

Evolution of orthologous tandemly arrayed gene clusters

Olivier Tremblay Savard^{*1}, Denis Bertrand^{*2} and Nadia El-Mabrouk^{*1}

¹Department of Computer Science (DIRO), University of Montreal, Montreal, Quebec, Canada ²Computational and Mathematical Biology, Genome Institute of Singapore, Singapore

Email: Olivier Tremblay Savard* - olivier.tremblay-savard@umontreal.ca; Denis Bertrand* - bertrandd@gis.a-star.edu.sg; Nadia El-Mabrouk* - mabrouk@iro.umontreal.ca;

* Corresponding author

Abstract

Background: Tandemly Arrayed Gene (TAG) clusters are groups of paralogous genes that are found adjacent on a chromosome. TAGs represent an important repertoire of genes in eukaryotes. In addition to tandem duplication events, TAG clusters are affected during their evolution by other mechanisms, such as inversion and deletion events, that affect the order and orientation of genes. The DILTAG algorithm developed in [1] allows to infer a set of optimal evolutionary histories explaining the evolution of a single TAG cluster, from an ancestral single gene, through tandem duplication (simple or multiple, direct or inverted), deletion and inversion events.

Results: We present a general methodology, which is an extension of DILTAG, for the study of the evolutionary history of a set of orthologous TAG clusters in multiple species. In addition to the speciation events reflected by the phylogenetic tree of the considered species, the evolutionary events that are taken into account are simple or multiple tandem duplications, direct or inverted, simple or multiple deletions, and inversions.

Conclusions: Our results obtained on simulated data sets showed a good performance in inferring the total number and size distribution of duplication events. A limitation of the algorithm is however in dealing with multiple gene deletions, as the algorithm is highly exponential in this case, and becomes quickly intractable.

1 Background

Gene duplication is a fundamental process in the evolution of species [2], especially in eukaryotes [3–8], where it is believed to play a leading role for the creation of novel gene functions. Several mechanisms are at the origin of gene duplications, among them tandem repeat through slippage during recombination. As this phenomenon is favoured by the presence of repetitive sequences, a single duplication can induce a chain reaction leading to further duplications, eventually creating large *Tandemly Arrayed Gene (TAG) clusters*: groups of paralogous genes that are adjacent on a chromosome. TAGs account for about one-third of the duplicated genes in eukaryotes [9]. In human, they represent about 15% of all genes [10]. In *Arabidopsis*, 17% of the total predicted genes are members of TAG clusters [11], and in maize, about 35% of the genes were predicted to belong to TAG clusters [12].

Deciphering the evolutionary history of a TAG cluster is important to provide new insights into the mechanisms of gene amplification, and to answer several questions regarding the nature and size of duplication and other evolutionary events that have shaped TAG clusters. In most biology-oriented studies, a gene tree is obtained by applying a classical phylogenetic method to an alignment of the amino acid sequences corresponding to the collected gene sequences, and a duplication scenario is proposed for the gene family, based on a careful analysis of this gene tree (see for example [9] for the study of the 22-kDA prolamin gene amplification in grass genomes). Although such manual analysis may be useful to propose amplification scenarios for families of limited size and simple organization, it is usually unable to infer more general evolutionary scenarios for TAG clusters affected, in addition to duplications, by other events such as segmental deletion, that may lead to gene loss, and rearrangements (such as inversions or inverted duplications), that may affect gene order and transcriptional orientations.

The *tandem-duplication model of evolution*, first introduced by Fitch in 1977 [13], assumes that, from a single ancestral gene at a given position in the chromosome, the locus grows through a series of consecutive duplications placing the newly created copy next to the original one. Such tandem duplications may be *simple* (duplication of a single gene) or *multiple* (simultaneous duplication of neighbouring genes). Based on this idea, a number of theoretical studies have considered the problem of reconstructing the tandem-duplication history of a TAG cluster [14–17]. However, due to rearrangements and losses, it is often impossible to reconstruct a duplication history for a TAG cluster [18], even from well-supported gene trees. In [19], we considered a generalization of the tandem-duplication model allowing for inversions. The model was then extended in [20] to the study of orthologous TAG clusters in different species. A similar work, considering more operations (translocations, fusions, fissions, duplications in tandem or not), but requiring

more preliminary information (gene and species trees with branch length) has also been done [21]. Various other heuristic and probabilistic methods have been developed for reconstructing a hypothetical ancestral sequence and a most parsimonious set of duplications (in tandem or not) and other evolutionary events leading to the observed gene cluster [22–25]. They are based on a preprocessing of a self-alignment dot-plot of a cluster, or the dot-plot of a pairwise-alignment of two clusters. Although these methods are useful to infer recent evolutionary events, they are less appropriate for longer timescales as alignment of the nonfunctional regions becomes impossible due to mutations continuously affecting each duplicated segment. In both of our previous cited methods [19, 20], only simple duplications were considered. This assumption, while allowing for exact algorithmic solutions, is an important limitation to its applicability (see for example [26]). For this reason, we have developed a more general heuristic, the DILTAG algorithm [1], allowing to infer a set of optimal evolutionary histories for a gene cluster in a single species, according to a general cost model involving variable length duplications, in tandem or inverted, deletions and inversions. Experiments on simulated data showed that the most recent evolutionary events can be inferred accurately when the exact gene trees are used. Despite the uncertainty associated with the deeper parts of the reconstructed histories, they can be used to infer the duplication size distribution with some precision. DILTAG has been used recently in [27] to infer an evolutionary scenario for the Maltase gene clusters in *Drosophila*.

A clear limitation of DILTAG is the fact that it is applicable only to a single cluster. The benefit of an extension to multiple species is obvious, as comparative genomics is clearly a more appropriate approach to infer loss and inversion events. In particular, considering an outgroup may help in choosing among many possible optimal evolutionary scenarios for a gene cluster.

