}$ that remains in G_1', the species and label of w, is the same as before, and $m_{\mathcal{G}_1,\mathcal{G}_2}(w) = m_{\mathcal{G}_1',\mathcal{G}_2}(w)$. Therefore, the contribution of w to both the d_{path} and d_{lbl} components are the same as before, and thus $d_{asym}(\mathcal{G}_1', \mathcal{G}_2) \leq d_{asym}(\mathcal{G}_1, \mathcal{G}_2)$. In the other direction, for $w \in V(G_2)$, we get that either $m_{\mathcal{G}_2,\mathcal{G}_1}(w)$ is unchanged after the contraction and w contributes to d_{path} and d_{lbl} just as before, or $m_{\mathcal{G}_2,\mathcal{G}_1}(w) = v$. In the latter case, $m_{\mathcal{G}_2,\mathcal{G}_1'}(w) = u$, and since $\mu_1(u) = \mu_1(v), l_1(u) = l_1(u)$, w costs the same as before. So $d_{asym}(\mathcal{G}_2, \mathcal{G}_1) = d_{asym}(\mathcal{G}_2, \mathcal{G}_1')$. ◀

Since Lemma 3 can be applied to any sequence of contractions, in either $\mathcal{G}_1$ or $\mathcal{G}_2$ by symmetry, we get the following.

▶ **Corollary 4.** *Let $\mathcal{G}_1 = (G_1, S, \mu_1, l_1), \mathcal{G}_2 = (G_2, S, \mu_2, l_2)$ be reconciled gene trees with the same leafset. Then $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) \geq d_{plr}(LR(\mathcal{G}_1), LR(\mathcal{G}_2))$.*

The above is sufficient to deduce that if $\simeq_d$ is interpreted as "being the same reconciled tree", then we have a semi-metric, unless $\alpha = 0$ or $\alpha = 1$ (see arxiv version for full proof).

▶ **Theorem 5.** *For any $\alpha \in (0, 1)$, d_{plr} is a semi-metric under $\simeq_d$.*

Proof sketch. Identity and symmetry are easy to show. For $\mathcal{G}_1 = (G_1, S, \mu_1, l_1)$ and $\mathcal{G}_2 = (G_2, S, \mu_2, l_2)$ with $\mathcal{G}_1 \not\simeq_d \mathcal{G}_2$, we need to argue $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) > 0$. Corollary 4 lets us assume that $\mathcal{G}_1, \mathcal{G}_2$ are least duplication-resolved trees, but not isomorphic. If G_1 and G_2 have the same topology, then there must be some $v \in V(G_1)$ whose correspondent $m_{\mathcal{G}_1,\mathcal{G}_2}(v)$ has either a different species or a different event label, as otherwise $\mathcal{G}_1$ and $\mathcal{G}_2$ would be isomorphic. In this case, $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) > 0$. If G_1 and G_2 have a different topology, then they must have some different clades and there is some $v \in V(G_1)$ such that its correspondent $v' = m_{\mathcal{G}_1,\mathcal{G}_2}(v)$ satisfies $L(v) \subsetneq L(v')$ (or if not, we swap the roles of $\mathcal{G}_1$ and $\mathcal{G}_2$ to guarantee this). If v and v' have a different species or label, we are done. If v and v' have the same species and label, then $v'' = m_{\mathcal{G}_2,\mathcal{G}_1}(v')$ must be a strict ancestor of v, which must have a different species or label than v by Lemma 1. This species or label also differs from v', and $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) > 0$. ◀

We next show that, despite being a semi-metric, the d_{plr} dissimilarity measure is not a metric since it does not satisfy the triangle inequality on non-binary gene trees, regardless of α. If α is a constant, it can even be far from satisfying the inequality.

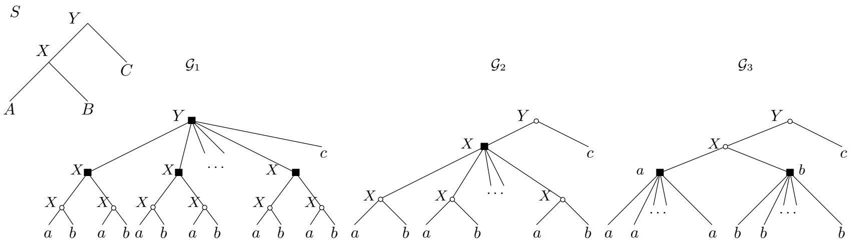

Figure 3 A species tree S and reconciled gene trees $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ that violate the triangle inequality.

▶ **Proposition 6.** *For any $\alpha \in [0, 1]$, possibly depending on the number of leaves of the gene trees, d_{plr} does not necessarily satisfy the triangle inequality under $\simeq_d$. This is true even if the gene trees use the lca-mapping.*

Moreover, for any fixed $\alpha < 1$, the quantity $d_{plr}(\mathcal{G}_1, \mathcal{G}_3)$ can be arbitrarily larger than $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) + d_{plr}(\mathcal{G}_2, \mathcal{G}_3)$.

