BMC Bioinformatics

PROCEEDINGS   Open Access

From event-labeled gene trees to species trees

Maribel Hernandez-Rosales[1,2*], Marc Hellmuth[3], Nicolas Wieseke[4,5], Katharina T Huber[6], Vincent Moulton[6], Peter F Stadler[1,2,7,8]

Abstract

Background: Tree reconciliation problems have long been studied in phylogenetics. A particular variant of the reconciliation problem for a gene tree T and a species tree S assumes that for each interior vertex x of T it is known whether x represents a speciation or a duplication. This problem appears in the context of analyzing orthology data.

Results: We show that S is a species tree for T if and only if S displays all rooted triples of T that have three distinct species as their leaves and are rooted in a speciation vertex. A valid reconciliation map can then be found in polynomial time. Simulated data shows that the event-labeled gene trees convey a large amount of information on underlying species trees, even for a large percentage of losses.

Conclusions: The knowledge of event labels in a gene tree strongly constrains the possible species tree and, for a given species tree, also the possible reconciliation maps. Nevertheless, many degrees of freedom remain in the space of feasible solutions. In order to disambiguate the alternative solutions additional external constraints as well as optimization criteria could be employed.

Background

The reconstruction of the evolutionary history of a gene family is necessarily based on at least three interrelated types of information. The true phylogeny of the investigated species is required as a scaffold with which the associated gene tree must be reconcilable. Orthology or paralogy of genes found in different species determines whether an internal vertex in the gene tree corresponds to a duplication or a speciation event. Speciation events, in turn, are reflected in the species tree.

The reconciliation of gene and species trees is a widely studied problem [1-10]. In most practical applications, however, neither the gene tree nor the species tree can be determined unambiguously.

Although orthology information is often derived from the reconciliation of a gene tree with a species tree (cf. e.g. TreeFam [11], PhyOP [12], PHOG [13], EnsemblCompara

GeneTrees [14], and MetaPhOrs [15]), recent benchmarks studies [16] have shown that orthology can also be inferred at similar levels of accuracy without the need to construct trees by means of clustering-based approaches such as OrthoMCL [17], the algorithms underlying the COG database [18,19], InParanoid [20], or ProteinOrtho [21]. In [22] we have therefore addressed the question: how much information about the gene tree, the species tree, and their reconciliation is already contained in the orthology relation between genes?

According to Fitch's definition [23], two genes are (co-) orthologous if their last common ancestor in the gene tree represents a speciation event. Otherwise, i.e., when their last common ancestor is a duplication event, they are paralogs. The orthology relation on a set of genes is therefore determined by the gene tree T and an "event labeling" that identifies each interior vertex of T as either a duplication or a speciation event. (We disregard here additional types of events such as horizontal transfer and refer to [22] for details on how such extensions might be incorporated into the mathematical framework.) One of

* Correspondence: maribel@bioinf.uni-leipzig.de
[1]Max-Planck-Institute for Mathematics in the Sciences, Leipzig, D-04103, Germany
Full list of author information is available at the end of the article

the main results of [22], which relies on the theory of symbolic ultrametrics developed in [24], is the following: a relation on a set of genes is an orthology relation (i.e., it derives from some event-labeled gene tree) if and only if it is a cograph (for several equivalent characterizations of cographs see [25]). Note that the cograph does not contain the full information on the event-labeled gene tree. Instead the cograph is equivalent to the gene tree's homomorphic image obtained by collapsing adjacent events of the same type [22]. The orthology relation thus places strong and easily interpretable constraints on the gene tree.

This observation suggests that a viable approach to reconstructing histories of large gene families may start from an empirically determined orthology relation, which can be directly adjusted to conform to the requirement of being a cograph. The result is then equivalent to an (usually incompletely resolved) event-labeled gene tree, which might be refined or used as constraint in the inference of a fully resolved gene tree. In this contribution we are concerned with the next conceptual step: the derivation of a species tree from an event-labeled gene tree. As we shall see below, this problem is much simpler than the full tree reconciliation problem. Technically, we will approach this problem by reducing the reconciliation map from gene tree to species tree to rooted triples of genes residing in three distinct species. This is related to an approach that was developed in [26] for addressing the full tree reconciliation problem.

Methods
Definitions and notation
Phylogenetic trees
A *phylogenetic tree T (on L)* is a rooted tree $T = (V, E)$, with leaf set $L \subseteq V$, set of directed edges E, and set of interior vertices $V^0 = V \backslash L$ that does not contain any vertices with in- and outdegree one and whose root $\rho_T \in V$ has indegree zero. In order to avoid uninteresting trivial cases, we assume that $|L| \geq 3$. The ancestor relation $\preceq_T$ on V is the partial order defined, for all $x, y \in V$, by $x \preceq_T y$ whenever y lies on the path from x to the root. If there is no danger of ambiguity, we will write $x \preceq y$ rather than $x \preceq_T y$. Furthermore, we write $x \prec y$ to mean $x \preceq y$ and $x \neq y$. For $x \in V$, we write $L(x) := \{y \in L | y \preceq x\}$ for the set of leaves in the subtree $T(x)$ of T rooted in x. Thus, $L(\rho_T) = L$ and $T(\rho_T) = T$. For $x, y \in V$ such that x and y are joined by an edge $e \in E$ we write $e = [y, x]$ if $x \prec y$. Two phylogenetic trees $T = (V, E)$ and $T' = (V', E')$ on L are said to be *equivalent* if there exists a bijection from V to V' that is the identity on L, maps ρ_T to ρ_T', and extends to a graph isomorphism between T and T'. A *refinement* of a phylogenetic tree T on L is a phylogenetic tree T' on L such that T can be obtained from T' by collapsing edges (see e.g. [27]). Suppose for

the remainder of this section that $T = (V, E)$ is a phylogenetic tree on L with root ρ_T. For a non-empty subset of leaves $A \subseteq L$, we define $\mathrm{lca}_T (A)$, or the *most recent common ancestor of A*, to be the unique vertex in T that is the greatest lower bound of A under the partial order $\preceq_T$. In case $A = \{x, y\}$, we put $\mathrm{lca}_T (x, y) := \mathrm{lca}_T (\{x, y\})$ and if $A = \{x, y, z\}$, we put $\mathrm{lca}_T (x, y, z) := \mathrm{lca}_T (\{x, y, z\})$. For later reference, we have, for all $x \in V$, that $x = \mathrm{lca}_T (L(x))$. Let $L' \subseteq L$ be a subset of $|L'| \geq 2$ leaves of T. We denote by $T (L') = T(\mathrm{lca}_T (L'))$ the (rooted) subtree of T with root $\mathrm{lca}_T (L')$. Note that $T(L')$ may have leaves that are not contained in L'. The *restriction* $T|_{L'}$ of T to L' is the phylogenetic tree with leaf set L' obtained from T by first forming the minimal spanning tree in T with leaf set L' and then by suppressing all vertices of degree two with the exception of ρ_T if ρ_T is a vertex of that tree. A phylogenetic tree T' on some subset $L' \subseteq L$ is said to be *displayed* by T (or equivalently that T *displays* T') if T' is equivalent with tree $T|_{L'}$. A set $\mathscr{T}$ of phylogenetic trees T each with leaf set L_T is called *consistent* if $\mathscr{T} = \emptyset$ or there is a phylogenetic tree T on $L = \cup_{T \in \mathscr{T}} L_T$ that *displays* $\mathscr{T}$, that is, T displays every tree contained in $\mathscr{T}$. Note that a consistent set of phylogenetic trees is sometimes also called compatible (see e.g. [27]).

It will be convenient for our 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 directed 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 $u \preceq_E x$. For edges $e = [u, v]$ and $f = [a, b]$ in T we put $e \preceq f$ if and only if $v \preceq b$.

Rooted triples
Rooted triples are phylogenetic trees on three leaves with precisely two interior vertices. Sometimes also called rooted triplets [28] they constitute an important concept in the context of supertree reconstruction [27,29] and will also play a major role here. Suppose $L = \{x, y, z\}$. Then we denote by $((x, y), z)$ the triple r with leaf set L for which the path from x to y does not intersect the path from z to the root ρ_r and thus, having. $\mathrm{lca}_r(x, y) \prec \mathrm{lca}_r(x, y, z)$ For T a phylogenetic tree, we denote by $\Re(T)$ the set of all triples that are displayed by T.

