Inferring Ancestral Gene Orders for a Family of Tandemly Arrayed Genes

Denis Bertrand*, bertrden@iro.umontreal.ca, (514) 343-6111#3479,
Mathieu Lajoie, lajoimat@iro.umontreal.ca, (514) 343-6111#3479,
Nadia El-Mabrouk mabrouk@iro.umontreal.ca, (514) 343-7481.

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 TAG cluster. However, the presence of different transcriptional orientations in many clusters highlights the fact that processes such as inversions also contribute to their evolution. Moreover, existing algorithms are restricted to the study of TAGs evolution in a single species (only paralogous genes are considered). To circumvent these limitations, we consider an evolutionary model for TAGs involving duplication, gene loss, inversion and speciation events. A general framework to infer ancestral gene orders that minimize the number of inversions in the whole evolutionary history is presented. At the methodological level, this paper integrates three approaches to genome evolution: the duplication tree reconstruction, the gene tree/species tree reconciliation theory, and the concept of inversion median used in order-based phylogeny reconstruction. An application on a cluster of olfactory receptor genes in 4 mammals is presented.

Keywords: gene family, gene order, inversion, tandem duplication, reconciliation.

1 Introduction

A multigene family is a set of genes that have evolved by duplication and speciation 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 on a single or many chromosomes. In this paper, we focus on clusters of tandemly arrayed genes (TAG): copies that are adjacent on the chromosome. TAGs have been shown to represent a large proportion of genes in mammalian

*The two first authors contributed equally to this work

genomes. In particular, they represent about 14-17% of all genes in human, mouse and rat (Shoja and Zhang, 2006). Clusters of TAGs may vary in size from two to hundreds genes, though small clusters are largely predominant (an average of 3 to 4 genes in mouse, rat and human) (Shoja and Zhang, 2006). They are involved in many functions of binding or receptor activities. In particular, the olfactory receptor genes constitute the largest multigene family in vertebrate genomes, with several hundred genes per species (Aloni *et al.*, 2006). Other families of TAGs include the HOX genes (Zhang and Nei, 1996), the immunoglobulin and T-cell receptor genes (Arden *et al.*, 1995), the MHC genes (Geraghty *et al.*, 1992) and the Zinc Finger genes (Shannon *et al.*, 2003).

TAGs are widely viewed as resulting from unequal recombination during meiosis (Fitch, 1977), generating clusters of similar genes with the same transcriptional orientation. When fixed in a genome, such duplicates increase the chance of giving rise to other mispairings, thus leading to other duplicates.

Several studies have considered the problem of inferring an evolutionary history for a TAG cluster (Tang *et al.*, 2002; Elemento *et al.*, 2002; Elemento and Gascuel, 2002; Jaitly *et al.*, 2002; Zhang *et al.*, 2003; Bertrand and Gascuel, 2005). 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. Such trees are called *duplication trees*. When a gene tree is already available for a TAG cluster, a linear-time algorithm can be used to check whether it is a duplication tree (Gascuel *et al.*, 2003; Zhang *et al.*, 2003). As the probability for an arbitrary gene tree to be a duplication tree is very low (2.10^{-5} for a random tree with 15 leaves (Gascuel *et al.*, 2003)), the fact that a gene tree is a duplication tree is a strong argument in favor of the tandem duplication model of evolution for the associated gene family. However, it is often impossible to reconstruct a duplication history for a TAG cluster (Gascuel *et al.*, 2005), even from well supported gene trees. This is due to the occurrence of other mechanisms, such as deletions and genomic rearrangements (Eichler and Sankoff, 2003), during the evolution of the gene family. In particular, Shoja and Zhang (2006) have observed that more than 25% of all neighboring pairs of TAGs in human, mouse and rat have non-parallel orientations. This highlights the fact that other mechanisms, such as inversions, should be considered in an evolutionary model of TAGs. In a previous publication (Lajoie *et al.*, 2007b), we have presented an algorithm allowing to find the minimum number of inversions involved in the evolutionary history of a TAG cluster, assuming single gene duplications.

An important restriction of the above models of evolution is the fact that they are limited to the analysis of a TAG cluster located in a single species and on a single chromosome. However, the increasing availability of complete genomic sequences and of many different TAG databases (Aloni *et al.*, 2006; Huntley *et al.*, 2006) 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 on different

TAG families such as the Zinc-Finger genes in human and mouse (Shannon *et al.*, 2003), and the olfactory receptor genes in various mammalian species (Aloni *et al.*, 2006). However, no rigorous approach has been developed so far to explain the non agreement between a gene tree of a TAG family and a duplication and speciation history.

This paper is the first attempt to account for tandem duplication, speciation, gene loss and inversion events in an evolutionary model of TAGs. Given the gene and species trees for a set of orthologous TAG clusters and their respective gene orders, we aim to infer the *ancestral* gene orders leading to a most parsimonious sequence of evolutionary events. Clearly, an important prerequisite is to have, as an input, a well supported gene tree. This is unrealistic in the framework of "concerted evolution", where all the members of a gene family are assumed to evolve in a concerted manner by repeated occurrences of gene conversions. Hopefully, evidences for many TAG families (e.g. MHC, immunoglobulin and olfactory receptor genes) is in favor of a "birth-and-death" model of evolution (Nei and Rooney, 2005), in which gene conversion is much less important than previously believed.

At the methodological level, this paper integrates three approaches to genome evolution: the duplication tree reconstruction, the gene tree/species tree reconciliation, and the concept of inversion median used in order-based phylogeny reconstruction. We proceed in two steps. First, ignoring gene orders, a classical gene tree/species tree reconciliation method is used to infer a "minimal" duplication, speciation and loss history in agreement with a known species tree (Page, 1994). Second, we infer the ancestral gene orders allowing to minimize the number of inversions required to obtain a valid duplication tree. This problem is related to the more classical one of inferring gene orders of the ancestral genomes in a species tree (Sankoff and Blanchette, 1998; Bourque and Pevzner, 2002; Moret *et al.*, 2002; Ma *et al.*, 2006).

This paper is organized as follows. We describe the evolutionary model in Section 2 and our optimization problem in Section 3. The general iterative method used for minimizing the inversions in a whole species tree is then presented in Section 4. The detailed algorithm used for a single branch is then presented in Section 5. In Section 6 we present an exact branch-and-bound algorithm and a heuristic to solve the median problem. In Section 7, we compare the running times and the accuracy of our algorithms on different simulated data sets. Finally, an application on a set of orthologous TAG clusters in four mammalian species is presented.

2 The evolutionary model

The classical model of evolution considered for TAGs is based on tandem duplications resulting from unequal recombination during meiosis, which together with point mutations are assumed to be the sole evolutionary mechanisms 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.

Consider a set of m orthologous TAG clusters located on m different genomes. We denote by $\mathcal{O} = \{O_1, O_2, \ldots, O_m\}$ the set of gene orders, 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, a gene tree can be inferred from the TAG sequences. In this paper, a *gene tree* T for a TAG family is a rooted binary tree with labeled leaves, where each label represents a gene copy. A leaf labeled 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 (see Figure 1(a)).

We denote by $d_{inv}(O_i, \widehat{O}_i)$ the inversion distance between two orders O_i and $\widehat{O}_i$ on the same set of genes. Such a distance can be computed using the original Hannenhalli and Pevzner (1999) algorithm, or any of the existing optimizations (Kaplan *et al.*, 2000; Bader *et al.*, 2001; Bergeron *et al.*, 2004).

