Evolution of Tandemly Arrayed Genes in Multiple Species

Mathieu Lajoie[1], Denis Bertrand[1], and Nadia El-Mabrouk[1]

DIRO - Université de Montréal - H3C 3J7 - Canada
{bertrden,lajoimat,mabrouk}@iro.umontreal.ca

Abstract. Tandemly arrayed genes (TAG) constitute a large fraction of most genomes and play important biological roles. They evolve through unequal recombination, which places duplicated genes next to the original ones (tandem duplications). Many algorithms have been proposed to infer a tandem duplication history for a TAGs cluster in a single species. However, the presence of different transcriptional orientations in most TAGs clusters highlight the fact that processes such as inversion also contribute to their evolution. This makes those algorithms inapplicable in many cases. To circumvent this limitation, we proposed an extended evolutionary model which includes inversions and presented a branch-and-bound algorithm allowing to infer a most parsimonious scenario of evolution for a given TAGs cluster. Here, we generalize this model to multiple species and present a general framework to infer ancestral gene orders that minimize the number of inversions in the whole evolutionary history. An application on a pair of human-rat TAGs clusters is presented.

1 Introduction

A multigene family is a set of genes that have evolved by duplication from a common ancestral gene, and share a similar sequence and usually a similar function. Members of a gene family in a given genome may appear in clusters, or scattered in a single or many chromosomes. In this paper, we focus on families of tandemly arrayed genes (TAG): copies that are adjacent on the chromosome. TAGs have been shown to represent a large proportion of genes in a genome. In particular, they represent about 14-17% of all genes in human, mouse and rat [24]. Clusters of TAGs may vary in length from two to hundreds genes, though small clusters are largely predominant (an average of 3 to 4 genes in mouse, rat and human) [24]. They are involved in many different functions of binding or receptor activities. In particular, the olfactory receptor genes constitute the largest multigene family in the vertebrate genome, with several hundred genes per species [21]. Other families of TAGs include the HOX genes [29], the immunoglobulin and T-cell receptor genes [1], the MHC genes [15] and the ZNF genes encoding for transcription factors [23].

TAGs are widely viewed as being generated solely by tandem duplications resulting from unequal recombination [13] or slipped strand mispairing [2]. Such mechanisms have the effect of generating sequences of repetitive units with the same transcriptional orientation. However, it is not infrequent to observe TAGs with different orientations. In particular, Shoja and Zhang [24] have observed that more than 25% of all neighboring pairs of TAGs in human, mouse and rat have non-parallel orientations. This underlines the fact that other rearrangement mechanisms such as inversions should be considered in an evolutionary model of TAGs.

Based on the unequal recombination model of evolution, a large number of studies have considered the problem of reconstructing a tandem duplication history of a TAGs family [4, 9, 10, 25]. These are essentially phylogenetic inference methods using the additional constraint that the resulting tree should induce a duplication history according to the given gene order. When a gene tree is already available for a gene family, a linear-time algorithm can be used to check whether it is a duplication tree [30]. However, it is often impossible to reconstruct a duplication history [14], due to other evolutionary events such as gene losses or genomic rearrangements [8]. In [5] we have considered an evolutionary model accounting for both tandem duplications and inversions. Given a

gene tree for a family of TAGs, we developed an algorithm allowing to find the minimum number of inversions in any possible evolutionary scenario for this family.

All the above methods are restricted to the analysis of TAGs located on a single chromosome (and thus in a single species). However, the increasing availability of complete genomic sequences and of many different TAGs databases [21, 27] makes it possible to study the evolution of gene families with members belonging to different species. Such a global evolutionary study may help deciphering the common origins of TAGs, highlighting the inter-species differences and identifying the genetic basis of species-specific features. Various phylogenetic studies have been conducted by biological groups on different TAGs families such as the Zinc-Finger transcription factors in human and mouse [23], and the olfactory receptor genes in various mammalian species [21]. However no rigorous approach has been developed so far to explain the non agreement between a given gene tree of a TAGs family and a duplication and speciation history.

In this paper, we consider an evolutionary model of TAGs accounting for duplication, speciation, gene loss and inversion events. This is a generalization of [5] to TAGs located on different genomes. More precisely, given a gene tree for a family of TAGs and their signed order on the genomes (chromosomes or clusters), we aim to find an evolutionary scenario involving the minimum number of inversions, and the corresponding gene orders of the ancestral genomes. The Fitch model allows for the simultaneous duplication of several gene copies, but there are now evidence that simple duplications are predominant over multiple duplications [4, 29]. As a first attempt, we only consider simple duplications.