Clearly, a set $\mathscr{R}$ of triples is *consistent* if there is a phylogenetic tree T on $X = \cup_{r \in \mathscr{R}} L(\rho_r)$ such that $\mathscr{R} \subseteq \mathscr{R}(T)$. Not all sets of triples are consistent of course. Given a triple set $\mathscr{R}$ there is a polynomial-time algorithm, referred to in [27] as BUILD, that either constructs a phylogenetic tree T that displays $\mathscr{R}$ or that recognizes that $\mathscr{R}$ is *inconsistent*, that is, not consistent [30]. Various practical implementations have been described starting with [30], improved variants are discussed in [31,32].

The problem of determining a maximum consistent subset $\mathscr{R}'$ of an inconsistent set of triples, on the other

hand is NP-hard and also APX-hard, see [33,34] and the references therein. We refer to [35] for an overview on the available practical approaches and further theoretical results.

The BUILD algorithm, furthermore, does not necessarily generate for a given triple set $\mathcal{R}$ a minimal phylogenetic tree T that displays $\mathcal{R}$, i.e., T may resolve multifurcations in an arbitrary way that is not implied by any of the triples in $\mathcal{R}$. However, the tree generated by BUILD is minor-minimal, i.e., if T' is obtained from T by contracting an edge, T' does not display $\mathcal{R}$ anymore. The trees produced by BUILD do not necessarily have the minimum number of internal vertices. Thus, depending on $\mathcal{R}$, not all trees consistent with $\mathcal{R}$ can be obtained from BUILD. Semple [36] gives an algorithm that produces all minor-minimal trees consistent with $\mathcal{R}$. It requires only polynomial time for each of the possibly exponentially many minor-minimal trees. The problem of constructing a tree consistent with $\mathcal{R}$ and minimizing the number of interior vertices is NP-hard and hard to approximate [37].

Event labeling, species labeling, and reconciliation map

A gene tree T arises through a series of events along a species tree S. We consider both T and S as phylogenetic trees with leaf sets L (the set of genes) and B (the set of species), respectively. We assume that $|L| \geq 3$ and $|B| \geq 1$. We consider only gene duplications and gene losses, which take place between speciation events, i.e., along the edges of S. Speciation events are modeled by transmitting the gene content of an ancestral lineage to each of its daughter lineages.

The true evolutionary history of a single ancestral gene thus can be thought of as a scenario such as the one depicted in Figure 1. Since we do not consider horizontal gene transfer or lineage sorting in this contribution, an evolutionary scenario consists of four components: (1) A true gene tree $\hat{T}$, (2) a true species tree $\hat{S}$, (3) an assignment of an event type (i.e., speciation •, duplication □, loss ⊗, or observable (extant) gene ⊙) to each interior vertex and leaf of $\hat{T}$, and (4) a map μ assigning every vertex of $\hat{T}$ to a vertex or edge of $\hat{S}$ in such a way that (a) the ancestor order of $\hat{T}$ is preserved, (b) a vertex of $\hat{T}$ is mapped to an interior vertex of $\hat{S}$ if and only if it is of type speciation, (c) extant genes of $\hat{T}$ are mapped to leaves of S. Alternatively, one could define $\hat{T}$ and $\hat{S}$ to be metric graphs (i.e., comprising edges that are real intervals glued together at the vertices) with a distance function that measures evolutionary time. In this picture, $\hat{\mu}$ is a continuous map that preserves the temporal order and satisfied conditions (b) and (c).

In order to allow $\hat{\mu}$ to map duplication vertices to a time point before the last common ancestor of all species in $\hat{S}$, we need to extend our definition of a species

tree by adding an extra vertex and an extra edge "above" the last common ancestor of all species. Note that strictly speaking $\hat{S}$ is not a phylogenetic tree anymore. In case there is no danger of confusion, we will from now on refer to a phylogenetic tree on B with this extra edge and vertex added as a species tree on B and to ρ_B as the root of B. Also, we canonically extend our notions of a triple, displaying, etc. to this new type of species tree.

The true gene tree $\hat{T}$ represents all extant as well as all extinct genes, all duplication, and all speciation events. Not all of these events are observable from extant genes data, however. In particular, extinct genes cannot be observed. The observable part $T = T (V, E)$ of $\hat{T}$ is the restriction of $\hat{T}$ to the leaf set L of extant genes, i.e., $T = \hat{T}|_L$.

Furthermore, we can observe a map $\sigma\colon L \to B$ that assigns to each extant gene the species in which it resides. Of course, for $x \in L$ we have $\sigma(x) = \hat{\mu}(x)$. Here B is the leaf set of the extant species tree, i.e., $B = \sigma(L)$. For ease of readability, we also put $\sigma(T') = \{\sigma(x)\colon x \in L(y)\}$ for any subtree T' of T with $T' = T (y)$ where $y \in V^\circ$. Alternatively, we will sometimes also write $\sigma(y)$ instead of $\sigma(T (y))$. Last but not least, for $Y \subseteq L$, we put $\sigma(Y) = \{\sigma(y)\colon y \in Y\}$.
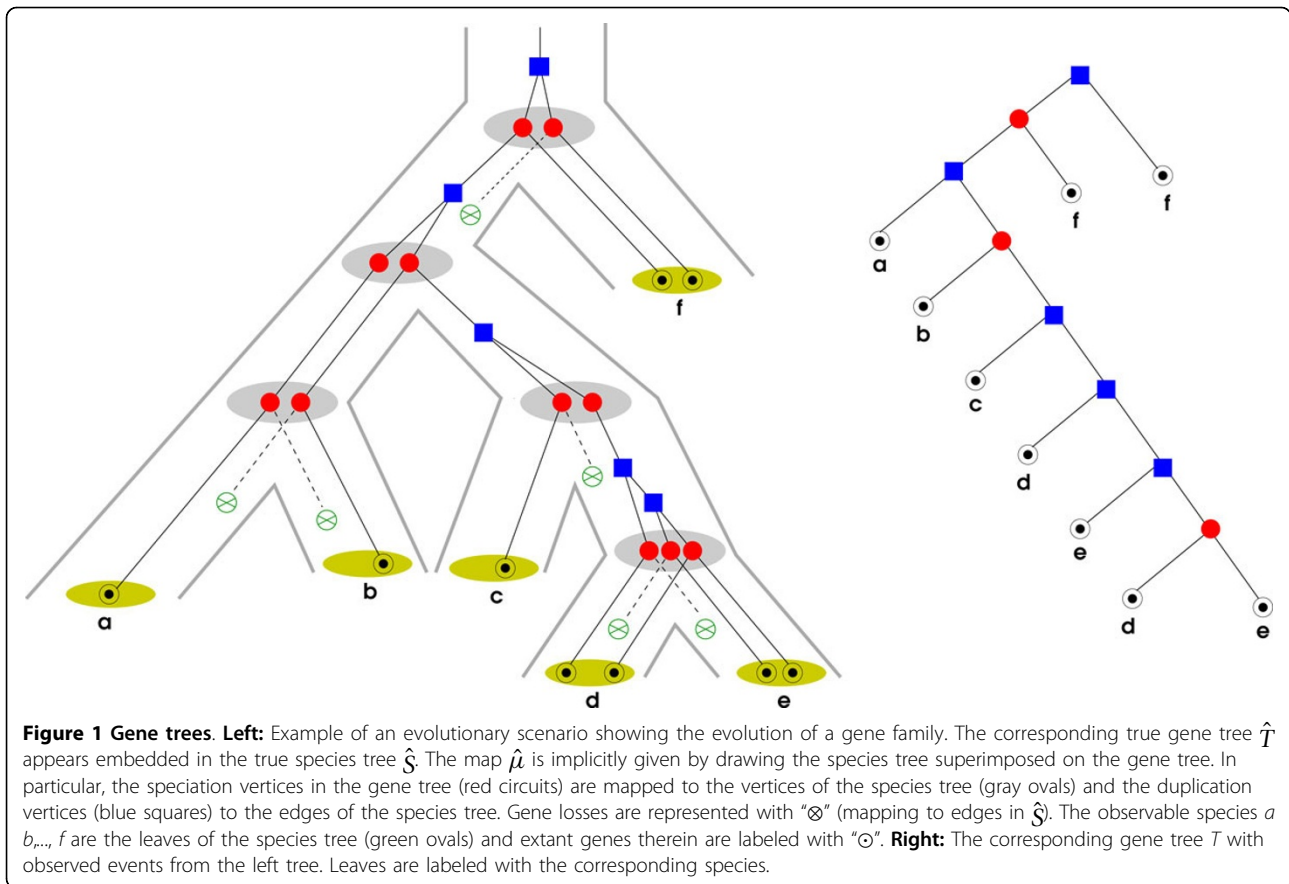
The observable part of the species tree $S = (W\ H)$ is the restriction $\hat{S}|_B$ of $\hat{S}$ to B. In order to account for duplication events that occurred before the first speciation event, the additional vertex $\rho_S \in W$ and the additional edge $[\rho s, \mathrm{lca}_s B] \in H$ must be part of S.

