Forbidden time travel

Characterization of time-consistent tree reconciliation maps

Nøjgaard, Nikolai; Geiß, Manuela; Merkle, Daniel; Stadler, Peter F.; Wieseke, Nicolas; Hellmuth, Marc

Published in:
17th International Workshop on Algorithms in Bioinformatics, WABI 2017

DOI:
10.4230/LIPIcs.WABI.2017.17

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Forbidden Time Travel: Characterization of Time-Consistent Tree Reconciliation Maps*

Nikolai Nøjgaard[1], Manuela Geiß[2], Daniel Merkle[3], Peter F. Stadler[4], Nicolas Wieseke[5], and Marc Hellmuth[6]

1 **Department of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany; and Department of Mathematics and Computer Science, University of Southern Denmark, Denmark**
2 **Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany**
3 **Department of Mathematics and Computer Science, University of Southern Denmark, Denmark**
4 **Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany; and Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany; and Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria; and Santa Fe Institute, Santa Fe, NM, USA**
5 **Parallel Computing and Complex Systems Group, Department of Computer Science, Leipzig University, Leipzig, Germany**
6 **Department of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany; and Saarland University, Center for Bioinformatics, Saarbrücken, Germany**

Abstract

Motivation: In the absence of horizontal gene transfer it is possible to reconstruct the history of gene families from empirically determined orthology relations, which are equivalent to *event-labeled* gene trees. Knowledge of the event labels considerably simplifies the problem of reconciling a gene tree T with a species trees S, relative to the reconciliation problem without prior knowledge of the event types. It is well-known that optimal reconciliations in the unlabeled case may violate time-consistency and thus are not biologically feasible. Here we investigate the mathematical structure of the event labeled reconciliation problem with horizontal transfer.

Results: We investigate the issue of time-consistency for the event-labeled version of the reconciliation problem, provide a convenient axiomatic framework, and derive a complete characterization of time-consistent reconciliations. This characterization depends on certain weak conditions on the event-labeled gene trees that reflect conditions under which evolutionary events are observable at least in principle. We give an $\mathcal{O}(|V(T)|\log(|V(S)|))$-time algorithm to decide whether a time-consistent reconciliation map exists. It does not require the construction of explicit timing maps, but relies entirely on the comparably easy task of checking whether a small auxiliary graph is acyclic.

Significance: The combinatorial characterization of time consistency and thus biologically feasible reconciliation is an important step towards the inference of gene family histories with horizontal transfer from orthology data, i.e., without presupposed gene and species trees. The fast algorithm to decide time consistency is useful in a broader context because it constitutes an attractive component for all tools that address tree reconciliation problems.

* Supported in part by the Danish Council for Independent Research, Natural Sciences, grants DFF-1323-00247 and DFF-7014-00041.

17th International Workshop on Algorithms in Bioinformatics (WABI 2017).
Editors: Russell Schwartz and Knut Reinert;
Article No. 17; pp. 17:1–17:12

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1998 ACM Subject Classification G.2.2 Graph Theory, G.2.3 Applications, F.2.2 Nonnumerical Algorithms and Problems

Keywords and phrases Tree Reconciliation, Horizontal Gene Transfer, Reconciliation Map, Time-Consistency, History of gene families

Digital Object Identifier 10.4230/LIPIcs.WABI.2017.17

1 Introduction

Modern molecular biology describes the evolution of species in terms of the evolution of the genes that collectively form an organism's genome. In this picture, genes are viewed as atomic units whose evolutionary history *by definition* forms a tree. The phylogeny of species also forms a tree. This species tree is either interpreted as a consensus of the gene trees or it is inferred from other data. An interesting formal manner to define a species tree independent of genes and genetic data is discussed e.g. in [7]. In this contribution, we assume that gene and species trees are given independently of each other. The relationship between gene and species evolution is therefore given by a reconciliation map that describes how the gene tree is embedded in the species tree: after all, genes reside in organisms, and thus at each point in time can be assigned to a species.

From a formal point of view, a reconciliation map μ identifies vertices of a gene tree with vertices and edges in the species tree in such a way that (partial) ancestor relations given by the genes are preserved by μ. Vertices in the species tree correspond to speciation events. Since in this situation genes are faithfully transmitted from the parent species into both (all) daughter species, some of the vertices in the gene tree correspond to speciation events. Other important events considered here are gene duplications, in which two copies of a gene keep residing in the same species, and horizontal gene transfer events (HGT). Here, the original remains in the parental species, while the offspring copy "jumps" into a different branch of the species tree. It is customary to define pairwise relations between genes depending on the event type of their last common ancestor [8, 11, 13].

Most of the literature on this topic assumes that both the gene tree and the species tree are known. The aim is then to find a mapping of the gene tree T into the species tree S and, at least implicitly, an event-labeling on the vertices of the gene tree T. Here we take a different point of view and assume that T and the types of evolutionary events on T are known. This setting has ample practical relevance because event-labeled gene trees can be derived from the pairwise orthology relation [15, 13]. These relations in turn can be estimated directly from sequence data using a variety of algorithmic approaches that are based on the pairwise best match criterion and hence do not require any *a priori* knowledge of the topology of either the gene tree or the species tree, see e.g. [19, 2, 17, 1].

Genes that share a common origin (homologs) can be classified into orthologs, paralogs, and xenologs depending whether they originated by a speciation, duplication or horizontal gene transfer (HGT) event [8, 13]. Recent advances in mathematical phylogenetics [10, 11, 15] have shown that the knowledge of these event-relations (orthologs, paralogs and xenologs) suffices to construct event-labeled gene trees and, in some case, also a species tree.

Conceptually, both the gene tree and species tree are associated with a timing of each event. Reconciliation maps must preserve this timing information because there are *biologically infeasible* event labeled gene trees that cannot be reconciled with any species tree. In the absence of HGT, biologically feasibility can be characterized in terms of certain triples (rooted binary trees on three leaves) that are displayed by the gene trees [16]. In contrast, the timing

information must be taken into account explicitly in the presence of HGT. In other words, there are gene trees with HGT that can be mapped to species trees only in such a way that some genes travels back in time.