The following is a formal definition of a Duplication, gene Loss, Inversion and Speciation history (DLIS-history) leading to an ordered gene tree $(T, \mathcal{O})$ (see Figure 1(b) for an illustration).

Definition 1 *A DLIS-history of $(T, \mathcal{O})$ is a sequence of ordered gene trees $\mathcal{H} = ((T^1, \mathcal{O}^1), (T^2, \mathcal{O}^2), \ldots, (T^h, \mathcal{O}^h))$ where:*

1. *T^1 is a tree consisting of a single leaf v and $\mathcal{O}^1 = \{O_1^1\} = \{(\pm v)\}$ is one of the two trivial orders.*

2. *For $1 \leq k < h$, there is a unique i such that exactly one of the four following situations holds:*

   a. Duplication event: *T^{k+1} is obtained from T^k by adding two children u and w to a leaf v belonging to genome i. Moreover $\mathcal{O}^{k+1}$ is obtained from $\mathcal{O}^k$ by replacing v by (u, w) in O_i^k, where u and w have the same sign as v.*

   b. Gene loss event: *T^{k+1} is obtained from T^k by removing a leaf v belonging to genome i. If v was the only leaf in O_i^k then $\mathcal{O}^{k+1} = \mathcal{O}^k \setminus \{O_i^k\}$, otherwise $\mathcal{O}^{k+1}$ is obtained from $\mathcal{O}^k$ by deleting v from O_i^k.*

   c. Inversion event: *$T^{k+1} = T^k$ and $d_{inv}(O_i^k, O_i^{k+1}) = 1$.*

   d. Speciation event: *T^{k+1} is obtained from T^k by adding two children u_j and w_j to each leaf v_j belonging to genome i. Moreover, $\mathcal{O}^{k+1}$ is obtained from $\mathcal{O}^k$ by replacing $O_i^k = (v_1, \ldots, v_t)$, by $\{(u_1, \ldots, u_t), (w_1, \ldots, w_t)\}$, where u_j and w_j have the same sign as v_j.*

3. *$(T, \mathcal{O}) = (T^h, \mathcal{O}^h)$.*

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

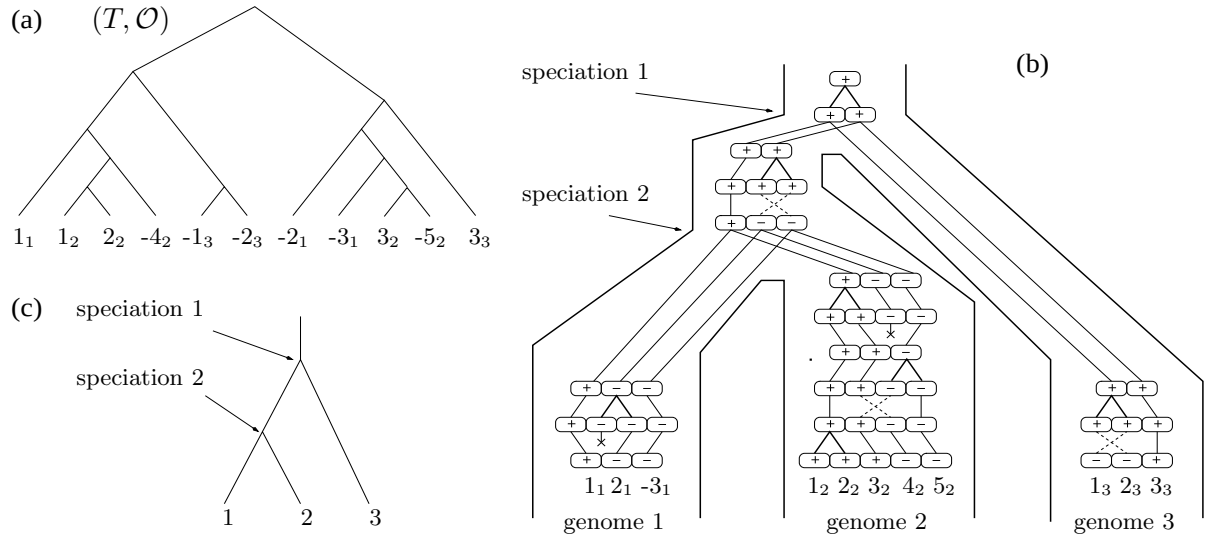

Figure 1: (a) An ordered gene tree (T, $\mathcal{O} = \{(1_1, -2_1, -3_1), (1_2, 2_2, 3_2, -4_2, -5_2), (-1_3, -2_3, 3_3)\}$). Genes are denoted as g_i meaning "the gth gene in genome i". (b) A DLIS-history for $(T, \mathcal{O})$. Duplications are indicated by bold lines, gene losses by 'X' and inversions by dashed lines. For clarity, we omitted successive identical configurations in each lineage. (c) The induced species tree for the three genomes.

3 An inference problem

Let $(T, \mathcal{O})$ be an ordered gene tree for a family of TAGs on m genomes. Suppose that a species tree S is already known for the m genomes. Then a natural problem is to find a DLIS-history of $(T, \mathcal{O})$ that is consistent with S. By Lemma 1, such a history exists. It follows from the existence of a duplication/speciation/loss history of T consistent with S in the general case of an unordered gene family. In this context, the *reconciliation* approach, first introduced by Goodman *et al.* (1979), and subsequently developed by many other authors (Page, 1994; Guigó *et al.*, 1996; Ma *et al.*, 2000; Bonizzoni *et al.*, 2005), allows to reconstruct such a history with a minimum number of duplication and/or loss events. The different reconciliation approaches are all based on a particular mapping (Least Common Ancestor mapping) from the nodes of T to the nodes of S, allowing to "embed" the gene tree into the species tree.

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 DLIS-history of $(T, \mathcal{O})$ consistent with S.*

Proof: Obtain a sequence of duplications, gene losses and speciations from the reconciliation of T and S. From that sequence, construct a DLIS-history $\mathcal{H}' = ((T^1, \mathcal{Q}^1), \ldots, (T^h = T, \mathcal{Q}^h))$ by applying the operations described in cases a, b and d of Definition 1. Then, obtain $\mathcal{H}$ from $\mathcal{H}'$ by performing any sequence of inversions transforming $\mathcal{Q}^h$ into $\mathcal{O}$ (case c in Definition 1). □

As the number of possible DLIS-histories consistent with S is unlimited, reasonable criteria should be considered. Here, we restrict ourselves to the most parsimonious DLIS-histories that are in agreement with a given reconciled tree.
We proceed in two steps:

1. We obtain a reconciled tree G from T and S. In the present study, we used the parsimony method of Zhang (1997), but any other method could be used (e.g. Arvestad *et al.* (2004) and Wapinski *et al.* (2007)).

2. We find the ancestral gene orders that minimize the total number of inversions involved in a DLIS-history of $(T, \mathcal{O})$. Formally, the problem considered in this step is the following:

MINIMUM-DLIS PROBLEM
Input: An ordered reconciled tree $(G, \mathcal{O})$.
Output: A gene order for each ancestral genome inducing a DLIS-history of minimum inversions.

In the rest of this paper, we focus on solving the MINIMUM-DLIS PROBLEM. We further introduce some precisions about the nodes of G and their implicit mapping to S (see Figure 2):

- A *duplication node* is an internal node which corresponds to a duplication event. It maps to a branch of S, i.e. the lineage in which the duplication occurred.

- A *speciation node* is an internal node which corresponds to an ancestral gene at the time of a speciation event. It maps to an internal node 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.