Proof sketch. Consider the three reconciled gene trees $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ illustrated in Figure 3. Suppose that the root of the gene tree in $\mathcal{G}_1$ has, as children, $k \geq 2$ copies of an $((a, b), (a, b))$ subtree. Then the gene tree of $\mathcal{G}_2$ has $2k$ copies of an (a, b) subtree. Assume that every (a, b) subtree of $\mathcal{G}_1$ maps uniquely to some (a, b) subtree of $\mathcal{G}_2$. One of the main ideas is that each $((a, b), (a, b))$ subtree of $\mathcal{G}_1$ generates a d_{lbl} cost when compared to $\mathcal{G}_3$, because the duplication is a speciation in $\mathcal{G}_3$. We can calculate that $d_{plr}(\mathcal{G}_1, \mathcal{G}_2) + d_{plr}(\mathcal{G}_2, \mathcal{G}_3) = 4 - \alpha$ and $d_{plr}(\mathcal{G}_1, \mathcal{G}_3) = k(1 - \alpha) + 2\alpha + 3$. Since we can choose k as large as desired, we can infer the lack of triangle inequality, and arbitrarily large gap for constant $\alpha < 1$. ◀

We observe that in the example from Figure 3, the triangle inequality is violated mainly because the trees are heavily imbalanced in terms of number of internal nodes. We could not find counter-examples in which all trees are *binary*.

4 Diameters

We now study the question of computing the *diameter* of d_{plr}, which is the maximum possible dissimilarity achievable over a given instance size. This can be useful in practice for normalization, since we can compare heterogeneous datasets by dividing obtained dissimilarities by the diameter. In the context of general trees, the diameter is usually the maximum dissimilarity among all pairs of trees with the same number of leaves n. In reconciled gene trees though, there are multiple ways to define the diameter. We may fix two numbers n, m, and find the maximum d_{plr} value among all species trees on n leaves and pairs of gene trees on m leaves. Or, we could decide to fix the species tree S, and find the gene trees over m leaves of maximum d_{plr} value with respect to S. Or, we could fix the species tree S, and for each species leaf $s \in L(S)$ also fix the number m_s of extant genes that belong to s, and find the most distant gene trees under these parameters.

Even the simplest forms of diameters are not trivial to determine. We thus provide initial results by determining the diameter in the case that the species tree S is fixed, and gene trees contain exactly one gene per species. Even though this assumption may not hold in practice, we hope that the bounds established here can be extended to broader classes of scenarios in the future. We leave the question of finding the theoretical values of the other diameters as open problems.

For a fixed species tree S, let G^S represent the set of all reconciled gene trees $\mathcal{G} = (G, S, \mu, l)$, such that for each $s \in L(S)$, exactly one leaf x of G satisfies $\mu(x) = s$. Since each leaf of G is uniquely identifiable by its species, we assume that all the elements of G^S have the same leaves and are pairwise-comparable. We define the *diameter for fixed* S as:

$$diam(d_{plr}, S) = \max_{\mathcal{G}_1, \mathcal{G}_2 \in \mathsf{G}^S} \left\{ d_{plr}(\mathcal{G}_1, \mathcal{G}_2) \right\}$$

In terms of d_{lbl}, in the worst case $d_{lbl}(\mathcal{G}_1, \mathcal{G}_2)$ is the number of internal nodes of the gene tree of $\mathcal{G}_1$, which occurs when all labels differ. We next characterize the maximum possible path distance. It is tempting to make every node of $\mathcal{G}_1$ map to a deepest leaf of S, and every node of $\mathcal{G}_2$ to the root of S, thereby maximizing $dist_S(\mu_1(v), \mu_2(m(v)))$ for every node v, but such an example may not satisfy the rules of reconciliation.

For a species tree S, let $H(S) = \sum\limits_{v \in V(S) \setminus L(S)} dist_S(v, r(S))$ be the sum of root-to-internal node distances.

▶ **Lemma 7.** *Let S be a species tree on $n \geq 1$ leaves. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two reconciled trees in G^S. Then $d_{path}(\mathcal{G}_1, \mathcal{G}_2) \leq H(S) \leq (n-1)(n-2)/2$.*

Proof sketch. Denote $\mathcal{G}_1 = (G_1, S, \mu_1, l_1)$ and $\mathcal{G}_2 = (G_2, S, \mu_2, l_2)$. For the first bound, we can show that the maximum d_{path} is achieved when, for each $v \in V(G_1) \setminus L(G_1)$ and corresponding $v' = m_{\mathcal{G}_1, \mathcal{G}_2}(v)$, v uses the lca-mapping and v' is mapped to $r(S)$ (or vice-versa). The intuition is that one gene tree maps genes as low as possible, and the other as high as possible. Because of the time-consistency conditions, the gene tree that puts all ancestral genes as low as possible is obtained by copying the species tree. In this case, each ancestral species in $V(S) \setminus L(S)$ has one ancestral gene that maps to it, and so the highest sum-of-path distances adds the path lengths from $r(S)$ to every internal node. The second bound is a standard proof by induction and is achieved by caterpillar species trees. ◀

We can now proceed to prove the following theorem.

▶ **Theorem 8.** *Let S be a species tree on $n \geq 2$ leaves. Then*

$$diam(d_{plr}, S) = 2\alpha \cdot H(S) + (1-\alpha)(2n-2).$$

Moreover, among all species trees with n leaves, the diameter is maximized when S is a caterpillar, in which case $diam(d_{plr}, S) = \alpha(n-1)(n-2) + (1-\alpha)(2n-2)$.