In this paper we present an extension of DILTAG to the study of a set of orthologous TAG clusters in multiple species. In other words, in addition to multiple duplications (in tandem or inverted), deletions and inversions events, the speciation events reflected by a given phylogenetic tree for the set of species are also taken into account. We develop MultiDILTAG, a heuristic algorithm that is shown on simulated data sets to be very accurate in inferring the total number and size distribution of duplication events.

2 Methods

2.1 Data

Preliminary to all the developments in this paper is the identification of m orthologous TAG clusters in m genomes of interest. In other words, given a gene family F of interest, a tandemly arrayed sequence (called

TAG cluster) of paralogous genes from F has already been identified in each genome, and such m TAG clusters have already been pointed out as orthologs. For example, gene orders and clusters orthology for the olfactory receptor gene family can be found in the HORDE database [28] for human and several other mammalian species.

We denote by $\mathcal{O} = \{O_1, O_2, \dots, O_m\}$ the set of m TAG clusters, i.e. for $1 \leq i \leq m$, O_i is the signed order of the family members in genome i . The sign $(+/-)$ of a gene represents its transcriptional orientation. In addition to the observed gene orders, we also assume that a gene tree is available for the *TAG family*, i.e. the set of genes contained in the m TAG clusters. A *gene tree* T for a TAG family is a rooted binary tree with labelled leaves, where each label represents an unsigned gene copy. A leaf labelled by a gene copy in genome i is said to *belong to genome* i . For conciseness, we make no difference between a leaf and its label. The pair (T, \mathcal{O}) is called the *ordered gene tree* for the gene family. Finally, we assume that the species tree, reflecting the speciation history of the m considered genomes, is also available. See Figure 1 for an example.

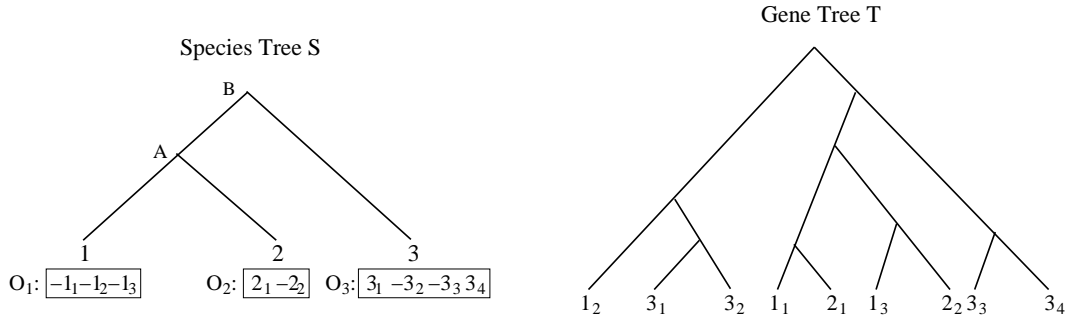


Figure 1: Left: The species tree for three genomes 1, 2, 3. Three orthologous TAG clusters $\{O_1, O_2, O_3\}$ are identified in the three genomes. The notation j_i denotes the i th gene on genome j . Right: A gene tree for the gene family.

2.2 The Evolutionary Model

Our evolutionary model is an extension of the one introduced by Fitch [13] for TAGs, which considers only tandem duplications resulting from unequal crossing-over during meiosis. However, TAGs are shaped during their evolution by other events affecting the gene order, orientation and content of the clusters. For example, Shoja and Zhang [10] have observed that more than 25% of all neighbouring pairs of TAGs in human, mouse and rat have non-parallel orientations. The Fitch model of evolution does not apply to such data. Our model extends the Fitch model of evolution by considering deletion events affecting gene content, as well as inversion and inverted duplication events affecting gene orientation.

Below is a formal definition of the evolutionary model considered in this paper. In this definition, a *cherry* of T is a pair of leaves (l, r) separated by a single vertex, called its *root*.

Definition 1: An *evolutionary history* for (T, \mathcal{O}) is a sequence of ordered gene trees

$((T^1, \mathcal{O}^1), (T^2, \mathcal{O}^2), \dots, (T^h, \mathcal{O}^h) = (T, \mathcal{O}))$, such that for each $1 \leq k \leq h$, $\mathcal{O}^k = \{O_1^k, \dots, O_i^k, \dots, O_{n_k}^k\}$ is a set of n_k gene orders corresponding to orthologous TAG clusters on n_k genomes, where :

1. T^1 is a tree consisting of a single leaf u , and $\mathcal{O}^1 = \{O_1^1\} = \{(\pm u)\}$.
2. For $1 \leq k < h$, there is a unique genome i such that $(T^{k+1}, \mathcal{O}^{k+1})$ can be obtained from (T^k, \mathcal{O}^k) by applying one of the following evolutionary events on (T^k, \mathcal{O}_i^k) :
 - (a) **Duplication:** A sub-sequence $(u_p, u_{p+1}, \dots, u_q)$ of \mathcal{O}_i^k is replaced by a sequence of new elements $(l_p, l_{p+1}, \dots, l_q, r_p, r_{p+1}, \dots, r_q)$, where, for each $p \leq x \leq q$, l_x and r_x have the same sign as u_x . Moreover, each leaf u_x in T^k is replaced by the cherry (l_x, r_x) .
 - (b) **Inverted-duplication:** A sub-sequence $(u_p, u_{p+1}, \dots, u_q)$ of \mathcal{O}_i^k is replaced by $(-(l_q), -(l_{q-1}), \dots, -(l_p), r_p, r_{p+1}, \dots, r_q)$ or $(l_p, l_{p+1}, \dots, l_q, -(r_q), -(r_{q-1}), \dots, -(r_i))$, where, for each $p \leq x \leq q$, l_x and r_x have the same sign as u_x . Moreover, each leaf u_x of T_k is replaced by the cherry (l_x, r_x) .
 - (c) **Inversion:** A sub-sequence $(u_p, u_{p+1}, \dots, u_q)$ of \mathcal{O}_i^k is replaced by $(-(u_q), -(u_{q-1}), \dots, -(u_p))$ and T^k remains unchanged.
 - (d) **Deletion:** A sub-sequence $(u_p, u_{p+1}, \dots, u_q)$ of \mathcal{O}_i^k is deleted, and the corresponding leaves (genes) are removed from T^k (each removed gene corresponds to a gene loss).
 - (e) **Speciation:** The complete order $\mathcal{O}_i^k = (u_1, \dots, u_t)$ is replaced by $\{(l_1, \dots, l_t), (r_1, \dots, r_t)\}$, where, for each $p \leq x \leq q$, l_x and r_x have the same sign as u_x . Moreover, each leaf u_x belonging to genome i is replaced by the cherry (l_x, r_x) .

