Paszek and Górecki *BMC Genomics* 2016, **17**(Suppl 1):15
DOI 10.1186/s12864-015-2308-4

BMC Genomics

PROCEEDINGS Open Access

# Genomic duplication problems for unrooted gene trees

Jarosław Paszek* and Paweł Górecki

*From* The Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016)
San Francisco, CA, USA. 11 - 13 January 2016

## Abstract

**Background:** Discovering the location of gene duplications and multiple gene duplication episodes is a fundamental issue in evolutionary molecular biology. The problem introduced by Guigó et al. in 1996 is to map gene duplication events from a collection of rooted, binary gene family trees onto theirs corresponding rooted binary species tree in such a way that the total number of multiple gene duplication episodes is minimized. There are several models in the literature that specify how gene duplications from gene families can be interpreted as one duplication episode. However, in all duplication episode problems gene trees are rooted. This restriction limits the applicability, since unrooted gene family trees are frequently inferred by phylogenetic methods.

**Results:** In this article we show the first solution to the open problem of episode clustering where the input gene family trees are unrooted. In particular, by using theoretical properties of unrooted reconciliation, we show an efficient algorithm that reduces this problem into the episode clustering problems defined for rooted trees. We show theoretical properties of the reduction algorithm and evaluation of empirical datasets.

**Conclusions:** We provided algorithms and tools that were successfully applied to several empirical datasets. In particular, our comparative study shows that we can improve known results on genomic duplication inference from real datasets.

**Keywords:** Genomic duplication, Duplication episode, Reconciliation, Unrooted gene tree, Species tree

## Background

Genomic duplication plays important role in evolution of life on Earth. This phenomenon have been extensively studied in the last decades for plant, bacterial and many other genomes [1–7]. Duplication events can involve individual genes, genomic segments or whole genomes. While the reconstruction of evolutionary history of individual genes is generally well established [8–13], still little is known on the inference of large genomic duplications that can span through thousands of genes families.

In this approach we propose to use the model of reconciliation in which a gene tree is reconciled with its species tree. The concept of reconciliation was introduced by Goodman [14] and formalized by Page [8] in the context of reconciling potential incongruence between a rooted gene family tree and its species tree. In this model, differences between gene and species trees are explained in terms of evolutionary events such as gene duplication, gene loss and speciation. Reconciliation can be interpreted as the embedding of a gene tree into a species tree where these evolutionary events, located in the species tree, induce a biologically consistent scenario [15]. Tree reconciliation has been extensively studied in recent decades in many theoretical and practical contexts including supertree inference, error correction and HGT detection [16–24]. In the process of reconciliation, which is relatively simple from computational point of view, each gene from a single gene family is mapped into the species tree and it is classified as a single gene duplication or related to speciation. However, the problem becomes much more complex, when a gene duplication is a part of large genomic duplications, called *multiple gene*

*Correspondence: jpaszek@mimuw.edu.pl
University of Warsaw, Institute of Informatics, Banacha 2, 02-097 Warsaw, Poland

*duplication episode*, in which parts of a genome are duplicated. In fact, it is known that a large duplication event is usually followed by many gene losses and gene rearrangements. In consequence, the reconstruction of large gene duplication events may be difficult.

The first approach to detect multiple gene duplication episodes from a collection of rooted gene trees was proposed by Guigó et al. [10]. In the model, for a given collection of rooted gene trees and a rooted species tree, the authors proposed heuristic to aggregate single gene duplication events into a large gene duplication. This approach was formalized and refined by Page and Cotton [25]. They formally defined the problem of *episode clustering* (EC) as the problem of locating the minimal number of locations in the species tree, where all duplications from the input gene trees can be placed. This model was applied in the context of the supertree problem by Fellows [26]. Burleigh et al. [27] and Bansal and Eulenstein [28] proposed the first polynomial time solutions for two types of the multiple gene duplication problems: the episode clustering (EC) and a more general variant of clustering called *minimum episodes* (ME). Finally, Luo et al. [29] proposed linear time and space algorithms to these problems.

While the classical reconciliation model is applicable to rooted trees only, most standard phylogenetic inference methods, like maximum likelihood, maximum parsimony or neighbour joining, infer unrooted gene family trees, and it is often difficult, to identify credible rootings. For example, outgroup rooting can result in incorrect rootings when evolutionary events cause heterogeneity in the gene trees, and rooting gene trees under the molecular clock assumption, or similarly by using midpoint rooting, also can result in error when there is a molecular rate variation throughout the tree [30, 31]. Tree reconciliation have been successfully extended to reconcile an unrooted gene tree with a rooted species tree by seeking a rooting of the unrooted gene tree that invokes the minimum number of evolutionary events such as gene duplications (D) or gene duplications and losses (DL), in the context of a given species tree [32, 33]. It is known that the rooting edges with minimal D or DL cost, induce a full subtree, called *plateau*, in the unrooted gene tree [34].

In this article we present the first solution to the open problem [27] of *unrooted episode clustering*, that is, the problem of episode clustering where the input consists of unrooted gene trees. We show that for a given set of unrooted gene trees and a species tree we can solve the unrooted episode clustering by reducing it to the rooted episode clustering problem that has a linear time complexity. Our solutions require a linear time preprocessing and a creation of at most $1 + 2^k$ collections of rooted gene trees, that is, instances of rooted EC Problem, where $k$ is the number of input gene trees having a special topology located in the plateau of the duplication cost (formally, the condition requires two stars S2 [32]). Usually $k$ represents a small fraction of the whole input, thus, this condition significantly reduces the complexity. In other words, we show that the problem of unrooted episode clustering is fixed parameter tractable. Finally, in a number of empirical computational experiments we show that despite the exponential worst case complexity our algorithm is able to resolve instances of the problem after the verification of at most two rooted datasets. In consequence, our solution can be efficiently applied to locate duplication clusters in collections of unrooted gene trees.

## Results

### Basic notation

A *species tree* is a rooted binary tree with leaves uniquely labeled by the names of species. Throughout this work, the species tree is fixed, therefore, we use $S$ to denote it. A *rooted gene tree* is a rooted binary tree with leaves labeled by the names of species. The set of species present in $T$ is denoted by $\mathcal{L}(T)$. The rooted tree $(T_1, T_2)$ has two subtrees $T_1$ and $T_2$ whose roots are children of the tree root. Additionally, for nodes $a$ and $b$, $a \preceq b$ means that $a$ and $b$ are on the same path from the root, with $b$ being closer to the root than $a$. We write $a \prec b$ if $a \preceq b$ and $a \neq b$. The root of a tree $T$ we denote by $\mathsf{root}(T)$.