- A leaf is an extant gene and maps to a leaf of S, i.e. the extant genome to which it belongs.

- A maximal set of speciation nodes or leaves mapped to the same node A of S is defined as the *gene content* of A. When this set is ordered, we denote it by O_A.

- Let ρ be a direct descendant of a speciation node r. Then, the subtree rooted at ρ is said to be *externally rooted* at r.

From now on, we consider the "embedded" representation of G in S. More precisely, a *branch* (R, L) *of* S will denote the set of subtrees in G connecting the gene contents of R and L (see Figure 2).

Suppose that the gene content is ordered for each node of S. Then there exists a DI-history (a history restricted to duplication and inversion events) with a minimum number of inversions explaining each branch of S, and the minimum number of inversions in any DLIS-history of $(G, \mathcal{O})$ (and its corresponding ancestral gene orders) is the sum of the inversions involved in those minimal DI-histories. Thus, our problem reduces to the one of finding the ancestral gene orders minimizing

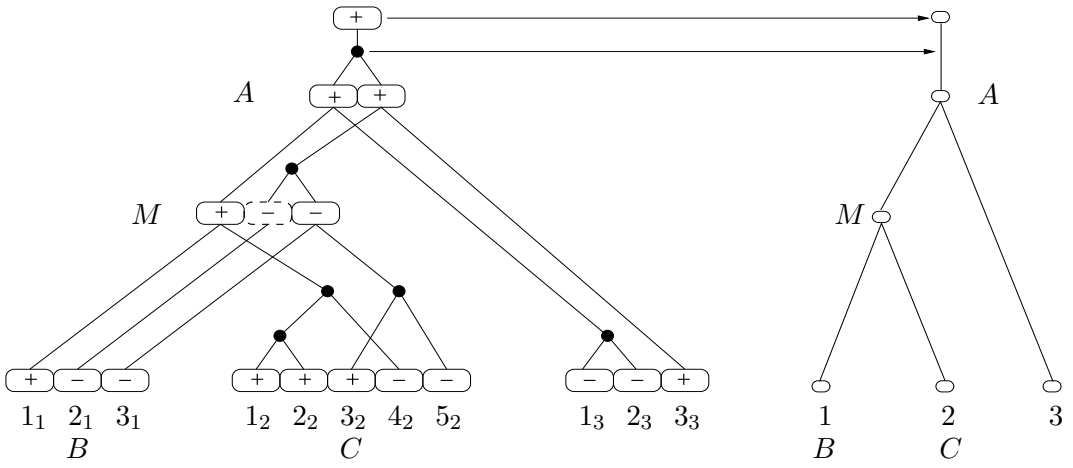

Figure 2: On the left, the ordered reconciled tree $(G, \mathcal{O})$ induced by the DLIS-history of Figure 1, with the corresponding ancestral gene orders. We see that each duplication node (black dot) in G implicitly maps to an *edge* of the species tree S (on the right), and each speciation node (box) to a *node* of S. The dashed gene in genome M has no descendants in lineage C, indicating a gene loss.

the sum of inversions involved in a DI-history of each branch of S. The formal definitions of a branch of S and a DI-history are given in Section 5.

4 A general method based on the median problem

The Minimum-DLIS problem is related to the more classical one of inferring the gene orders at the internal nodes of a species tree, where each leaf is labeled by an ordered sequence of genes (see for example Sankoff and Blanchette (1998), Bourque and Pevzner (2002) and Moret *et al.* (2002)). After fixing the ancestral gene contents, which is an intricate problem in the general case of unequal gene content and gene paralogy, the problem is to find the ancestral gene orders minimizing a given genomic distance.

Although the case of an ordered reconciled tree G has the additional constraint of tandem duplications, the two problems are related, suggesting a similar global approach summarized below.

1. Begin with an initial order O_M for each internal node M of S.

2. Traverse S in a depth-first manner. For each subtree consisting of a branch (A, M), where A is the immediate ancestor of M, and two sister branches (M, B) and (M, C) (see Figure 2), ignore the assigned order for M, and reconstruct an order that minimizes the value:

$$DI(O_A, O_M) + DI(O_M, O_B) + DI(O_M, O_C),$$

where $DI(O_R, O_L)$ is the minimum number of inversions in a DI-history explaining the branch (R, L). This step consists in solving the well known *median problem*.

3. Iterate Step 2. a given number of times, or until convergence to a local minimum.

In case no duplication and no gene loss have occurred during the evolution of the gene family along the branches from A to B and C, the DI value becomes the inversion distance, and the median problem formulated in Step 2 is just the *inversion median problem*, which has been proved to be NP-hard (Caprara, 2003). Therefore, the "generalized median problem" considered here is also NP-hard.

A rigorous definition and computation of $DI(O_R, O_L)$ for a branch (R, L) is given in the next section. We then present an exact algorithm and a heuristic to solve the median problem in Section 6. Finally, we present a heuristic allowing to begin with appropriate initial orders.

5 The generalized Minimum-DI problem

5.1 Definitions

We consider the problem of minimizing the number of inversions involved in a history explaining a given branch (R, L) of S when the gene contents are ordered. Formally, such a branch is called an *ordered forest* and is defined as follows:

Definition 2 *An* ordered forest (F, O_R, O_L) *is a forest of* n *gene trees* $F = \{T_1, T_2, \ldots, T_n\}$ *externally rooted at* $O_R = (r_1, r_2, \ldots, r_n)$*, with an order* O_L *on its leaves.*

We now formally define the notion of a DI-history explaining a given branch (R, L) of S. It is a generalization of the definition introduced in our previous paper (Lajoie *et al.*, 2007b) for a single ordered gene tree.

Definition 3 *A* DI-history *of an ordered forest* (F, O_R, O_L) *is a sequence of ordered forests* $\mathcal{H} = ((F^1, O_R, O_L^1), (F^2, O_R, O_L^2), \ldots, (F^h, O_R, O_L^h))$ *such that:*

1. F^1 *is a set of single leaf gene trees externally rooted at* O_R *and ordered as* $O_L = O_R$*.*
2. *For* $1 \leq k < h$*, exactly one the following situations holds:*
   a. Duplication event: F^{k+1} *is obtained from* F^k *by adding two children* u *and* w *to one of its leaf* v*, and* O_L^{k+1} *is obtained from* O_L^k *by replacing* v *by* (u, w)*, where* u *and* w *have the same sign as* v*.*
   b. Inversion event: $F^{k+1} = F^k$ *and* $d_{inv}(O_L^k, O_L^{k+1}) = 1$*.*
3. $(F, O_R, O_L) = (F^h, O_R, O_L^h)$*.*

From Definition 3, we also introduce the notion of a *duplication history*, which is simply a DI-history restricted to duplication events. A duplication history gives rise to a duplication forest, defined as follows.

Definition 4 *A* duplication forest *is an ordered forest* $(F = \{T_1, \ldots, T_n\}, O_R, O_L)$ *containing only* duplication trees*, and such that for every pair* (r_i, r_j) *in* O_R*, if* r_i *precede* r_j*, then all the leaves of* T_i *precedes all the leaves of* T_j *in* O_L*. Moreover, for any* $1 \leq i \leq n$*, the leaves of* T_i *have the same sign as* r_i*.*

The following theorem is a generalization of the result we obtained for a single ordered gene tree in one species (Lajoie *et al.*, 2007b).