This paper is organized as follows. After describing the evolutionary models in Section 2, we present the general problem in Section 3. It is related to the more classical one of inferring the gene order of the ancestral genomes in a species tree minimizing a given genomic distance [20, 22]. In Section 3.1, we present an algorithm to infer the most parsimonious scenario of inversion on a single branch of the species tree. In Section 3.2, we present a simple iterative method used to infer the ancestral gene orders minimizing the total number of inversions in a species tree. It is based on the median problem, for which we propose a branch-and-bound algorithm in Section 3.3. Finally, in Section 4, we test the algorithm's time-efficiency on simulated data, and present an application on a pair of human-rat TAGs clusters.

2 The evolutionary model

The classical model of evolution considered for TAGs is based on tandem duplications resulting from unequal recombination during meiosis. The later is assumed to be the sole evolutionary mechanism (except point mutations) acting on sequences. Formally, from a single ancestral gene at a given position in the chromosome, the locus grows through a series of consecutive duplications placing the created copy next to the original one. Such *tandem duplications* may be *simple* (duplication of a single gene) or multiple (simultaneous duplication of neighboring genes). In this paper, we only consider simple duplications. From now on, a *duplication* will refer to a simple tandem duplication.

The former model of evolution applies only on a family of TAGs all located on the same chromosome and having the same transcriptional orientation. In particular, it is inadequate to describe the evolution of a TAGs family containing members on both DNA strands. To circumvent this limitation, we have proposed, in [5], an extended model of duplications including inversions. In this paper, we further extend the model to account for several genomes.

Consider a family of TAGs located on m different genomes. In addition to the TAGs orders on each genome, all we can infer from the gene sequences is a gene tree representing the global evolution of the gene family. Formally, an *ordered gene tree* is a set $(T, \mathcal{O})$, where T is a gene tree of the TAGs family and $\mathcal{O} = (O_1, O_2, \cdots O_m)$ where O_i is the signed order of the family members

in genome i, for $1 \leq i \leq m$. Thereafter, the transcriptional orientations of the genes in an ordered gene tree $(T, \mathcal{O})$ are specified by signs $(+/-)$ in each O_i. We denote by $d_{inv}(O_i, O_j)$ the inversion distance between the two signed permutations O_i and O_j. Such a distance can be computed using the original Hannenhalli and Pevzner algorithm[16], or any of the existing optimizations [3, 18, 26].

An ordered gene tree $(T, \mathcal{O})$ can always be explained by a history $\mathcal{H}$ involving duplication, gene loss, speciation and inversion events (DLSI history), as stated in Lemma 1 below. We say that $\mathcal{H}$ is a DLSI history of $(T, \mathcal{O})$, and that $(T, \mathcal{O})$ is *compatible* with $\mathcal{H}$. Hereafter, we begin by formally defining a DLSI history (see Figure 1 for an illustration).

Definition 1. *Let $\mathcal{H} = ((T^1, \mathcal{O}^1), \cdots (T^k, \mathcal{O}^k), \cdots (T^{n-1}, \mathcal{O}^{n-1}), (T^n, \mathcal{O}^n))$ be a sequence of n ordered gene trees. For each k, $1 \leq k \leq n$, we denote by m_k the number of genomes represented in T^k, and by O_i^k the gene order in genome i, for $1 \leq i \leq m_k$.*

We say that $\mathcal{H}$ is a DLSI history if and only if:

1. *$T^1 = v$ is the single leaf gene tree and $\mathcal{O}^1 = (O_1^1) = (\pm v)$ is one of the two trivial orders;*
2. *For $1 < k < n$, one of the four following situations hold:*
   a. Duplication event: *There is an i, $1 \leq i \leq m_k$, such that T^{k+1} is obtained from T^k by adding two children u and w to a leaf v belonging to genome i. Moreover:*
      - *$m_{k+1} = m_k$;*
      - *$\mathcal{O}^{k+1}$ is obtained from $\mathcal{O}^k$ by replacing $\pm v$ by $(\pm u, \pm w)$ in O_i^k.*
   b. Gene loss event: *There is an i, $1 \leq i \leq m_k$, such that T^{k+1} is obtained from T^k by removing a leaf v belonging to genome i. Moreover: If v was the only leaf in O_i^k, $m_{k+1} = m_k - 1$. Otherwise, O_i^{k+1} is obtained from O_i^k by deleting v and $m_{k+1} = m_k$.*
   c. Inversion event: *$T^{k+1} = T^k$ and there is an i, $1 \leq i \leq m_k$, such that $d_{inv}(O_i^k, O_i^{k+1}) = 1$. Moreover, $O_j^{k+1} = O_j^k$ for $j \neq i$ and $m_{k+1} = m_k$.*
   d. Speciation event: *There is an i, $1 \leq i \leq m_k$, such that T^{k+1} is obtained from T^k by adding two children u and w to each leaf v belonging to genome i. Moreover:*
      - *$m_{k+1} = m_k + 1$;*
      - *$\mathcal{O}^{k+1}$ is obtained from $\mathcal{O}^k$ by dedoubling the order O_i^k.*