There have been several attempts in the literature to handle this issue, see e.g. [6] for a review. In [18, 5] a *single* HGT adds timing constraints to a time map for a reconciliation to be found. Time-consistency is then subsequently defined based on the existence of a topological order of the digraph reflecting all the time constraints. In [20] NP-hardness was shown for finding a parsimonious time-consistent reconciliation based on a definition for time-consistency that essentially is based on considering *pairs* of HGTs. However, the latter definitions are explicitly designed for *binary* gene trees and do not apply to non-binary gene trees, which are used here to model incomplete knowledge of the exact gene phylogenies. Different algorithmic approaches for tackling time-consistency exist [6] such as the inclusion of time-zones known for specific evolutionary events. It is worth noting that *a posteriori* modifications of time-inconsistent solutions will in general violate parsimony [18]. So-far, no results have become available to determine the *existence* of time-consistent reconciliation maps given the (undated) species tree and the event-labeled gene tree.

Here, we introduce an axiomatic framework for time-consistent reconciliation maps and characterize for given event-labeled gene trees and species trees whether there exists a time-consistent reconciliation map. We provide an algorithm that constructs a time-consistent reconciliation map if one exists. The algorithms are implemented in `C++` using the boost graph library and are freely available at `https://github.com/Nojgaard/tc-recon`. In addition, the proofs and additional information on this paper are provided at this url.

2 Notation and Preliminaries

We consider *rooted trees* $T = (V, E)$ *(on* L_T*)* with root $\rho_T \in V$ and leaf set $L_T \subseteq V$. A vertex $v \in V$ is called a *descendant* of $u \in V$, $v \preceq_T u$, and u is an *ancestor* of v, $u \succeq_T v$, if u lies on the path from ρ_T to v. As usual, we write $v \prec_T u$ and $u \succ_T v$ to mean $v \preceq_T u$ and $u \neq v$. The partial order $\succeq_T$ is known as the *ancestor order* of T; the root is the unique maximal element w.r.t $\succeq_T$. If $u \preceq_T v$ or $v \preceq_T u$ then u and v are *comparable* and otherwise, *incomparable*. We consider edges of rooted trees to be directed away from the root, that is, the notation for edges (u, v) of a tree is chosen such that $u \succ_T v$. If (u, v) is an edge in T, then u is called *parent* of v and v *child* of u. It will be convenient for the discussion below to extend the ancestor relation $\preceq_T$ on V to the union of the edge and vertex sets of T. More precisely, for the edge $e = (u, v) \in E$ we put $x \prec_T e$ if and only if $x \preceq_T v$ and $e \prec_T x$ if and only if $u \preceq_T x$. For edges $e = (u, v)$ and $f = (a, b)$ in T we put $e \preceq_T f$ if and only if $v \preceq_T b$. For $x \in V$, we write $L_T(x) := \{y \in L_T \mid y \preceq_T x\}$ for the set of leaves in the subtree $T(x)$ of T rooted in x.

For a non-empty subset of leaves $A \subseteq L$, we define $\mathrm{lca}_T(A)$, or the *least common ancestor of* A, to be the unique $\preceq_T$-minimal vertex of T that is an ancestor of every vertex in A. In case $A = \{u, v\}$, we put $\mathrm{lca}_T(u, v) := \mathrm{lca}_T(\{u, v\})$. We have in particular $u = \mathrm{lca}_T(L_T(u))$ for all $u \in V$. We will also frequently use that for any two non-empty vertex sets A, B of a tree, it holds that $\mathrm{lca}(A \cup B) = \mathrm{lca}(\mathrm{lca}(A), \mathrm{lca}(B))$.

A *phylogenetic tree* is a rooted tree such that no interior vertex in $v \in V \setminus L_T$ has degree two, except possibly the root. If L_T corresponds to a *set of genes* $\mathbb{G}$ or *species* $\mathbb{S}$, we call a phylogenetic tree on L_T *gene tree* or *species tree*, respectively. In this contribution we will **not** restrict the gene or species trees to be binary, although this assumption is made implicitly or explicitly in much of the literature on the topic. The more general setting allows

us to model incomplete knowledge of the exact gene or species phylogenies. Of course, all mathematical results proved here also hold for the special case of binary phylogenetic trees.

In our setting a gene tree $T = (V, E)$ on $\mathbb{G}$ is equipped with an *event-labeling* map $t : V \cup E \to I \cup \{0, 1\}$ with $I = \{\bullet, \square, \triangle, \odot\}$ that assigns to each interior vertex v of T a value $t(v) \in I$ indicating whether v is a speciation event ($\bullet$), duplication event ($\square$) or HGT event ($\triangle$). It is convenient to use the special label $\odot$ for the leaves x of T. Moreover, to each edge e a value $t(e) \in \{0, 1\}$ is added that indicates whether e is a *transfer edge* (1) or not (0). Note, only edges (x, y) for which $t(x) = \triangle$ might be labeled as transfer edge. We write $\mathcal{E} = \{e \in E \mid t(e) = 1\}$ for the set of transfer edges in T. We assume here that all edges labeled "0" transmit the genetic material vertically, that is, from an ancestral species to its descendants.

We remark that the restriction $t_{|V}$ of t to the vertex set V coincides with the "symbolic dating maps" introduced in [4]; these have a close relationship with cographs [10, 12, 14]. Furthermore, there is a map $\sigma : \mathbb{G} \to \mathbb{S}$ that assigns to each gene the species in which it resides. The set $\sigma(M)$, $M \subseteq \mathbb{G}$, is the set of species from which the genes M are taken. We write $(T; t, \sigma)$ for the gene tree $T = (V, E)$ with event-labeling t and corresponding map σ.