Let $T = \langle V_T, E_T \rangle$ be a rooted gene tree such that $\mathcal{L}(T) \subseteq \mathcal{L}(S)$. *The least common ancestor (lca) mapping*, $\mathsf{M}_T : V_T \rightarrow V_S$, is defined as follows. If $v$ is a leaf in $T$ then $\mathsf{M}_T(v)$ is the leaf in $S$ labeled by the label of $v$. When $v$ is an internal node in $T$ having two children $a$ and $b$, then $\mathsf{M}_T(v)$ is the least common ancestor of $\mathsf{M}_T(a)$ and $\mathsf{M}_T(b)$ in $S$. An internal node $g \in V_T$ is called a *duplication* if $\mathsf{M}_T(g) = \mathsf{M}_T(a)$ for a child $a$ of $g$. *The duplication cost*, denoted by $\mathsf{D}(T, S)$, is the total number of duplications in $T$. Each non-duplication node of $T$ we call a *speciation*. The total number of *gene losses* required to reconcile $T$ and $S$ can be defined by: $\mathsf{L}(T, S) = 2\mathsf{D}(T, S) + \sum_{g \text{ is internal}, a,b \text{ children of } g} (\|\mathsf{M}_T(a), \mathsf{M}_T(b)\| - 2)$, where $\|a, b\|$ is the number of edges on the path connecting $a$ and $b$ in $S$. Finally, we can define the *duplication-loss cost* of reconciling a rooted gene tree $T$ and a species tree $S$ as follows: $\mathsf{DL}(T, S) = \mathsf{D}(T, S) + \mathsf{L}(T, S)$ [34]. Examples of the reconciliation are depicted in Fig. 1.

### Unrooted reconciliation

*The unrooted gene tree* is an undirected acyclic connected graph in which each node has degree 1 (leaves) or 3 (internal nodes), and the leaves are labeled by the names of species. For an unrooted gene tree $G = \langle V_G, E_G \rangle$ and an edge $e \in E_G$, by $G_e$, we denote the rooting of $G$ obtained from $G$ by placing the root on $e$. Such a rooting induces the duplication cost $\mathsf{D}(G_e, S)$. We call *D-minimal*, the rooting or edges having the minimal duplication cost. It follows from the theory of unrooted reconciliation [32, 34] that

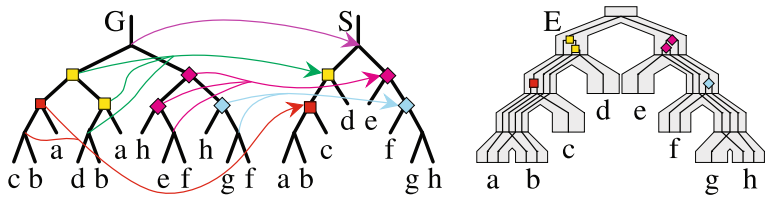


**Fig. 1** Gene and species tree reconciliation. *Left*: the lca-mapping between a gene tree $G$ and a species tree $S$ shown for internal nodes. The decoration of nodes indicates gene duplication events. *Right*: an embedding of $G$ into $S$. To reconcile these trees 6 gene duplications and 17 gene losses (not shown) are required, i.e., $\mathrm{D}(G,S) = 6$ and $\mathrm{DL}(G,S) = 23$. See also $G_{4*}$ in Fig. 3

the set of D-minimal edges, called *D-plateau*, is a full subtree of $G$. The same property holds for the DL-plateau, that is, the set of edges with the minimal duplication-loss cost. We use a similar notation for DL-minimal edges, rootings and so on. The most important property of these plateaus is below.

**Theorem 1** (From [34]). *DL-plateau is a subgraph of D-plateau.*

Without loss of generality we assume that every root of a gene tree is mapped into the root of $S$, denoted by $\top$, and both trees are non-trivial. An edge $e = \langle v, w\rangle$ of $G$ is *empty* if the root of $G_e$ is a speciation, i.e., $\mathsf{M}_{G_e}(v) \neq \top \neq \mathsf{M}_{G_e}(w)$. We call $e$ *double* if $\mathsf{M}_{G_e}(v) = \top = \mathsf{M}_{G_e}(w)$. Otherwise, $e$ is called *single*. A single edge $e$ is called *$v$-incoming* or *$w$-outgoing* if $\mathsf{M}_{G_e}(v) \neq \top = \mathsf{M}_{G_e}(w)$.

Let $v$ be an internal node of $G$, then a *star* with a *center* $v$ consists of three edges, denoted by $e_a$, $e_b$ and $e_c$, sharing $v$ and incident to nodes $a$, $b$ and $c$, respectively (see Fig. 2). The are several types of possible star topologies based on the above classification of edges: the $S1$ star has one $v$-incoming edge and two $v$-outgoing edges, the $S2$ star has exactly two $v$-outgoing edges and one empty edge, the $S3$ star has two $v$-outgoing edges and one double edge, the $S4$ star all 3 edges are double, and the $S5$ star has one $v$-outgoing edge and two double edges. The star topologies are depicted in Fig. 2.

**Theorem 2** (Adopted from [32]). *For a given unrooted gene tree G, we have*

- *either G has exactly one empty edge or G has at least one double edge,*
- *if the DL-plateau of G consists of exactly one edge, then this edge is either empty or double, and all other edges are single.*
- *if the DL-plateau of G has more than one edge, then it contains all edges present in stars S4 and S5, and all other edges are single.*

Note that if a gene has an empty edge, then it has at most two stars S2 (see examples in Fig. 3).

### Episode clustering problems

To model gene duplication episodes we allow to relocate a gene duplication from its lca-mapping location to one of its ancestors. In other words, we introduce mappings representing evolutionary scenarios that can differ from the scenario defined by the lca-mapping. Additionally, we require that the total number of gene duplications is minimal. To ensure biological correctness of such mappings, we introduce several conditions, e.g., time order preservation.

A mapping $\mathsf{F}_G\colon V_G \to V_S$ is called *valid* if the following conditions are satisfied:

- $\mathsf{F}_G(a) \preceq \mathsf{F}_G(b)$ if $a \preceq b$ (time consistency),
- $\mathsf{F}_G(a) = \mathsf{M}_G(a)$ for any speciation node $a$ (fixed speciations),
- $\mathsf{F}_G(a) \succeq \mathsf{M}_G(a)$ for any duplication node $a$ (duplication can be raised),
- $\mathsf{F}_G(a) \prec \mathsf{M}_G(b)$ for any speciation node $b$ such that $a \prec b$ (fixed number of gene duplications).

It can be shown that every valid mapping uniquely defines an evolutionary scenario represented by a DLS-tree [15]. Additionally, every DLS-tree obtained from a