Theorem 1 *For any DI-history of* (F, O_R, O_L) *with* i *inversions, there exists a duplication forest* $(F, O_R, \widehat{O}_L)$ *such that* $d_{inv}(O_L, \widehat{O}_L) \leq i$*.*

Proof: Let $\mathcal{H}^k = ((F^1, O_R, O_L^1), (F^2, O_R, O_L^2), \ldots, (F^k, O_R, O_L^k))$ be a DI-history of (F, O_R, O_L). We prove the theorem by induction on k:

- Base case: If $k = 1$, then $\mathcal{H}^1 = ((F^1, O_R, O_L^1))$ is a DI-history with no duplication and no inversion. Clearly $(F^1, O_R, \widehat{O}_L) = (F^1, O_R, O_L^1)$ is a duplication forest and $d_{inv}(O_L^1, \widehat{O}_L) = 0$.
- Induction step:
  Let $\mathcal{H}^{k+1} = ((F^1, O_R, O_L^1), \ldots, (F^k, O_R, O_L^k), (F^{k+1}, O_R, O_L^{k+1}))$ be a DI-history involving i inversions. From Definition 3, there are two possibilities:
  - If $F^{k+1} \neq F^k$, then the last event is a duplication, i.e. there is a leaf v of a tree of F^k that was replaced by two consecutive leaves u, w of the same sign in O_L^{k+1}. By the induction hypothesis, there exists a duplication forest $(F^k, O_R, \widehat{O}_L^k)$ such that $d_{inv}(O_L^k, \widehat{O}_L^k) \leq i$. Suppose v is positive in $\widehat{O}_L^k$. If v is also positive in O_L^k, we define $\widehat{O}_L^{k+1}$ as the order obtained by replacing $+v$ by $(+u, +w)$ in $\widehat{O}_L^k$. Otherwise, v is negative in O_L^k and we obtain $\widehat{O}_L^{k+1}$ by replacing $+v$ by $(+w, +u)$ in $\widehat{O}_L^k$. It follows that $d_{inv}(O_L^{k+1}, \widehat{O}_L^{k+1}) = d_{inv}(O_L^k, \widehat{O}_L^k) \leq i$ and $(F^{k+1}, O_R, \widehat{O}_L^{k+1})$ is a duplication forest. The case where v is negative in $\widehat{O}_L^k$ is treated similarly.
  - If $F^{k+1} = F^k$, then the last event is an inversion and $\mathcal{H}^k$ involves $i - 1$ inversions. By the induction hypothesis, there exists a duplication forest $(F^k, O_R, \widehat{O}_L^k)$ such that $d_{inv}(O_L^k, \widehat{O}_L^k) \leq i - 1$. Then we have $d_{inv}(O_L^{k+1}, \widehat{O}_L^{k+1}) \leq d_{inv}(O_L^k, \widehat{O}_L^k) + 1 \leq i$, where $\widehat{O}_L^{k+1} = \widehat{O}_L^k$. □

The following result immediately follows from Theorem 1.

Corollary 1 *Let* (F, O_R, O_L) *be an ordered forest and* $(F, O_R, \widehat{O}_L)$ *be a duplication forest such that* $d_{inv}(O_L, \widehat{O}_L) = i$ *is minimal over all* $\widehat{O}_L$*. Then, there exists a DI-history of* (F, O_R, O_L) *with exactly* i *inversions. Moreover,* i *is the minimum number of inversions in a DI-history of* (F, O_R, O_L)*.*

Corollary 1 allows to reformulate the problem as follows:

GENERALIZED-MINIMUM-DI PROBLEM
Input: An ordered forest (F, O_R, O_L).
Output: An order $\widehat{O}_L$ on the leaves of F such that $(F, O_R, \widehat{O}_L)$ is a duplication forest and $d_{inv}(O_L, \widehat{O}_L)$ is minimal.

Given a branch (R, L) of S and the orders O_R and O_L, the minimum number of inversions involved in a DI-history of the branch (R, L) is denoted as $DI(O_R, O_L)$. In the following section, we present an algorithm for solving the GENERALIZED-MINIMUM-DI PROBLEM.

5.2 A Branch-and-Bound algorithm

The algorithm is a generalization of the one we presented in a previous paper (Lajoie *et al.*, 2007b). Given an ordered gene tree (T, O), the goal was to find an order $\widehat{O}$ such that $(T, \widehat{O})$ is a duplication tree and $d_{inv}(O, \widehat{O})$ is minimal.

Ordered gene tree: As mentioned by Gascuel *et al.* (2005), simple duplication trees are equivalent to binary search trees. Therefore, to enumerate all the orders $\widehat{O}$ such that $(T, \widehat{O})$ is a duplication tree, we associated a binary variable b_i to each internal node i of T. By setting $b_i = 0$, we make the left descendant leaves of i *smaller* than the right ones in $\widehat{O}$, whereas by setting $b_i = 1$ we makes them *larger*. If we assign these values by a post-order traversal of T, then each b_i value induces an adjacency between two of its descendant leaves in $\widehat{O}$.

Hence, $(T, \widehat{O})$ is a duplication tree iff $\widehat{O}$ is defined by an assignment of all the binary variables in T, and all its genes have the same sign ($+$ or $-$). If n is the number of leaves in T, this leads to 2^n distinct orders.

To avoid computing $d_{inv}(O, \widehat{O})$ for every possible order $\widehat{O}$, we considered a branch-and-bound strategy based on the following property: $d_{inv}(O, \widehat{O}) \geq n + 1 - c$, where n is the number of genes and c is the number of cycles in the *breakpoint graph* (Hannenhalli and Pevzner, 1999) of O and $\widehat{O}$. In this graph, each edge corresponds to an adjacency in one of the two orders (see Figure 3). The general idea is to bound c as we progressively add the edges induced by the assignment of a given b_i. More precisely, if at a given step we have e cycles and p remaining edges, we know that $c \leq e + p$ since each remaining edge can create at most one cycle. Therefore, we can use the following lower bound in a branch-and-bound strategy:

$$d_{inv}(O, \widehat{O}) \geq n + 1 - e - p.$$

Ordered forest: Generalization to an ordered forest (F, O_R, O_L) is straightforward. Indeed, let $(T_1, T_2, \ldots, T_t)$ be the trees in F ordered according to the order O_R of their external roots. From