Proof sketch. By Lemma 7, the d_{path} component is at most $2\alpha H(S)$ if we consider both d_{asym} directions. The $(1-\alpha)(2n-2)$ term is because the d_{lbl} is at most the number of nodes in the two gene trees, which is twice $n-1$ (since we have one gene per species). The upper bound is achieved if $\mathcal{G}_1$ is a copy of S, with all speciations and that uses the lca-mapping, and $\mathcal{G}_2$ is a copy of S, with all duplications and all gene tree nodes mapped to $r(S)$. ◀

On the labeled RF distances

We now take a brief detour into another distance designed to compare reconciliations, namely the labeled Robinson-Foulds distances as presented in [11, 12], of which there are two variants. These distances are used in the next section and we briefly discuss upper bounds on their diameters. An edge of a tree is *internal* if none of its endpoints is a leaf. *labeled tree* is a pair $\mathcal{T} = (T, l)$ where T is an unrooted tree without degree two nodes, and $l : V(T) \setminus L(T) \to X$ assigns some label from some set X to each internal node (one can think of the label set as

$X = \{spec, dup\}$). A *label-flip* is an operation that changes the label of an internal node. An *extension* is the reverse of a contraction: it takes a node v and a non-empty subset X of its neighbors, creates a new node w, deletes the edges $\{vx : x \in X\}$, then adds the edges $\{wx : x \in X\}$ along with vw, such that the latter must be internal. A *labeled contraction* is an operation that contracts an internal edge uv satisfying $l(u) = l(v)$, and a *labeled extension* is an extension of v that creates node w with $l(w) = l(v)$.

Given two labeled trees $\mathcal{T}_1 = (T_1, l_1), \mathcal{T}_2 = (T_2, l_2)$, the *ELRF distance* [11] between $\mathcal{T}_1$ and $\mathcal{T}_2$ is the minimum number of labeled contractions, labeled extensions, and label-flips required to transform $\mathcal{T}_1$ into $\mathcal{T}_2$.

The *LRF distance* [12] is the minimum number of contractions, extensions, and label-flips required to transform $\mathcal{T}_1$ into $\mathcal{T}_2$ (note that the authors use the notion of node deletions and insertions, but are stated in [12] to be the same as contractions and extensions).

For an integer $n \geq 3$, the diameter of the ELFR (resp. LRF) distance is the largest possible distance among all possible labeled trees with n leaves. These diameters were not discussed in the literature. We provide bounds which we believe to be tight, under the assumption that the label set consists of two elements $X = \{spec, dup\}$.

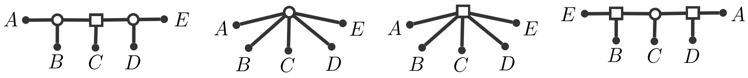

Figure 4 An example of two labeled trees (left and right), with $n = 5$ leaves and two internal edges, which both need to be contracted. To achieve this under the ELRF distance, we can perform $\lfloor (n-2)/2 \rfloor = 1$ relabeling to make every label a circle (not shown), then contract every internal edge to obtain a star tree (second drawing). We can then change the remaining label, and reverse the operations to obtain the right tree. This takes $7 = 3n - 8$ operations.

▶ **Proposition 9.** *For $n \geq 3$ and label set X of size 2, the ELRF diameter is at most $3n - 8$. Furthermore, the LRF diameter is at most $2n - 5$.*

The intuition is that we can always contract all $n - 3$ internal edges of the first tree. In ELRF, we may have to relabel half of the $n - 2$ internal nodes to do this, so using $n - 3 + (n - 2)/2$ operations to reach a star tree (in the proof we show that this bound can be achieved while also attaining any desired label at the root of the star, with some case handling required for odd versus even n). This has to be reversed, leading to $3n - 8$. In LRF, we can just contract all $n - 3$ internal edges directly, possibly relabel the internal node of the star tree, then extend. It is possible that these bounds are tight. Consider Figure 4 for the ELRF distance. If we generalize this pattern, it would appear that we need to flip $\lfloor (n-2)/2 \rfloor$ nodes, do $n - 3$ mandatory contractions, flip the central node, and reverse the process. This results in the upper bound $3n - 8$. For LRF, one can think of a pair of trees with no label in common, that require the maximum of $2n - 6$ contractions and extensions, plus a label flip. However, proving that such examples cannot be handled better is not trivial, and since these distances are not the focus of the paper, we reserve those for future work.

5 Methods

We compared the distribution of the PLR semi-metric against the classical Robinson-Foulds (RF) and its ELRF and LRF variants. To this end, we designed and implemented a stepwise procedure to simulate reconciled trees. The software tool to compute d_{plr} is available as open source at: `https://pypi.org/project/parle/`.

5.1 Simulation of reconciliations

The existing programs for simulation of reconciliations like AsymmeTree or SaGePhy [60, 38] operate in a top-bottom fashion by first simulating ancestral genes/species followed by a birth-death process generating speciation, duplication, and loss events among others. This procedure does not guarantee trees with a fixed set of genes, whereas the PLR, LRF, and ELRF metrics require trees with the same set of leaves. To fulfill this requirement, we designed Algorithm 2, which takes as input a species tree S, as well as a set of genes Γ and the assignment of species $\sigma : \Gamma \to L(S)$, then builds a reconciled gene tree over leafset Γ in a bottom-up fashion. At each iteration it picks pairs of genes $x', x'' \in \Gamma$ and substitutes them with a newly created node x, being the parent of the chosen genes. Finally, x is associated with an event and mapped to the species tree in Line 7. Algorithm 2 uses the lca-mapping μ for the generated gene trees. It is known that this map satisfies time-consistency, and that a node x with children x', x'' can be a speciation if and only if $\mu(x) \notin \{\mu(x'), \mu(x'')\}$[27]. If this is not satisfied, the algorithm assigns $l(x) = dup$, and otherwise chooses $l(x) \in \{dup, spec\}$, which guarantees the *speciations separate species* condition.

Algorithm 2 considers a probability distribution P of picking $x', x'' \in \Gamma$. In our implementation, this probability decays exponentially w.r.t. the distance between the species where x' and x'' reside, in other words, the larger $d = dist_S(\mu(x'), \mu(x''))$ is, the smaller the chance of choosing x', x''. In particular, we use the probability $e^{-0.7d}$. This approach is intended to prevent close elements in the gene tree from being mapped to distant nodes in the species tree, such a setting causes most of the inner nodes in the gene tree to be mapped near the root of the species tree, which would in turn create many *dup* nodes.