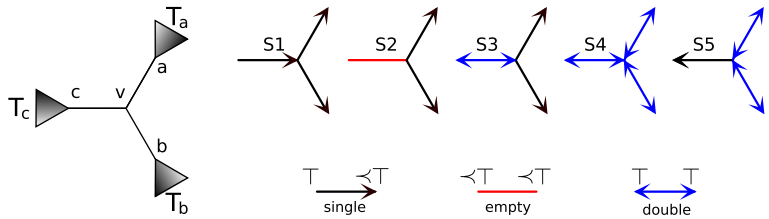


**Fig. 2** Types of stars. Star topology with the center $v$, types of edges and stars

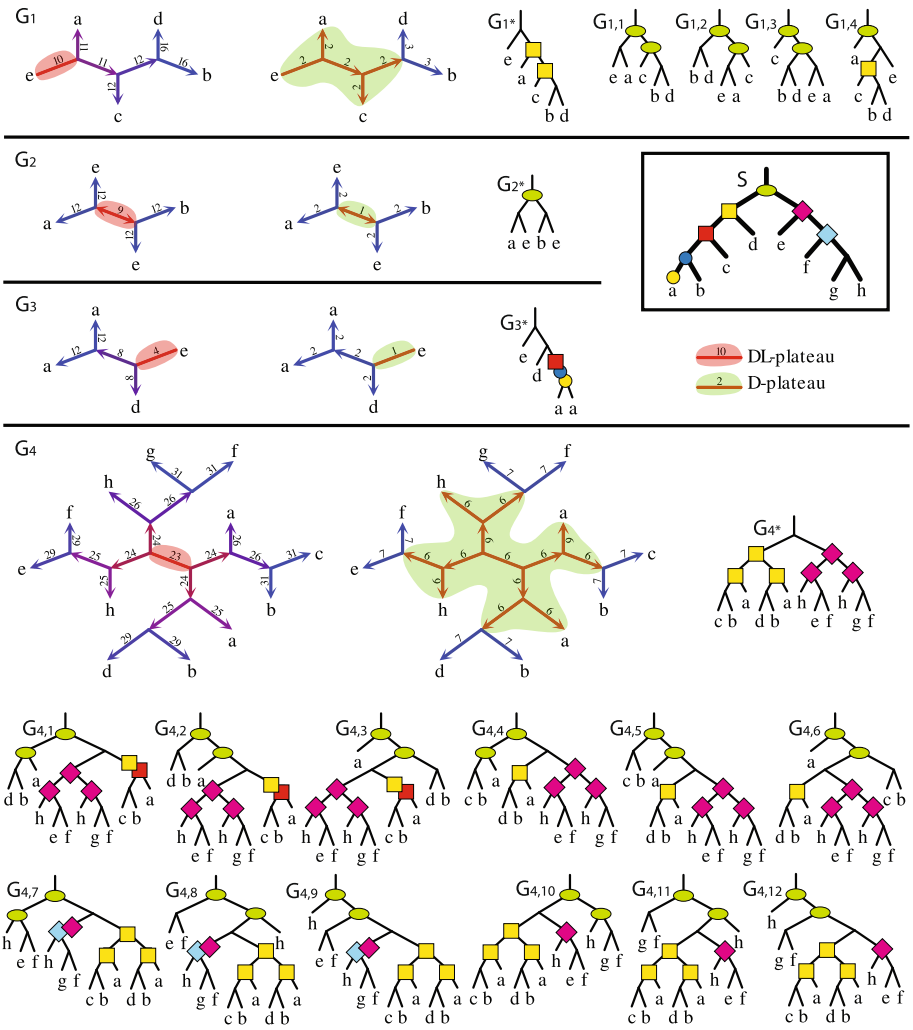


**Fig. 3** An example of unrooted episode clustering. A species tree $S$ and four unrooted gene trees $G_1, G_2, G_3, G_4$ with all D-minimal rootings. For every gene tree two star topologies are shown: one for the duplication-loss cost (*left*) and one for the duplications cost (*right*). Every edge of a gene tree is decorated with the corresponding cost of rooting. Every duplication node in rootings of gene trees is decorated by all possible locations (i.e., valid mappings) of its duplication cluster from optimal solutions of single-UEC. Note that the rooting $G_{4*}$, whose lca-mappings are shown in Fig. 1, has two duplications at $(c,(b,a))$ and $(h,(f,g))$ that are raised (here) to create two duplications clusters. Let $\{G_2, G_4\}$ be an instance of UEC Problem. Then, the $\top$-cluster, that is present in $G_{2*}$, contributes to the optimal solution. In such a case, the solution is induced by one of the two instances of EC problem: $\{G_{2*}, G_{4,1}\}$ or $\{G_{2*}, G_{4,7}\}$. This property is proved in Theorem 5 and in Lemma 6

valid mapping can be transformed into the optimal evolutionary scenario (i.e., lca-based scenario), by a sequence of TMOVE (i.e., lowering duplication) transformations. Please refer to [15] for more details on formal modeling of evolutionary scenarios. Observe, that the above model is more general than the model from [28].

We denote by $\mathsf{Dup}(T)$, the set of all duplication nodes in $T$. Let $G_1, G_2, \ldots, G_n$ be a collection of rooted gene trees. Assume that, for every $i \in \{1,2,\ldots,n\}$, $\mathsf{F}_i$ is a valid mapping between $G_i$ and the species tree $S$. Every element $s \in \bigcup_i \mathsf{F}_i(\mathsf{Dup}_{G_i})$ denotes the location of multiple gene duplication events in $S$. Such locations will be called *duplication episodes*. A *duplication cluster* for $s$ is the set of all gene duplications present in $G_i$'s that are mapped to $s$. By $\top$-cluster we denote the duplication cluster whose elements are mapped to $\top$.

**Problem 1** (Rooted Episode Clustering (EC)). *Given a collection of rooted gene trees $G_1, G_2, \ldots, G_n$ and a species tree $S$. Compute the minimal number of duplication episodes, denoted by $\mathsf{EC}(G_1, G_2, \ldots, G_n, S)$, in the set of all valid mappings $\mathsf{F}_1, \mathsf{F}_2, \ldots, \mathsf{F}_n$ such that $\mathsf{F}_i\colon V_{G_i} \to V_S$.*

This problem can be solved in linear-time and space [29]. In this article we solve the following problem.

**Problem 2** (Unrooted Episode Clustering (UEC)). *Given a collection of unrooted gene trees $G_1, G_2, \ldots, G_n$ and a species tree $S$. Compute the minimal $\mathsf{EC}(T_1, T_2, \ldots, T_n, S)$ in the set of rooted gene trees $\{T_1, T_2, \ldots, T_n\}$ such that $T_i$ is a rooting obtained from $G_i$ by placing the root on the edge from the D-plateau.*

Observe, that we allow rootings only in the D-plateau. Otherwise, the total number of gene duplications is not minimal. By single-UEC we denote the problem UEC for a single unrooted gene tree, i.e., when $n = 1$. Every edge in an unrooted gene tree that induces the optimal solution for single-UEC will be called *optimal* (for single-UEC). For convenience, we use $\mathsf{EC}(T_1, T_2, \ldots, T_n)$ instead of $\mathsf{EC}(T_1, T_2, \ldots, T_n, S)$.

### Episodes in a gene tree with an empty edge

In this Section we solve single-UEC problem for the case when the input gene tree has one empty edge.

Let $v$ be a center of the star that contains the only DL-plateau edge in a gene tree $G$. This star induces three rooted subtrees $T_a$, $T_b$ and $T_c$ rooted at neighbours $a$, $b$ and $c$, respectively, as indicated in Fig. 2. Let $\mathbb{1}$ be the indicator function, that is, $\mathbb{1}(p)$ is 1 if $p$ is satisfied and 0 otherwise.