Any DLSI history $\mathcal{H}$ induces a unique species tree S obtained from the speciation events of $\mathcal{H}$. We say that $\mathcal{H}$ is *consistent with S* (see Figure 1).

Let $(T, \mathcal{O})$ be an ordered gene tree for a family of TAGs on m genomes. Suppose that a species tree is already known for the m genomes. Then a natural problem is to find a DLSI history of $(T, \mathcal{O})$ that is consistent with S. The existence of such a history is stated in Lemma 1. It follows from the existence of a duplication/speciation/loss history in the more general case of non-ordered gene families generated by general duplications, e.g. not necessarily in tandem. More precisely, given a gene tree T for a set of (unsigned) genes located on m genomes, and a species tree S for these genomes, the classical reconciliation approach [7, 11, 19] infers a duplication/speciation/loss history, involving a minimum number of gene losses, that has led to the gene tree T. It is based on a particular mapping (the LCA mapping) from the vertices of T to the vertices of S. Moreover, it infers the gene contents in the ancestral species preceding each speciation event.

Lemma 1. *Given an ordered gene tree $(T, \mathcal{O})$ on m genomes and a species tree S for the m genomes, there is at least one DLSI history $\mathcal{H}$ of $(T, \mathcal{O})$ consistent with S.*

Proof. Obtain a sequence of duplications (not necessarily in tandem), gene losses and speciations from the reconciliation of T and S. From that sequence, construct a DLSI history $\mathcal{H}' = ((T^1, \mathcal{Q}^1), \cdots, (T^n = T, \mathcal{Q}^n)$ by applying the corresponding rules in definition 1 (case a,b or d). Then, obtain $\mathcal{H}$ from $\mathcal{H}'$ by appending the inversions required to transform $\mathcal{Q}^n$ in $\mathcal{O}$ (case c in definition 1) □

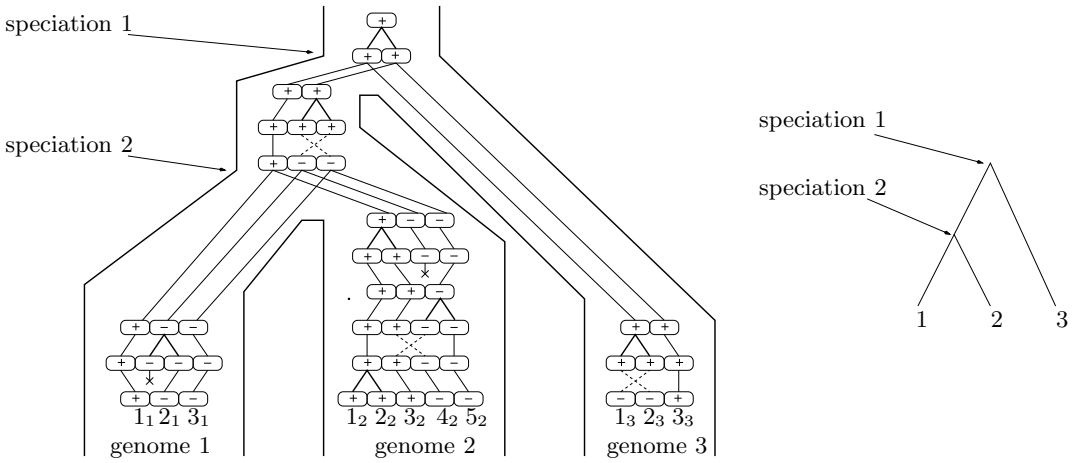

Fig. 1. A DLSI history. Transcriptional orientations are indicated by signs, duplications by bold lines, gene losses by 'X' and inversions by dashed lines. The resulting TAGs orders are denoted as i_k meaning "gene i in genome k". For clarity, we omitted successive identical configurations in each lineage. The induced species tree for the three genomes is given right.

From the general case of Definition 1, we also introduce the following restricted evolutionary history: A *Duplication/Inversion* history (*DI history*) is an evolutionary history of TAGs in one species involving only duplication and inversion events (case a and c of Definition 1).