Any evolutionary history \mathcal{H} for (T, \mathcal{O}) induces a unique species tree S obtained from the speciation events of \mathcal{H} . We say that \mathcal{H} is *consistent with* S .

Finally, a *simple-event*, will refer to an event acting on a single gene. For example, a simple-deletion will refer to the deletion of a single gene. A simple-deletion event is also referred to as a *loss event*. Moreover, a *general-duplication* will refer to a duplication that does not necessarily place the duplicated genes next to the original copies (not necessarily in tandem). An example of an evolutionary history is given in Figure 2.

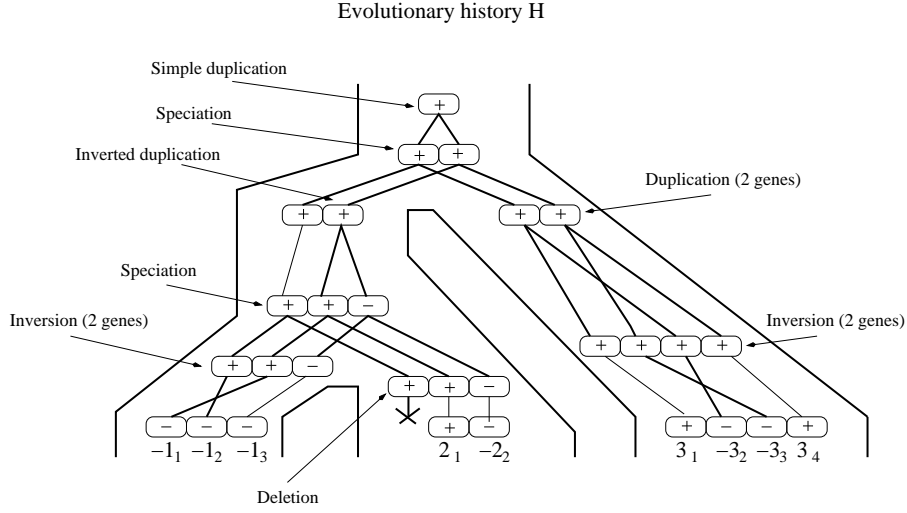


Figure 2: An evolutionary history leading to the gene tree of Figure 1. This history is consistent with the species tree S of Figure 1.

We are now ready to formulate our optimization problem:

MINIMUM-EVOLUTION PROBLEM:

Input: An ordered gene tree (T, \mathcal{O}) and a species tree S .

Output: A most parsimonious evolutionary history \mathcal{H} for (T, \mathcal{O}) consistent with S .

The “most parsimonious” constraint given above can be most naturally expressed in terms of number of events. Alternatively, a cost can be associated to each event depending on the type and size of the event (i.e. number of genes affected by this event), and the “most parsimonious” history would be the history of minimum cost, where the cost of a history is simply the sum of costs associated with its events. This later approach is the one considered in [1].

2.3 The DILTAG method

The DILTAG algorithm [1] allows to infer a most parsimonious history of duplications, inverted-duplications, inversions and deletions (i.e. all events introduced in Definition 1 except speciation) acting on a single gene to produce a given TAG cluster represented by a given ordered gene tree (T, \mathcal{O}) . DILTAG proceeds by exploring a “history graph” (search space), where vertices correspond to ordered gene trees and edges correspond to evolutionary events. More precisely, an edge from (T^i, \mathcal{O}^i) to (T^j, \mathcal{O}^j) is defined if and only if (T^i, \mathcal{O}^i) can be transformed into (T^j, \mathcal{O}^j) through one event, and each edge is weighted by the cost of its corresponding event. This graph is actually simplified into a finite graph,

without loss of information, by considering deletions only in combination with duplication events. The history graph is constructed backwards, i.e. starting at vertex (T, O) , and constructing edges in their opposite direction (backward-edges) by exploring the neighbourhood of each vertex.

It is shown in [1] that, given a vertex representing an ordered gene tree (T, O) , its *duplication* and *inverted-duplication* neighbourhoods are both linear (in the size of T) in space, where as its *inversion*, *duplication-with-deletion* and *inverted-duplication-with-deletion* neighbourhoods are all quadratic in space. However the size of the whole search-space is clearly exponential, which makes an exhaustive search through the whole graph impossible for gene trees of reasonable size. A greedy heuristic is therefore developed that only conserves, in a queue, the most promising partial evolutionary histories obtained after exploring a given depth of the history-graph.

The input of DILTAG is an ordered gene tree (T, O) with n leaves, and the output is a set of shortest backward-paths in the history graph from (T, O) to a tree containing a single vertex. For the purpose of our new Multi-DILTAG algorithm, it is easy to modify DILTAG in order to reach an ancestral genome with g genes, for any $1 \leq g \leq n$: simply stop the procedure as soon as we attain the right number of genes. Notice that the attained ancestor is ordered, i.e. defined by an ordered sequence of g genes. It can be seen as an ordered tree (T', O') with T' being reduced to a set of g vertices and no edges. An ordered tree with no edges will be confounded with a gene order.

In Section 2.5, the input and output of DILTAG will be as follows:

Input: An ordered gene tree (T, O) and a number g of ancestral genes;

Output: The cost of a shortest backward-path from (T, O) to an ancestral genome with g genes, together with the *solution graph* composed by the actual set of shortest paths, and the *solution set* of ancestral gene orders attained.