The evolutionary scenario also implies an *event labeling* map $t\colon V \to \{\bullet, \square,\ \odot\}$ that assigns to each interior vertex v of T a value $t(v)$ indicating whether v is a speciation event ($\cdot$) or a duplication event ($\square$). It is convenient to use the special label $\odot$ for the leaves x of T. We write (T,t) for the event-labeled tree. We remark that t was introduced as "symbolic dating map" in [24]. It is called *discriminating* if, for all edges $\{u, v\} \in E$, we have $t(u) \neq t(v)$ in which case (T,t) is known to be in 1-1-correspondence to a cograph [22]. Note that we will in general not require that t is discriminating in this contribution. For $T = (V, E)$ a gene tree on L, B a set of species, and maps t and σ as specified above, we require however that μ and σ must satisfy the following compatibility property:

(C) Let $z \in V$ be a speciation vertex, i.e., $t(z) = \cdot$, and let T' and T'' be subtrees of T rooted in two distinct children of z. Then $\sigma (T') \cap \sigma (T'') = \varnothing$.

Note the we do not require the converse, i.e., from the disjointness of the species sets $\sigma (T')$ and $\sigma(T'')$ we do **not** conclude that their last common ancestor is a speciation vertex.

For $x, y \in L$ and $z = \mathrm{lca}_T (x, y)$ it immediately follows from condition (C) that if $t(\mathrm{lca}_T (x, y)) = \bullet$ then $\sigma(x) \neq \sigma(y)$ since, by assumption, x and y are leaves in distinct subtrees below z. Equivalently, two distinct genes $x \neq y$

Figure 1 Gene trees. **Left:** Example of an evolutionary scenario showing the evolution of a gene family. The corresponding true gene tree $\hat{T}$ appears embedded in the true species tree $\hat{S}$. The map $\hat{\mu}$ is implicitly given by drawing the species tree superimposed on the gene tree. In particular, the speciation vertices in the gene tree (red circuits) are mapped to the vertices of the species tree (gray ovals) and the duplication vertices (blue squares) to the edges of the species tree. Gene losses are represented with "⊗" (mapping to edges in $\hat{S}$). The observable species a b,..., f are the leaves of the species tree (green ovals) and extant genes therein are labeled with "⊙". **Right:** The corresponding gene tree T with observed events from the left tree. Leaves are labeled with the corresponding species.

in L for which $\sigma(x) = \sigma(y)$ holds, that is, they are contained in the same species of B, must have originated from a duplication event, i.e., $t(\mathrm{lca}_T (x, y)) = \square$. Thus we can regard σ as a proper vertex coloring of the cograph corresponding to (T, t).

Let us now consider the properties of the restriction of $\hat{\mu}$ to the observable parts T of $\hat{T}$ and S of $\hat{S}$. Consider a speciation vertex x in $\hat{T}$. If x has two children y' and y'' so that $L(y')$ and $L(y'')$ are both non-empty then $x = \mathrm{lca}_{\hat{T}}(z', z'')$ for all $z' \in L(y')$ and $z'' \in L(y'')$ and hence, $x = \mathrm{lca}_T (L(y')\cup(L(y''))$. In particular, x is an observable vertex in T. Furthermore, we know that $\sigma(L(y')) \cap \sigma(L(y'')) = \emptyset$, and therefore, $\hat{\mu}(x) = \mathrm{1ca}_S(\sigma(L(y') \cup L(y''))$. Considering all pairs of children with this property this can be rephrased as $\hat{\mu}(x) = \mathrm{1ca}_{\hat{S}}(\sigma(L(x)))$. On the other hand, if x does not have at least two children with this property, and hence the corresponding speciation vertex cannot be viewed as most recent common ancestor of the set of its descendants in S, then x is not a vertex in the restriction $T = \hat{T}|_L$ of $\hat{T}$ to the set L of the extant genes. The restriction μ of $\hat{\mu}$ to the observable tree T therefore satisfies the properties used below to define reconciliation maps.

Definition 1. *Suppose that B is a set of species, that $S = (W, H)$ is a phylogenetic tree on B, that $T = (V, E)$ is a gene*

tree with leaf set L and that $\sigma : L \rightarrow B$ and $t : V \rightarrow \{\bullet, \square, \odot\}$ are the maps described above. Then we say that S is a species tree for (T, t, σ) *if there is a map $\mu : V \rightarrow W \cup H$ such that, for all $x \in V$:*

(i) If $t(x) = \odot$ then $\mu(x) = \sigma(x)$.
(ii) If $t(x) = \bullet$ then $\mu(x) \in W \setminus B$.
(iii) If $t(x) = \square$ then $\mu(x) \in H$.
(iv) Let $x, y \in V$ with $x \prec_T y$. We distinguish two cases:

> *1. If $t(x) = t(y) = \square$ then $\mu(x) \preccurlyeq_S \mu(y)$ in S.*
> *2. If $t(x) = t(y) = \bullet$ or $t(x) \neq t(y)$ then $\mu(x) \prec_S \mu(y)$ in S.*

(v) If $t(x) = \bullet$ then $\mu(x) = \mathrm{lca}_S(\sigma(L(x)))$
We call μ the reconciliation map from (T, t, σ) *to S.*

We note that $\mu^{-1}(\rho_S) = \emptyset$ holds as an immediate consequence of property *(v)*, which implies that no speciation node can be mapped above $\mathrm{lca}_S(B)$, the unique child of ρ_S.