3 An inference problem

In the present study, we are given an ordered gene tree $(T, \mathcal{O})$ of a TAGs family on m genomes, and we assume that a species tree S is known for the m genomes (in the case of an unknown species tree, we can take advantage of our algorithm presented in [6] that infers a speciation tree leading to the minimum number of gene losses).

As the number of possible DLSI histories of $(T, \mathcal{O})$ consistent with S is unlimited, we restrict ourselves to finding a most parsimonious sequence of evolutionary events leading to the observed ordered gene tree. Moreover, if the gene tree T is not compatible with a history of duplication and speciation consistent with S, then the reconciliation method allows to create a new gene tree T' that is compatible with a duplication and speciation history consistent with S. This is done by performing a minimum number of subtrees insertions in T, where each subtree insertion represents a gene loss in an ancestral genome. However, when the gene order is important (in the case of an ordered gene tree), as no order is known for the hypothetical "inserted" genes, the only information resulting from the reconciliation analysis that is of interest is the set of ancestral genes, and the localization of gene losses. In other words, we construct, T'' from T' by keeping the roots of the inserted subtrees. Formally, a *reconciled gene tree* is a tree where each internal node has exactly one or two children, and an *ordered reconciled gene tree* is just a reconciled gene tree with ordered leaves (see Figure 2).

It follows from the previous developments that the problem of interest reduces to the one of finding the ancestral gene orders that minimize the total number of inversions involved in a DLSI history of $(T, \mathcal{O})$. Formally, the considered problem is the following:

Minimum-DLSI problem
Input: An ordered reconciled gene tree $(T, \mathcal{O})$.
Output: A gene order for each ancestral genome inducing a history of minimum inversions.

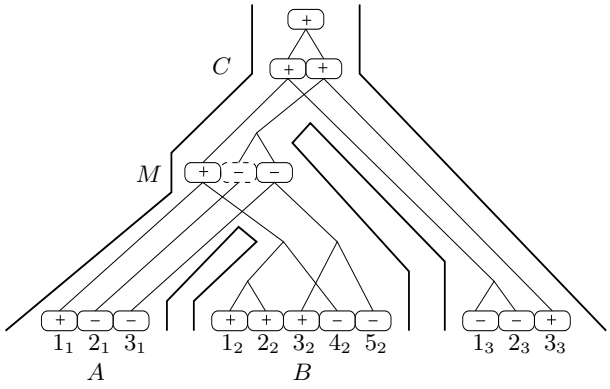

Fig. 2. The ordered reconciled gene tree induced by the DLSI history of Figure 1, with an arbitrary gene order at each ancestral genome preceding each speciation event. The gene tree is "embedded" in the species tree. The dashed gene in genome M indicates a loss of its descendants in lineage B.

A DLSI history $\mathcal{H}$ can be seen as a set of DI histories embedded in the species tree S. Each such individual history correspond to a *branch* of the species tree (see Figure 2). The next section focuses on a single branch of the species tree.

3.1 The generalized Minimum-DI problem

This problem is a generalization of the Minimum-DI problem we presented in [5], which consists in finding the minimum number of inversions required to explain a single "rooted" ordered gene tree. Here the goal is to find the minimum number of inversions required to explain an *ordered forest of gene trees*, associated to a given branch of the species tree (see Figure 2 and 3a). Formal definitions follow.

Definition 2. *An* ordered forest of gene trees (F, R, O) *is a set of* n *gene trees* $F = \{T_1, T_2, \cdots T_n\}$ *rooted at* $R = \{r_1, r_2 \cdots r_n\}$ *(r_i is the ancestral gene that gave rise to T_i) with an order O on the set of leaves of F. When an ancestral order O_R is imposed on R, we use the notation (F, O_R, O).*

Definition 3. *Let $O_R = (r_1, r_2 \cdots r_n)$ be an ordered sequence of roots, and $O_1 = (o_1, o_2, \cdots o_n)$ be an ordered sequence of genes such that, for each $1 \leq i \leq n$, o_i is a direct descendant of r_i. A* partial DI history *rooted at O_R is a sequence of ordered forests of gene trees $\mathcal{H} = ((F_1, O_R, O_1), ..., (F_{k-1}, O_R, O_{k-1}), (F_k, O_R, O_k))$ where (F_1 is just a set of single leaf gene trees, and for $0 < i < k$:*

1. Inversion event: *If $F_{i+1} = F_i$, then $d_{inv}(O_i, O_{i+1}) = 1$.*
2. Duplication event: *If $F_{i+1} \neq F_i$, then F_{i+1} is obtained from F_i by adding two children u and w to one of its leaf v, and O_{i+1} is obtained from O_i by replacing v by (u, w), where u and w have the same sign as v.*

Moreover, a partial duplication history *is a partial DI history restricted to duplication events.*

A partial duplication history gives rise to a duplication forest, defined as follows.

