Duplication and Inversion History of a Tandemly Repeated Genes Family

Mathieu Lajoie*, DIRO, Université de Montréal, H3C 3J7, Montréal QC, Canada,

Denis Bertrand, DIRO, Université de Montréal, H3C 3J7, Montréal QC, Canada

Nadia El-Mabrouk, DIRO, Université de Montréal, H3C 3J7, Montréal QC, Canada

Olivier Gascuel, LIRMM, UMR 5506, CNRS et Université Montpellier 2, Montpellier France

Abstract

Given a phylogenetic tree for a family of tandemly repeated genes and their *signed* order on the chromosome, we aim to find the minimum number of inversions compatible with an evolutionary history of this family. This is the first attempt to account for inversions in an evolutionary model of tandemly repeated genes. We present a branch-and-bound algorithm that finds the exact solution, and a polynomial-time heuristic based on the breakpoint distance. We show, on simulated data, that those algorithms can be used to improve phylogenetic inference of tandemly repeated gene families. An application on a published phylogeny of KRAB zinc finger genes is presented.

Keywords: gene family, gene order, inversion, duplication, phylogeny.

*The two first authors contributed equally to this work

1 Introduction

A large fraction of most genomes consists of repetitive DNA sequences. In mammals, up to 60% of the DNA is repetitive. A large proportion of such repetitive sequences is organized in tandem: copies of a same basic unit that are adjacent on the chromosome. The duplicated units can be small (from 10 to 200 bps) as it is the case of micro- and minisatellites, or very large (from 1 to 300 kb) and potentially contain several genes. The formation of those large duplicated sequences is widely assumed to be due to unequal recombination.

Many gene families are organized in tandem, including HOX genes (Zhang and Nei, 1996), immunoglobulin and T-cell receptor genes (Arden *et al.*, 1995), MHC genes (Geraghty *et al.*, 1992) and olfactory receptor genes (Glusman *et al.*, 2001). Reconstructing the duplication history of each gene family is important to understand the functional specificity of each copy, and to provide new insights into the mechanisms and determinants of gene duplication, often recognized as major generators of novelty at the genome level.

Both the linear order among tandemly repeated sequences, and the knowledge of the biological mechanisms responsible for their generation, suggest a simple model of evolution by duplication. This model, first described by Fitch (1977), introduces tandem duplication trees as phylogenies constrained by the unequal recombination mechanism. The main features of this model can be grasped from the examples of Figure 1. Figure 1(a) shows the duplication tree of the 13 Antennapedia-class homeobox genes (Zhang and Nei, 1996) which contains only simple duplication events (duplication of a segment containing only one gene). Starting from the unique ancestral gene, this series of events has produced the extant locus containing the 13 linearly ordered contemporary genes. As described by Elemento and Gascuel (2005), trees that contain only simple duplication events are equivalent to binary search trees with labeled leaves. Fitch model also allows for the simultaneous duplication of several gene copies, as observed in the duplication tree of the

9 variable genes of the human T cell receptor Gamma (TRGV) locus (Elemento *et al.*, 2002) (see Figure 1(b)). This duplication tree contains a double duplication where two adjacent genes have been simultaneously duplicated. Figure 1

Based on this model, a number of recent studies have considered the problem of reconstructing the tandem duplication tree of a gene family (Benson and Dong, 1999; Tang *et al.*, 2001; Elemento *et al.*, 2002; Elemento and Gascuel, 2002; Jaitly *et al.*, 2002; Zhang *et al.*, 2003; Bertrand and Gascuel, 2005; Elemento and Gascuel, 2005). These are essentially phylogenetic inference methods which compute the duplication tree that best explains the evolution of a gene family. When a phylogeny is already available, a linear-time algorithm can be used to check whether it is a duplication tree (Gascuel *et al.*, 2003; Zhang *et al.*, 2003). However, even for gene families that have evolved through tandem duplications, it is often impossible to reconstruct a duplication history (Gascuel *et al.*, 2005). This can be explained by the fact that the duplication model is oversimplified, and other evolutionary events have occurred, such as gene losses or genomic rearrangements.

Evidence of gene inversion is observed in many tandemly repeated gene families, such as zinc finger (ZNF) genes, where gene copies have different transcriptional orientations (Shannon *et al.*, 2003). Although genome rearrangement with inversions has received large attention in the last decade (Hannenhalli and Pevzner, 1999; El-Mabrouk, 2000; Kaplan *et al.*, 2000; Siepel, 2002; Bergeron *et al.*, 2004), inversions have never been considered in the context of reconstructing a duplication history from a gene tree. In the case of general segmental duplications (not necessarily in tandem), potential gene losses have been considered to explain the non congruence between a gene tree and a species tree (Guigó *et al.*, 1996; Page and Charleston, 1997; Ma *et al.*, 1998; Chen *et al.*, 2000). Similarly, in the case of tandem duplication, the non-congruence between a gene tree and an observed gene order can be naturally explained by introducing the possibility of segmental inversions.