We illustrate this definition by means of an example in Figure 2 and remark that it is consistent with the definition of reconciliation maps for the case when the event labeling t on T is not known [38]. Continuing with our notation from Definition 1 for the remainder of this section, we easily derive their axiom set as
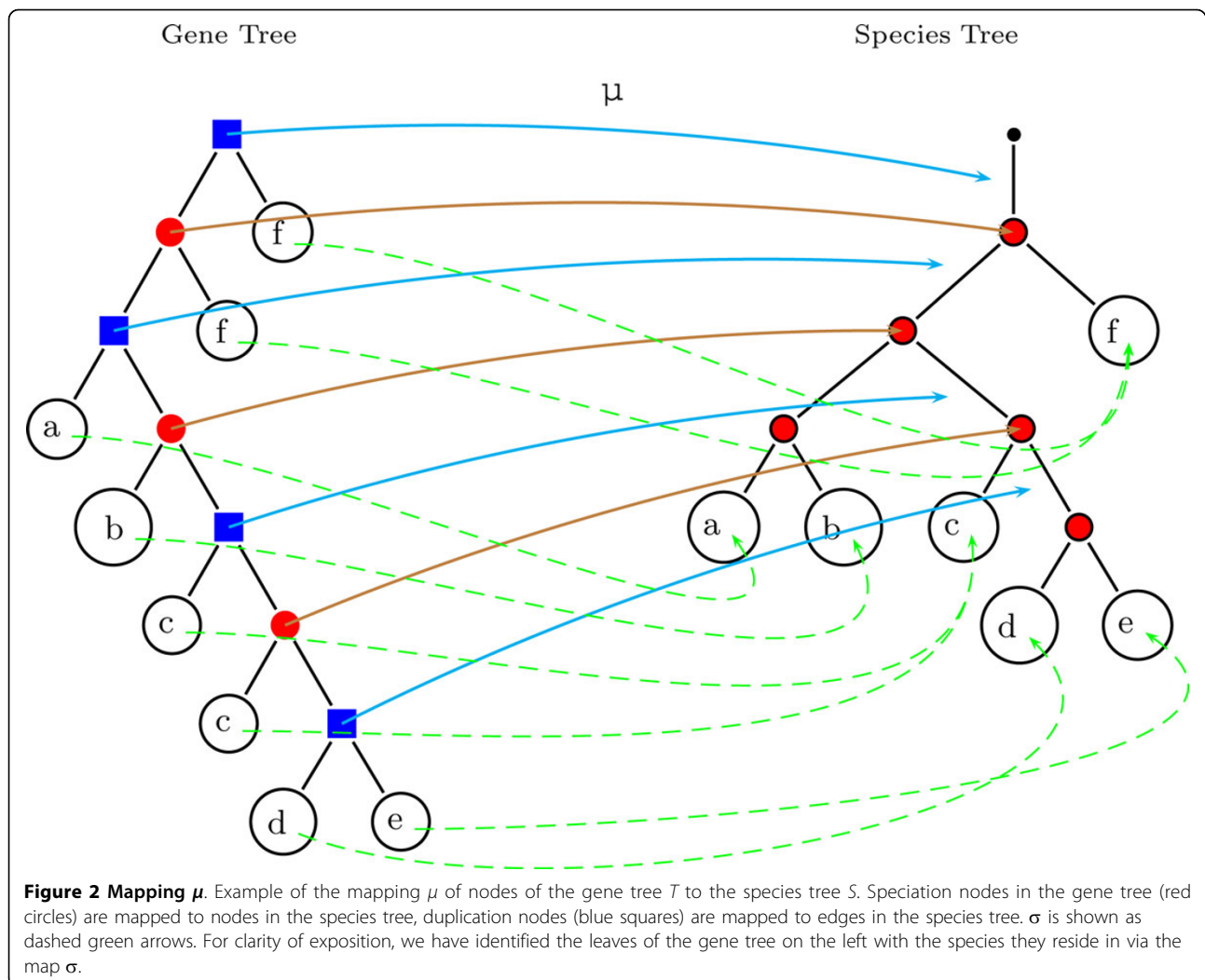
Figure 2 Mapping μ. Example of the mapping μ of nodes of the gene tree T to the species tree S. Speciation nodes in the gene tree (red circles) are mapped to nodes in the species tree, duplication nodes (blue squares) are mapped to edges in the species tree. σ is shown as dashed green arrows. For clarity of exposition, we have identified the leaves of the gene tree on the left with the species they reside in via the map σ.

Lemma 2. *If μ is a reconciliation map from (T,t,σ) to S and L is the leaf set of T then, for all $x \in V$.*

(D1) $x \in L$ implies $\mu(x) = \sigma(x)$.

(D2.a) $\mu(x) \in W$ implies $\mu(x) = \mathrm{lca}_S(\sigma(L(x)))$.

(D2.b) $\mu(x) \in H$ implies $\mathrm{lca}_S(\sigma(L(x))) \prec_S \mu(x)$.

(D3) Suppose $x, y \in V$ such that $x \prec_T y$. If $\mu(x), \mu(y) \in H$ then $\mu(x) \preceq_S \mu(y)$; otherwise $\mu(x) \prec_S \mu(y)$.

Proof. Suppose $x \in V$. Then (D1) is equivalent to *(i)* and the fact that $t(x) = \odot$ if and only if $x \in L$. Conditions *(ii)* and *(v)* together imply (D2.a). If $\mu(x) \in H$ then x is duplication vertex of T. From condition *(iv)* we conclude that $\mathrm{lca}_S(\sigma(L(x))) \preceq_S \mu(x)$. Since $\mathrm{lca}_S(\sigma(L(x))) \in W$, equality cannot hold and so (D2.b) follows. (D3) is an immediate consequence of *(iv)*. □

For T a gene tree, B a set of species and maps σ and t as above, our goal is now to characterize (1) those (T,t,σ) for which a species tree on B exists and (2) species trees on B that are species trees for (T,t,σ).

Results and discussion
Results

Unless stated otherwise, we continue with our assumptions on B, (T,t,σ), and S as stated in Definition 1. We start with the simple observation that a reconciliation map from (T,t,σ) to S preserves the ancestor order of T and hence T imposes a strong constraint on the relationship of most recent common ancestors in S:

Lemma 3. *Let $\mu : V \to W \cup H$ be a reconciliation map from (T,t,σ) to S. Then*

$$\mathrm{lca}_S(\mu(x), \mu(y)) \preceq_S \mu(\mathrm{lca}_T(x,y)) \qquad (1)$$

holds for all $x, y \in V$.

Proof. Assume that x and y are distinct vertices of T. Consider the unique path P connecting x with y. P is uniquely subdivided into a path P' from x to $\mathrm{lca}_T(x,y)$ and a path P'' from $\mathrm{lca}_T(x,y)$ to y. Condition (iv) implies that the images of the vertices of P' and P''

under μ, resp., are ordered in S with regards to $\preccurlyeq_S$ and hence are contained in the intervals Q'' and Q'' that connect $\mu(\text{lca}_T (x, y))$ with $\mu(x)$ and $\mu(y)$, respectively. In particular $\mu(\text{lca}_T (x, y))$ is the largest element (w.r.t. $\preccurlyeq_S$) in the union of $Q' \cup Q''$ which contains the unique path from $\mu(x)$ to $\mu(y)$ and hence also $\text{lca}_S(\mu(x), \mu(y))$. □

Equation (1) is well known to hold for gene tree/species reconciliation in the absence of a prescribed event labeling in T.

Since a phylogenetic tree (in the original sense) T is uniquely determined by its induced triple set $\Re(T)$, it is reasonable to expect that all the information on the species tree(s) for (T,t, σ) is contained in the images of the triples in $\Re(T)$ (or more precisely their leaves) under σ. However, this is not the case in general as the situation is complicated by the fact that not all triples in $\Re(T)$ are informative about a species tree that displays T. The reason is that duplications may generate distinct paralogs long before the divergence of the species in which they eventually appear. To address this problem, we associate to (T,t, σ) the set of triples

$$\mathfrak{G} = \mathfrak{G}(T, t, \sigma) = \{r \in \Re(T) | t(\text{lca}_T(r)) = \bullet \text{ and } \sigma(x) \neq \sigma(y), \text{ for all } x,y \in L(r) \text{ pairwise distinct}\}. \quad (2)$$

As we shall see below, $\mathfrak{G}(T, t, \sigma)$ contains all the information on a species tree for (T,t, σ) that can be gleaned from (T,t, σ).

Lemma 4. *If μ is a reconciliation map from (T,t, σ) to S and $((x,y),z) \in \mathfrak{G}(T, t, \sigma)$ then S displays $((\sigma(x), \sigma(y)), \sigma(z))$.*

Proof. Put $\mathfrak{G} = \mathfrak{G}(T, t, \sigma)$ and recall that L denotes the leaf set of T. Let $\{x, y, z\} \in \binom{L}{3}$ and assume w.l.o.g. that $((x,y),z) \in \mathfrak{G}$. First consider the case that $t(\text{lca}_T (x, y)) = \bullet$. From condition *(v)* we conclude that $\mu(\text{lca}_T (x, y)) = \text{lca}_S(\sigma(x), \sigma(y))$ and $\mu(\text{lca}_T (x, y, z)) = \text{lca}_S(\sigma(x), \sigma(y), \sigma(z))$. Since, by assumption, $\text{lca}_T(x,y) \prec \text{lca}_T(x, y, z)$, we have as a consequence of condition *(iv)* that $\mu(\text{lca}_T(x,y)) \prec \mu(\text{lca}_T(x, y, z))$. From $\text{lca}_T (x, z) = \text{lca}_T (y, z) = \text{lca}_T (x, y, z)$ we conclude that S must display $((\sigma(x), \sigma(y)), \sigma(z))$ as S is assumed to be a species tree for (T,t, σ).