Definition 4. *A* duplication forest *is an ordered forest of gene trees (F, O_R, O) which contains only duplications trees, and such that for every pair of roots r_i, r_j in R, if r_i precedes r_j in O_R, then all the leaves of T_i precedes all the leaves of T_j in O. Moreover, the leaves of each T_k in F, must have the same sign as r_k.*

The following theorem is a generalization of the result obtained in [5] for a single ordered gene tree.

Theorem 1. *Let (F, O_R, O) be an ordered forest of gene trees and (F, O_R, O') be a duplication forest such that $d_{inv}(O, O') = i$ is minimum. Then there exists a partial DI history of (F, O_R, O) with exactly i inversions. Moreover, i is the minimum number of inversions involved in any partial DI history of (F, O_R, O).*

Proof. The proof uses arguments similar to those considered in [5], and will be detailed in a full version of this extended abstract □

Theorem 1 allows us to formulate the problem as follows:

Generalized-Minimum-DI problem
Input: An ordered forest of gene trees (F, O_R, O),
Output: An order O' on the leaves of F such that (F, O_R, O') is a duplication forest and $d_{inv}(O, O')$ is minimal.

For a branch represented by the forest (F, O_R, O), we denote by $DI(O_R, O)$ the minimal $d_{inv}(O, O')$ defined above, and we call it the *minimum DI value*.

A Branch-and-Bound algorithm: The algorithm is a generalization of the one we presented in [5]. Given an ordered gene tree (T, O), the goal was to find an order O' minimizing the distance $d_{inv}(O, O')$ that is *compatible with* T, i.e. such that (T, O') is a duplication tree. As mentioned in [14], the considered duplication trees are equivalent to binary search trees. Therefore, to enumerate all the orders compatible with T, we associated a binary variable b_i to each internal node i of T as follows: each b_i defines an order relation between the left and right descendant leaves of i, i.e. by setting b_i to 0 (respec. 1), we make all the left descendants smaller than the right ones (respec. all left descendants are larger than the right ones). Then an order O' is compatible with T iff it is defined by an assignment of all the binary variables b_i in T, and all its genes have the same sign (+ or −). If n is the number of leaves of T, this leads to 2^n distinct orders O' compatible with T.

To avoid computing $d_{inv}(O, O')$ for each order O', we considered a branch-and-bound strategy. The idea was to compute a lower bound on $d_{inv}(O, O')$ as we progressively define a partial order O^*, by updating the breakpoint graph of (O, O^*) [16]. The b_i values must be defined in a depth-first manner according to T (see [5] for more details).

Generalization to an ordered forest of gene trees (F, O_R, O) is straightforward. Indeed, let $(T_1, T_2, \cdots T_n)$ be the set of trees of F ordered according to the order O_R of their roots. Then an order O' *compatible with* (F, O_R), i.e. such that (F, O_R, O') is a partial duplication tree, is the concatenation of n orders $(o'_1, o'_2, \cdots o'_n)$ such that o'_i is compatible with T_i. Therefore, similarly to the preceding case, an order O' is compatible with (F, O_R) iff it is defined by an assignment of all the binary variables b_i in F, and for each $1 \leq i \leq n$, all the genes belonging to T_i have the same sign as r_i (see Figure 3). The same branch-and-bound strategy can then be used to explore the space of all possible orders.

3.2 A general method using the median problem

The Minimum-DLSI problem is related to the more classical one of inferring the gene orders of the hypothetical ancestral genomes represented by the internal nodes of a species tree. In this case, each species is characterized by a given gene order, and the problem is to find the ancestral gene