In total, we generated 9 sets of random reconciliations, obtained as follows. First, we generated three species trees S_n, where n is the number of leaves: S_{10}, S_{25}, and S_{50}, using the `AsymmeTree` package [60] under the *innovations model* as described in [34]. For each species tree S_i we generated the gene sets $\Gamma_{i,5}$, $\Gamma_{i,10}$, and $\Gamma_{i,20}$, together with the assignments of species $\sigma_{i,5}$, $\sigma_{i,10}$, and $\sigma_{i,20}$ in such a way that for the set $\Gamma_{i,j}$ each species $y \in L(S_i)$ contains at least one gene and at most j genes. Considering this restriction, the number of genes for each species was chosen with uniform probability.

Algorithm 2 Simulation of random reconciliation scenarios.

```
1  function generate_random_scenario($S, \Gamma, \sigma$)
       // $S$ is a species tree, $\Gamma$ is the set of genes, $\sigma$ is a map from $\Gamma$
          to their species.
2      Initialise $\mathcal{G} = (G, S, \mu, l)$ with $L(G) = \Gamma$ such that $\mu$ maps every leaf to their
        corresponding species in $L(S)$ according to $\sigma$
3      while $|\Gamma| > 1$ do
4          Pick two genes $x', x''$ in $\Gamma$ according to a probability distribution $P$.
5          Create a new node $x$ as the parent of $x'$ and $x''$.
           // Set reconciliation map and label of the new node.
6          Set $\mu(x) = lca_S(\mu(x'), \mu(x''))$
7          if $\mu(x) \in \{\mu(x'), \mu(x'')\}$ then $l(x) = dup$
8          else choose $l(x)$ from $\{dup, spec\}$ with uniform probability
9          $\Gamma \leftarrow (\Gamma \setminus \{x', x''\}) \cup \{x\}$                    // Update set of genes
10     return $\mathcal{G}$
```

Distance distribution and normalization

Given a set $R_{i,j}$ of random reconciliations generated from S_i and $\Gamma_{i,j}$, we computed the PLR, ELRF, LRF, and RF measures for each pair of different reconciliations. We set $|R_{i,j}| = 50$, resulting in 1225 total comparisons per set of random reconciliations. As argued in Section 2.2, the parameter α of PLR is aimed to balance the quadratic-versus-linear components of the distance. Following this analysis, we set $\alpha = 1/i$ for the dataset $R_{i,j}$. Furthermore, to address the impact of α on the metric we also used the values 0, 0.25, 0.5, 0.75, and 1.

We normalized the distances obtained to have a fair comparison between the distribution of the different measures. We used two strategies, first, we normalized PLR by the theoretical diameter of the distance, while ELFR by its upper bond, and second by the empirical max normalization, which consists of dividing each computed value of a measure by the maximum encountered in the dataset for that measure.

5.2 Computational results

Comparisons with max-normalization

Each subplot of Figure 5 shows four distributions comparing the PLR, ELRF, LRF, and RF metrics represented in blue, light orange, green, and red, respectively.

The ELRF, LRF, and RF distributions exhibit right-skewness, indicating that many data points cluster towards higher values. This skewness suggests a higher frequency of larger distances, a common trait among these metrics. Notably, the RF metric often shows smaller distances because it ignores label changes, whereas the ELRF and LRF metrics yield almost identical values, performing very similarly, as expected.

In contrast, the PLR distribution is centered around its mean, displaying a broader spread of measurements. This symmetric distribution indicates that the PLR metric has a greater variability in distance measurements, highlighting its sensitivity, that is, a balanced penalization of all the elements of an evolutionary scenario. This contrasts with the more concentrated and nearly identical distributions of ELRF, LRF, and RF.

The theoretical diameter is hard to reach

Figure 6 presents the distribution of the ELRF distance and the PLR distance for various values of the parameter α. We omit the plots for LRF and RF distances since they closely resemble the ELRF distributions, as discussed in the previous section.

The first two rows in Figure 6 compare trees with fewer duplications than speciations, while the subsequent rows involve trees with an equal or greater number of duplications compared to speciations. The PLR measure is normalized by the theoretical diameter introduced here, whereas the ELRF is normalized by its upper bound. Note that ELRF consistently has higher values than PLR and that these values are significantly far from the theoretical diameter. The shape of the PLR distribution remains largely unchanged as α increases, likely due to the diminishing contribution of the linear component relative to the quadratic component as α grows. On the right side of the figure, we observe the frequency of speciation and duplication events in our simulated reconciled trees, as well as their least duplication-resolved (LDR) counterparts. Notably, when there are more speciations than duplications, the PLR measure increases but still remains far from the theoretical diameter.