Now suppose that $t(\text{lca}_T (x, y)) = \square$ and therefore, $\mu(\text{lca}_T (x, y)) \in H$. Moreover, $\mu(\text{lca}_T (x, y, z)) \in W$ holds. Hence, Lemma 3 and property *(iv)* together imply that $\text{lca}_S(\sigma(x), \sigma(y)) \prec_S \mu(\text{lca}_T(x,y)) \prec_S \mu(\text{lca}_T(x,y,z))$. Thus, we again obtain that the triple $((\sigma(x), \sigma(y)), \sigma(z))$ is displayed by S. □

It is important to note that a similar argument cannot be made for triples in $\Re(T)$ rooted in a duplication vertex of T as such triplets are in general not displayed by a species tree for (T,t, σ). We present the generic counterexample in Figure 3. To state our main result (Theorem 6), we require a further definition.

Definition 5. *For (T,t, σ), we define the set*

$$\mathfrak{S} = \mathfrak{S}(T, t, \sigma) = \{((a,b),c) | \exists (x,y,z) \in \mathfrak{G}(T,t,\sigma) \text{ with } \sigma(x) = a, \sigma(y) = b, \text{ and } \sigma(z) = c\} \quad (3)$$

As an immediate consequence of Lemma 4, $\mathfrak{S}(T, t, \sigma)$ must be displayed by any species tree for (T, t, σ) with leaf set B.

Theorem 6. *Let S be a species tree with leaf set B. Then there exists a reconciliation map μ from (T,t, σ) to S whenever S displays all triples in $\mathfrak{S}(T, t, \sigma)$.*

Proof. Recall that L is the leaf set of $T = (V, E)$. Put $S = (W, H)$ and $\mathfrak{S} = \mathfrak{S}(T, t, \sigma)$. We first consider the subset $G := \{x \in V | t(x) \in \{\bullet, \odot\}\}$ of V comprising of the leaves and speciation vertices of T.

We explicitly construct the map $\mu : G \to W$ as follows. For all $x \in V$, we put

(M1) $\mu(x) = \sigma(x)$ if $t(x) = \odot$,

(M2) $\mu(x) = \text{lca}_S(\sigma(L(x)))$ if $t(x) = \bullet$.

Note that alternative (M1) ensures that μ satisfies Condition *(i)*. Also note that in view of the simple consequence following the statement of Condition *(C)* we have for all $x \in V$ with $t(x) = \bullet$ that there are leaves $y', y'' \in L(x)$ with $\sigma(y') \neq \sigma(y'')$. Thus $\text{lca}_S(\mu(L(x)) \in W \setminus B$, i.e. μ satisfies Condition *(ii)*. Also note that, by definition, alternative (M2) ensures that μ satisfies Condition *(v)*.

Claim: If $x, y \in G$ with $x \prec_T y$ then $\mu(x) \prec_S \mu(y)$.

Since y cannot be a leaf of T as $x \prec_T y$ we have $t(y) = \bullet$. There are two cases to consider, either $t(x) = \bullet$ or $t(x) = \odot$. In the latter case $\mu(x) = \sigma(x) \in B$ while $\mu(y) \in W \setminus B$ as argued above. Since $x \in L(y)$ we have $\mu(x) \prec_S \mu(y)$, as desired.
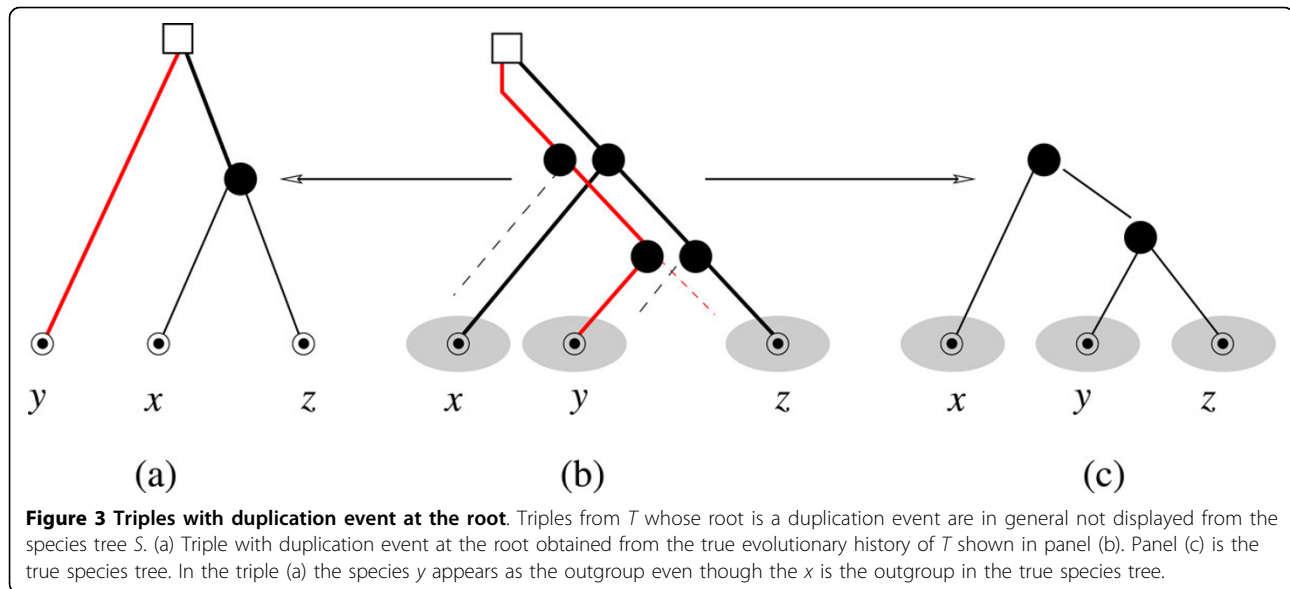
Now suppose $t(x) = \bullet$. Again by the simple consequence following Condition *(C)*, there are leaves $x', x'' \in L(x)$ with $a = \sigma(x') \neq \sigma(x'') = b$. Since $x \prec_T y$ and $t(y) = \bullet$, by Condition *(C)*, we conclude that $c = \sigma(y') \notin \sigma(L(x))$ holds for all $y' \in L(y) \setminus L(x)$. Thus, $((a, b), c) \in \mathfrak{S}$. But then $((a, b), c)$ is displayed by S and therefore $\text{lca}_S(a, b) \prec_S \text{lca}_S(a, b, c)$.. Since this holds for all triples $((x', x''), y') \in \mathfrak{G}$ with $x', x'' \in L(x)$ and $y' \in L(y) \setminus L(x)$ we conclude $\mu(x) = \text{lca}_S(\sigma(L(x))) \prec_S \text{lca}_S(\sigma(L(x)) \cup \sigma(L(y)\setminus L(x))) = \text{lca}_S(\sigma(L(y))) = \mu(y)$, establishing the claim. It follows immediately that μ also satisfies Condition *(iv.2)* if x and y are contained in G.

Next, we extend the map μ to the entire vertex set V of T using the following observation. Let $x \in V$ with $t(x) = \square$. We know by Lemma 3 that $\mu(x)$ is an edge $[u, v] \in H$ so that $\text{lca}_S(\sigma(L(x))) \preccurlyeq_S v$. Such an edge exists for $v = \text{lca}_S(\sigma(L(x)))$ by construction. Every speciation vertex $y \in V$ with $x \prec_T y$ therefore necessarily maps above this edge, i.e., $u \preccurlyeq_S \mu(y)$ must hold. Thus we set

(M3) $\mu(x) = [u, \text{lca}_S(\sigma(L(x)))]$ if $t(x) = \square$.

which now makes μ a map from V to $W \cup H$.

By construction, Conditions *(iii)*, *(iv.2)* and *(v)* are thus satisfied by μ. On the other hand, if there is a speciation vertex y between two duplication vertices x and x' of T,

Figure 3 Triples with duplication event at the root. Triples from *T* whose root is a duplication event are in general not displayed from the species tree *S*. (a) Triple with duplication event at the root obtained from the true evolutionary history of *T* shown in panel (b). Panel (c) is the true species tree. In the triple (a) the species *y* appears as the outgroup even though the *x* is the outgroup in the true species tree.

i.e., $x \prec_T y \prec_T x'$, then $\mu(x) \prec_S \mu(x')$. Thus μ also satisfies Condition *(iv.1)*.