**Lemma 1.** *Let $a_0, a_1, a_2, \ldots, a_{n+1}$ (for $n \geq 0$) be the path of D-plateau nodes connecting $v = a_0$ and $a_{n+1} \in T_a$ in $G$. Let $G_n$ be the D-minimal rooting induced by the edge $\langle a_n, a_{n+1}\rangle$. If $e_* = \langle v, c\rangle$ is empty then*

$$\mathsf{EC}(G_n) = \mathsf{EC}(T_1, T_2, \ldots, T_{n+1}, T_b, T_c) + \mathbb{1}(\mathsf{root}(T_i) \notin \mathsf{Dup}(G_n) \text{ for all } i),$$

*where $T_1, T_2, \ldots, T_{n+1}$ are subtrees of $T_a$ such that $T_a = (T_1, (T_2, \ldots, (T_n, T_{n+1}) \ldots))$ and the root of $T_{n+1}$ is $a_{n+1}$ (see Figs. 2 and 4).*

*Proof.* First we show that $v$ is a speciation node in $G_n$. It follows from the fact that $v$ is a center of S2 star and $\langle v, b\rangle$ is single. Thus, $\mathsf{M}_n(v) = \top$, $\mathsf{M}_n(c) \prec \top$ and $\mathsf{M}_n(b) \prec \top$, where $\mathsf{M}_n$ is the lca-mapping for $G_n$. From the fact that $\mathsf{M}_n(v) = \top$ we conclude that all nodes on the path connecting the parent of $v$ with the root in $G_n$ are mapped to $\top$, therefore, they are duplications.

Lets consider the number of duplication clusters in $G_n$. We have the $\top$-cluster composed of the duplication nodes $a_1, a_2, \ldots, a_n, \mathsf{root}(G_n)$ mapped to $\top$. Both $T_c$ and $T_b$ in $G_n$ are under speciation node $v$ so their clusters are disjoint with the $\top$-cluster. Finally, if the root of some $T_i$ is a duplication then its cluster can be merged with the $\top$-cluster. Therefore, the $\top$-cluster contributes to $\mathsf{EC}(G_n)$ only if the root of $T_i$ is a speciation for every $i$. Now, it is easy to conclude the final formula. □

**Lemma 2.** *Under the assumptions from the previous lemma, we have*

$$\mathsf{EC}(G_n) = \mathsf{EC}(G_*) + \mathbb{1}(b \in \mathsf{Dup}(G_*) \text{ and } \mathsf{root}(T_i) \notin \mathsf{Dup}(G_*) \text{ for all } i),$$

*where $G_*$ is the rooting induced the empty edge $e_* = \langle v, c\rangle$ (see Fig. 4).*

*Proof.* Both rootings $G_n$ and $G_*$ are D-minimal. Hence, $\mathsf{D}(G_*, S) = \mathsf{D}(G_n, S)$ and, in consequence, the number of duplication nodes in $A = \{a_1, a_2, \ldots, a_n, v, \mathsf{root}(G_*)\}$ in $G_*$ and $B = \{a_1, a_2, \ldots, a_n, v, \mathsf{root}(G_n)\}$ in $G_n$ are equal. It follows from the properties of star S2, that in $G_n$ node $v$ is a speciation mapped to $\top$. Hence, all predecessors of $v$ are duplications in $G_n$. Thus, we have exactly $n + 1$ duplications in $B$. On the other hand, by star S2, $\mathsf{root}(G_*)$ is a speciation, therefore all remaining nodes in $A$ are duplications.

We conclude that $G_n$ has the $\top$-cluster containing duplications from $A$, and $G_*$ has a cluster (mapped below $\top$) containing duplications from $B$, respectively. These two

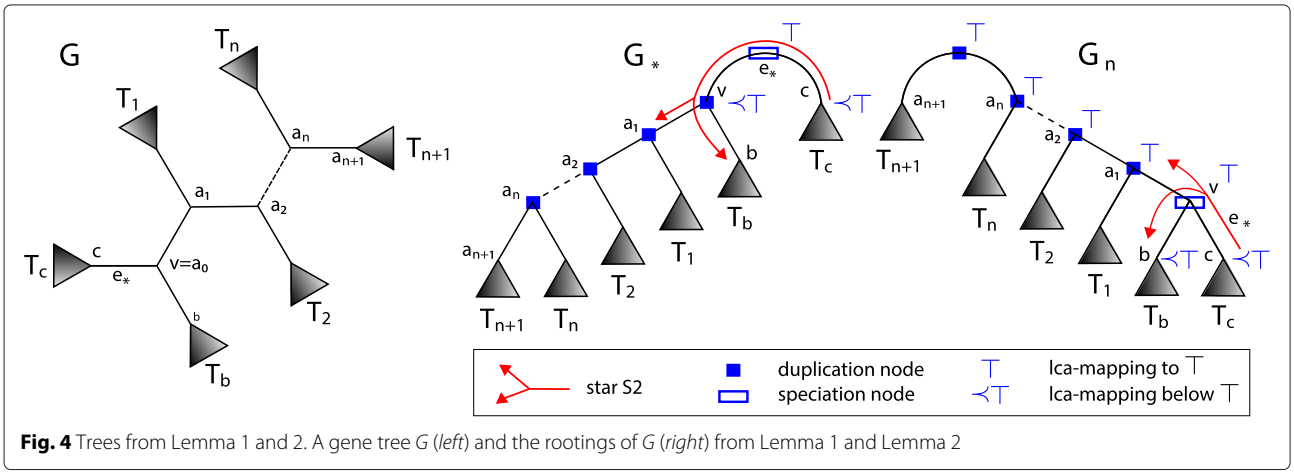


**Fig. 4** Trees from Lemma 1 and 2. A gene tree $G$ (*left*) and the rootings of $G$ (*right*) from Lemma 1 and Lemma 2

clusters we call *high clusters*. If the root of one of $T_i$'s is a duplication, then it can be merged with the high cluster in both rootings. Otherwise, if every root of these subtrees is a speciation then the high cluster is disjoint with clusters from $T_1, T_2, \ldots, T_{n+1}$. Moreover, if $b$ is a duplication then the high cluster contains $b$ in $G_*$. However, in $G_n$ the cluster of $b$ will be disjoint with the $\top$-cluster due to the speciation node $v$. Combining the above observations we obtain our formula. □

Lemma 1 and Lemma 2 complete the case of empty rootings. We proved that rooting on empty edge has the best EC.

#### Episodes in a gene tree with a double edge

We start with two technical lemmas on the properties of the plateaus.

**Lemma 3.** *If the DL-plateau consists of exactly one double edge then the D-plateau and the DL-plateau are equal.*

*Proof.* Let $\langle v, a\rangle$ be the DL-plateau edge (see Fig. 2). It follows from the property of star S3 that both $v$ and $a$ are mapped to $\top$ in the DL-minimal rooting and their children (if present) are mapped below $\top$. Hence, the root is a duplication, while $v$ and $a$ are speciation nodes. Now, it is easy to show that rooting on edge $\langle v, b\rangle$ (or $\langle v, c\rangle$) induces one additional gene duplication at $v$. We conclude that the only edge with the minimal duplication cost is $\langle v, a\rangle$. □

We write that a node $g$ from unrooted gene tree $G$ is a *super-duplication*, if $g$ is a duplication in every rooting of $G$. Please recall, that the plateau is a subtree of a gene tree, thus a leaf of the D-plateau may refer to an internal node of a gene tree. For example, in Fig. 3, the D-plateau of $G_1$ has four leaves: one is an internal node of $G_1$ and others, labeled $a, c, e$, are leaves of $G_1$.