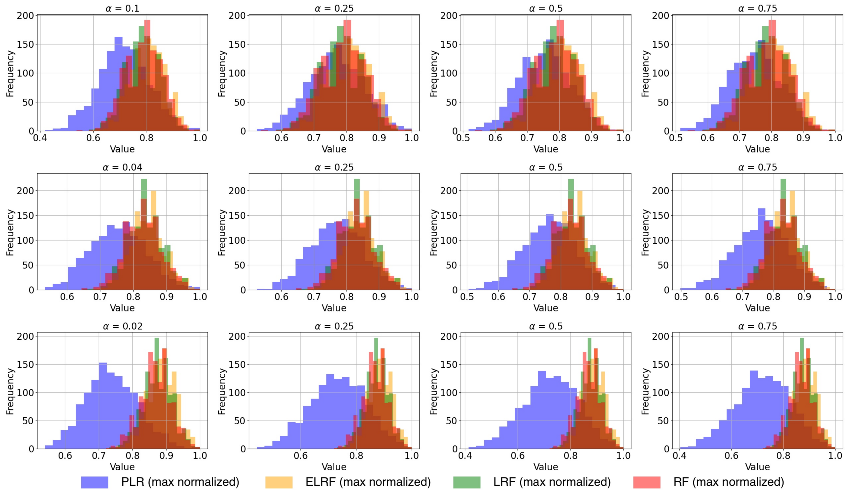

Figure 5 Distributions of the PLR, ELRF, LRF, and RF metrics for datasets $\Gamma_{10,20}$, $\Gamma_{25,10}$, and $\Gamma_{50,5}$, from top to bottom rows, respectively, and alpha values from the set $\{\frac{1}{n}, 0.25, 0.5, 0.75\}$, with n as number of species. Each row corresponds to a dataset, while each column represents a different value of α. The x-axis represents max-normalized values ranging from 0 to 1, and the y-axis is the frequency of these values. The PLR measure in purple shows a centered and symmetric distribution with a broader spread. The ELRF, LRF, and RF metrics, shown in light orange, green, and red, respectively, exhibit right-skewed distributions towards the higher end of the scale.

Figure 7 illustrates important differences between the measures, since we can observe two different scenarios: 1) where ELRF is significantly smaller than PLR, suggesting that reconciliations may be completely different even when gene tree topologies are similar; and 2) conversely, PLR may be significantly small when the ELRF is large, suggesting that different gene tree topologies could have similar reconciliations.

6 Discussion

In this work, we have underscored the unique attributes of PLR, a novel semi-metric designed for comparing reconciled gene trees within a fixed species tree framework. Unlike existing metrics such as RF, LRF, and ELRF, which primarily focus on tree topology, PLR incorporates all elements of an evolutionary scenario: a species tree, gene trees, speciation/duplication labeling and a mapping from gene trees to species tree. This broader scope provides a more holistic measure of dissimilarity between reconciled gene trees, offering researchers a nuanced understanding of evolutionary relationships.

One notable advantage of PLR is its flexibility, particularly regarding the parameter α, which allows users to balance the quadratic and linear components of the distance according to their specific research context. This flexibility enhances the metric's applicability across diverse evolutionary scenarios, providing researchers with a customizable tool for reconciliation analysis. Additionally, our experiments reveal that PLR exhibits a symmetric and broadly spread distribution around its mean, indicating sensitivity to variations in reconciliations and finer granularity in distinguishing between different tree pairs. Despite its strengths, PLR does have some limitations. For instance, while the flexibility of α is advantageous, it also introduces a degree of subjectivity into the metric's application, as users must determine the appropriate value for their specific context. Moreover, our theoretical analysis highlights a large theoretical diameter for PLR, which is seldom reached in practice. Tighter bounds

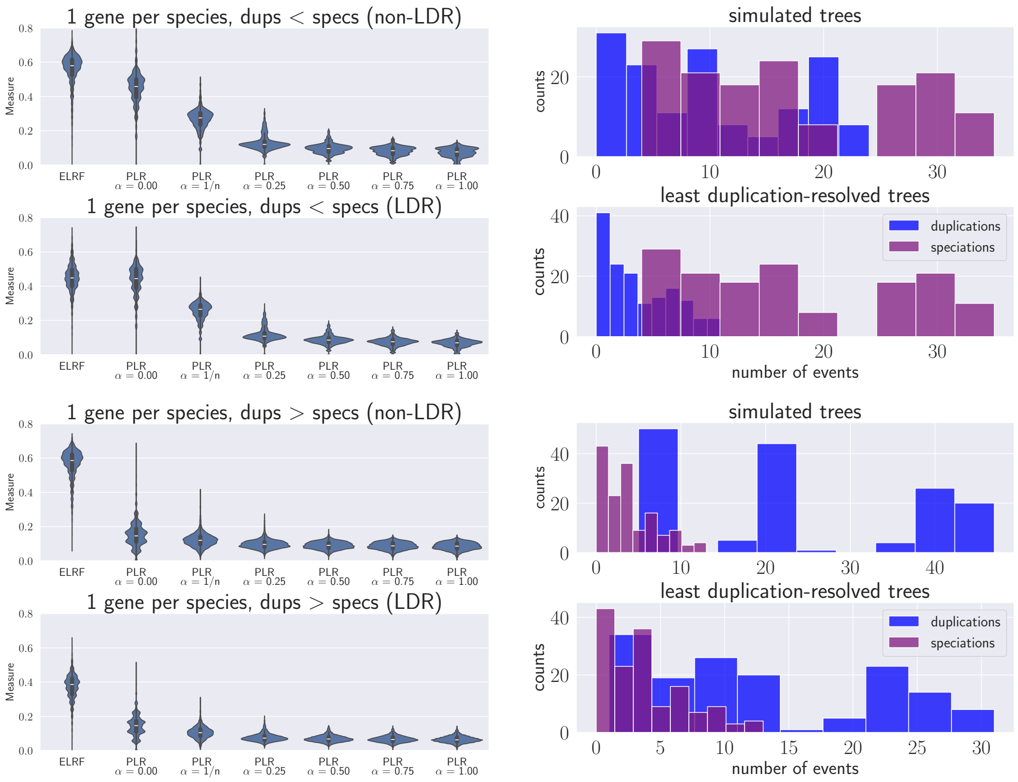

Figure 6 Comparison of the distribution of ELRF and PLR measures with different values for the parameter α, and different proportions of duplication/speciation events. The measures are shown for both the least duplication-resolved trees (LDR) and non-LDR. All the plots consider reconciliations with 10, 25, and 50 species. The parameter $\alpha = 1/n$ aims to balance the linear-vs-quadratic components of the distance, where n is the number of species. Note that the biggest change in the distribution of the PLR measure happens for small values of α.

are needed to improve practical applicability and interpretability. One of the key strengths of PLR is its computational efficiency, with an $O(n)$ time complexity. This efficiency is particularly beneficial for analyzing large datasets or trees, where computational resources and time are critical constraints.