In this paper, our goal is to infer an evolutionary history of a gene family accounting for both tandem duplications and inversions. As the number of such possible evolutionary

histories can be very large, we restrict ourselves to finding the minimum number of inversions required to explain a given ordered phylogeny. 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 (Zhang and Nei, 1996; Bertrand and Gascuel, 2005). As a first attempt, we only consider simple duplications.

After describing the evolutionary models in Section 2 and the optimization problem in Section 3, we present a branch-and-bound algorithm in Section 4. Then, in Section 5, we present a similar problem based on the breakpoint distance. This variant has a polynomial-time solution and can be used as an accurate heuristic to solve our original problem. Finally, in Section 6, we compare the time efficiencies of the two algorithms and show, using simulated data, there usefulness to improve phylogenetic inference. An application on a KRAB zinc finger gene family is presented.

2 The Evolutionary Model

2.1 Duplication Model

This model, first introduced by Fitch (1977), is based on unequal recombination during meiosis. The later is assumed to be the sole evolutionary mechanism, with point mutations, acting on sequences. However, the model is robust to gene conversion, as long as the phylogenetic signal remains strong enough to reconstruct the correct tree, which seems a realistic assumption for many tandemly repeated gene families. Indeed, from a single sequence, the locus grows through a series of consecutive duplications, giving rise to a sequence of n adjacent copies of homologous genes *having the same transcriptional orientation*. We denote by $O = (l_1, \cdots l_n)$ the observed ordered sequence of extant gene copies.

A *tandem duplication history* (or just *duplication history* for brevity) is the sequence of tandem duplications that have generated O. It can be represented by a rooted tree with n ordered leaves corresponding to the n ordered genes, in which internal nodes correspond to duplication events (Figure 2(a)). Duplications may be *simple* (duplication of a single gene) or multiple (simultaneous duplication of neighboring genes). In our duplication/inversion model, we consider only simple duplications. As mentioned previously, simple duplications are believed to be predominant over multiple duplications (Zhang and Nei, 1996; Bertrand and Gascuel, 2005), and there are examples of simple duplication trees in the literature, such as the one presented in Figure 1. Figure 2

In a real duplication history, the time intervals between consecutive duplications are known, and the internal nodes are ordered from top to bottom according to the moment they occurred in the course of evolution. However, in the absence of a molecular clock mode of evolution, it is impossible to recover the order of duplication events. All we can infer from gene sequences is a phylogeny with ordered leaves (Figure 2(c)). Formally, an *ordered phylogeny* is a pair (T, O) where T is a phylogeny and O is the ordered sequence

of its leaves.

If an ordered phylogeny (T, O) can be explained by a duplication history $\mathcal{H}$, we say that (T, O) is *compatible* with $\mathcal{H}$, and that $\mathcal{H}$ *is a duplication history of* (T, O). If (T, O) is compatible with at least one duplication history, it is called a *duplication tree*. Choosing appropriate roots for unrooted duplication trees is discussed in (Gascuel *et al.*, 2005).

In the rest of this paper, a *duplication tree* will refer to a *simple rooted duplication tree*, that is a rooted duplication tree compatible with at least one history involving only simple duplications (see Figure 2(d)). Unless otherwise stated, all the phylogenies are rooted.

2.2 A Duplication/Inversion Model

Many tandemly repeated gene families contain members in both transcriptional orientations. The actual duplication model is thus inadequate to describe their evolution. To circumvent this limitation, we propose an extended model of duplication which includes inversions. Hereafter, the transcriptional orientations of the genes in a *signed* ordered phylogeny (T, O) are specified by signs $(+/-)$ in O. We denote by $d_{inv}(O_i, O_j)$ the inversion distance between the two signed orders O_i and O_j. Note that a signed ordered phylogeny (T, O) cannot be a duplication tree unless all the genes in O have the same sign (although this is not a sufficient condition).

Definition 1 *A* simple duplication/inversion history *(or just* dup/inv history*) of length* k *is an ordered sequence* $\mathcal{H}_k = ((T_1, O_1), ..., (T_{k-1}, O_{k-1}), (T_k, O_k))$ *where :*

1. *Every* (T_i, O_i) *is a* signed *ordered phylogeny.*

2. $T_1 = v$ *is a single leaf phylogeny and* $O_1 = (\pm v)$.

3. *For* $0 < i < k$,

   - *if* $T_{i+1} = T_i$, *then* $d_{inv}(O_i, O_{i+1}) = 1$. *This corresponds to one inversion event.*

- *if $T_{i+1} \neq T_i$, then T_{i+1} is obtained from T_i by adding two children u and w to one of its leaves 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. This corresponds to a simple duplication event.*

3 An Inference Problem

A signed ordered phylogeny is not necessarily compatible with a duplication history. The following lemma shows that additional inversions can always be used to infer a possible evolutionary history for the gene family.