It follows that μ is a reconciliation map from (T,t,σ) to *S*. □

Corollary 7. *Suppose that S is a species tree for (T,t, σ) and that L and B are the leaf sets of T and S, respectively. Then a reconciliation map μ from (T,t, σ) to S can be constructed in O(|L||B|).*

Proof. In order to find the image of an interior vertex *x* of *T* under μ, it suffices to determine $\sigma(L(x))$ (which can be done for all *x* simultaneously, e.g. by bottom up transversal of *T* in $O(|L||B|)$ time) and $\text{lca}_S(\sigma(L(x)))$. The latter task can be solved in linear time using the idea presented in [39] to calculate the lowest common ancestor for a group of nodes in the species tree. □

We remark that given a species tree *S* on *B* that displays all triples in $\mathfrak{S}(T,t,\sigma)$, there is no freedom in the construction of a reconciliation map on the set $\{x \in V | t(x) \in \{\bullet, \odot\}\}$. The duplication vertices of *T*, however, can be placed differently, resulting in possibly exponentially many reconciliation maps from (T,t,σ) to *S*.

Lemma 4 implies that consistency of the triple set $\mathfrak{S}(T,t,\sigma)$ is necessary for the existence of a reconciliation map from (T,t,σ) to a species tree on *B*. Theorem 6, on the other hand, establishes that this is also sufficient. Thus, we have

Theorem 8. *There is a species tree on B for (T,t, σ) if and only if the triple set $\mathfrak{S}(T, t, \sigma)$ is consistent.*

We remark that a related result is proven in [26, Theorem.5] for the full tree reconciliation problem starting from a forest of gene trees.

It may be surprising that there are no strong restrictions on the set $\mathfrak{S}(T,t,\sigma)$ of triples that are implied by the fact that they are derived from a gene tree (T,t,σ).
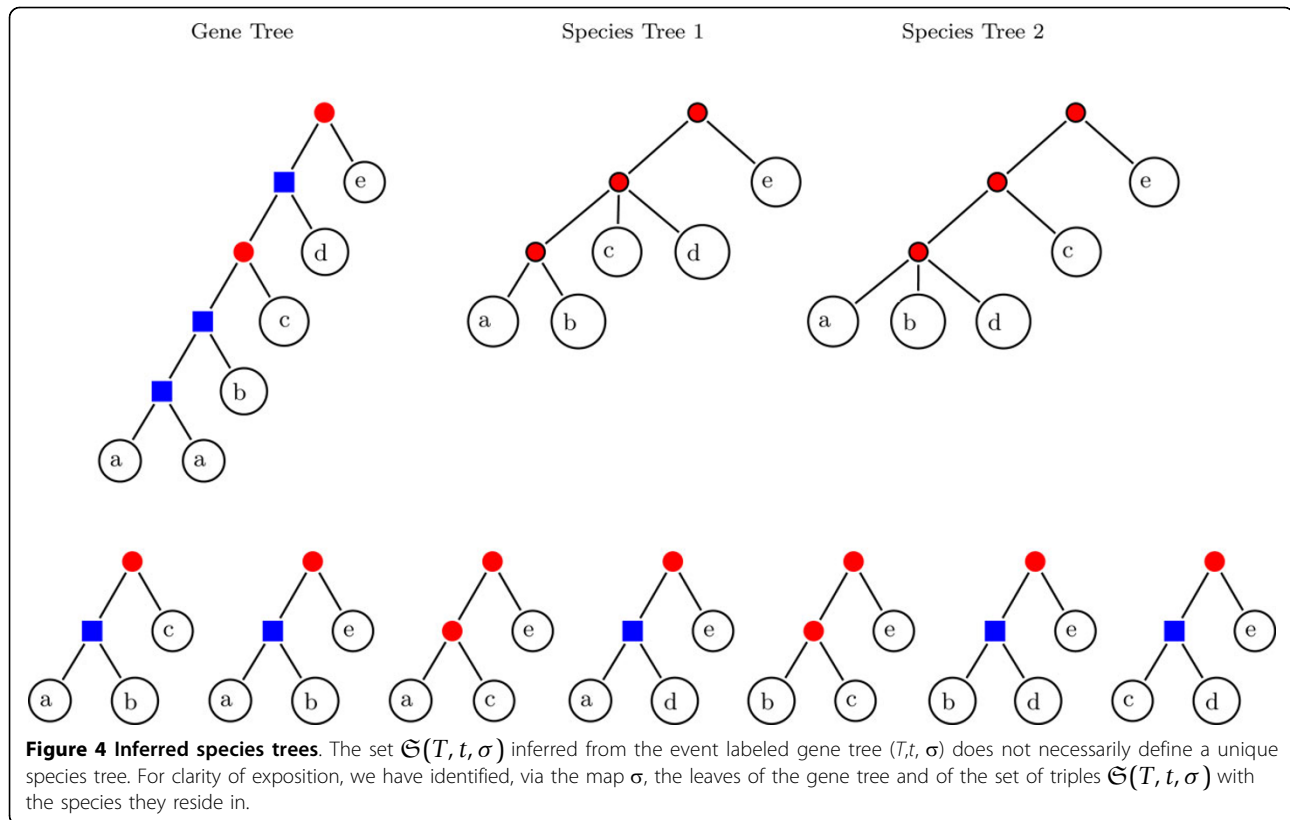
Theorem 9. *For every set $\mathfrak{x}$ of triples on some finite set B of size at least one there is a gen e tree T = (V, E) with leaf set L together with an event map $t : V \rightarrow \{\bullet, \square, \odot\}$ and a map $\sigma : L \rightarrow B$ that assigns to every leaf of T the species in B it resides in, such that $\mathfrak{x} = \mathfrak{S}(T, t, \sigma)$.*

Proof. Irrespective of whether $\mathfrak{x}$ is consistent or not we construct the components of the required 3-tuple (T,t,σ) as follows: To each triple $r_k = ((x_{k1}, x_{k2}), x_{k3}) \in \mathfrak{x}$ we associate a triple $T_k = ((a_{k1}, a_{k2}), a_{k3})$ via a map $\sigma_k : L_k = \{a_{k1}, a_{k2}, a_{k3}\} \rightarrow \{x_{k1}, x_{k2}, x_{k3}\}$ with $\sigma(a_{ki}) = x_{ki}$ for $i = 1, 2, 3$ where we assume that for any two distinct triples $r_k, r_l \in \mathfrak{x}$ we have that $\sigma_k(L_k) \cap \sigma_l(L_l) = \emptyset$. Then we obtain $T = (V, E)$ by first adding a single new vertex ρ_T to the union of the vertex sets of the triples T_k and then connecting ρ_T to the root ρ_k of each of the triples T_k. Clearly, *T* is a phylogenetic tree on $L = \cup_{r_k \in \mathfrak{x}} L(\rho_k)$. Next, we define the map $t : V \rightarrow \{\bullet, \square, \odot\}$ by putting $t(\rho_T) = \square$, $t(a) = \odot$ for all $a \in L$ and $t(a) = \bullet$ for all $a \in V - (L \cup \{\rho_T\})$. Finally, we define the map $\sigma : L \rightarrow B$ by putting, for all $a \in L$, $\sigma(a) = \sigma_k(a)$ where $a \in L_k$. Clearly $\mathfrak{S}(T, t, \sigma) = \mathfrak{x}$. □

We remark that the gene tree constructed in the proof of Theorem 9 can be made into a binary tree by splitting the root ρ_T into a series of duplication and loss events so that each subtree is the descendant of a different paralog. Since by Theorem. 9 there are no restrictions on the possible triple sets $\mathfrak{S}(T, t, \sigma)$, it is clear that *S* will in general not be unique. An example is shown in Figure 4.