Finally, we need the following definition for the subsequent developments: given two vertices x and y of the oriented history graph, if there is an edge oriented from x to y (there is an evolutionary event transforming x into y), then we say that y is a *predecessor* of x .

2.4 A two step method for multiple species

Back to our evolutionary model on multiple species, we aim to find a most parsimonious evolutionary history for (T, \mathcal{O}) that is consistent with S . This problem has been considered in [20], but in the more

restricted case of *simple-duplications*, and *no inverted-duplications*. A two step methodology has been considered:

1. **Reconciliation Step:** Ignoring gene orders, infer a history of *simple-general-duplication*, *simple-deletion* and *speciation* for T consistent with S , by using a reconciliation approach [29]. Conceptually, a *reconciliation* R between a gene tree T and a species tree S is a tree accounting for the evolutionary history of the species and all genes of the gene family, including lost and missing gene copies, by simple general duplication, speciation and loss. R can be “embedded” into S , reflecting the duplication and deletion events leading to the observed tree T . Such embedding allows to infer the number of genes at the speciation nodes of S , as well as the evolutionary relationships between ancestral gene copies. A reconciliation between the gene tree T and the species tree S of Figure 1 is given in Figure 3. Notice that this reconciliation does not lead to the observed gene order.
2. **Minimization Step:** Reinserting the gene order and sign information \mathcal{O} on leaves of S , infer the order and sign of genes at internal nodes of S allowing to minimize the total number of events involved in a history of (T, \mathcal{O}) .

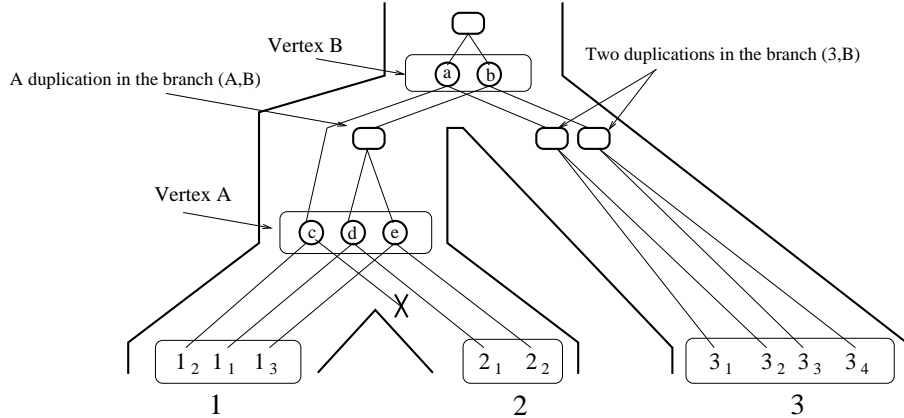


Figure 3: A reconciliation R between the gene tree and the species tree of Figure 1. Ovals are speciation vertices and squares are duplication vertices. The crossed line indicates a gene loss. The genome sets of internal nodes are $G(A) = \{c, d, e\}$ and $G(B) = \{a, b\}$. As for the pre-speciation sets, we have $PG(A) = G(B) = \{a, b\}$, $PG(1) = G(A) = \{c, d, e\}$, $PG(3) = G(B) = \{a, b\}$, but $PG(2) = \{d, e\}$, which is different from $G(A)$ due to the loss of gene c .

We use the same two-steps methodology here. As for the first step, any existing reconciliation method can be used. In particular, the so called Lowest Common Ancestor (LCA) mapping between a gene tree and a species tree, formulated in [30,31] and widely used [31–40], defines a reconciliation tree R that minimizes

both the simple-general-duplication and simple-deletion events.

In the following developments, we will consider the “embedded” representation of a reconciliation tree R into the species tree S . More precisely:

- A *leaf* of R is an extant gene and maps to a leaf of S , i.e. the extant genome to which it belongs.
- A *duplication vertex* of R is an internal vertex which corresponds to a duplication event. It maps to a branch of S , i.e. the lineage in which the duplication occurred (see Figure 3).
- A *speciation vertex* of R is an internal vertex which corresponds to an ancestral gene at the time of a speciation event. It maps to an internal vertex of S , i.e. the ancestral genome to which it belongs. It has either one child (in the case of a gene loss), or two children each belonging to a different lineage. The set of speciation vertices mapping to a vertex A of S is the *genome set* $G(A)$ of A . If A is not the root, let B be the father of A . Then the *pre-speciation genome set* $PG(A)$ of A is the subset of $G(B)$ containing the vertices of $G(B)$ with a child in the branch (A, B) , in other words, the genes in $G(B)$ that have not been lost after speciation on the branch going to A . We have $|PG(A)| \leq |G(B)|$ (see Figure 3).

Considering now the **Minimization step**, if only *simple-duplications* are allowed, the problem has been shown in [20] to be equivalent to the one of finding gene orders at internal nodes of S minimizing a global inversion distance. In this context, the evolutionary model can be reduced to the one where all duplications occur first, followed by all inversions. The problem is then to find the minimum number of inversions allowing to obtain a forest of simple duplication trees. As simple duplication trees are equivalent to binary search tree, we can enumerate all binary search trees. All these simplifications and shortcuts do not hold any more for simultaneous duplications and deletions of multiple genes. In the following section, we focus on the **Minimization Step**.

2.5 MultiDILTAG: Extension of DILTAG to multiple species

Our algorithm is a generalization of DILTAG that proceeds with the whole species tree S and produces a solution set for each internal vertex, and a solution graph with additional speciation edges. Figure 4 illustrates the algorithm execution at each internal vertex A of S .