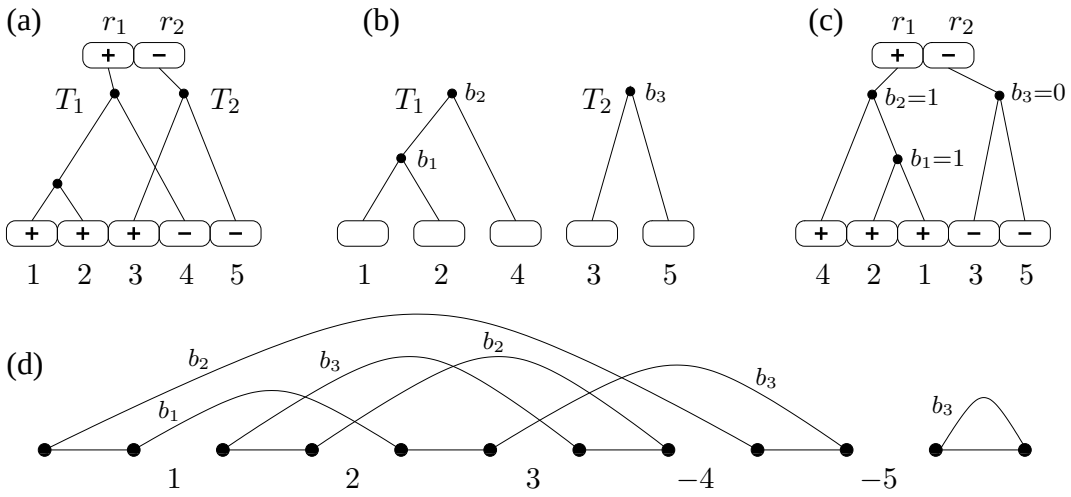

Figure 3: (a) The ordered forest corresponding to the branch (M, B) of the tree in Figure 2 ($F = \{T_1, T_2\}, O_R = (r_1, -r_2), O_L = (1, 2, 3, -4, -5)$). (b) The gene trees T_1 and T_2, with an arbitrary left/right orientation of the children at each internal node. (c) The duplication forest $(F, O_R, \widehat{O}_L = (4, 2, 1, -3, -5))$ induced by an assignment of the b_i variables. (d) The breakpoint graph of $\widehat{O}_L$ and O_L, with each curved edge labeled by the b_i inducing it, according to this assignment sequence: $(b_1 = 1, b_2 = 1, b_3 = 0)$.

Definition 4 and the discussion above, it is clear that $(F, O_R, \widehat{O}_L)$ is a duplication forest iff $\widehat{O}_L$ is the concatenation of the n orders $(\widehat{o}_1, \widehat{o}_2, \ldots, \widehat{o}_t)$ respectively defined by an assignment of the binary variables in $T_1, T_2, \ldots, T_t$, and for each $1 \leq j \leq t$, all the genes belonging to $\widehat{o}_j$ have the same sign as r_j. Consequently, we can enumerate the orders $\widehat{O}_L$ as above and the same bound can be used (see Figure 3 for an example).

6 The median problem

To formally define the median problem, we need to extend the notion of an ordered forest (Definition 2) by allowing the orders to be defined only on the leaves or only on the external roots of the trees. A *leaf-ordered forest* will be denoted as (F_{RL}, R, O_L) and a *root-ordered forest* as (F_{RL}, O_R, L).

The median problem is formulated as follows. Given a root-ordered forest (F_{AM}, O_A, M) and two leaf-ordered forests (F_{MB}, M, O_B) and (F_{MC}, M, O_C) (M is the set of ancestral genes generating both B and C), the goal is to find an order O_M minimizing the *median score*:

$$S(O_M) = DI(O_A, O_M) + DI(O_M, O_B) + DI(O_M, O_C)$$

6.1 A branch-and-bound algorithm

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

Property 1 *Let (F_{RL}^*, O_R^*, O_L^*) be an ordered forest obtained from (F_{RL}, O_R, O_L) by removing a tree rooted at the last element of O_R, or the leaf corresponding to the last element of O_L. Then:*

$$DI(O_R^*, O_L^*) \leq DI(O_R, O_L)$$

From the property above, it follows that:

$$S(O_M) \geq S(O_M^*) = DI(O_A^*, O_M^*) + DI(O_M^*, O_B^*) + DI(O_M^*, O_C^*)$$

An exact branch-and-bound strategy for solving the median problem is sketched below. **Algorithm BBM-DI** (Branch-and-Bound for the Median with DI distance):

1. Consider an initial candidate O_M. Define the empty orders O_M^*, O_A^*, O_B^* and O_C^*.
2. Add a gene $\pm g_M \in M$ to the end of O_M^*. Then, insert the descendants of g_M in O_B^* and O_C^* according to their positions and signs in O_B and O_C. Moreover, if g_M is the descendant of a gene $g_A \in A$ that is not yet in O_A^*, insert g_A in O_A^* according to its position and sign in O_A.
3. If $S(O_M^*) < S(O_M)$:
   - If $S(O_M^*)$ contains less then n genes, then return to Step 2.
   - Else $O_M \leftarrow O_M^*$.
4. Backtrack to Step 2 and consider another gene g_M (or sign) for the last position of O_M^*. When all the genes have been considered, backtrack one position left. When all the positions have been tried, stop and return O_M.

This branch-and-bound approach can be used with medians containing up to a dozen of genes (see Execution time in Section 7.1). For larger instances, we next present a fast and simple heuristic which yields good approximations when the number of inversions is low.

6.2 A simple heuristic for the median problem

The idea is to consider an initial order and optimize the median score locally by successive applications of *transposition* or *transversion*[1] on that order. It is similar to the exact algorithm of Siepel

[1] A transposition followed by an inversion.

and Moret (2001) except that our neighborhood is different, and we keep only the best candidates at each step. A local optimum is reached when no move can improve the median score. The algorithm is sketched bellow.

Algorithm LSM-DI (Local Search for the Median with the DI distance):

1. Consider an initial candidate median O_M. Set $S_{min} \leftarrow S(O_M)$.
2. For each of the $O(n^3)$ neighbor O_i of O_M:
   (a) Compute $S(O_i) = DI(O_A, O_i) + DI(O_i, O_B) + DI(O_i, O_C)$.
   (b) If $S(O_i) < S(O_M)$, then push O_i on the priority queue. Moreover, if $S(O_i) < S_{min}$, then set $S_{min} \leftarrow S(O_i)$.
3. As long as the priority queue contains an order O_i such that $S(O_i) = S_{min}$, set $O_M \leftarrow O_i$, remove O_i and return to Step 2.
4. Output O_M.

6.3 Getting the initial orders

The success of the above methods depends strongly on the choice of the initial candidates O_M. Our solution is to use a greedy version of the algorithm described in Section 5.2, but generalized to the *whole* reconciled tree G. More precisely, for each duplication node v of G, we set b_i to the value that maximizes the total number of cycles in the breakpoint graphs of the m extant genomes. Once all the b_i are defined, it is straightforward to obtain the orders in the ancestral genomes.

7 Results

7.1 Simulated data

Execution time

We measured the execution time of our general method for inferring ancestral orders (Section 4) using either the branch-and-bound (**BBM-DI**) or the heuristic (**LSM-DI**) for solving the median problem. Algorithms were implemented in C++ and run on a typical Linux workstation.

The ordered gene trees where obtained by simulating DLIS-histories consistent with balanced species trees with 2 or 4 leaves. The number of genes in the resulting genomes (extant or ancestral) depends uniquely on their depth in the species tree. Starting from the root which contains a unique ancestral gene, this number becomes respectively n, $\lfloor 3n/2 \rfloor$ and $\lfloor 9n/4 \rfloor$ as we reach depth 1, 2 and 3 (depth 3 applies only to species trees with 4 leaves). Inversion events are distributed evenly among the branches of the species trees and their cutting-points are chosen randomly.

Results are presented in Figure 4. We observe that the execution time of BBM-DI depends significantly on the number of inversions and rapidly becomes impractical. In contrast, LSM-DI can be used on relatively important datasets within reasonable time (100 seconds on average for a median of 30 genes with 12 inversions).