Removal of the transfer edges from $(T; t, \sigma)$ yields a forest $T_{\overline{\mathcal{E}}} := (V, E \setminus \mathcal{E})$ that inherits the ancestor order on its connected components, i.e., $\preceq_{T_{\overline{\mathcal{E}}}}$ iff $x \preceq_T y$ and x, y are in same subtree of $T_{\overline{\mathcal{E}}}$ [20]. Clearly $\preceq_{T_{\overline{\mathcal{E}}}}$ uniquely defines a root for each subtree and the set of descendant leaf nodes $L_{T_{\overline{\mathcal{E}}}}(x)$.

In order to account for duplication events that occurred before the first speciation event, we need to add an extra vertex and an extra edge "above" the last common ancestor of all species in the species tree $S = (V, E)$. Hence, we add an additional vertex to V (that is now the new root ρ_S of S) and the additional edge $(\rho_S, \mathrm{lca}_S(\mathbb{S}))$ to E. Strictly speaking S is not a phylogenetic tree in the usual sense, however, it will be convenient to work with these augmented trees. For simplicity, we omit drawing the augmenting edge $(\rho_S, \mathrm{lca}_S(\mathbb{S}))$ in our examples.

3 Observable Scenarios

The true history of a gene family, as it is considered here, is an arbitrary sequence of speciation, duplication, HGT, and gene loss events. The applications we envision for the theory developed, here, however assume that the gene tree and its event labels are inferred from (sequence) data, i.e., $(T; t, \sigma)$ is restricted to those labeled trees that can be constructed at least in principle from observable data. The issue here are gene losses that may completely eradicate the information on parts of the history. Specifically, we require that $(T; t, \sigma)$ satisfies the following three conditions:

(O1) Every internal vertex v has degree at least 3, except possibly the root which has degree at least 2.

(O2) Every HGT node has at least one transfer edge, $t(e) = 1$, and at least one non-transfer edge, $t(e) = 0$;

(O3) (a) If x is a speciation vertex, then there are at least two distinct children v, w of x such that the species V and W that contain v and w, resp., are incomparable in S.
(b) If (v, w) is a transfer edge in T, then the species V and W that contain v and w, resp., are incomparable in S.

Condition (O1) ensures that every event leaves a historical trace in the sense that there are at least two children that have survived in at least two of its subtrees. If this were not the case, no evidence would be left for all but one descendant tree, i.e., we would have no

evidence that event v ever happened. We note that this condition was used e.g. in [16] for scenarios without HGT. Condition (O2) ensures that for an HGT event a historical trace remains of both the transferred and the non-transferred copy. If there is no transfer edge, we have no evidence to classify v as a HGT node. Conversely, if all edges were transfers, no evidence of the lineage of origin would be available and any reasonable inference of the gene tree from data would assume that the gene family was vertically transmitted in at least one of the lineages in which it is observed. In particular, Condition (O2) implies that for each internal vertex there is a path consisting entirely of non-transfer edges to some leaf. This excludes in particular scenarios in which a gene is transferred to a different "host" and later reverts back to descendants of the original lineage without any surviving offspring in the intermittent host lineage. Furthermore, a speciation vertex x cannot be observed from data if it does not "separate" lineages, that is, there are two leaf descendants of distinct children of x that are in distinct species. However, here we only assume to have the weaker Condition (O3.a) which ensures that any "observable" speciation vertex x separates at least locally two lineages. In other words, if all children of x would be contained in species that are comparable in S or, equivalently, in the same lineage of S, then there is no clear historical trace that justifies x to be a speciation vertex. In particular, most-likely there are two leaf descendants of distinct children of x that are in the same species even if only $T_{\overline{\mathcal{E}}}$ is considered. Hence, x would rather be classified as a duplication than as a speciation upon inference of the event labels from actual data. Analogously, if $(v, w) \in \mathcal{E}$ then v signifies the transfer event itself but w refers to the next (visible) event in the gene tree T. Given that (v, w) is a HGT-edge in the observable part, in a "true history" v is contained in a species V that transmits its genetic material (maybe along a path of transfers) to a contemporary species Z that is an ancestor of the species W containing w. Clearly, the latter allows to have $V \succeq_S W$ which happens if the path of transfers points back to the descendant lineage of V in S. In this case the transfer edge (v, w) must be placed in the species tree such that $\mu(v)$ and $\mu(w)$ are comparable in S. However, then there is no evidence that this transfer ever happened, and thus v would be rather classified as speciation or duplication vertex.

It can be shown that (O1), (O2) and (O3) imply Lemma 1 as well as two important properties (Σ1) and (Σ2) of event labeled species trees that play a crucial role for the results reported here.

▶ **Lemma 1.** *Let $\mathcal{T}_1, \ldots, \mathcal{T}_k$ be the connected components of $T_{\overline{\mathcal{E}}}$ with roots $\rho_1, \ldots, \rho_k$, respectively. If* (O2) *holds, then, $\{L_{T_{\overline{\mathcal{E}}}}(\rho_1), \ldots, L_{T_{\overline{\mathcal{E}}}}(\rho_k)\}$ forms a partition of $\mathbb{G}$.*

(Σ1) If $t(x) = \bullet$ then there are distinct children v, w of x in T such that $\sigma(L_{T_{\overline{\mathcal{E}}}}(v)) \cap \sigma(L_{T_{\overline{\mathcal{E}}}}(w)) = \emptyset$.

Intuitively, (Σ1) is true because within a component $T_{\overline{\mathcal{E}}}$ no genetic material is exchanged between non-comparable nodes. Thus, a gene separated in a speciation event necessarily ends up in distinct species in the absence of horizontal transfer. It is important to note that we do not require the converse: $\sigma(L_{T_{\overline{\mathcal{E}}}}(y)) \cap \sigma(L_{T_{\overline{\mathcal{E}}}}(y')) = \emptyset$ does **not** imply $t(\mathrm{lca}_T(L_{T_{\overline{\mathcal{E}}}}(y) \cup L_{T_{\overline{\mathcal{E}}}}(y')) = \bullet$, that is, the last common ancestor of two sets of genes from different species is not necessarily a speciation vertex.