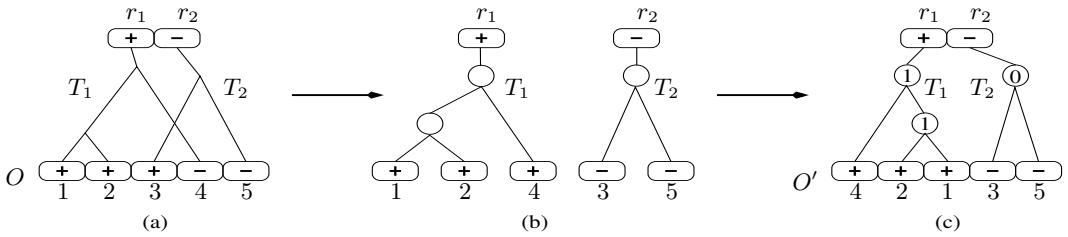

Fig. 3. (a) The ordered forest of gene trees corresponding to the branch (M, B) of the tree in Figure 2 ($F = \{T_1, T_2\}, O_R = (r_1, r_2), O = (1, 2, 3, -4, -5)$). (b) The gene trees in a), with an arbitrary left/right orientation of the children at each internal node. (c) The ordered forest of duplication trees (F, O_R, O') induced by an assignment of the b_i variables in b). The resulting order is $O' = (4, 2, 1, -3, -5)$, and $d_{inv}(O, O') = 3$.

orders minimizing a given genomic distance. The two distances that have been considered in the literature are the breakpoint and inversion distances [20, 22].

Although the case of ordered gene trees is more involved due to the fact that the considered duplication are in tandem, the two problems are related, suggesting a similar global approach summarized below.

1. Begin with an arbitrary order for each internal node of the species tree;
2. Traverse the tree in a depth-first manner. For each subtree consisting of two sister branches (M, A) and (M, B) and a branch (C, M) where C is the immediate ancestor of M (see Figure 2), ignore the assigned order of M, and reconstruct an order that minimizes the value:
$$DI(O_M, O_A) + DI(O_M, O_B) + DI(O_C, O_M).$$
3. Iterate step 2. a given number of times, or after convergence to a minimizing configuration.

Step 1. can be improved by the use of a heuristic that will be detailed in a full version of this extended abstract. Step 2. can be seen as a generalization of the median problem in the context of reconstructing ancestral gene orders of a phylogenetic tree.

To formally define the median problem, we need to extend the notion of an ordered forest of gene trees by allowing the order to be defined either for the leaves or for the roots of the trees. An ordered forest of gene trees defined by a set of trees F_{XY}, a set of roots X and a set of leaves Y will be denoted as (F_{XY}, X, O_Y) and called a *leaf-ordered forest of gene trees* if an order O_Y is defined on the leaves, by (F_{XY}, O_X, Y) and called a *root-ordered forest of gene trees* if an order O_X is defined on the roots, and by (F_{XY}, O_X, O_Y) and called a *fully-ordered forest of gene trees* if an order is defined for both the leaves and the roots.

The median problem is formulated as follows. Given two leaf-ordered forest of gene trees (F_{MA}, M, O_A) and (F_{MB}, M, O_B) (M is the set of ancestral genes generating both A and B) and a root-ordered forest of gene trees (F_{CM}, O_C, M), the goal is to find an order O_M minimizing the value:
$$DI(O_M, O_A) + DI(O_M, O_B) + DI(O_C, O_M)$$
The following section focuses on the median problem.

3.3 A Branch-and-Bound algorithm for the median problem

To avoid considering each of the $2^n n!$ possible signed orders O_M, where n is the number of genes of M, we consider a branch-and-bound strategy. The idea is to compute a lower bound on $DI(O_M, O_A)$, $DI(O_M, O_B)$ and $DI(O_C, O_M)$ as we progressively extend the prefixes O_M^* of M. This is justified by the following property.

Property 1. Let (F^*_{XY}, O^*_X, O^*_Y) be a fully-ordered forest of gene trees obtained from (F_{XY}, O_X, O_Y) by removing the tree rooted at the last element of O_X, or the leaf corresponding to the last element of O_Y. Then:

$$DI(O^*_X, O^*_Y) \leq DI(O_X, O_Y)$$

This bound can be used when we progressively construct the median candidate order O_M. The branch-and-bound strategy is explained below.

1. Consider an initial upper bound for the median problem and the empty orders O^*_M, O^*_A, O^*_B and O^*_C.
2. Construct O^*_M by adding a gene g_M at the end of O^*_M, and construct O^*_A and O^*_B by inserting the genes of O_A and O_B that are descendant of g_M in the right positions. Moreover, if g_M is the descendant of a gene g_C that is not in O^*_C, then construct O^*_C by inserting this gene, otherwise O^*_C is unchanged.
3. Compute $DI(O^*_M, O^*_A)$, $DI(O^*_M, O^*_B)$ and $DI(O^*_C, O^*_M)$, using the branch-and-bound algorithm described in Section 3.1.
4. If $DI = DI(O^*_M, O^*_A) + DI(O^*_M, O^*_B) + DI(O^*_C, O^*_M)$ is lower than the current upper bound then: if O_M is of size n then replace the current upper bound by DI, otherwise go back to step 2.
5. If DI is larger than the current upper bound or O_M is of size n, then stop extending O_M, and consider another possible gene for the last position of O_M, or backtrack to the preceding position if all genes have been considered for the last position.