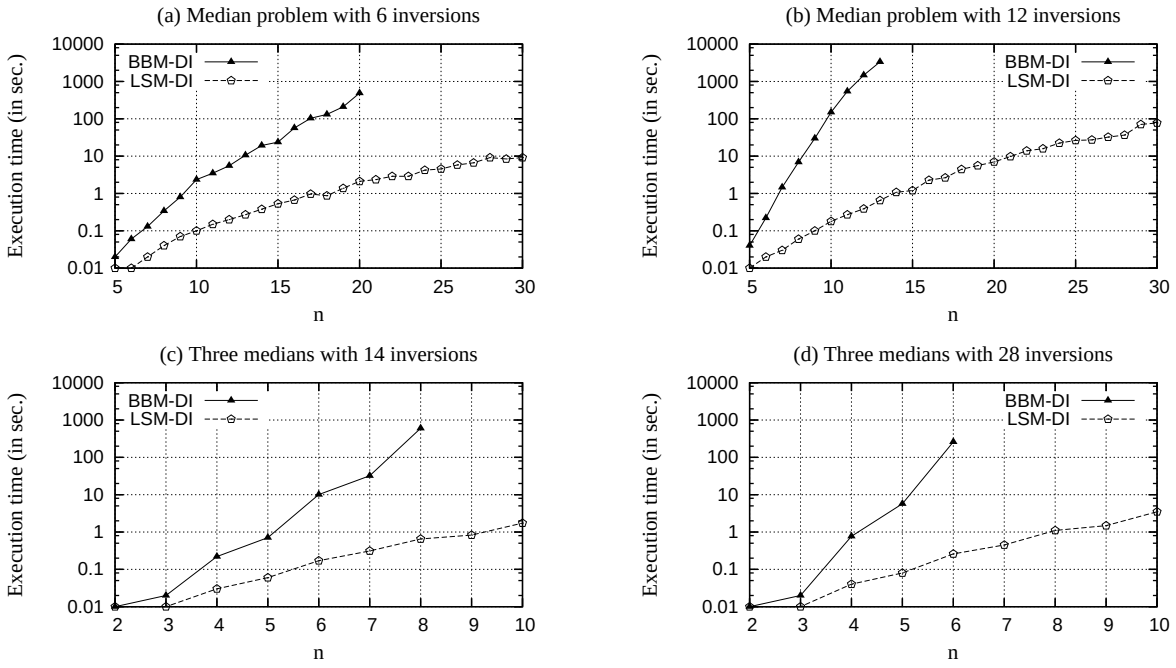

Figure 4: Average execution time in seconds on simulated data (50 replicates). (a and b) One ancestral genome with n genes and two extant genomes each containing $\lfloor 3n/2 \rfloor$ genes. (c and d) Three ancestral genomes containing respectively n, $\lfloor 3n/2 \rfloor$ and $\lfloor 3n/2 \rfloor$ genes, and four extant genomes each containing $\lfloor 9n/4 \rfloor$ genes.

Algorithms Accuracy

We measured the accuracy of our general method for inferring ancestral orders on simulated data, using either the BBM-DI or LSM-DI for solving the median problem. Ordered gene trees were obtained as described above.

Accuracy is evaluated based on two criteria: the inferred number of inversions, and the inferred gene orders. Evaluation of the gene order is based on the percentage of adjacencies shared between the inferred order and the actual one. An inferred adjacency (a, b) is shared iff (a, b) or $(-b, -a)$ is in the actual order.

We observe in Figure 5(a and b) that the number of inversions inferred by LSM-DI is very close to the global minimum found by BBM-DI. However, the probability that this global minimum corresponds to the reality decreases when the actual number of inversions increases in the DLIS-

history. The same is observed for the percentage of correctly inferred adjacencies (see Figure 5(c and d)).

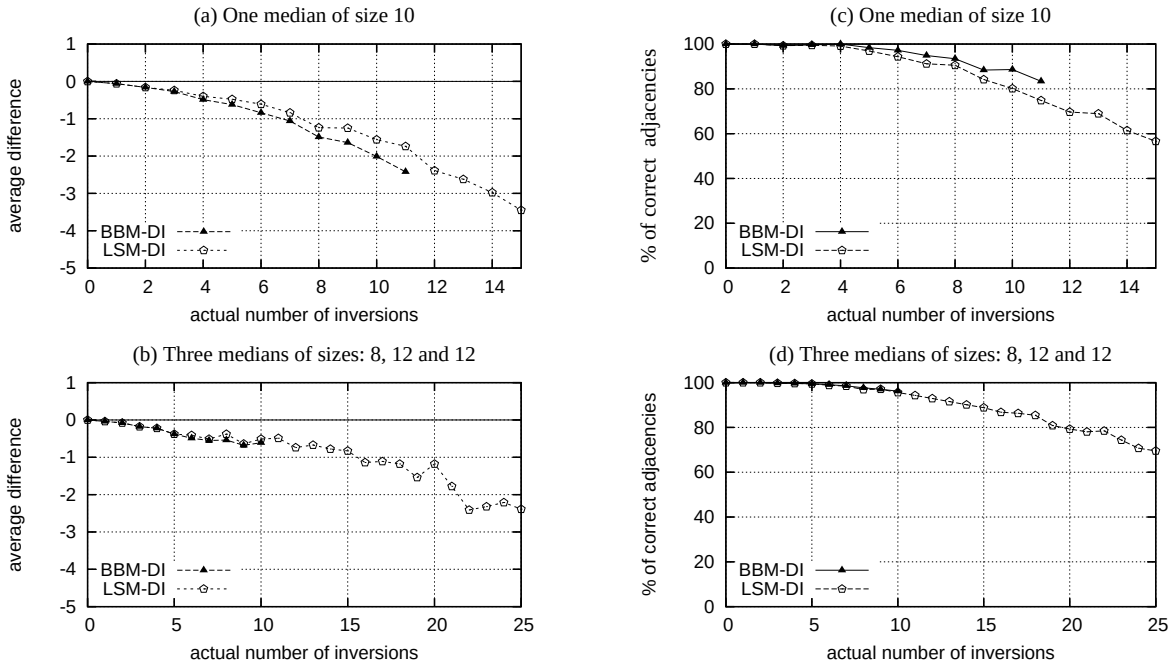

Figure 5: Comparison between BBM-DI and LSM-DI (100 replicates). (a and b) Average difference between the inferred number of inversions and the actual one (*inferred minus actual*). (c and d) Percentage of shared adjacencies between the inferred order and the actual one.

Effect of gene losses

To evaluate the effect of gene losses on the accuracy of LSM-DI, we generated appropriate DLIS-histories using a protocol similar to the one described above. Gene losses were distributed randomly among the branches of the species trees. Results are shown for correctly reconciled trees (80% and 60% respectively for 8 and 16 gene losses) in Figure 6. We see that gene losses have very little effect on the accuracy of our heuristic when the reconciled gene tree is correct.

Effect of double duplication (model deviation)

Recall that our DLIS model allows only simple tandem duplications. To measure the robustness of our inference method against model deviations, we simulated the evolution of orthologous clusters with a limited number of *double duplications*[2] (DD). As expected, we observe in Figure 7 that

[2]A double duplication simultaneously copies two adjacent genes as a single unit. For example, $O^k = (g_1, g_2, g_3, g_4)$ that becomes $O^{k+1} = (g_1, g_2, g_3, g'_2, g'_3, g_4)$.