Now consider a transfer edge $(v, w) \in \mathcal{E}$, i.e., $t(v) = \triangle$. Then $T_{\overline{\mathcal{E}}}(v)$ and $T_{\overline{\mathcal{E}}}(w)$ are subtrees of distinct connected components of $T_{\overline{\mathcal{E}}}$. Since HGT amounts to the transfer of genetic material *across* distinct species, the genes v and w must be contained in distinct species X and Y, respectively. Since no genetic material is transferred between contemporary species X' and Y' in $T_{\overline{\mathcal{E}}}$, where X' and Y' is a descendant of X and Y, respectively we derive

(Σ2) If $(v, w) \in \mathcal{E}$ then $\sigma(L_{T_{\overline{\mathcal{E}}}}(v)) \cap \sigma(L_{T_{\overline{\mathcal{E}}}}(w)) = \emptyset$.

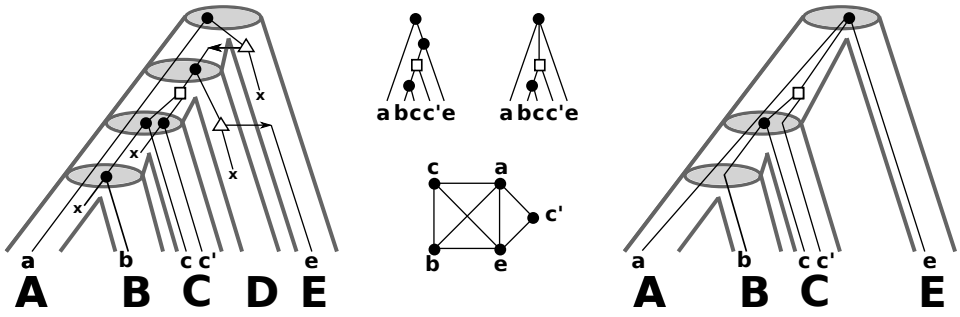

Figure 1 *Left:* A "true" evolutionary scenario for a gene tree with leaf set $\mathbb{G}$ evolving along the tube-like species trees is shown. The symbol "x" denotes losses. All speciations along the path from the root ρ_T to the leaf a are followed by losses and we omit drawing them.
Middle: The observable gene tree is shown in the upper-left. The orthology graph $G = (\mathbb{G}, E)$ (edges are placed between genes x, y for which $t(\mathrm{lca}(x, y)) = \bullet$) is drawn in the lower part. This graph is a cograph and the corresponding *non-binary* gene tree T on $\mathbb{G}$ that can be constructed from such data is given in the upper-right part (cf. [10, 11, 13] for further details).
Right: Shown is species trees S on $\mathbb{S} = \sigma(\mathbb{G})$ with reconciled gene tree T. The reconciliation map μ for T and S is given implicitly by drawing the gene tree T within S. Note, this reconciliation is not consistent with DTL-scenarios [20, 3]. A DTL-scenario would require that the duplication vertex and the leaf a are incomparable in S. for further details.

From here on we simplify the notation a bit and write $\sigma_{T_{\overline{\varepsilon}}}(u) := \sigma(L_{T_{\overline{\varepsilon}}}(u))$. We are aware of the fact that condition (O3) cannot be checked directly for a given event-labeled gene tree. In contrast, (Σ1) and (Σ2) are easily determined. Hence, in the remainder of this paper we consider the more general case, that is, gene trees that satisfy (O1), (O2), (Σ1) and (Σ2).

4 Time-Consistent Reconciliation Maps

The problem of reconciliation between gene trees and species tree is formalized in terms of so-called DTL-scenarios in the literature [20, 3]. This framework, however, usually assumes that the event labels t on T are unknown, while a species tree S is given. The "usual" DTL axioms, furthermore, explicitly refer to binary, fully resolved gene and species trees. We therefore use a different axiom set here that is a natural generalization of the framework introduced in [16] for the HGT-free case:

▶ **Definition 2.** Let $T = (V, E)$ and $S = (W, F)$ be phylogenetic trees on $\mathbb{G}$ and $\mathbb{S}$, resp., $\sigma : \mathbb{G} \to \mathbb{S}$ the assignment of genes to species and $t : V \cup E \to \{\bullet, \square, \triangle, \odot\} \cup \{0, 1\}$ an event labeling on T. A map $\mu : V \to W \cup F$ is a *reconciliation map* if for all $v \in V$ it holds that:

(M1) *Leaf Constraint.* If $t(v) = \odot$, then $\mu(v) = \sigma(v)$.

(M2) *Event Constraint.*

- **(i)** If $t(v) = \bullet$, then $\mu(v) = \mathrm{lca}_S(\sigma_{T_{\overline{\varepsilon}}}(v))$.
- **(ii)** If $t(v) \in \{\square, \triangle\}$, then $\mu(v) \in F$.
- **(iii)** If $t(v) = \triangle$ and $(v, w) \in \mathcal{E}$, then $\mu(v)$ and $\mu(w)$ are incomparable in S.

(M3) *Ancestor Constraint.*
Suppose $v, w \in V$ with $v \prec_{T_{\overline{\varepsilon}}} w$.

- **(i)** If $t(v), t(w) \in \{\square, \triangle\}$, then $\mu(v) \preceq_S \mu(w)$,
- **(ii)** otherwise, i.e., at least one of $t(v)$ and $t(w)$ is a speciation $\bullet$, $\mu(v) \prec_S \mu(w)$.

We say that S is a *species tree for* $(T; t, \sigma)$ if a reconciliation map $\mu : V \to W \cup F$ exists.