Initially, the solution set of each leaf is reduced to the gene order observed at that leaf, and the solution graph is reduced to the set of vertices defined by the ordered gene trees at the leaves. We then extend the solution graph by exploring S bottom-up, and for each internal vertex A , we compute a solution set \mathcal{S}_A by

performing DILTAG respectively on the left branch (A_l, A) and right branch (A_r, A) of S (with A_l and A_r being respectively the left and right child of A), and taking, as potential orders at A , the union of genome sets $PG(A_l)$ and $PG(A_r)$ obtained respectively in the left and right branch. However, due to gene losses, gene orders in $PG(A_l)$ do not necessarily have the same number of genes than gene orders in $PG(A_r)$. We therefore consider all possible extensions of gene orders, by reinserting lost copies in any possible way, and take the union of all sets obtained as the solution set \mathcal{S}_A . We then define a single “speciation edge” in the solution graph from each vertex representing a gene order in \mathcal{S}_A to each vertex representing a gene order in $PG(A_l) \cup PG(A_r)$. As the only evolutionary events likely to have occurred on these edges of the history graph are inversions and deletions, we label each speciation edge (x, y) by the minimum Inversions+Deletions (ID) distance allowing to transform x into y . In the literature, the problem of computing the ID-distance between two permutations has already been considered, and a polynomial-time algorithm exists [41, 42].

More precisely, the MultiDILTAG algorithm traverses the tree bottom-up, and for each internal node A proceeds as follows:

1. For each of $s \in \{l, r\}$, execute DILTAG on each element of \mathcal{S}_{A_s} , and stops as soon as the attained gene order contains $|PG(A_s)|$ genes. The set of all ancestral gene orders obtained (output of DILTAG) form an initial *pre-speciation* set \mathcal{PS}_{A_s} , further truncated as follows: if MIN is the minimum cost obtained over all elements of \mathcal{S}_{A_s} , we remove from \mathcal{PS}_{A_s} all elements O that are not attained with the cost MIN . Moreover, we remove from the partial current solution graph all the predecessors of O that are not linked to another element of \mathcal{PS}_{A_s} by a minimum-cost path.
2. For each of $s \in \{l, r\}$, construct the set \mathcal{PS}'_{A_s} by replacing each gene order O of \mathcal{PS}_{A_s} by the set of all possible orders obtained from O by inserting the genes lost on the branch (A, A_s) .
3. Compute $\mathcal{S}_A = \mathcal{PS}'_{A_l} \cup \mathcal{PS}'_{A_r}$. The solution graph is extended by adding one vertex per each element of \mathcal{S}_A .
4. Let $O \in \mathcal{S}_A$, and suppose, w.l.o.g. that $O \in \mathcal{PS}'_{A_l}$. Then complete the solution graph by constructing an oriented “speciation edge” from O to the vertex corresponding to its originating order in A_l , and an oriented edge from O to the vertex corresponding to each element of \mathcal{PS}_{A_r} giving rise to the minimum ID-distance with O .

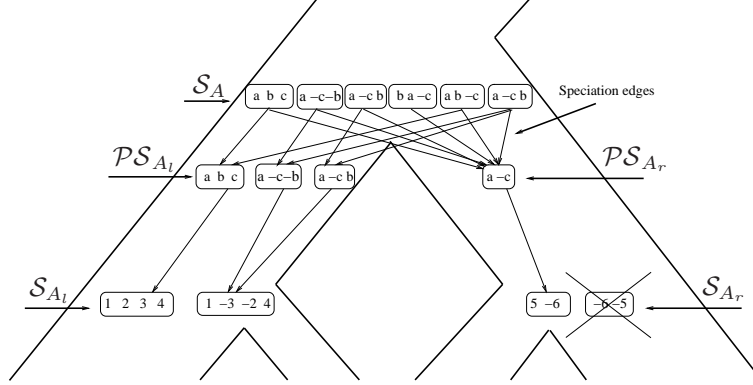


Figure 4: Computation of the solution set \mathcal{S}_A at the internal node A of a species tree S by MultiDILTAG. DILTAG is executed on each of the two branches (A_l, A) and (A_r, A) , leading to the two pre-speciation sets $\mathcal{P}\mathcal{S}_{A_l}$ and $\mathcal{P}\mathcal{S}_{A_r}$. In each branch, only minimum-cost paths are kept, which explains the removal of one gene order (indicated by a cross) from \mathcal{S}_{A_r} . The gene missing from the gene order of $\mathcal{P}\mathcal{S}_{A_r}$ is reinserted in all possible positions (and all possible signs, which is not shown), and the resulting set is added to $\mathcal{P}\mathcal{S}_{A_l}$ to form \mathcal{S}_A . Appropriate speciation edges are then added from the elements of \mathcal{S}_A to the elements of $\mathcal{P}\mathcal{S}_{A_l} \cup \mathcal{P}\mathcal{S}_{A_r}$.

3 Results and Discussion

We implemented our algorithm and applied it to simulated data sets to evaluate its execution time and precision in terms of the number and size distribution of the inferred duplications. Ordered gene trees were generated by simulating evolutionary histories consistent with balanced species trees of 2, 4 or 8 leaves.

Unless stated otherwise, the size of each event was sampled according to a geometric distribution of parameter $p = 0.5$, truncated by the number of genes in the order before the event. All the results shown below are averaged over 50 replicates.

Similarly to the DILTAG algorithm, we define the penalty cost of an event e of size m (acting on a segment of m genes) as $\alpha_e + m\beta_e$, where α_e is the opening cost and β_e the extension cost of e . Our results were obtained with the same values used in [1] to test the DILTAG algorithm, namely:

- $\alpha_{t-dup} = 100$; $\beta_{t-dup} = 1$,
- $\alpha_{del} = 500$; $\beta_{del} = 1$,
- $\alpha_{i-dup} = 100$; $\beta_{i-dup} = 1$,
- $\alpha_{inv} = 500$; $\beta_{inv} = 1$.