Looking ahead, future directions for PLR include refining the theoretical bounds of its diameter. An important theoretical problem that remains open is determining whether *binary* gene trees satisfy the triangle inequality. Additionally, developing metrics between gene trees with different leaf sets would significantly broaden its applicability. Incorporating alternative methods for matching ancestral genes, such as those proposed by Lin et al. [44], or using asymmetric cluster affinity as suggested by Wagle [65], could further enhance the metric's accuracy and relevance.

In conclusion, PLR represents a significant advancement in the comparison of reconciled gene trees, offering a detailed and flexible measure of dissimilarity. Its computational efficiency and comprehensive event consideration make it a valuable tool for evolutionary studies, with potential for further refinement and application in future research.

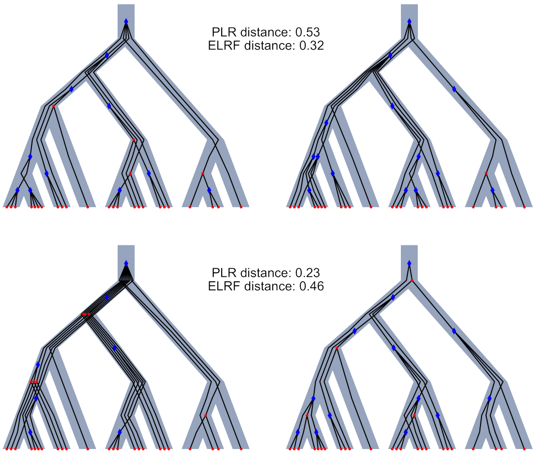

Figure 7 Examples of distance between reconciliations and gene trees, plotted using `REvolutionH-tl` [55]. The reconciliations have 10 species and 24 genes, with $\alpha = 1/10$. The upper row has a large PLR value but a small ELRF distance. In contrast, the bottom row shows trees when PLR is small even when ELRF is big. In this example, we set $\alpha = 1/10$.

References

1 Örjan Åkerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–5719, 2009.

2 Yoann Anselmetti, Nadia El-Mabrouk, Manuel Lafond, and Aïda Ouangraoua. Gene tree and species tree reconciliation with endosymbiotic gene transfer. *Bioinformatics*, 37(Supplement_1):i120–i132, 2021.

3 Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren, and Bengt Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics-Oxford*, 19(1):7–15, 2003.

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

5 Mukul S Bansal and Oliver Eulenstein. The multiple gene duplication problem revisited. *Bioinformatics*, 24(13):i132–i138, 2008.

6 Mukul S Bansal, Manolis Kellis, Misagh Kordi, and Soumya Kundu. Ranger-dtl 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216, 2018.

7 Bérénice Batut, David P Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution: a tool to test evolutionary scenarios. In *BMC bioinformatics*, volume 14, pages 1–11. Springer, 2013.

8 Michael A Bender and Martin Farach-Colton. The lca problem revisited. In *LATIN 2000: Theoretical Informatics: 4th Latin American Symposium, Punta del Este, Uruguay, April 10-14, 2000 Proceedings 4*, pages 88–94. Springer, 2000.

9 Paola Bonizzoni, Gianluca Della Vedova, and Riccardo Dondi. Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical computer science*, 347(1-2):36–53, 2005.

10 Bastien Boussau and Celine Scornavacca. Reconciling gene trees with species trees. *Phylogenetics in the genomic era*, pages 3–2, 2020.

11 Samuel Briand, Christophe Dessimoz, Nadia El-Mabrouk, Manuel Lafond, and Gabriela Lobinska. A generalized robinson-foulds distance for labeled trees. *BMC Genomics*, 21(S10), November 2020. doi:10.1186/s12864-020-07011-0.

12 Samuel Briand, Christophe Dessimoz, Nadia El-Mabrouk, and Yannis Nevers. A linear time solution to the labeled robinson–foulds distance problem. *Systematic Biology*, 71(6):1391–1403, 2022.

13 J Gordon Burleigh, Mukul S Bansal, Andre Wehe, and Oliver Eulenstein. Locating multiple gene duplications through reconciled trees. In *Research in Computational Molecular Biology: 12th Annual International Conference, RECOMB 2008, Singapore, March 30-April 2, 2008. Proceedings 12*, pages 273–284. Springer, 2008.

14 Yao-ban Chan, Vincent Ranwez, and Céline Scornavacca. Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *Journal of theoretical biology*, 432:1–13, 2017.

15 Chris Conow, Daniel Fielder, Yaniv Ovadia, and Ran Libeskind-Hadas. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5:1–10, 2010.

16 Adrián A Davín, Théo Tricou, Eric Tannier, Damien M de Vienne, and Gergely J Szöllősi. Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*, 36(4):1286–1288, 2020.

17 Mattéo Delabre, Nadia El-Mabrouk, Katharina T Huber, Manuel Lafond, Vincent Moulton, Emmanuel Noutahi, and Miguel Sautie Castellanos. Reconstructing the history of syntenies through super-reconciliation. In *Comparative Genomics: 16th International Conference, RECOMB-CG 2018, Magog-Orford, QC, Canada, October 9-12, 2018, Proceedings 16*, pages 179–195. Springer, 2018.