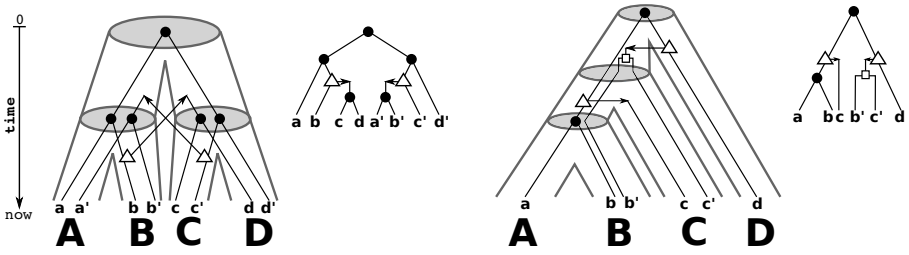

Figure 2 Shown are two (tube-like) species trees with reconciled gene trees. The reconciliation map μ for T and S is given implicitly by drawing the gene tree (upper right to the respective species tree) within the species tree. In the left example, the map μ is unique. However, μ is not time-consistent and thus, there is no time consistent reconciliation for T and S. In the example on the right hand side, μ is time-consistent.

It can be shown that the DTL axioms and the notation used here as in Definition 2 are equivalent in the case of binary trees. In Figure 1 an example of a biologically plausible reconciliation of non-binary trees that is valid w.r.t. Definition 2 is shown, however, it does not satisfy the conditions of a DTL-scenario.

Condition (M1) ensures that each leaf of T, i.e., an extant gene in $\mathbb{G}$, is mapped to the species in which it resides. Conditions (M2.i) and (M2.ii) ensure that each inner vertex of T is either mapped to a vertex or an edge in S such that a vertex of T is mapped to an interior vertex of S if and only if it is a speciation vertex. Condition (M2.i) might seem overly restrictive, an issue to which we will return below. Condition (M2.iii) satisfies condition (O3) and maps the vertices of a transfer edge in a way that they are incomparable in the species tree, since a HGT occurs between distinct (co-existing) species. It becomes void in the absence of HGT; thus Definition 2 reduces to the definition of reconciliation maps given in [16] for the HGT-free case. Importantly, condition (M3) refers only to the connected components of $T_{\overline{\mathcal{E}}}$ since comparability w.r.t. $\prec_{T_{\overline{\mathcal{E}}}}$ implies that the path between x and y in T does not contain transfer edges. It ensures that the ancestor order $\preceq_T$ of T is preserved along all paths that do not contain transfer edges.

We will make use of the following bound that effectively restricts how close to the leafs the image of a vertex in the gene tree can be located.

▶ **Lemma 3.** *If $\mu : (T; t, \sigma) \to S$ satisfies (M1) and (M3), then $\mu(u) \succeq_S \mathrm{lca}_S(\sigma_{T_{\overline{\mathcal{E}}}}(u))$ for any $u \in V(T)$.*

Proof. If u is a leaf, then by Condition (M1) $\mu(u) = \sigma(u)$ and we are done. Thus, let u be an interior vertex. By Condition (M3), $z \preceq_S \mu(u)$ for all $z \in \sigma_{T_{\overline{\mathcal{E}}}}(u)$. Hence, if $\mu(u) \prec_S \mathrm{lca}_S(\sigma_{T_{\overline{\mathcal{E}}}}(u))$ or if $\mu(u)$ and $\mathrm{lca}_S(\sigma_{T_{\overline{\mathcal{E}}}}(u)))$ are incomparable in S, then there is a $z \in \sigma_{T_{\overline{\mathcal{E}}}}(u)$ such that z and $\mu(u)$ are incomparable; contradicting (M3). ◀

Condition (M2.i) implies in particular the weaker property "(M2.i') if $t(v) = \bullet$ then $\mu(v) \in W$". In the light of Lemma 3, $\mu(v) = \mathrm{lca}_S(\sigma_{T_{\overline{\mathcal{E}}}}(v))$ is the lowest possible choice for the image of a speciation vertex. Clearly, this restricts the possibly exponentially many reconciliation maps for which $\mu(v) \succ_S \mathrm{lca}_S(\sigma_{T_{\overline{\mathcal{E}}}}(v))$ for speciation vertices v is allowed to only those that satisfy (M2.i). However, the latter is justified by the observation that if v is a speciation vertex with children u, w, then there is only one unique piece of information given by the gene tree to place $\mu(v)$, that is, the unique vertex x in S with children y, z such that $\sigma_{T_{\overline{\mathcal{E}}}}(u) \subseteq L_S(y)$ and $\sigma_{T_{\overline{\mathcal{E}}}}(w) \subseteq L_S(z)$. The latter arguments easily generalizes to the case that v has more than two children in T. Moreover, any *observable* speciation node $v' \succ_T v$ closer to the root

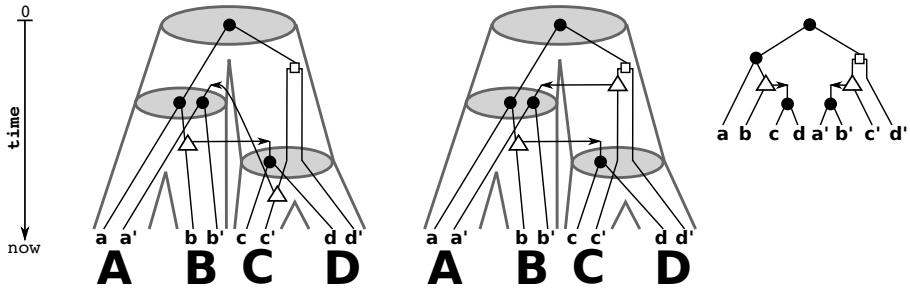

Figure 3 Shown are a gene tree $(T; t, \sigma)$ (right) and two identical (tube-like) species trees S (left and middle). There are two possible reconciliation maps for T and S that are given implicitly by drawing T within the species tree S. These two reconciliation maps differ only in the choice of placing the HGT-event either on the edge $(\mathrm{lca}_S(C, D), C)$ or on the edge $(\mathrm{lca}_S(\{A, B, C, D\}), \mathrm{lca}_S(C, D))$. In the first case, it is easy to see that μ would not be time-consistent, i.e., there are no time maps τ_T and τ_S that satisfy (C1) and (C2). The reconciliation map μ shown in the middle is time-consistent.

than v must be mapped to a node ancestral to $\mu(v)$ due to (M3.ii). Therefore, we require $\mu(v) = x = \mathrm{lca}_S(\sigma_{T_{\overline{\varepsilon}}}(v))$ here.