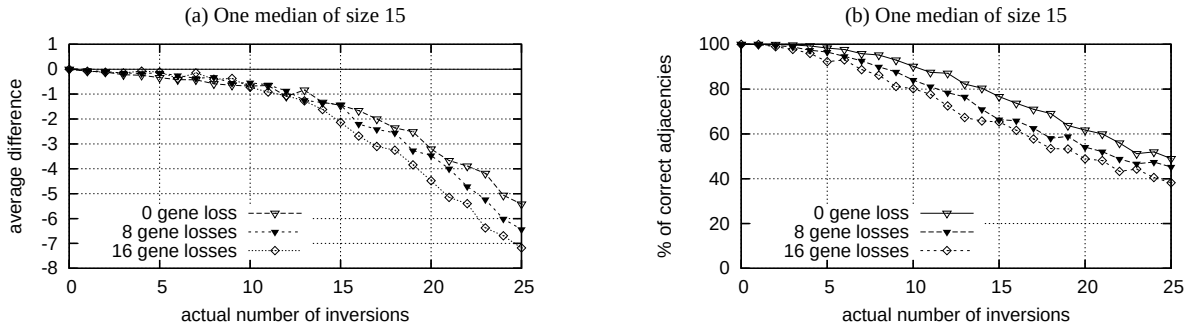

Figure 6: Accuracy of LSM-DI on ordered gene trees resulting from DLIS-histories with different numbers of gene losses (500 replicates). (a) Average difference between the inferred number of inversions and the actual one (*inferred minus actual*). (c) Percentage of shared adjacencies between the inferred order and the actual one.

LSM-DI largely overestimates the number of inversions when DD are introduced, especially when few inversions really occurred (one DD can produce as much as 3 *false* inversions). However, the effect on the percentage of correctly inferred adjacencies is much less important.

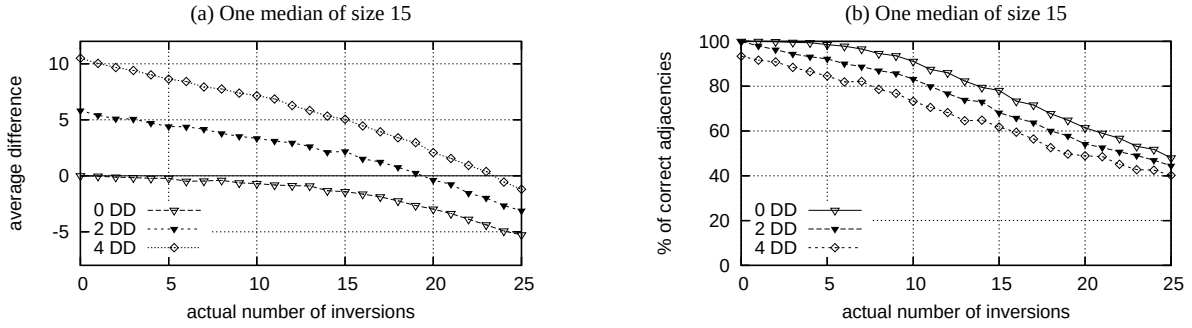

Figure 7: Accuracy of LSM-DI on ordered gene trees generated from DLIS-histories with different numbers of double duplications (500 replicates). (a) Average difference between the inferred number of inversions and the actual one (*inferred minus actual*). (c) Percentage of shared adjacencies between the inferred order and the actual one.

7.2 Application on biological data

The olfactory receptor (OR) gene family contains several hundred members in mammalian genomes, scattered in about 50 genomic clusters. We used our general method with LSM-DI to infer ancestral gene orders for one of these clusters, which is located on chr14@21.2 in the human genome. Four orthologous clusters were used in our study: human chr14@21.2; rat chr15@27.9; mouse chr14@47.5; opossum scaffold_19262@4.7. Protein sequences, gene orders and clusters orthology

were all obtained from CLIC#35 in the HORDE database (Aloni *et al.*, 2006). Human OR6Y1 gene was used as an outgroup. Sequences were aligned with ClustalW (Thompson *et al.*, 1994) and the gene trees with the largest posterior probability were obtained with MrBayes (Ronquist and Huelsenbeck, 2003), using the Jones-Taylor-Thornton substitution matrix (Jones *et al.*, 1992) and $500,000$ MCMC iterations.

The 16 most probable trees have a cumulative posterior probability of 0.8. For each of them, we obtained a reconciled gene tree with the RECONCILE software (Sennblad *et al.*, 2007) and used our general algorithm to infer ancestral gene orders and the corresponding number of inversions. The most parsimonious DLIS-histories were obtained with the fourth ($p = 0.09$) and the sixth ($p = 0.05$) most probable trees returned by MrBayes. Both involve a single inversion and no gene loss. Other trees involve 4.7 gene losses and 1.8 inversions on average. The fourth tree is presented in Figure 8. According to this tree, a unique inversion event occurred before the divergence between eutheria and marsupialia. We point out that this scenario differs slightly with the one we obtained previously by considering only the human and rat clusters which involves an additional gene loss (Lajoie *et al.*, 2007a).

This simple application gives an example of a TAG cluster which is very likely to have evolved in agreement with our model of evolution.

8 Conclusion

We have presented a general framework for studying the evolution of tandemly arrayed gene families in multiple genomes. It is the first formal approach to integrate inversion and speciation events in a tandem duplication model of evolution.

Our study has been placed in the context of a known species tree. In the case of an unknown species tree, an alternative method for constructing the preliminary reconciled tree should be considered. Different methods have been developed in the literature based on different measures: the *duplication cost model* (Ma *et al.*, 2000), the *mutation cost model* (Ma *et al.*, 2000) and the *minimum loss model* (Chauve *et al.*, 2007).

The methods we presented allow to infer ancestral gene orders minimizing (locally) the number of inversions for *a given reconciled tree*. However, this tree is not guaranteed to provide the minimum number of inversions for *any* DLIS-history compatible with the species tree. Finding a DLIS-history of minimum inversions remains an open problem.

We point out the difficulty of measuring the accuracy of the phylogenetic methods used to infer the gene trees, especially for TAG 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. Different strategies could be used to cope with these problems. For example, regions subject to gene conversions could be identified and excluded from the phylogenetic analysis, and the gene contexts could be considered. Pseudogenes could also be treated separately with more

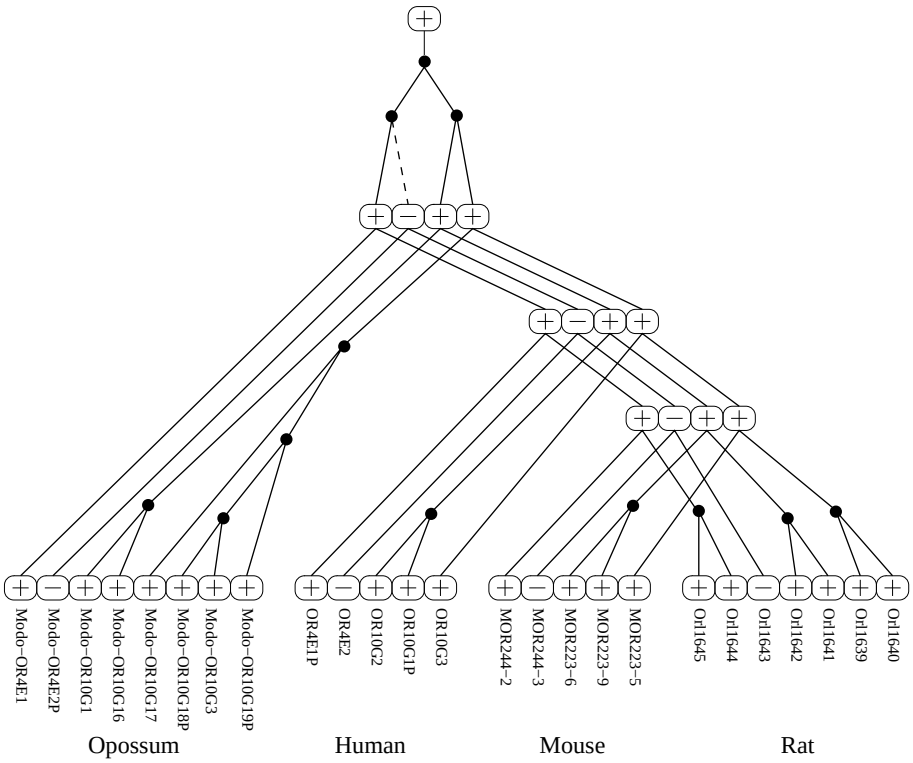

Figure 8: The ancestral gene orders inferred by our general method using LSM-DI on the orthologous clusters of CLIC#35. Transcriptional orientations are indicated by signs. The unique inversion event is indicated by the dashed edge. No gene loss was inferred by the reconciliation process.

appropriate methods and models, or ultimately discarded from the analysis. Despite these efforts, it would remain difficult to infer the correct gene tree for several TAG families.