18 Riccardo Dondi, Manuel Lafond, and Celine Scornavacca. Reconciling multiple genes trees via segmental duplications and losses. *Algorithms for Molecular Biology*, 14:1–19, 2019.

19 Jean-Philippe Doyon, Celine Scornavacca, K Yu Gorbunov, Gergely J Szöllősi, Vincent Ranwez, and Vincent Berry. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *Comparative Genomics: International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9-11, 2010. Proceedings 8*, pages 93–108. Springer, 2010.

20 Dannie Durand, Bjarni V Halldórsson, and Benjamin Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. In *Research in Computational Molecular Biology: 9th Annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005. Proceedings 9*, pages 250–264. Springer, 2005.

21 Manuela Geiß, Marcos E González Laffitte, Alitzel López Sánchez, Dulce I Valdivia, Marc Hellmuth, Maribel Hernández Rosales, and Peter F Stadler. Best match graphs and reconciliation of gene trees with species trees. *Journal of mathematical biology*, 80(5):1459–1495, 2020.

22 Manuela Geiß, Marcos E. González Laffitte, Alitzel López Sánchez, Dulce I. Valdivia, Marc Hellmuth, Maribel Hernández Rosales, and Peter F. Stadler. Best match graphs and reconciliation of gene trees with species trees. *Journal of Mathematical Biology*, 80(5):1459–1495, January 2020. doi:10.1007/s00285-020-01469-y.

23 Pablo A Goloboff, Joan S Arias, and Claudia A Szumik. Comparing tree shapes: beyond symmetry. *Zool. Scr.*, 46(5):637–648, September 2017.

24 Morris Goodman, John Czelusniak, G William Moore, Alejo E Romero-Herrera, and Genji Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163, 1979.

25 Pawel Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. In *Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 316–325, 2004.

26 Paweł Górecki, Natalia Rutecka, Agnieszka Mykowiecka, and Jarosław Paszek. Unifying duplication episode clustering and gene-species mapping inference. *Algorithms for Molecular Biology*, 19(1):1–20, 2024.

27 Paweł Górecki and Jerzy Tiuryn. Dls-trees: a model of evolutionary scenarios. *Theoretical computer science*, 359(1-3):378–399, 2006.

28 Damir Hasić and Eric Tannier. Gene tree species tree reconciliation with gene conversion. *Journal of mathematical biology*, 78(6):1981–2014, 2019.

29 Marc Hellmuth, Maribel Hernandez-Rosales, Katharina T. Huber, Vincent Moulton, Peter F. Stadler, and Nicolas Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *Journal of Mathematical Biology*, 66(1–2):399–420, March 2012. `doi:10.1007/s00285-012-0525-x`.

30 Maribel Hernandez-Rosales, Marc Hellmuth, Nicolas Wieseke, Katharina T Huber, Vincent Moulton, and Peter F Stadler. From event-labeled gene trees to species trees. In *BMC bioinformatics*, volume 13, pages 1–11. Springer, 2012.

31 Katharina T. Huber, Vincent Moulton, Marie-France Sagot, and Blerina Sinaimeri. Geometric medians in reconciliation spaces of phylogenetic trees. *Information Processing Letters*, 136:96–101, August 2018. `doi:10.1016/j.ipl.2018.04.001`.

32 Edwin Jacox, Cedric Chauve, Gergely J Szöllősi, Yann Ponty, and Celine Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.

33 Edwin Jacox, Mathias Weller, Eric Tannier, and Celine Scornavacca. Resolution and reconciliation of non-binary gene trees with transfers, duplications and losses. *Bioinformatics*, 33(7):980–987, 2017.

34 Stephanie Keller-Schmidt and Konstantin Klemm. A model of macroevolution as a branching process based on innovations. *Advances in Complex Systems*, 15(07):1250043, 2012.

35 Misagh Kordi and Mukul S Bansal. Exact algorithms for duplication-transfer-loss reconciliation with non-binary gene trees. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 297–306, 2016.

36 Misagh Kordi, Soumya Kundu, and Mukul S Bansal. On inferring additive and replacing horizontal gene transfers through phylogenetic reconciliation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 514–523, 2019.

37 Esaie Kuitche, Manuel Lafond, and Aïda Ouangraoua. Reconstructing protein and gene phylogenies using reconciliation and soft-clustering. *Journal of bioinformatics and computational biology*, 15(06):1740007, 2017.

38 Soumya Kundu and Mukul S Bansal. SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics*, 35(18):3496–3498, February 2019. `doi:10.1093/bioinformatics/btz081`.

39 Manuel Lafond, Krister M Swenson, and Nadia El-Mabrouk. An optimal reconciliation algorithm for gene trees with polytomies. In *Algorithms in Bioinformatics: 12th International Workshop, WABI 2012, Ljubljana, Slovenia, September 10-12, 2012. Proceedings 12*, pages 106–122. Springer, 2012.

40 Manuel Lafond, Krister M Swenson, and Nadia El-Mabrouk. Error detection and correction of gene trees. *Models and algorithms for genome evolution*, pages 261–285, 2013.

41 Bret R Larget, Satish K Kotha, Colin N Dewey, and Cécile Ané. Bucky: gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 2010.

42 Lei Li and Mukul S Bansal. Simultaneous multi-domain-multi-gene reconciliation under the domain-gene-species reconciliation model. In *Bioinformatics Research and Applications: 15th International Symposium, ISBRA 2019, Barcelona, Spain, June 3–6, 2019, Proceedings 15*, pages 73–86. Springer, 2019.