If S is a species tree for the gene tree (T, t, σ) then there is no freedom in the construction of a reconciliation map μ on the set $\{x \in V(T) \mid t(x) \in \{\bullet, \odot\}\}$. The duplication and HGT vertices of T, however, can be placed differently. As a consequence there is a possibly exponentially large set of reconciliation maps from (T, t, σ) to S.

From a biological point of view, however, the notion of reconciliation used so far is too weak. In the absence of HGT, subtrees evolve independently and hence, the linear order of points along each path from root to leaf is consistent with a global time axis. This is no longer true in the presence of HGT events, because HGT events imply additional time-consistency conditions. These stem from the fact that the appearance of the HGT copy in a distant subtree of S is concurrent with the HGT event. To investigate this issue in detail, we introduce time maps and the notion of time-consistency, see Figures 2 – 4 for illustrative examples.

▶ **Definition 4** (Time Map)**.** The map $\tau_T : V(T) \to \mathbb{R}$ is a time map for the rooted tree T if $x \prec_T y$ implies $\tau_T(x) > \tau_T(y)$ for all $x, y \in V(T)$.

▶ **Definition 5.** A reconciliation map μ from $(T; t, \sigma)$ to S is *time-consistent* if there are time maps τ_T for T and τ_S for S for all $u \in V(T)$ satisfying the following conditions:
(C1) If $t(u) \in \{\bullet, \odot\}$, then $\tau_T(u) = \tau_S(\mu(u))$.
(C2) If $t(u) \in \{\square, \triangle\}$ and, thus $\mu(u) = (x, y) \in E(S)$, then $\tau_S(y) > \tau_T(u) > \tau_S(x)$.

Condition (C1) is used to identify the time-points of speciation vertices and leaves u in the gene tree with the time-points of their respective images $\mu(u)$ in the species trees. In particular, all genes u that reside in the same species must be assigned the same time point $\tau_T(u) = \tau_S(\sigma(u))$. Analogously, all speciation vertices in T that are mapped to the same speciation in S are assigned matching time stamps, i.e., if $t(u) = t(v) = \bullet$ and $\mu(u) = \mu(v)$ then $\tau_T(u) = \tau_T(v) = \tau_S(\mu(u))$.

To understand the intuition behind (C2) consider a duplication or HGT vertex u. By construction of μ it is mapped to an edge of S, i.e., $\mu(u) = (x, y)$ in S. The time point of u must thus lie between time points of x and y. Now suppose $(u, v) \in \mathcal{E}$ is a transfer edge. By construction, u signifies the transfer event itself. The node v, however, refers to the next (visible) event in the gene tree. Thus $\tau_T(u) < \tau_T(v)$. In particular, $\tau_T(v)$ must not be

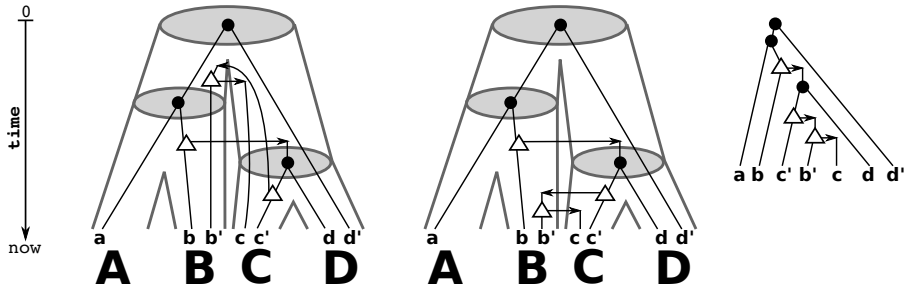

Figure 4 Shown are a gene tree $(T; t, \sigma)$ (right) and two identical (tube-like) species trees S (left and middle). There are two possible reconciliation maps for T and S that are given implicitly by drawing T within the species tree S. The left reconciliation maps each gene tree vertex as high as possible into the species tree. However, in this case only the middle reconciliation map is time-consistent.

misinterpreted as the time of introducing the HGT-duplicate into the new lineage. While this time of course exists (and in our model coincides with the timing of the transfer event) it is not marked by a visible event in the new lineage, and hence there is no corresponding node in the gene tree T.

W.l.o.g. we fix the time axis so that $\tau_T(\rho_T) = 0$ and $\tau_S(\rho_S) = -1$. Thus, $\tau_S(\rho_S) < \tau_T(\rho_T) < \tau_T(u)$ for all $u \in V(T) \setminus \{\rho_T\}$.

Clearly, a necessary condition to have biologically feasible gene trees is the existence of a reconciliation map μ. However, not all reconciliation maps are time-consistent, see Fig. 2.

▶ **Definition 6.** An event-labeled gene tree $(T; t, \sigma)$ is *biologically feasible* if there exists a time-consistent reconciliation map from $(T; t, \sigma)$ to some species tree S.

As a main result of this contribution, we provide simple conditions that characterize (the existence of) time-consistent reconciliation maps and thus, provides a first step towards the characterization of biologically feasible gene trees.

▶ **Theorem 7.** *Let μ be a reconciliation map from $(T; t, \sigma)$ to S. There is a* time-consistent *reconciliation map from $(T; t, \sigma)$ to S if and only if there are two time-maps τ_T and τ_S for T and S, respectively, such that the following conditions are satisfied for all $x \in V(S)$:*
(D1) *If $\mu(u) = x$, for some $u \in V(T)$ then $\tau_T(u) = \tau_S(x)$.*
(D2) *If $x \preceq_S \mathrm{lca}_S(\sigma_{T_{\bar{\varepsilon}}}(u))$ for some $u \in V(T)$ with $t(u) \in \{\square, \triangle\}$, then $\tau_S(x) > \tau_T(u)$.*
(D3) *If $\mathrm{lca}_S(\sigma_{T_{\bar{\varepsilon}}}(u) \cup \sigma_{T_{\bar{\varepsilon}}}(v)) \preceq_S x$ for some $(u, v) \in \mathcal{E}$, then $\tau_T(u) > \tau_S(x)$.*

From the algorithmic point of view it is desirable to design methods that allow to check whether a reconciliation map is time-consistent. Moreover, given a gene tree T and species tree S we wish to decide whether there exists a time-consistent reconciliation map μ, and if so, we should be able to construct μ.