**Lemma 4.** *If the DL-plateau has a double edge then*

- *every leaf of the D-plateau is a speciation in every rooting from the D-plateau,*
- *and every internal node of the D-plateau is a super-duplication.*

*Proof.* For the first part of the proof, let us assume that $v$ is a leaf of the D-plateau. By using the notation from Fig. 2, let $v$ be a center of a star such that $\langle v, a\rangle$ belongs to the D-plateau. Assume that $v$ is a duplication in every D-minimal rooting. Then, the D-minimal rooting $G_{\langle v,a\rangle}$ has one duplication in $v$. The edge $\langle v, b\rangle$ does not belong to D-plateau, therefore, the rooting $G_{\langle v,b\rangle}$ has at least one more duplication than $G_{\langle v,a\rangle}$. Hence, $G_{\langle v,b\rangle}$ has two duplications in $v$ and in the root. Moreover, the root of $G_{\langle v,a\rangle}$ is not a duplication. However, this is possible only when $T_a$ and $T_v$ are mapped below $\top$, thus the $\langle v, a\rangle$ is an empty edge, which is a contradiction with Theorem 2. This completes the first part of the proof.

Next, if the DL-plateau consists of exactly one double edge, then, by Lemma 3 the property holds trivially. Now, we assume that the DL-plateau has more than one edge. We show that every internal node $v$ of the DL-plateau is a super-duplication. From Theorem 2 we know that $v$ is incident to at least two double edges. Hence, in any rooting at least one of its children is mapped to $\top$. We conclude that $v$ is a duplication mapped to $\top$.

Let us consider a path $p = v_1, v_2, \ldots, v_n$ ($n > 1$) connecting an internal node $v_1$ from the DL-plateau with a leaf $v_n$ from the D-plateau. We show that the first $n - 1$ nodes on $p$ are duplications for every rooting placed on this path. It follows from the first part of this proof that $v_1$ is a super-duplication mapped to $\top$. Hence, when rooting at $\langle v_{n-1}, v_n\rangle$, we have $n$ gene duplications: for $v_1, v_2, \ldots, v_{n-1}$ and one for the root. All edges from $p$ are elements of the D-plateau, thus moving the root to other edges on $p$ will preserve the total number of gene duplications.

It should be clear that the same holds when choosing other root positions. We omit the details. □

In the next lemma we show that rootings at edges of the D-plateau induce the same EC cost.

**Lemma 5.** *If an unrooted gene tree G has no empty edge then for any D-minimal rooting of G denoted by $G_*$*

$$\mathsf{EC}(G_*) = \mathsf{EC}(T_1, T_2, \ldots, T_n) + 1,$$

*where $T_1, T_2, \ldots, T_n$ are the rooted subtrees of G obtained from G by removing all internal nodes of the D-plateau.*

*Proof.* It follows from Lemma 4 and its proof that all internal nodes of the D-plateau are present in the $\top$-cluster in the clustering with minimal number of clusters. This cluster is separated from other duplication clusters by speciation nodes located on the border of the D-plateau. Thus, the clusters induced by optimal solution to EC for $G_*$ are the clusters induced by optimal solution to EC of $T_1, T_2, \ldots, T_n$ plus the $\top$-cluster. □

#### Solutions

Now we present solutions to our unrooted episode clustering problem.

**Theorem 3** (Solution to single-UEC)**.** *For any gene tree G, an edge e is optimal for* single-UEC*, if either e is empty or e is in the D-plateau and G has a double edge.*

*Proof.* The first part of the proof follows immediately from Lemma 2 and the second part from Lemma 5. □

**Theorem 4.** *For a collection of unrooted gene trees $G_1, G_2, \ldots, G_n$, if every gene tree has a double edge then rooting every gene tree on an edge from the D-plateau yields the optimal solution for UEC.*

*Proof.* Assume that $n = 2$ and let $G'_1$ and $G'_2$ be two D-plateau rootings of $G_1$ and $G_2$, respectively. It should be clear that $\mathsf{EC}(G'_1, G'_2) = \mathsf{EC}(T)$, where $T = (G'_1, G'_2)$. Next, by Lemma 5, $\mathsf{EC}(T)$ is independent on the choice of rooting of $G_1$ and $G_2$, as long as the rootings are in the D-plateau. Therefore, we conclude that $\mathsf{EC}(T)$ is the solution to UEC Problem for $G_1$ and $G_2$. This observation can be easily generalized by induction to any $n$. □

Note that we cannot generalize the property stated in Theorem 4 to gene trees with empty edges. The example is shown in Fig. 3. Consider the dataset $\{G_1, G_2\}$. $G_1$ has five D-minimal rootings, while $G_2$ has exactly one. In $G_{2*}$ we have one $\top$-cluster, therefore $G_{2*}$ with $G_{1*}$, i.e., the empty edge rooting of $G_1$, have two duplication clusters. However, the best clusterings for $\{G_1, G_2\}$ having exactly one cluster are obtained for $G_{1,1}$, $G_{1,2}$ or $G_{1,3}$. On the other hand, the best clusterings can be also obtained for empty edge rootings, e.g. $\{G_{1,*}, G_{4,*}\}$ with cost 2 for the input $\{G_1, G_4\}$. From these examples, we see that the empty edges have different properties than double edges in the context of UEC, and we cannot generalize Theorem 4 to empty edges.

**Theorem 5** (Candidate rootings for UEC)**.** *For a collection of unrooted gene trees $\mathcal{G}$, the solution to UEC is induced by a rooting edge e of $G \in \mathcal{G}$ satisfying:*

- *(U1) if G has a double edge, then e is any D-minimal edge in G,*
- *(U2) if G has an empty edge, then e is an element of star S2.*

*Proof.* If some $G \in \mathcal{G}$ has a double edge then the property follows from Theorem 4 and Lemma 5. For gene trees with an empty edge $e_*$ we show that any D-minimal rooting of the edge that is not adjacent to $e_*$ can be equivalently replaced by a rooting adjacent to $e_*$. By using the notation from Fig. 2, let $T_a = (T_{a'}, T_{a''})$ such that $a'$ and $a''$ are the roots of $T_{a'}$ and $T_{a''}$, respectively. We show that the rooting $G_{\langle v,a\rangle}$ denoted by $G_a$ (see Fig. 5) has the same duplication episodes as the rooting $G_{a'}$ obtained for the edge $\langle a, a'\rangle$. In both rootings $v$ is a speciation, therefore the structure of clusters present in $T_b$ and $T_c$ is the same in both rootings. The edge $\langle v, a\rangle$ is $a$-incoming, thus the roots are duplications mapped to $\top$. From the fact that $\langle a, a'\rangle$ is in the D-plateau we have that $a$ is a duplication. Thus, every root and $a$ induce the $\top$-cluster. Finally, if $a''$ is a duplication node, then in both rootings it will be a member of the $\top$-cluster. We proved these two adjacent rootings have the same structure of clusters. Therefore, it is sufficient to choose the rooting $G_a$ instead of $G_{a'}$. This proof can be naturally extended by induction to any edge from the D-plateau. □

We conclude that for a gene tree $G$ we have at most 5 candidates for rootings. For instance, $G_4$ has two stars S2 in the D-plateau, therefore we have 5 candidate rootings: the empty edge rooting $G_{4,*}$ and the rootings of adjacent edges $G_{4,1}$, $G_{4,4}$, $G_{4,7}$ and $G_{4,10}$. Note that the clusters from $G_{4,1}$ are equivalent to clusters from $G_{4,2}$ and $G_{4,3}$. Similar property holds for other candidates.