3.1 Execution time

Our algorithm was implemented in C++ and run on a typical Linux workstation. Figure 5 shows the execution time of Multi-DILTAG. The left diagram shows results for balanced species trees of 2, 4 and 8

leaves. The depth d of the extant genomes for trees with 2, 4 and 8 leaves are respectively 2, 3 and 4. We generated histories with n single, n double tandem duplications (simultaneous duplication of 2 genes) and 2 inversions on each branch of the species tree. At each step in the curves, n is incremented by 1 and thus the number of genes in each extant genome is equal to $3dn + 1$.

Figure 5 right then shows the effect of introducing deletions. Only histories with 2 extant genomes were generated, and we plotted the running times for simulated histories containing no deletion, 1 deletion of size 1 and 1 deletion of size 2.

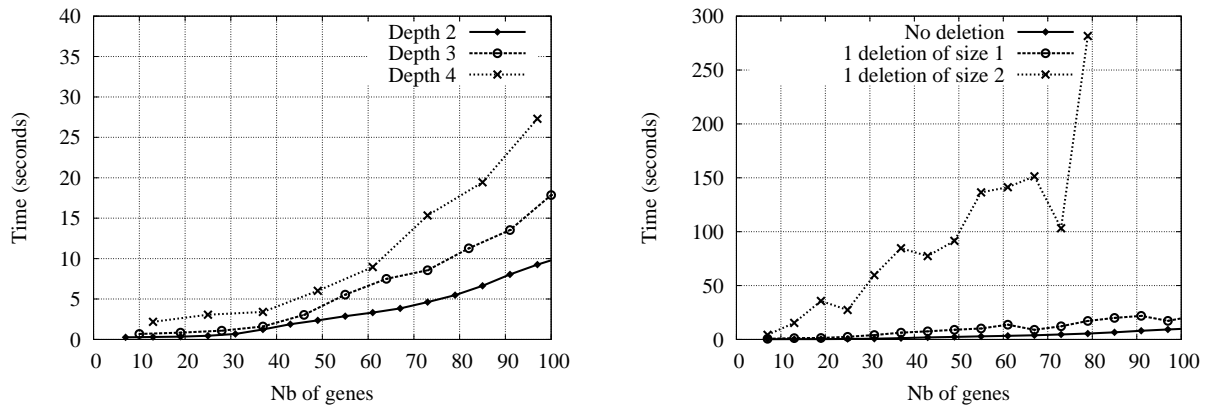


Figure 5: Left: Execution time of Multi-DILTAG on genomes containing a fixed number of genes on all the leaf genomes. Balanced species trees of maximum depth 2 (2 leaves), 3 (4 leaves) and 4 (8 leaves) were generated. Right: Execution time of Multi-DILTAG on species trees with two leaves, for simulated history with no deletion (the same curve as the one indicated by a plain black line on the left diagram), 1 deletion of size 1 and 1 deletion of size 2.

Clearly the execution time of Multi-DILTAG is exponential in the number of genes of extant genomes.

Nevertheless, it is possible to get results in under 30 seconds for a family of approximately 100 genes in 8 species. On the other hand, deletions of size greater than 1 slows down Multi-DILTAG dramatically. The idea of considering all possible extensions of gene orders, by reinserting lost copies in any possible way, results in an exponential number of orders in the number of copies to reinsert and the size of the orders in which we make the insertions.

3.2 Number of duplications

We now evaluate the ability of Multi-DILTAG to infer the correct number of total duplications (direct + inverted). We simulated evolutionary histories containing as many duplications as inverted duplications with 2 (Figure 6 left) and 4 (Figure 6 right) extant genomes. In both cases, we plotted the total number of duplications inferred for histories generated with 0 %, 33 % and 50 % of inversions.

More precisely, for each x (x axis), we generate a history with a total of $x/2$ duplications and $x/2$ inverted duplications together with 0, $x/2$ or x inversions, respectively leading to the curves for 0 %, 33 % and 50 % of inversions. The total number of events performed for each value of x is distributed evenly on the branches of the species tree.

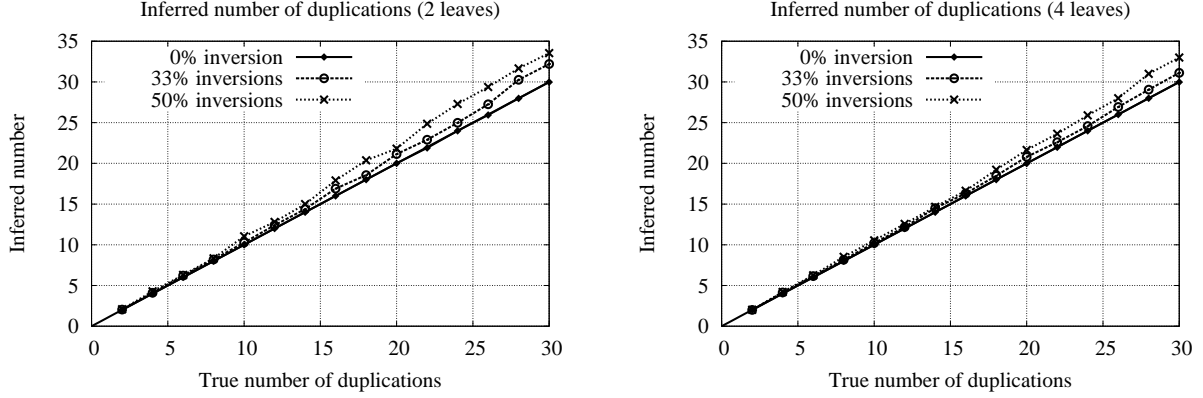


Figure 6: Inferred number of duplications (direct + inverted) for histories containing duplications and respectively 0 %, 33 % and 50 % of inversions. Left: Two extant genomes. Right: Four extant genomes.

As we see, Multi-DILTAG is almost perfect in inferring the total number of duplications when there are no inversions. The presence of inversions induces a small overestimation in the inferred number of duplications. As noticed in [1], this can be explained by the size limit of the DILTAG priority queue used to explore the search space and the chosen cost configuration, which may lead to choosing a history with more duplications in order to infer less inversions.