To this end, observe that any constraints given by Definition 4, Theorem 7 (D2)–(D3), and Definition 5 (C2) can be expressed as a total order on $V(S) \cup V(T)$, while the constraints (C1) and (D1) together suggest that we can treat the preimage of any vertex in the species tree as a "single vertex". In fact we can create an auxiliary graph in order to answer questions that are concerned with time-consistent reconciliation maps.

▶ **Definition 8.** Let μ be a reconciliation map from $(T; t, \sigma)$ to S. The *auxiliary graph* A is defined as a directed graph with a vertex set $V(A) = V(S) \cup V(T)$ and an edge-set $E(A)$ that is constructed as follows:

Algorithm 1 Check existence and construct time-consistent reconciliation map

Precondition: $S = (W, F)$ is a species tree for $T = (V, E)$.

$\ell \leftarrow$ `ComputeLcaSigma`$((T; t, \sigma), S)$
$\mu(u) \leftarrow \emptyset$ for all $u \in V$ ▷ "$\emptyset$" means uninitialized
for all $u \in V$ **do**
  if $t(u) \in \{\bullet, \odot\}$ **then** $\mu(u) \leftarrow \ell(u)$
  else $\mu(u) \leftarrow (p(\ell(u)), \ell(u))$ ▷ $p(\ell(u))$ denotes the parent of $\ell(u)$
Compute the auxiliary graph A_2
if A_2 contains a cycle **then return** *"No time-consistent reconciliation map exists."*
Let $\tau : V(A_2) \to \mathbb{R}$ such that if $(x, y) \in E(A_2)$ then $\tau(x) < \tau(y)$
  ▷ W.l.o.g. we can assume that $\tau(x) \neq \tau(y)$ for all $x, y \in V(A_2)$
$\tau_S \leftarrow$ A time map such that $\tau_S(x) = \tau(x)$ for all $x \in W$
$\tau_T \leftarrow$ A time map such that $\tau_T(u) = \tau(\mu(u))$ if $t(u) \in \{\bullet, \odot\}$, otherwise $\tau_T(u) = \tau(u)$ for all $u \in V$.
for $u \in V$ where $t(u) \in \{\square, \triangle\}$ **do**
  while it does not hold that $\tau_S(x) < \tau_T(u) < \tau_S(y)$ for $(x, y) = \mu(u)$ **do**
    $\mu(u) \leftarrow (p(x), x)$
return μ

(A1) For each $(u, v) \in E(T)$ we have $(u', v') \in E(A)$, where

$$u' = \begin{cases} \mu(u) & \text{if } t(u) \in \{\odot, \bullet\} \\ u & \text{otherwise} \end{cases}, \quad v' = \begin{cases} \mu(v) & \text{if } t(v) \in \{\odot, \bullet\} \\ v & \text{otherwise} \end{cases},$$

(A2) For each $(x, y) \in E(S)$ we have $(x, y) \in E(A)$.
(A3) For each $u \in V(T)$ with $t(u) \in \{\square, \triangle\}$ we have $(u, lca_S(\sigma_{T_{\overline{\varepsilon}}}(u))) \in E(A)$.
(A4) For each $(u, v) \in \mathcal{E}$ we have $(lca_S(\sigma_{T_{\overline{\varepsilon}}}(u) \cup \sigma_{T_{\overline{\varepsilon}}}(v)), u) \in E(A)$
(A5) For each $u \in V(T)$ with $t(u) \in \{\triangle, \square\}$ and $\mu(u) = (x, y) \in E(S)$ we have $(x, u) \in E(A)$ and $(u, y) \in E(A)$.

We define A_1 and A_2 as the subgraphs of A that contain only the edges defined by (A1), (A2), (A5) and (A1), (A2), (A3), (A4), respectively.

We note that the edge sets defined by conditions (A1) through (A5) are not necessarily disjoint. The mapping of vertices in T to edges in S is considered only in condition (A5). The following two theorems are the key results of this contribution.

▶ **Theorem 9.** *Let μ be a reconciliation map from $(T; t, \sigma)$ to S. The map μ is time-consistent if and only if the auxiliary graph A_1 is a directed acyclic graph (DAG).*

▶ **Theorem 10.** *Assume there is a reconciliation map μ from $(T; t, \sigma)$ to S. There is a time-consistent reconciliation map, possibly different from μ, from $(T; t, \sigma)$ to S if and only if the auxiliary graph A_2 (defined on μ) is a DAG.*

Naturally, Theorems 9 or 10 can be used to devise algorithms for deciding time-consistency. To this end, the efficient computation of $\mathrm{lca}_S(\sigma_{T_{\overline{\varepsilon}}}(u))$ for all $u \in V(T)$ is necessary. This can be achieved with Algorithm 2 in $O(|V(T)| \log(|V(S)|))$. More precisely, we have the following statement.

▶ **Lemma 11.** *For a given gene tree $(T = (V, E); t, \sigma)$ and a species tree $S = (W, F)$, Algorithm 2 correctly computes $\ell(u) = \mathrm{lca}_S(\sigma_{T_{\overline{\varepsilon}}}(u))$ for all $u \in V(T)$ in $O(|V| \log(|W|))$ time.*

Let S be a species tree for $(T; t, \sigma)$, that is, there is a valid reconciliation between the two trees. Algorithm 1 describes a method to construct a time-consistent reconciliation map for $(T; t, \sigma)$ and S, if one exists, else "No time-consistent reconciliation map exists" is returned. First, an arbitrary reconciliation map μ that satisfies the condition of Def. 2 is computed. Second, Theorem 10 is utilized and it is checked whether the auxiliary graph A_2 is not a DAG in which case no time-consistent map μ exists for $(T; t, \sigma)$ and S. Finally, if A_2 is a DAG, then we continue to adjust μ to become time-consistent.