Next, we show that the condition U2 can be improved.

**Lemma 6.** *Under the assumptions from Theorem 5. Let the set of clusters induced by the solution to UEC contains $\top$-cluster. Then, the condition (U2) from Theorem 5 can be refined as follows:*

- *(U2') if $e_*$ is the empty edge in G, then e is one among at most two non-adjacent edges such that $e = \langle x, y\rangle$ is adjacent to $e_*$ and $\mathsf{M}_*(x) = \mathsf{M}_*(y)$, where $\mathsf{M}_*$ is the lca-mapping for $G_*$.*

*Proof.* Let $G$ be a gene tree with an empty edge. Let $e_a$ be that edge from (U2'). By using the notation from Fig. 5, we compare the rooting $G_*$ and $G_{\langle v,a\rangle}$, denoted here by $G_a$. We have the following clusters in $G_*$: the cluster $C$ that contains $c$ (if $c$ is a duplication) and the cluster $X$ that contains $v$ (it follows from the proof of Lemma 2 that $v$ is a duplication node). Thus, $X = \{v\} \cup A \cup B$ where $A$ and $B$ denote duplications from $T_a$ and $T_b$, respectively. Note that $C$ has the same contribution to EC in both rootings, which follows from the property that valid mappings of $C$ are the same in both rootings. In $G_a$, $A$ is a subset of the $\top$-cluster whose contribution to EC is already incorporated (by the assumption). The node $v$ is a duplication in $G_*$. Hence, without loss of generality we assume that $M_*(a) = M_*(v)$, i.e., the rooting edge $\langle v, a\rangle$ satisfies the condition from (U2').

We have two cases depending on whether $B$ is empty. If $B$ is empty then $G_a$ has "better" composition of clusters than in $G_*$, i.e., one cluster less then in $G_*$ and other clusters has the same valid mappings. Otherwise, both rootings are equivalent if $M_*(b) = M_*(v)$ ($B$ in $G_a$ has the same valid mappings as $X$ in $G_*$), or again $G_a$ has a better structure of clusters than $G_*$ if $M_*(b) \prec M_*(v)$ (valid mappings of $X$ in $G_*$ are included in valid mappings of $B$ in $G_a$). Similarly, we show that $G_a$ is also better than $G_{\langle v,b\rangle}$ (see also rootings of $G_4$ in Fig. 3).

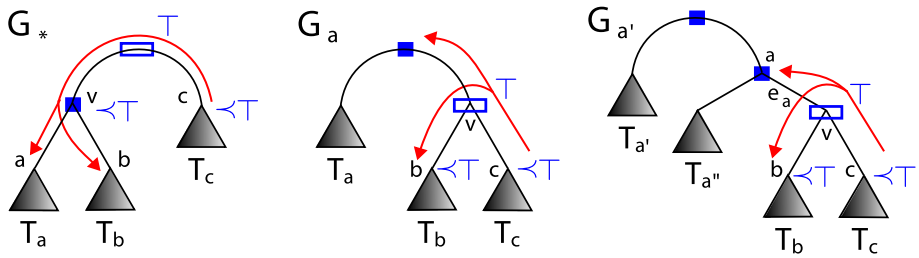


**Fig. 5** Trees from Theorem 5 and Lemma 6. The rootings of $G$ from Theorem 5 and Lemma 6. We use the notation $G_a$ instead of $G_{\langle v,a\rangle}$. See Fig. 4 for a legend of the symbols used

We proved that among three rootings from the star S2 we can choose one candidate. The second edge is obtained from the second star S2 (sharing the empty edge) if it is present in the gene tree (see Theorem 2). □

From the last lemma we have at most two candidates for any gene tree from the input collection. For example, the candidate rooting $G_{4,1}$ has more flexible valid mappings than $G_{4,4}$, e.g. the duplication cluster of $((c,b),a)$ in $G_{4,1}$ has larger range of possible mappings than the duplication cluster of $((d,b),a)$ in $G_{4,4}$, while the remaining two clusters have the same locations in the species tree. Hence, for the dataset $\{G_3, G_4\}$, if the $\top$-cluster is present in solution to UEC, we have two candidates $G_{4,1}$ and $G_{4,7}$ (which is more flexible than $G_{4,10}$). Note, that the clustering costs 3 is obtained by rootings $G_{3,*}$ and $G_{4,1}$ (or $G_{4,2}$, $G_{4,3}$).

### Algorithms

Algorithm 1 presents the solution to UEC problem. The correctness of this algorithm follows from Theorem 5 and Lemma 6. Algorithm 1 has two phases. In the first phase for every gene tree a set of candidate rootings is prepared with respect to the conditions (U1) and (U2'). To find optimal rootings we use a linear time algorithm (procedure FindOptEdge) based on greedy descent method that search a double or an empty edge in a gene tree [32]. Based on condition U2', we divide possible solutions into two categories depending on the presence of $\top$-cluster in an optimal clustering. If the $\top$-cluster is not present then every gene tree has an empty edge (in line 10). Otherwise, we check every possible variant of rooting candidates. Note that from Lemma 6, a gene tree has two candidates if and only if the gene tree has two stars S2 that are included in the D-plateau. Thus, the overall time complexity depends on the presence of such trees in the input. From this observation we conclude the following result.

**Theorem 6.** *The time complexity of Algorithm 1 is* $O(2^k(\sum_i |G_i| + |S|))$, *where* $k$ *is the number of input gene trees having two stars S2 that are included in the D-plateau.*

**Algorithm 1** Unrooted Episode Clustering 1

1: **Input** A binary species tree $S$, a collection of unrooted gene trees $G_1, G_2, \ldots, G_n$.
2: **Output** Minimal $\mathsf{EC}(T_1, T_2, \ldots, T_n, S)$ in the set of all rootings $T_i$ of $G_i$ such that $T_i$ is a rooting obtained from $G_i$ by placing the root on the edge from the D-plateau.
3: For every $i$ compute the set of candidate rooting edges $R_i$:
4: $e_*$:=FindOptEdge($G_i$)
5: **If** $e_*$ is double: $R_i := \{e_*\}$
6: **If** $e_*$ is empty **then for** $x \in e_*$ such that $x$ is not a leaf.
7: **Let** $c$ be a child of $x$ in $G_*$ such that $\langle x, c\rangle$ is D-minimal and not adjacent to any edge from $R_i$ and $M_*(c) = M_*(x)$
8: $R_i := R_i \cup \{\langle x, c\rangle\}$.
9: **If** every $G_i$ has an empty edge then $\alpha := \mathsf{EC}(T_1, T_2, \ldots, T_n)$, where $T_i$ is the empty edge rooting of $G_i$ **else** $\alpha := +\infty$.
10: $\beta = \min_{e_i \in R_i} \mathsf{EC}(G_{e_1}, G_{e_2}, \ldots, G_{e_n})$.
11: **Return** $\min\{\alpha, \beta\}$.
12: **Function FindOptEdge(*G*)**
13: **Let** $m_{x,y} = M_{G_{\langle x,y\rangle}}(x)$ *// can be computed in* $O(|G|)$ *steps [32].*
14: **Let** $v$ be a node from $V_G$ and let $\top$ by the lca-mapping of some rooting of $G$.
15: **While** there exists a node $w$ adjacent with $v$ such that $m_{w,v} = \top \neq m_{v,w}$
16: **do:** set $v := w$ (star S1).
17: **Return** $\langle v, w\rangle$ such that $\langle v, w\rangle$ an empty or double edge i.e., $m_{v,w} = \top = m_{w,v}$ or $m_{v,w} \neq \top \neq m_{w,v}$.