Results for simulated gene trees

In order to determine empirically how much information on the species tree we can hope to find in event labeled gene trees, we simulated species trees together

Figure 4 Inferred species trees. The set $\mathfrak{S}(T, t, \sigma)$ inferred from the event labeled gene tree (T,t, σ) does not necessarily define a unique species tree. For clarity of exposition, we have identified, via the map σ, the leaves of the gene tree and of the set of triples $\mathfrak{S}(T, t, \sigma)$ with the species they reside in.

with corresponding event-labeled gene trees with different duplication and loss rates. Approximately 150 species trees with 10 to 100 species were generated according to the "age model" [40]. These trees are balanced and the edge lengths are normalized so that the total length of the path from the root to each leaf is 1. For each species tree, we then simulated a gene tree as described in [41], with duplication and loss rate parameters $r \in 0[1]$ sampled uniformly. Events are modeled by a Poisson distribution with parameter $r \cdot \ell$, where ℓ is the length of an edge as generated by the age model. Losses were additionally constrained to retain at least one copy in each species, i.e., $\sigma(L) = B$ is enforced. After determining the triple set $\mathfrak{S}(T, t, \sigma)$ according to Theorem 6, we used BUILD [27] (see also [42]) to compute the species tree. In all cases BUILD returns a tree that is a homomorphic contraction of the simulated species tree. The difference between the original and the reconstructed species tree is thus conveniently quantified as the difference in the number of interior vertices. Note that in our situation this is the same as the split metric [27].
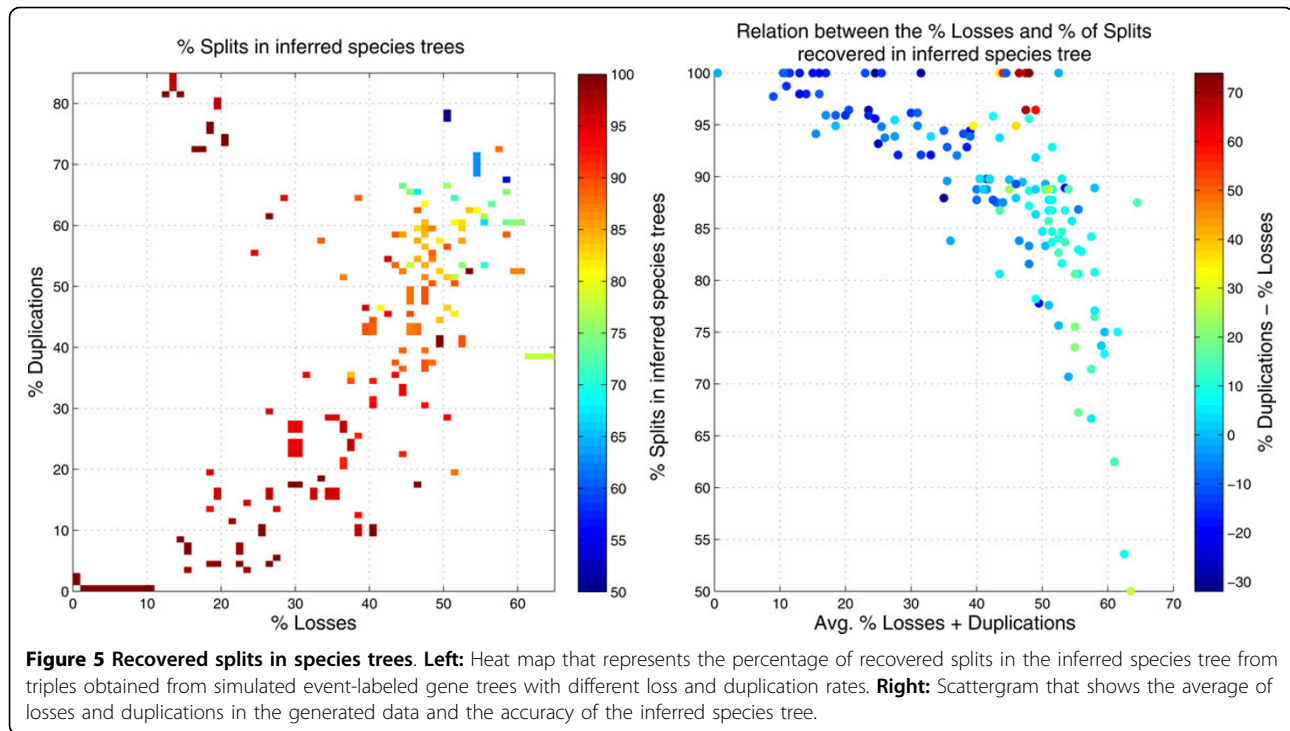
The results are summarized in Figure 5. Not surprisingly, the recoverable information decreases in particular with the rate of gene loss. Nevertheless, at least 50% of the splits in the species tree are recoverable even at very high

loss rates. For moderate loss rates, in particular when gene losses are less frequent than gene duplications, nearly the complete information on the species tree is preserved. It is interesting to note that BUILD does not incorporate splits that are not present in the input tree, although this is not mathematically guaranteed.

Discussion

Event-labeled gene trees can be obtained by combining the reconstruction of gene phylogenies with methods for orthology detection. Orthology alone already encapsulates partial information on the gene tree. More precisely, the orthology relation is equivalent to a homomorphic image of the gene tree in which adjacent vertices denote different types of events. We discussed here the properties of reconciliation maps μ from a gene tree T along with an event labelling map t and a gene to species assignment map σ to a species tree S. We show that (T,t) event labeled gene trees for which a species tree exists can be characterized in terms of the set σ of triples that is easily constructed from a subset of triples of T. Simulated data shows, furthermore, that such trees convey a large amount of information on the underlying species tree, even if the gene loss rate is high.

It can be expected that for real-life data the tree T contains errors so that $\mathfrak{S} := \mathfrak{S}(T, t, \sigma)$ may not be

Figure 5 Recovered splits in species trees. **Left:** Heat map that represents the percentage of recovered splits in the inferred species tree from triples obtained from simulated event-labeled gene trees with different loss and duplication rates. **Right:** Scattergram that shows the average of losses and duplications in the generated data and the accuracy of the inferred species tree.

consistent. In this case, an approximation to the species tree could be obtained e.g. from a maximum consistent subset of $\mathfrak{S}$. Although (the decision version of) this problem is NP-complete [43,44], there is a wide variety of practically applicable algorithms for this task, see [35,45]. Even if $\mathfrak{S}$ is consistent, the species tree is usually not uniquely determined. Algorithms to list all trees consistent with $\mathfrak{S}$ can be found e.g. in [46,47]. A characterization of triple sets that determine a unique tree can be found in [48]. Since our main interest is to determine the constraints imposed by (T, t, σ) on the species tree S, we are interested in a least resolved tree S that displays all triples in $\mathfrak{S}$. The BUILD algorithm and its relatives in general produce minor-minimal trees, but these are not guaranteed to have the minimal number of interior nodes. Finding a species tree with a minimal number of interior nodes is again a hard problem [37]. At least, the vertex minimal trees are among the possibly exponentially many minor minimal trees enumerated by Semple's algorithms [36].

For a given species tree S, it is rather easy to find a reconciliation map μ from (T, t, σ) to S. A simple solution μ is closely related to the so-called LCA reconciliation: every node x of T is mapped to the last common ancestor of the species below it, $lca_S(\sigma(L(x)))$ or to the edge immediately above it, depending on whether x is speciation or a duplication node. While this solution is unique for the speciation nodes, alternative mappings are possible for the duplication nodes. The set of

possible reconciliation maps can still be very large despite the specified event labels. If the event labeling t is unknown, there is a reconciliation from any gene tree T to any species tree S, realized in particular by the LCA reconciliation, see e.g. [26,38]. The reconciliation then defines the event types. Typically, a parsimony rule is then employed to choose a reconciliation map in which the number of duplications and losses is minimized, see e.g. [1,4,5,9]. In our setting, on the other hand, the event types are prescribed. This restricts the possible reconciliation maps so that the gene tree cannot be reconciled with an arbitrary species tree any more. Since the observable events on the gene tree are fixed, the possible reconciliations cannot differ in the number of duplications. Still, one may be interested in reconciliation maps that minimize the number of loss events. An alternative is to maximize the number of duplication events that map to the same edge in S to account for whole genome and chromosomal duplication events [9].

Conclusions

Our approach to the reconciliation problem via event-labeled gene trees opens up some interesting new avenues to understanding orthology. In particular, the results in this contribution combined with those in [22] concerning cographs should ultimately lead to a method for automatically generating orthology relations that takes into account species relationships without having

to explicitly compute gene trees. This is potentially very useful since gene tree estimation is one of the weak points of most current approaches to orthology analysis.