43 Qiuyi Li, Celine Scornavacca, Nicolas Galtier, and Yao-Ban Chan. The multilocus multispecies coalescent: a flexible new model of gene family evolution. *Systematic Biology*, 70(4):822–837, 2021.

44 Yu Lin, Vaibhav Rajan, and Bernard ME Moret. A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1014–1022, 2011.

45 Jingyi Liu, Ross Mawhorter, Nuo Liu, Santi Santichaivekin, Eliot Bush, and Ran Libeskind-Hadas. Maximum parsimony reconciliation in the dtlor model. *BMC bioinformatics*, 22:1–22, 2021.

46 Alitzel López Sánchez, José Antonio Ramírez-Rafael, Alejandro Flores-Lamas, Maribel Hernández-Rosales, and Manuel Lafond. PARLE: Path Analysis, Reconciliation, and Label Evaluation. Software, version 0.0.2. (visited on 2024-08-19). URL: https://pypi.org/project/parle/.

47 V Makarenkov and B Leclerc. Comparison of additive trees using circular orders. *J. Comput. Biol.*, 7(5):731–744, 2000.

48 Diego Mallo, Leonardo de Oliveira Martins, and David Posada. Simphy: phylogenomic simulation of gene, locus, and species trees. *Systematic biology*, 65(2):334–344, 2016.

49 Tamara Munzner, François Guimbretière, Serdar Tasiran, Li Zhang, and Yunhong Zhou. TreeJuxtaposer. In *ACM SIGGRAPH 2003 Papers*, New York, NY, USA, July 2003. ACM.

50 Nikolai Nøjgaard, Manuela Geiß, Daniel Merkle, Peter F Stadler, Nicolas Wieseke, and Marc Hellmuth. Time-consistent reconciliation maps and forbidden time travel. *Algorithms for Molecular Biology*, 13:1–17, 2018.

51 Nikolai Nøjgaard, Manuela Geiß, Daniel Merkle, Peter F. Stadler, Nicolas Wieseke, and Marc Hellmuth. Time-consistent reconciliation maps and forbidden time travel. *Algorithms for Molecular Biology*, 13(1), February 2018. doi:10.1186/s13015-018-0121-8.

52 Roderic DM Page and JA Cotton. Vertebrate phylogenomics: reconciled trees and gene duplications. In *Biocomputing 2002*, pages 536–547. World Scientific, 2001.

53 Jarosław Paszek and Paweł Górecki. Efficient algorithms for genomic duplication models. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5):1515–1524, 2017.

54 Pere Puigbò, Santiago Garcia-Vallvé, and James O McInerney. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics*, 23(12):1556–1558, June 2007.

55 José Antonio Ramírez-Rafael, Annachiara Korchmaros, Katia Aviña-Padilla, Alitzel López Sánchez, Andrea Arlette España-Tinajero, Marc Hellmuth, Peter F. Stadler, and Maribel Hernández-Rosales. Revolutionh-tl: Reconstruction of evolutionary histories tool. In Celine Scornavacca and Maribel Hernández-Rosales, editors, *Comparative Genomics*, pages 89–109, Cham, 2024. Springer Nature Switzerland.

56 Matthew D Rasmussen and Manolis Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research*, 22(4):755–765, 2012.

57 Santi Santichaivekin, Qing Yang, Jingyi Liu, Ross Mawhorter, Justin Jiang, Trenton Wesley, Yi-Chieh Wu, and Ran Libeskind-Hadas. empress: a systematic cophylogeny reconciliation tool. *Bioinformatics*, 37(16):2481–2482, 2021.

58 H M Savage. The shape of evolution: systematic tree topology. *Biol. J. Linn. Soc. Lond.*, 20(3):225–244, November 1983.

59 David Schaller, Marc Hellmuth, and Peter F Stadler. Asymmetree: a flexible python package for the simulation of complex gene family histories. *Software*, 1(3):276–298, 2022.

60 David Schaller, Marc Hellmuth, and Peter F Stadler. AsymmeTree: A flexible python package for the simulation of complex gene family histories. *Software*, 1(3):276–298, August 2022.

61 David Schaller, Manuel Lafond, Peter F Stadler, Nicolas Wieseke, and Marc Hellmuth. Indirect identification of horizontal gene transfer. *Journal of mathematical biology*, 83(1):10, 2021.

62 Celine Scornavacca, Joan Carles Pons Mayol, and Gabriel Cardona. Fast algorithm for the reconciliation of gene trees and lgt networks. *Journal of theoretical biology*, 418:129–137, 2017.

63 Maureen Stolzer, Han Lai, Minli Xu, Deepa Sathaye, Benjamin Vernot, and Dannie Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415, 2012.

64 Benjamin Vernot, Maureen Stolzer, Aiton Goldman, and Dannie Durand. Reconciliation with non-binary species trees. *Journal of computational biology*, 15(8):981–1006, 2008.

65 Sanket Wagle, Alexey Markin, Paweł Górecki, Tavis K. Anderson, and Oliver Eulenstein. Asymmetric cluster-based measures for comparative phylogenetics. *Journal of Computational Biology*, 31(4):312–327, April 2024. doi:10.1089/cmb.2023.0338.

66 Samson Weiner and Mukul S Bansal. Improved duplication-transfer-loss reconciliation with extinct and unsampled lineages. *Algorithms*, 14(8):231, 2021.

67 Yi-Chieh Wu, Matthew D Rasmussen, Mukul S Bansal, and Manolis Kellis. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome research*, 24(3):475–486, 2014.

68 Louxin Zhang. On a mirkin-muchnik-smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4(2):177–187, 1997.

69 Louxin Zhang. From gene trees to species trees ii: Species tree inference by minimizing deep coalescence events. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1685–1691, 2011.