4 Results

4.1 Branch-and-bound efficiency

To measure the efficiency of our branch-and-bound algorithm, we simulated 1,000 DSI histories, each involving i inversions an a unique speciation event, leading to two contemporary genomes (TAGs clusters) of 15 genes and an implicit median containing k genes. Table 1 contains the execution times (on a standard PC) and the average fraction of the search space explored for different values of i and k.

We observe that the algorithm performance depends significantly on the number of inversions and on the ancestral order size. Nevertheless, it can be used on realistic datasets within reasonable time (45 seconds on average for a history implying an ancestral order of 12 genes and a total of 8 inversions).

Table 1. Execution time (in minutes) for the 1,000 ordered forest of gene tree / Average fraction of the search space explored during the branch-and-bound.

	Median size			
	6 genes	8 genes	10 genes	12 genes
4 inversions	3 / 2×10^{-3}	4 / 3×10^{-5}	5 / 2×10^{-7}	9 / 9×10^{-10}
6 inversions	8 / 5×10^{-3}	14 / 1×10^{-4}	35 / 1×10^{-6}	95 / 8×10^{-9}
8 inversions	15 / 1×10^{-2}	46 / 4×10^{-4}	182/ 6×10^{-6}	723 / 5×10^{-8}

4.2 Application on biological data

As a first application, we used our branch-and-bound algorithm to infer an ancestral gene order for a pair of human and rat olfactory TAGs clusters. The results are shown in Fig 4. We see that this dataset is compatible with an optimal DLSI history containing only one inversion event, that occurred before the human-rat speciation.

The human cluster is located on chr14@21.2 and the rat cluster on chr15@27.9. Protein sequences and gene orders were obtained from the HORDE database(CLIC #35) [21]. The sequences were aligned with clustalW [28] and the gene tree generated with MrBayes, [12] using the Jones substitution matrix [17] and performing 1,000,000 MCMC iterations.

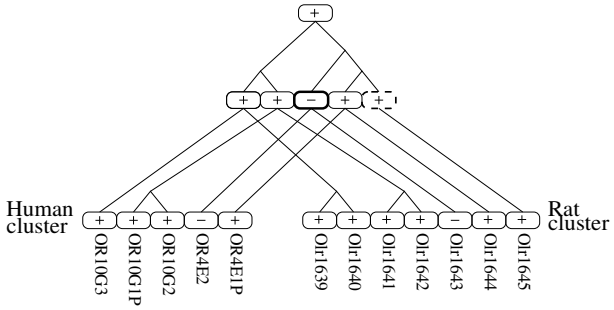

Fig. 4. The ordered reconciled gene tree obtained for the pair of olfactory receptor TAGs clusters, and the inferred ancestral gene order at the time of human-rat speciation. Transcriptional orientations are indicated by signs. The unique inversion occurred before human-rat speciation and is indicated by a black contour. The rightmost gene in the ancestral TAGs cluster (dashed contour) has its unique descendant in the rat TAGs cluster, indicating a gene loss in the human lineage after the speciation.

This first "simple" application only aims to give an example of a TAGs cluster which is very likely to have evolved in agreement with our model of evolution restricted to simple duplications and inversions, demonstrating its validity.

5 Conclusion

We have presented a formal approach to infer the ancestral gene orders inducing a most parsimonious scenario of inversions in the evolution of a TAGs family in multiple species.

The next important step would be the extension of the model to multiple duplications. However, gene losses are no longer independent from the duplication events in this case [10]. Inferring a tandem duplication tree with multiple duplications and gene losses remains an open problem, even when inversions are not taken into account and only one species is considered.

In addition to our model being restricted to simple duplications, the main problem we face with the inference of TAGs evolutionary histories is the difficulty to obtain a reliable gene tree for some families: Events such as gene conversions and unequal crossover can create "mosaic" genes that share more than one ancestor, and pseudogenization is a frequent process. Nevertheless, different strategies could be used to cope with these problems and produce biological knowledge from the present model. For example, the gene tree inference can be facilitated by excluding the pseudogenes of the analysis, and the signal noise can be reduced by choosing closely related species and excluding the period of time that precedes the first speciation from the analysis.