Thus, from theoretical point of view UEC is fixed parameter tractable. Later we show that $k$ usually represents a small fraction (up to 5 %) of the whole input. For the cases when $2^k$ is still too large for efficient computation, we propose Algorithm 2, in which we first solve the instance of UEC for the collection of gene trees that have a unique candidate. Clearly, if there are rootings

**Algorithm 2** Unrooted Episode Clustering 2

1: **Input/output** The same as in Algorithm 1.
2: **Let** $\delta = \mathsf{EC}(G'_1, G'_2, \ldots, G'_{n'})$ computed by Algorithm 1, such that $\{G'_1, G'_2, \ldots, G'_{n'}\}$ is the set of all input gene trees having a unique candidate rooting edge (i.e., $|R_i| = 1$).
3: **If** $\delta = \alpha$ **Return** $\alpha$, where $\alpha$ is from the 9th line of Algorithm 1 computed for the whole input.
4: **For** every $e_1 \in R_1, e_2 \in R_2, \ldots, e_n \in R_n$ (i.e. candidate rootings of the whole input)
5: If $\mathsf{EC}(G_{e_1}, G_{e_2}, \ldots, G_{e_n}) = \delta$ **then Return** $\delta$.
6: **Return** the minimal EC value computed in lines 3 and 6.

of the whole input that have the same cost, then this cost is optimal. The overall complexity of Algorithm 2 is the same as Algorithm 1, however, for large datasets this strategy appeared to be successful after checking just one additional candidate set (in lines 2–4).

### Experiments

We performed several computational experiments on three empirical datasets.

*Guigó dataset* consists of 53 rooted gene trees from 16 Eukaryotes from [10]. This dataset was evaluated with 71 species trees from [35], known to have the total minimal duplication cost. *Génolevures* is a dataset of 4144 gene trees [33] from nine yeast genomes [36] and two species trees: one from [37] and the second one having the lowest duplication-loss cost computed by Fasturec [38]. The third dataset *TreeFam*, spanning 25 mostly animal species, consists of 1274 curated gene family trees from TreeFam v7.0 [39]. The species tree for TreeFam is based on NCBI taxonomy.

We implemented our algorithms and the algorithms for the rooted variant of EC Problem (based on [29]). In our experiments the rooting candidates were used to compare the results for UEC with the model of mappings (for rooted gene trees) proposed in [28].

We performed two series of 74 computational experiments, one for our model and one with the model described in [28]. The total running time of our program was about 7 minutes on a standard PC workstation. For every dataset we were able to find solutions to UEC by testing at most two rooted instances of input gene trees (see Algorithm 2). The summary of experiments is depicted in Table 1.

For the Guigó dataset we found four duplication clusters, while for the rooted model from [28] we located five clusters. The difference can be explained by the properties of our model that is more flexible: the input trees are unrooted and the model of valid mappings is more generic. Observe that this dataset has unique rooting candidates ($k = 0$).

Génolevures is the most complex dataset due to its size and potentially large parameter $k$. Despite these properties, Algorithm 2 located 17 clusters for the filtered input with all unique rooting candidates. In other words, in this filtered dataset a duplication cluster is present in every node of the species tree. Obviously, the whole input dataset has the same property. The same holds for the model from [28].

In TreeFam we located 45 clusters for the filtered dataset with unique rooting candidates. Then, Algorithm 2 found the solution having the same cost for the whole dataset (see Fig. 6). The same result was obtained for the model from [28] (see Table 1).

## Conclusions

In this article we presented the first solution to the open problem of the duplication episode clustering for case when the input collection is composed of unrooted gene trees. By using theoretical properties of the unrooted reconciliation we proved that the problem has nice mathematical and computational properties. From practical point of view, we were able to provide efficient algorithms and tools that were successfully applied to locate duplication clusters in real datasets.

From the computational point of view the complexity of our algorithms depends on the parameter $k$, i.e., in the worst case EC Problem has to be solved $2^k$ times in order to find a solution to UEC. Even if $k$ usually represents a small fraction of the whole input it can be still large, e.g. $k > 100$ for the yeast dataset, which may prohibit computation of all possible variants. Here we proposed a solution, that is based on the observation that the clustering induced from the input gene trees having unique candidates (that is, without $k$ gene trees with non-unique variants), usually represents an optimal solution

**Table 1** Experimental results

| Set | # Species trees | # Leaves | # Gene trees | $k$ | Our model | | Model [28] | |
|---|---|---|---|---|---|---|---|---|
| | | | | | EC | % Locations | EC | % Locations |
| Guigó | 71 | 16 | 53 | 0 | 4 | 12,9 % | 5 | 16,1 % |
| Génolevures | 1 [37] | 9 | 4144 | 55 | 17 | 100 % | 17 | 100 % |
| | 1 [38] | 9 | 4144 | 156 | 17 | 100 % | 17 | 100 % |
| TreeFam | 1 | 28 | 1274 | 67 | 45 | 81,8 % | 45 | 81,8 % |

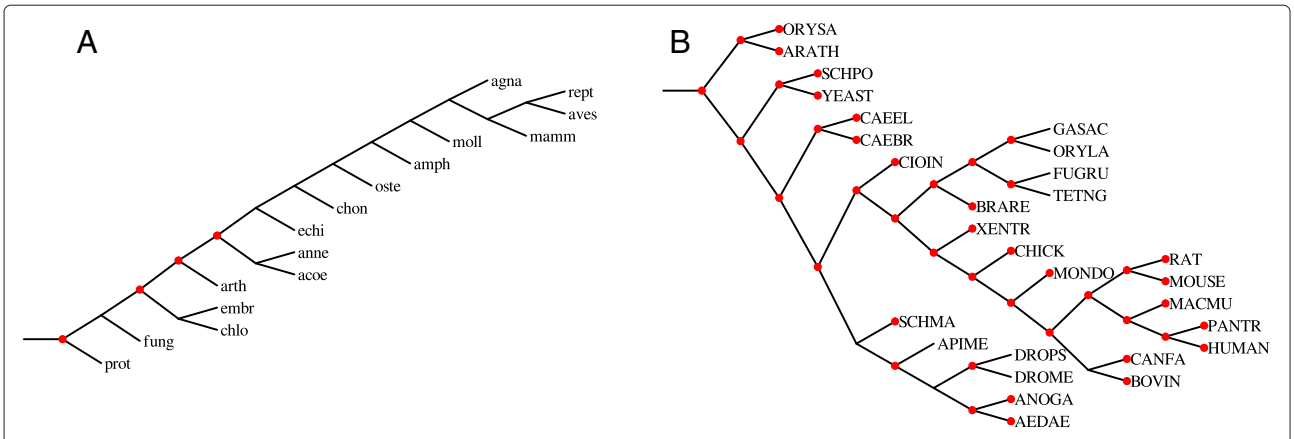


**Fig. 6** Duplication clusters in empirical datasets. Duplication clusters (*marked by red circles*) inferred from experiments. **a** Guigó species tree (chosen from 71 species trees from [35] as the most biologically reasonable [40]). **b** TreeFam species tree based on NCBI taxonomy

for the whole input. Thus, the strategy that we applied in Algorithm 2, i.e., first cluster easy part and then try to incorportate the hard one by using already identified clusters, appeared to be successful even for potentially complex datasets.