▶ **Theorem 12.** *Let $S = (W, F)$ be species tree for the gene tree $(T = (V, E); t, \sigma)$. Algorithm 1 correctly determines whether there is a time-consistent reconciliation map μ and in the positive case, returns such a μ in $O(|V| \log(|W|))$ time.*

5 Outlook and Summary

We have characterized here whether a given event-labeled gene tree $(T; t, \sigma)$ and species tree S can be reconciled in a time-consistent manner in terms of two auxiliary graphs A_1 and A_2 that must be DAGs. These are defined in terms of given reconciliation maps. This condition yields an $O(|V| \log(|W|))$-time algorithm to check whether a given reconciliation map μ is time-consistent, and an algorithm with the same time complexity for the construction of a time-consistent reconciliation maps, provided one exists.

Our results depend on three conditions on the event-labeled gene trees that are motivated by the fact that event-labels can be assigned to internal vertices of gene trees only if there is observable information on the event. The question which event-labeled gene trees are actually observable given an arbitrary, true evolutionary scenario deserves further investigation in future work. Here we have used conditions that arguable are satisfied when gene trees are inferred using sequence comparison and synteny information. A more formal theory of observability is still missing, however.

Our results provide an efficient way of deciding whether a *given* pair of gene and species tree can be time-consistently reconciled. There are, however, in general exponentially many putative species trees. This begs the question whether there is *at least one* species tree S for a gene tree and if so, how to construct S. "Informative triples" extracted from the gene tree answer this question in the absence of HGT [16]. It is plausible that this idea can be generalized to our current setting to provide at least a partial characterization [9].

Acknowledgment. We thank the organizers of the 32nd TBI Winterseminar 2017 in Bled (Slovenia), where the authors participated, met and basically drafted the main ideas of this paper, while drinking a cold and tasty red Union, or was it a green Laško?

References

1 A. M. Altenhoff, B. Boeckmann, S. Capella-Gutierrez, D. A. Dalquen, T. DeLuca, K. Forslund, J. Huerta-Cepas, B. Linard, C. Pereira, L. P. Pryszcz, F. Schreiber, A. S. da Silva, D. Szklarczyk, C. M. Train, P. Bork, O. Lecompte, C. von Mering, I. Xenarios, K. Sjölander, L. J. Jensen, M. J. Martin, M. Muffato, T. Gabaldón, S. E. Lewis, P. D. Thomas, E. Sonnhammer, and C. Dessimoz. Standardized benchmarking in the quest for orthologs. *Nature Methods*, 13:425–430, 2016.

2 A. M. Altenhoff and C. Dessimoz. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol.*, 5:e1000262, 2009.

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

4 S. Böcker and A. W. M. Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Adv. Math.*, 138:105–125, 1998.

5 M. A. Charleston. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math Biosci.*, 149(2):191–223, 1998.

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

7 A. Dress, V. Moulton, M. Steel, and T. Wu. Species, clusters and the 'tree of life': A graph-theoretic perspective. *J. Theor. Biol.*, 265:535–542, 2010.

8 W. M. Fitch. Homology: a personal view on some of the problems. *Trends Genet.*, 16:227–231, 2000.

9 M. Hellmuth. Biologically feasible gene trees, reconciliation maps and informative triples, 2017. (submitted) arXiv:1701.07689.

10 M. Hellmuth, M. Hernandez-Rosales, K. T. Huber, V. Moulton, P. F. Stadler, and N. Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *J. Math. Biology*, 66(1-2):399–420, 2013.

11 M. Hellmuth, P. F. Stadler, and N. Wieseke. The mathematics of xenology: Di-cographs, symbolic ultrametrics, 2-structures and tree- representable systems of binary relations. *Journal of Mathematical Biology*, 2016. DOI: 10.1007/s00285-016-1084-3.

12 M. Hellmuth and N. Wieseke. On symbolic ultrametrics, cotree representations, and cograph edge decompositions and partitions. In Dachuan et al., editor, *Proceedings COCOON 2015*, pages 609–623, Cham, 2015. Springer International Publishing.

13 M. Hellmuth and N. Wieseke. From sequence data including orthologs, paralogs, and xenologs to gene and species trees. In Pierre Pontarotti, editor, *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods*, pages 373–392, Cham, 2016. Springer.

14 M. Hellmuth and N. Wieseke. On tree representations of relations and graphs: Symbolic ultrametrics and cograph edge decompositions. *J. Comb. Opt.*, 2017. doi:DOI10.1007/s10878-017-0111-7.

15 M. Hellmuth, N. Wieseke, M. Lechner, H.-P. Lenhof, M. Middendorf, and P. F. Stadler. Phylogenomics with paralogs. *Proceedings of the National Academy of Sciences*, 112(7):2058–2063, 2015. doi:10.1073/pnas.1412770112.

16 M. Hernandez-Rosales, M. Hellmuth, N. Wieseke, K. T. Huber, V. Moulton, and P. F. Stadler. From event-labeled gene trees to species trees. *BMC Bioinformatics*, 13(Suppl 19):S6, 2012.

17 M. Lechner, M. Hernandez-Rosales, D. Doerr, N. Wieseke, A. Thévenin, J. Stoye, R. K. Hartmann, S. J. Prohaska, and P. F. Stadler. Orthology detection combining clustering and synteny for very large datasets. *PLoS ONE*, 9(8):e105015, 08 2014.

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

19 A. C. J. Roth, G. H. Gonnet, and C. Dessimoz. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, 9:518, 2008.

20 A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):517–535, 2011.