Author details

[1]Max-Planck-Institute for Mathematics in the Sciences, Leipzig, D-04103, Germany. [2]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, D-04107, Germany. [3]Center for Bioinformatics, Saarland University, Saarbrücken, D-66041, Germany. [4]High Throughput Bioinformatics, Faculty of Mathematics and Computer Science, Friedrich Schiller Universität Jena, Jena, D-07743, Germany. [5]Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Leipzig, D04103, Germany. [6]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK. [7]Inst. f. Theoretical Chemistry, University of Vienna, Vienna, A-1090, Austria. [8]Santa Fe Institute, Santa Fe, NM, 87501, USA.

Authors' contributions

All authors contributed to the development of the theory. MHR and NW produced the simulated data. All authors contributed to writing, reading, and approving the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 19 December 2012

References

1. Guigó R, Muchnik I, Smith TF: **Reconstruction of ancient molecular phylogeny.** *Mol Phylogenet Evol* 1996, **6**:189-213.
2. Page RD, Charleston MA: **From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem.** *Mol Phylogenet Evol* 1997, **7**:231-240.
3. Arvestad L, Berglund AC, Lagergren J, Sennblad B: **Bayesian gene/species tree reconciliation and orthology analysis using MCMC.** *Bioinformatics* 2003, **19**:i7-i15.
4. Bonizzoni P, Della Vedova G, Dondi R: **Reconciling a gene tree to a species tree under the duplication cost model.** *Theor Comp Sci* 2005, **347**:36-53.
5. Górecki P, J T: **DSL-trees: A model of evolutionary scenarios.** *Theor Comp Sci* 2006, **359**:378-399.
6. Hahn MW: **Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution.** *Genome Biol* 2007, **8**:R141.
7. Bansal MS, Eulenstein O: **The multiple gene duplication problem revisited.** *Bioinformatics* 2008, **24**:i132-i138.
8. Chauve C, Doyon JP, El-Mabrouk N: **Gene family evolution by duplication, speciation, and loss.** *J Comput Biol* 2008, **15**:1043-1062.
9. Burleigh JG, Bansal MS, Wehe A, Eulenstein O: **Locating large-scale gene duplication events through reconciled trees: implications for identifying ancient polyploidy events in plants.** *J Comput Biol* 2009, **16**:1071-1083.
10. Larget BR, Kotha SK, Dewey CN, Ane C: **BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis.** *Bioinformatics* 2010, **26**:2910-2911.
11. Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, Dehal P, Wang J, Durbin R: **TreeFam: a curated database of phylogenetic trees of animal gene families.** *Nucleic Acids Res* 2006, **34**:D572-D580.
12. Goodstadt L, Ponting CP: **Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human.** *PLoS Comput Biol* 2006, **2**:e133.
13. Datta RS, Meacham C, Samad B, Neyer C, Sjölander K: **Berkeley PHOG: PhyloFacts orthology group prediction web server.** *Nucl Acids Res* 2009, **37**:W84-W89.
14. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35**:D610-617.
15. Pryszcz LP, Huerta-Cepas J, Gabaldón T: **MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score.** *Nucleic Acids Res* 2011, **39**:e32.
16. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS Comput Biol* 2009, **5**:e1000262.
17. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178-2189.
18. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28**:33-36.
19. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, **36**:D13-D21.
20. Berglund AC, Sjölund E, Ostlund G, Sonnhammer EL: **InParanoid 6: eukaryotic ortholog clusters with inparalogs.** *Nucleic Acids Res* 2008, **36**:D263-D266.
21. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ: **Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis.** *BMC Bioinformatics* 2011, **12**:124.
22. Hellmuth M, Hernandez-Rosales M, Huber KT, Moulton V, Stadler PF, Wieseke N: **Orthology Relations, Symbolic Ultrametrics, and Cographs.** *J Math Biol* 2012.
23. Fitch WM: **Homology: a personal view on some of the problems.** *Trends Genet* 2000, **16**:227-231.
24. Böcker S, Dress AWM: **Recovering symbolically dated, rooted trees from symbolic ultrametrics.** *Adv Math* 1998, **138**:105-125.
25. Brandstädt A, Le VB, Spinrad JP: *Graph Classes: A Survey* SIAM Monographs on Discrete Mathematics and Applications, Philadephia: Soc. Ind. Appl. Math; 1999.
26. Chauve C, El-Mabrouk N: **New Perspectives on Gene Family Evolution: Losses in Reconciliation and a Link with Supertrees.** *LNCS* 2009, **5541**:46-58.
27. Semple C, Steel M: *Phylogenetics, Volume 24 of Oxford Lecture Series in Mathematics and its Applications* Oxford, UK: Oxford University Press; 2003.
28. Dress AWM, Huber KT, Koolen J, Moulton V, Spillner A: *Basic Phylogenetic Combinatorics* Cambridge: Cambridge University Press; 2011.
29. Bininda-Emonds O: *Phylogenetic Supertrees* Dordrecht, NL: Kluwer Academic Press; 2004.
30. Aho AV, Sagiv Y, Szymanski TG, Ullman JD: **Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions.** *SIAM J Comput* 1981, **10**:405-421.
31. Rauch Henzinger M, King V, Warnow T: **Constructing a Tree from Homeomorphic Subtrees, with Applications to Computational Evolutionary Biology.** *Algorithmica* 1999, **24**:1-13.
32. Jansson J, Ng JHK, Sadakane K, Sung WK: **Rooted maximum agreement supertrees.** *Algorithmica* 2005, **43**:293-307.
33. Byrka J, Gawrychowski P, Huber KT, Kelk S: **Worst-case optimal approximation algorithms for maximizing triplet consistency within phylogenetic networks.** *J Discr Alg* 2010, **8**:65-75.

34. van Iersel L, Kelk S, Mnich M: **Uniqueness, intractability and exact algorithms: reflections on leve l-k phylogenetic networks.** *J Bioinf Comp Biol* 2009, **7**:597-623.
35. Byrka J, Guillemot S, Jansson J: **New results on optimizing rooted triplets consistency.** *Discr Appl Math* 2010, **158**:1136-1147.
36. Semple C: **Reconstructing minimal rooted trees.** *Discr Appl Math* 2003, **127**:489-503.
37. Jansson J, Lemence RS, Lingas A: **The Complexity of Inferring a Minimally Resolved Phylogenetic Supertree.** *SIAM J Comput* 2012, **41**:272-291.
38. Doyon JP, Chauve C, Hamel S: **Space of Gene/Species Trees Reconciliations and Parsimonious Models.** *J Comp Biol* 2009, **16**:1399-1418.
39. Zhang L: **On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies.** *J Comput Biol* 1997, **4**:177-187.
40. Keller-Schmidt S, Tuğrul M, Eguíluz VM, Hernández-Garcíi E, Klemm K: **An Age Dependent Branching Model for Macroevolution.** *Tech Rep* 2010, 1012.3298v1, arXiv.
41. Hernandez-Rosales M, Wieseke N, Hellmuth M, Stadler PF: **Simulation of Gene Family Histories.** *Tech Rep* Univ. Leipzig; 2011, 12-017 [http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/12-017.pdf].
42. Aho AV, Sagiv Y, Szymanski TG, Ullman JD: **Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions.** *SIAM J Comput* 1981, **10**:405-421.
43. Jansson J: **On the Complexity of Inferring Rooted Evolutionary Trees.** *Electronic Notes Discr Math* 2001, **7**:50-53.
44. Wu BY: **Constructing the Maximum Consensus Tree from Rooted Triples.** *J Comb Optimization* 2004, **8**:29-39.
45. He YJ, Huynh TN, Jansson J, Sung WK: **Inferring phylogenetic relationships avoiding forbidden rooted triplets.** *J Bioinform Comput Biol* 2006, **4**:59-74.
46. Ng MP, Wormald NC: **Reconstruction of rooted trees from subtrees.** *Discr Appl Math* 1996, **69**:19-31.
47. Constantinescu M, Sankoff D: **An efficient algorithm for supertrees.** *J Classification* 1995, **12**:101-112.
48. Bryant D, Steel M: **Extension Operations on Sets of Leaf-Labeled Trees.** *Adv Appl Math* 1995, **16**:425-453.