References

1. B. Arden, S.P. Clark, D. Kabelitz, and T.W. Mak. Human T-cell receptor variable gene segment families. *Immunogenetics*, 42(6):455–500, 1995.
2. G. Benson and L. Dong. Reconstructing the duplication history of a tandem repeat. In *Proceedings of Intelligent Systems in Molecular Biology (ISMB1999), Heidelberg, Germany*, pages 44–53. AAAI, 1999.
3. A. Bergeron, J. Mixtacki, and J. Stoye. Reversal distance without hurdles and fortresses. In *15th Symposium on Combinatorial Pattern Matching*, volume 3109 of *LNCS*, pages 388 - 399. Springer-Verlag, 2004.
4. D. Bertrand and O. Gascuel. Topological rearrangements and local search method for tandem duplication trees. *IEEE Transactions on Computational Biology and Bioinformatics*, pages 15–28, 2005.
5. D. Bertrand, M. Lajoie, N. El-Mabrouk, and O. Gascuel. Evolution of tandemly repeated sequences through duplication and inversion. In *Fourth RECOMB International Workshop on Comparative Genomics*, volume 4205 of *LNBI*, pages 129-140. Springer, 2006.
6. C. Chauve, J.F. Doyon, and N. El-Mabrouk. Inferring a duplication, speciation and loss history from a gene tree. In *Fifth RECOMB International Workshop on Comparative Genomics*, 2007. submitted.
7. J.A. Cotton and R.D.M. Page. Going nuclear: gene family evolution and vertebrate phylogeny reconcilied. In *Proc. R. Soc. Lond. B*, volume 269, pages 1555-1561, 2002.
8. E. Eichler and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793-797, 2003.
9. O. Elemento and O. Gascuel. A fast and accurate distance-based algorithm to reconstruct tandem duplication trees. *Bioinformatics*, 18:92–99, 2002.
10. O. Elemento, O. Gascuel, and M-P. Lefranc. Reconstructing the duplication history of tandemly repeated genes. *Molecular Biology and Evolution*, 19:278–288, 2002.
11. O. Eulenstein, B. Mirkin, and M. Vingron. Comparison of annotating duplication, tree mapping, and copying as methods to compare gene trees with species trees. In *Mathematical hierarchies and biology*, volume 37 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, 1997.
12. J.P. Huelsenbeck F. Ronquist. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–4, 2003.
13. W.M. Fitch. Phylogenies constrained by cross-over process as illustrated by human hemoglobins in a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I. *Genetics*, 86:623–644, 1977.
14. O. Gascuel, D. Bertrand, and O. Elemento. Reconstructing the duplication history of tandemly repeated sequences. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 205–235. OUP, 2005.
15. D.E. Geraghty, B.H. Koller, J.A. Hansen, and H.T. Orr. The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. *Journal of Immunology*, 149(6):1934–1946, 1992.
16. S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *J. ACM*, 48:1–27, 1999.
17. D. Jones, W. Taylor, and J. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3):275–282, 1992.
18. H. Kaplan, R. Shamir, and R. E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29:880–892, 2000.
19. B. Ma, M. Li, and L. Zhang. From gene trees to species trees. *SIAM J. Comput.*, 30(3):729-752, 2000.
20. B. Moret, J. Tang, L. Wang, and T. Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *Journal of Computer and System Science*, 65(3):508–525, 2002.
21. R. Aloni R, T. Olender, and D. Lancet. Ancient genomic architecture for mammalian olfactory receptor clusters. *Genome Biology*, 7(10):R88, 2006.
22. D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5:555–570, 1998.
23. M. Shannon, A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. Differential expension of Zinc- Finger transcription factor loci in homologous human and mouse gene clusters. *Genome Research*, 13:1097 - 1110, 2003.
24. V. Shoja and L. Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Molecular Biology and Evolution*, 23(11):2134- 2141, 2006.
25. M. Tang, M.S. Waterman, and S. Yooseph. Zinc finger gene clusters and tandem gene duplication. In *Proceedings of International Conference on Research in Molecular Biology (RECOMB2001)*, pages 297–304, 2001.
26. G. Tesler. GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492 - 493, 2002.
27. S. Huntley *et al.* A comprehensive catalogue of human krab-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Research*, 16:669–677, 2006.
28. J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673 - 4680, 1994.
29. J. Zhang and M. Nei. Evolution of antennapedia-class homeobox genes. *Genetics*, 142(1):295–303, 1996.
30. L. Zhang, B. Ma, L. Wang, and Y. Xu. Greedy method for inferring tandem duplication history. *Bioinformatics*, 19:1497–1504, 2003.