Notice that the overestimation is a little bit more pronounced in Figure 6 left. This can be easily explained by the fact that there are less branches in the balanced species tree containing 2 extant genomes than in the one of 4 extant genomes. Therefore, for the same total number of duplications, more inversions are present on each branch of the smallest species tree.

3.3 Duplication size distribution

Finally, we measure the accuracy of Multi-DILTAG for inferring the duplication size distribution. Histories containing 2 (Figure 7 left) and 4 (Figure 7 right) extant genomes were generated. In both cases, 4 tandem duplications, 1 inverted duplication, 1 inversion and 1 deletion of size 1 or 2 were simulated on each branch of the corresponding balanced species tree.

Clearly, Multi-DILTAG is able to infer the duplication size distribution very accurately for both data sets.

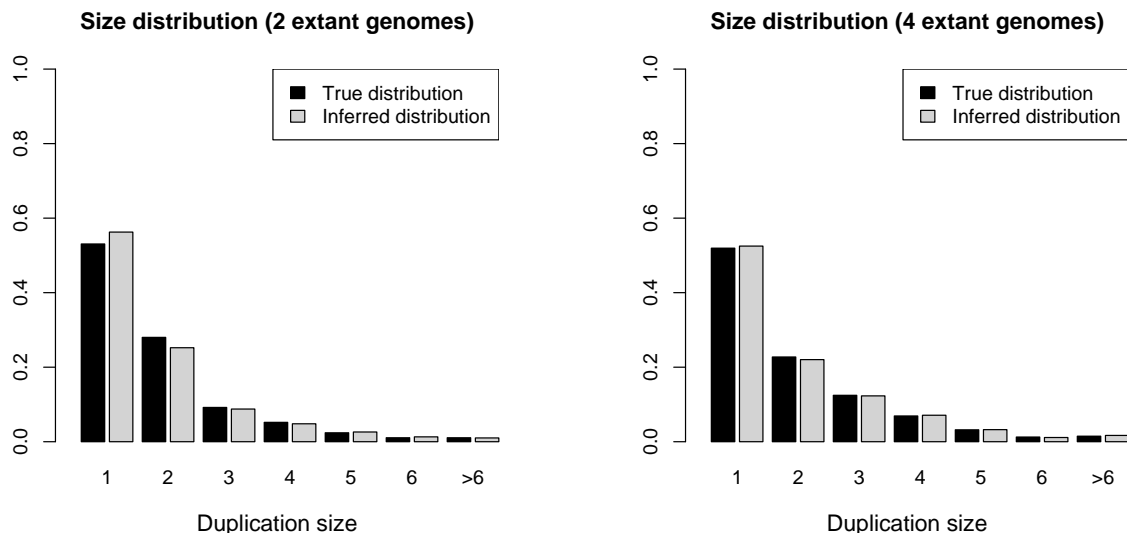


Figure 7: Comparison between the true and the inferred duplication size distribution. Histories were generated with 4 tandem duplications, 1 inverted duplication, 1 inversion and 1 deletion of size 1 or 2 on each branch of the species tree. Left: Two extant genomes. Right: Four extant genomes.

We can only observe a slight overestimation of duplications of size 1 and underestimation of duplications of size 2.

4 Conclusions

We presented Multi-DILTAG, a generalization of DILTAG for the study of the evolutionary history of a set of orthologous TAG clusters in multiple species, with an evolutionary model allowing for simple or multiple tandem duplications, direct or inverted, simple or multiple deletions, and inversion events. Our results showed that our algorithm is very robust in inferring the number and size distribution of duplications. A short-term future work will therefore concern the application of Multi-DILTAG to biological data sets. In particular, we will study the number and size of duplication events among the different branches of the mammalian species tree that can explain the many observed olfactory receptor TAG clusters in these species.

However, a clear limitation of Multi-DILTAG is the time complexity of the approach taken to deal with deleted genes. An important future work will be to develop a fast heuristic to find an optimal set of extensions of gene orders without reinserting the lost copies in any possible way.