In this context, the minimum number of inversions for a TAG family could be used as an additional criteria for the comparison of different candidate gene trees (Lajoie *et al.*, 2007b). Here again, results should be interpreted carefully since the actual model is limited to simple duplications. Although they are believed to be predominant, multiple duplications also occur in TAG evolution.

An important improvement would thus be the extension of our model to multiple duplications. This poses many challenges since 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.

Acknowledgments

This work was supported by grants from the "Fonds Québécois de la Recherche sur la Nature et les Technologies" (D.B. and N.E.M.), the Natural Sciences and Engineering Research Council of Canada (N.E.M.) and the Canadian Institutes of Health Research (M.L.).

References

Aloni, R., Olender, T., and Lancet, D., 2006. Ancient genomic architecture for mammalian olfactory receptor clusters. *Genome Biology* 7, R88.

Arden, B., Clark, S., Kabelitz, D., and Mak, T., 1995. Human t-cell receptor variable gene segment families. *Immunogenetics* 42, 455–500.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B., 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology (RECOMB04)*, 326–335. ACM.

Bader, D., Moret, B., and Yan, M., 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology* 8, 483–491.

Bergeron, A., Mixtacki, J., and Stoye, J., 2004. Reversal distance without hurdles and fortresses. In *15th Symposium on Combinatorial Pattern Matching (CPM04)*, *LNCS*, volume 3109, 388–399. Springer-Verlag.

Bertrand, D. and Gascuel, O., 2005. Topological rearrangements and local search method for tandem duplication trees. *IEEE Transactions on Computational Biology and Bioinformatics* 15–28.

Bonizzoni, P., Vedova, G., and Dondi, R., 2005. Reconciling gene trees to a species tree. *Theoretical Computer Science* 347, 36–53.

Bourque, G. and Pevzner, P., 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research* 12, 26–36.

Caprara, A., 2003. The reversal median problem. *Journal on Computing* 15, 93–113.

Chauve, C., Doyon, J., and El-Mabrouk, N., 2007. Inferring a duplication, speciation and loss history from a gene tree. In *Fifth RECOMB International Workshop on Comparative Genomics*, 45–57. Springer-Verlag.

Eichler, E. and Sankoff, D., 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* 301, 793–797.

Elemento, O. and Gascuel, O., 2002. A fast and accurate distance-based algorithm to reconstruct tandem duplication trees. *Bioinformatics* 18, 92–99.

Elemento, O., Gascuel, O., and Lefranc, M.-P., 2002. Reconstructing the duplication history of tandemly repeated genes. *Molecular Biology and Evolution* 19, 278–288.

Fitch, W., 1977. 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* 86, 623–644.

Gascuel, O., Bertrand, D., and Elemento, O., 2005. Reconstructing the duplication history of tandemly repeated sequences. In Gascuel, O., ed., *Mathematics of Evolution and Phylogeny*, 205–235. OUP.

Gascuel, O., Hendy, M., Jean-Marie, A., and McLachlan, S., 2003. The combinatorics of tandem duplication trees. *Systematic Biology* 52, 110–118.

Geraghty, D., Koller, B., Hansen, J., and Orr, H., 1992. The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. *Journal of Immunology* 149, 1934–1946.

Goodman, M., Czelusniak, J., Moore, G., Romero-Herrera, A., and Matsuda, G., 1979. Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology* 28, 132–168.

Guigó, R., Muchnik, I., and Smith, T., 1996. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution* 6, 189–213.

Hannenhalli, S. and Pevzner, P. A., 1999. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM* 48, 1–27.

Huntley, S., Baggott, D., Hamilton, A., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L., 2006. 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.

Jaitly, D., Kearney, P., Lin, G., and Ma, B., 2002. Methods for reconstructing the history of tandem repeats and their application to the human genome. *Journal of Computer and System Sciences* 65, 494–507.

Jones, D., Taylor, W., and Thornton, J., 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8, 275–282.

Kaplan, H., Shamir, R., and Tarjan, R. E., 2000. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing* 29, 880–892.

Lajoie, M., Bertrand, D., and El-Mabrouk, N., 2007a. Evolution of tandemly arrayed genes in multiple species. In *Fifth RECOMB International Workshop on Comparative Genomics*, 98–109. Springer-Verlag.

Lajoie, M., Bertrand, D., El-Mabrouk, N., and Gascuel, O., 2007b. Duplication and inversion history of a tandemly repeated genes family. *Journal of Computational Biology* 462–478.

Ma, B., Li, M., and Zhang, L., 2000. From gene trees to species trees. *SIAM Journal on Computing* 30, 729–752.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D., and Miller, W., 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Research* 16, 1557–1565.

Moret, B., Tang, J., Wang, L., and Warnow, T., 2002. Steps toward accurate reconstructions of phylogenies from gene-order data. *Journal of Computer and System Science* 65, 508–525.

Nei, M. and Rooney, A., 2005. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetic* 39, 121–152.

Page, R., 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* 43, 58–77.

Ronquist, F. and Huelsenbeck, J., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–4.

Sankoff, D. and Blanchette, M., 1998. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology* 5, 555–570.

Sennblad, B., Schreil, E., Sonnhammer, A. B., Lagergren, J., and Arvestad, L., 2007. Primetv: a viewer for reconciled trees. *BMC Bioinformatics* 148.

Shannon, M., Hamilton, A., Gordon, L., Branscomb, E., and Stubbs, L., 2003. Differential expansion of Zinc- Finger transcription factor loci in homologous human and mouse gene clusters. *Genome Research* 13, 1097–1110.

Shoja, V. and Zhang, L., 2006. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Molecular Biology and Evolution* 23, 2134–2141.

Siepel, A. and Moret, B., 2001. Finding an optimal inversion median: Experimental results. In *Proceedings of the 2nd International Workshop on Algorithms in Bioinformatics(WABI2003)*, Lecture Notes in Computer Science, 189–203. Springer.

Tang, M., Waterman, M., and Yooseph, S., 2002. Zinc finger gene clusters and tandem gene duplication. *Journal of Computational Biology* 429–446.

Thompson, J., Higgins, D., and Gibson, T., 1994. 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, 4673–4680.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A., 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 23, i549–i558.

Zhang, J. and Nei, M., 1996. Evolution of antennapedia-class homeobox genes. *Genetics* 142, 295–303.

Zhang, L., 1997. On a mirkin-muchnik-smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology* 4, 177–187.

Zhang, L., Ma, B., Wang, L., and Xu, Y., 2003. Greedy method for inferring tandem duplication history. *Bioinformatics* 19, 1497–1504.