Our computational experiments show that the duplication clusters are usually located in large parts of the species tree especially when the input dataset consists of thousands of gene trees. To provide more detailed information on the duplication clusters, we plan to study minimal episode problem (ME) which is a natural extension of the episode clustering problem. In the future we plan to extend the episode clustering problem by using other types of valid mappings.

Our software for solving unrooted episode clustering problem is publicly available at http://www.mimuw.edu.pl/jpaszek/uec.php.

#### Abbreviations

D: gene duplication; DL: Gene duplication and loss; EC: episode clustering for rooted gene trees; lca: least common ancestor; UEC: episode clustering for unrooted gene trees.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

JP and PG contributed equally to the writing of the paper. Both authors read and approved the final manuscript. JP implemented algorithms and performed all computational experiments.

#### Acknowledgements

We would like to thank the three reviewers for their detailed comments that allowed us to improve our paper. JP and PG were supported by the grant of NCN #2011/01/B/ST6/02777. JP was supported by the DSM funding for young researchers of the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw.

#### Declarations

The publication costs for this article were funded by the Polish Ministry of Science and Higher Education funding for Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw.

This article has been published as part of BMC Genomics Volume 17 Supplement 1, 2016: Selected articles from the Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/17/S1.

Published: 11 January 2016

#### References

1. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 2004;428:617–24.
2. Guyot R, Keller B. Ancestral genome duplication in rice. Genome. 2004;47(3):610–4.
3. Vision TJ, Brown DG, Tanksley SD. The origins of genomic duplications in Arabidopsis. Science. 2000;290(5499):2114–7.
4. Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, et al. Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science. 2014;343(6166):88–91.
5. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444(7116):171–8.
6. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, et al. Widespread genome duplications throughout the history of flowering plants. Genome Res. 2006;16(6):738–49.
7. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10(10):725–32.
8. Page RDM. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol. 1994;43(1):58–77.
9. Mirkin B, Muchnik I, Smith TF. A biologically consistent model for comparing molecular phylogenies. J Comput Biol. 1995;2(4):493–507.
10. Guigó R, Muchnik IB, Smith TF. Reconstruction of ancient molecular phylogeny. Mol Phylogenet Evol. 1996;6(2):189–213.
11. Page RDM. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol Phylogenet Evol. 2000;14:89–106.
12. Arvestad L, Berglund AC, Lagergren J, Sennblad B. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics. 2003;19(Suppl 1):i7–15.
13. Bonizzoni P, Della Vedova G, Dondi R. Reconciling a gene tree to a species tree under the duplication cost model. Theor Comput Sci. 2005;347:36–53.
14. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by Cladograms Constructed from Globin sequences. Syst Zool. 1979;28(2):132–63.
15. Górecki P, Tiuryn J. DLS-trees: a model of evolutionary scenarios. Theor Comput Sci. 2006;359:378–99.
16. Arvestad L, Lagergren J, Sennblad B. The gene evolution model and computing its associated probabilities. J ACM. 2009;56(2):1–44.

17. Doyon JP, Chauve C, Hamel S. Space of gene/species tree reconciliations and parsimonious models. J Comput Biol. 2009;16:1399–1418.
18. Durand D, Halldórsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol. 2006;13(2):320–35.
19. Hallett MT, Lagergren J. Efficient Algorithms for Lateral Gene Transfer Problems. In: Proceedings of the Fifth Annual International Conference on Computational Biology. RECOMB '01. New York, NY, USA: ACM; 2001. p. 149–156.
20. In: Bourque G, El-Mabrouk N, editors. Comparative Genomics, RECOMB 2006 International Workshop, RCG 2006, Montreal, Canada, September 24–26, 2006, Proceedings. vol. 4205 of Lect Notes Comput Sc. Berlin, Germany: Springer; 2006.
21. Ma B, Li M, Zhang L. From gene trees to species trees. SIAM J Comput. 2000;30(3):729–52.
22. Sjostrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J. A Bayesian method for analyzing lateral gene transfer. Syst Biol. 2014;63(3):409–20.
23. Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics. 2012;28(18):i409–15.
24. Zhang L. From gene trees to species trees II: species tree inference by minimizing deep coalescence events. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(6):1685–91.
25. Page RDM, Cotton JA. Vertebrate phylogenomics: reconciled trees and gene duplications. Pac Symp Biocomput. 2002;536–47.
26. Fellows M, Hallet M, Stege U. On the Multiple Gene Duplication Problem. In: 9th International Symposium on Algorithms and Computation (ISAAC'98), Lecture Notes in Computer Science 1533. Taejon, Korea: Springer Berlin Heidelberg; 1998. p. 347–356.
27. Burleighm JG, Bansal MS, Wehe A, Eulenstein O. Locating Multiple Gene Duplications through Reconciled Trees In: Vingron M, Wong L, editors. RECOMB. vol. 4955 of Lect Notes Comput Sc. Berlin, Germany: Springer; 2008. p. 273–284.
28. Bansal MS, Eulenstein O. The multiple gene duplication problem revisited. Bioinformatics. 2008;24(13):i132–8.
29. Luo CW, Chen MC, Chen YC, Yang RWL, Liu HF, Chao KM. Linear-time algorithms for the multiple gene duplication problems. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(1):260–5.
30. Holland BR, Penny D, Hendy MD. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. Syst Biol. 2003;52:229–38.
31. Huelsenbeck JP, Bollback JP, Levine AM. Inferring the Root of a Phylogenetic Tree. Syst Biol. 2002;51(1):32–43.
32. Górecki P, Tiuryn J. Inferring phylogeny from whole genomes. Bioinformatics. 2007;23(2):e116–22.
33. Górecki P, Eulenstein O. Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. BMC Bioinformatics. 2012;13(Suppl 10):S14.
34. Górecki P, Eulenstein O, Tiuryn J. Unrooted tree reconciliation: a unified approach. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(2):522–36.
35. Chang W, Górecki P, Eulenstein O. Exact solutions for species Tree Inference from discordant gene trees. J Bioinform Comput Bio. 2013;11(5): 1342005.
36. Sherman DJ, Martin T, Nikolski M, Cayla C, Souciet JL, Durrens P. Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. Nucleic Acids Res. 2009;37(suppl 1):D550–4.
37. Dujon B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet. 2006;22(7):375–87.
38. Górecki P, Eulenstein O. GTP supertrees from unrooted gene trees: linear time algorithms for NNI based local searches. Lect Notes Comput Sc. 2012;7292:83–105.
39. Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, et al. TreeFam 2008 Update. Nucleic Acids Res. 2008;36:D735–40.
40. Page RDM, Charleston MA. Reconciled trees and incongruent gene and species trees In: Mirkin B, McMorris FR, Roberts FS, Rzhetsky A, editors. Mathematical Hierarchies in Biology, American Mathematical Society, Providence, Rhode Island; 1997. p. 57–70.