References

1. Lajoie M, Bertrand D, El-Mabrouk N: **Inferring the Evolutionary History of Gene Clusters from Phylogenetic and Gene Order Data.** *Molecular Biology and Evolution* 2010, **27**:761-772.
2. Ohno S: *Evolution by gene duplication*. Berlin: Springer 1970.
3. Blomme T, Vandepoele K, Bodt SD, Silmillion C, Maere S, van de Peer Y: **The gain and loss of genes during 600 millions years of vertebrate evolution.** *Genome Biology* 2006, **7**:R43.
4. Cotton J, Page R: **Rates and patterns of gene duplication and loss in the human genome.** *Proceedings of the Royal Society of London. Series B* 2005, **272**:277-283.
5. Eichler E, Sankoff D: **Structural dynamics of eukaryotic chromosome evolution.** *Science* 2003, **301**:793-797.
6. Hahn M, Han M, Han SG: **Gene family evolution across 12 *Drosophila* genomes.** *PLoS Genetics* 2007, **3**:e197.
7. Lynch M, Conery J: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290**:1151-1155.
8. Wapinski I, Pfeffer A, Friedman N, Regev A: **Natural history and evolutionary principles of gene duplication in fungi.** *Nature* 2007, **449**:54-61.
9. Zhou L, Huang B, Meng X, Wang G, Wang F, Xu Z, Song R: **The amplification and evolution of orthologous 22-kDa α -prolamin tandemly arrayed genes in *coix*, sorghum and maize genomes.** *Plant Molecular Biology* 2010, **74**:631- 643.
10. Shoja V, Zhang L: **A Roadmap of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat.** *Molecular Biology and Evolution* 2006, **23**:2134-2141.
11. The Arabidopsis Genome Initiative: **Analysis of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796- 815.
12. Messing J, Bharti A, Karlowski W, Gundlach H, Kim H, Yu Y, Wei F, Fuks G, Soderlund C, Mayer K: **Sequence composition and genome organization of maize.** *Proceedings of the National Academy of Sciences USA* 2004, **101**:14349- 14354.
13. Fitch W: **Phylogenies constrained by cross-over process as illustrated by human hemoglobins and a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I.** *Genetics* 1977, **86**:623-644.
14. Bertrand D, Gascuel O: **Topological rearrangements and local search method for tandem duplication trees.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005, **2**:15-28.
15. Elemento O, Gascuel O, Lefranc MP: **Reconstructing the duplication history of tandemly repeated genes.** *Molecular Biology and Evolution* 2002, **19**:278-288.
16. Tang M, Waterman M, Yooseph S: **Zinc finger gene clusters and tandem gene duplication.** In *Research in Molecular Biology (RECOMB 2001)* 2001:297-304.
17. Zhang L, Ma B, Wang L, Xu Y: **Greedy method for inferring tandem duplication history.** *Bioinformatics* 2003, **19**:1497-1504.
18. Gascuel O, Bertrand D, Elemento O: **Reconstructing the duplication history of tandemly repeated sequences.** In *Mathematics of Evolution and Phylogeny*. Edited by Gascuel O, Oxford 2005:205-235.
19. Lajoie M, Bertrand D, El-Mabrouk N, Gascuel O: **Duplication and Inversion History of a Tandemly Repeated Genes Family.** *Journal of Computational Biology* 2007, **14**(4):462-478.
20. Bertrand D, Lajoie M, El-Mabrouk N: **Inferring Ancestral Gene Orders for a Family of Tandemly Arrayed Genes.** *Journal of Computational Biology* 2008, **15**(8):1063-1077.
21. J. Ma and A. Ratan and B. J. Raney and B. B. Sush and L. Zhang and W. Miller and D. Haussler. *Journal of Computational Biology* 2008, **15**(8).
22. Zhang Y, Song G, Vinar T, Green E, Siepel A, Miller W: **Reconstructing the evolutionary history of complex human gene clusters.** In *Research in Computational Molecular Biology. (RECOMB 2008), Volume 4955 of Lecture Notes in Computer Science*. Edited by M Vingron, L Wong, Springer 2008:29-49.

23. Zhang Y, Song G, Hsu C, Miller W: **Simultaneous history reconstruction for complex gene clusters in multiple species.** In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 2009:162–173.
24. Vinař T, Brejová B, Song G, Siepel A: **Reconstructing Histories of Complex Gene Clusters on a Phylogeny.** *Journal of Computational Biology* 2010, **17**:1267–1269.
25. Song G, Zhang L, Vinař T, Miller W: **Inferring the recent duplication history of a gene cluster.** In *Comparative Genomics, Volume 5817 of Lecture Notes in Computer Science*. Edited by Ciccarelli F, Miklós I, Springer 2009.
26. LaRue R, Jonsson S, Silverstein K, Lajoie M, Bertrand D, El-Mabrouk N, Hötzel I, Andresdottir V, Smith T, Harris R: **The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals.** *BMC Molecular Biology* 2008, **9**:104.
27. Gabrisko M, Janecek S: **Characterization of Maltese Clusters in the Genus Drosophila.** *Journal of Molecular Evolution* 2011, **72**:104–118.
28. Aloni R, Olender T, Lancet D: **Ancient genomic architecture for mammalian olfactory receptor clusters.** *Genome Biology* 2006, **7**:R88.
29. Goodman M, Czelusniak J, Moore G, Romero-Herrera A, Matsuda G: **Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences.** *Systematic Zoology* 1979, **28**:132–163.
30. Guigó R, Muchnik I, Smith T: **Reconstruction of Ancient Molecular Phylogeny.** *Molecular Phylogenetics and Evolution* 1996, **6**:189–213.
31. Page R, Charleston M: **Reconciled Trees and Incongruent Gene and Species Trees.** *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 1997, **37**:57–70.
32. Bonizzoni P, Della Vedova G, Dondi R: **Reconciling a gene tree to a species tree under the duplication cost model.** *Theoretical Computer Science* 2005, **347**:36–53.
33. Durand D, Haldórsson B, Vernot B: **A hybrid micro-macroevoolutionary approach to gene tree reconstruction.** *Journal of Computational Biology* 2006, **13**:320–335.
34. Eulenstein O, Mirkin B, Vingron M: **Comparison of annotating duplication, tree mapping, and copying as methods to compare gene trees with species trees.** *Mathematical hierarchies and biology; DIMACS Series Discrete Math. Theoret. Comput. Sci.* 1997, **37**:71–93.
35. Gorecki P, Tiuryn J: **DLS-trees: a model of evolutionary scenarios.** *Theoretical Computer Science* 2006, **359**:378–399.
36. Ma B, Li M, Zhang L: **From gene trees to species trees.** *SIAM Journal on Computing* 2000, **30**:729–752.
37. Page R: **Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas.** *Systematic Biology* 1994, **43**:58–77.
38. Page R: **GeneTree: comparing gene and species phylogenies using reconciled trees.** *Bioinformatics* 1998, **14**:819–820.
39. Zhang L: **On Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies.** *Journal of Computational Biology* 1997, **4**:177–188.
40. Chauve C, El-Mabrouk N: **New perspectives on gene family evolution: losses in reconciliation and a link with supertrees.** In *Research in Molecular Biology (RECOMB 2009), Volume 5541 of Lecture Notes in Computer Science*. Edited by Batzoglou S, Springer 2009:46–58.
41. El-Mabrouk N: **Genome Rearrangement by Reversals and Insertions/Deletions of Contiguous Segments.** In *Proceedings of the Eleventh Annual Symposium on Combinatorial Pattern Matching (CPM 2000), Volume 1848 of Lecture Notes in Computer Science*. Edited by Giancarlo R, Sankoff D 2000:222–234.
42. Marron M, Swenson K, Moret B: **Genomic distances under deletions and insertions.** In *Proc. 9th Int'l Combinatorics and Computing Conf. (COCOON'03), Volume Lecture Notes in Computer Science 2697*, Springer Verlag 2003:537–547.