Lemma 1 *A signed ordered phylogeny (T, O) is compatible with at least one simple duplication/inversion history.*

Proof. According to Definition 1, obtain a duplication tree (T, O') by successive duplication events. Then, transform O' into O by applying the required inversions □

As the number of possible dup/inv histories explaining (T, O) can be very large, we restrict ourselves to finding the minimum number of events involved in such evolutionary histories. More precisely, as the number of simple duplications is fixed by T, we are interested in finding the minimum number of inversions involved in a dup/inv history. This parsimony approach relies on the biological assumption that the quantified events are scarce, which seems to be the case with genomic inversions. The next theorem shows that if i is the minimum number of inversions needed to transform O into O' such that (T, O') is a duplication tree, any dup/inv history of (T, O) contains at least i inversions.

Theorem 1 *Let (T, O) be a signed ordered phylogeny. For any dup/inv history $\mathcal{H}$ with i inversions leading to (T, O), there exists a duplication tree (T, O') such that $d_{inv}(O, O') \leq i$.*

Proof by induction.

- Base case: Let $\mathcal{H}_1 = (T_1, O_1)$ be a dup/inv history with no duplication or inversion. Clearly $(T, O') = (T_1, O_1)$ is a duplication tree.
- Induction step (on the number k of events):
  Let $\mathcal{H}_{k+1} = ((T_1, O_1), ..., (T_k, O_k), (T_{k+1}, O_{k+1}))$ be a dup/inv history involving $k+1$

events and i inversions and $\mathcal{H}_k = ((T_1, O_1), ..., (T_k, O_k))$. From Definition 1, there are two possibilities:

- If $T_{k+1} = T_k$, then the last event is an inversion, and $\mathcal{H}_k$ is a dup/inv history involving $i-1$ inversions. By induction hypothesis, there exists a duplication tree (T_k, O'_k) such that $d_{inv}(O_k, O'_k) \leq i-1$. Let O_{k+1} be the order obtained from O_k by applying the last inversion. Then we have $d_{inv}(O_{k+1}, O'_k) \leq d_{inv}(O_k, O'_k) + 1 \leq i$.
- If $T_{k+1} \neq T_k$, the last event is a duplication, that is a leaf v of (T_k, O_k) is replaced by two consecutive leaves (u, w) in (T_{k+1}, O_{k+1}). Let (T_k, O'_k) be the duplication tree associated to $\mathcal{H}_k$ and suppose that all elements of O'_k are positive. If v has positive sign in O_k, we obtain O'_{k+1} by replacing v in O'_k by (u, w). Otherwise, v has negative sign in O_k and we obtain O'_{k+1} by replacing v in O'_k by (w, u). Thus, $d_{inv}(O_{k+1}, O'_{k+1}) = d_{inv}(O_k, O'_k) \leq i$ and (T_{k+1}, O'_{k+1}) is a duplication tree. The case where the elements of O'_k have a negative sign is similar □

Corollary 1 *Let (T, O) be a signed ordered phylogeny and (T, O') a duplication tree such that $d_{inv}(O, O') = i$ is minimum. There exists a dup/inv history $\mathcal{H}$ for (T, O) with exactly i inversions, which is optimal.*

Proof. The existence of $\mathcal{H}$ for (T, O) with exactly i inversions follows directly from the proof of Lemma 1. The number i of inversions in $\mathcal{H}$ must be optimal, otherwise, from Theorem 1, it would contradicts the hypothesis that $d_{inv}(O, O') = i$ is minimum □

Corollary 1 allows to reformulate our problem in the following way :

Minimum-Inversion Duplication problem

Input: A signed ordered phylogeny (T, O),

Output: An order O' such that (T, O') is a duplication tree and $d_{inv}(O, O')$ is minimal.

4 A Branch-and-Bound Algorithm

We begin by briefly summarizing the Hannenhalli-Pevzner method (Hannenhalli and Pevzner, 1999), as it will be used in our approach.

4.1 Hannenhalli-Pevzner (HP) Algorithm

Given two signed orders O, O' of size n on the same set of genes, the problem is to find the minimal number $d_{inv}(O, O')$ of inversions required to transform O to O' (or similarly O' to O). As the orders considered in this paper do not represent a whole chromosome, but rather a cluster of tandemly repeated genes, we can always consider them as linear (not circular), with a leftmost and rightmost gene.

The algorithm is based on a bicolored graph, called the *breakpoint graph*, constructed from the two signed orders as follows: if gene x of O has a positive sign, replace it by the pair $x_t x_h$, and if it is negative, by $x_h x_t$. Then the vertices of the graph are just the x_t and the x_h for all genes x plus two additional vertices, s and f, which represent the two extremities of the order. The graph contains two classes of edges: the real and desired edges (as named in (Setubal and Meidanis, 1997)). Any two vertices which are adjacent in O, other than x_t and x_h deriving from the same x, are connected by a *real edge*, and any two adjacent in O', by a *desired edge* (see Figure 3(c)).

This graph decomposes naturally into a set of c disjoint color-alternating cycles. An important property of the graph is its decomposition into components, where a *component* is a maximal set of "crossing" cycles.

Based on this graph, the inversion distance can be computed according to the following formula (Hannenhalli and Pevzner, 1999):

$$d_{inv}(O, O') = n + 1 - c + h + f,$$

where h and f are quantities related to the presence of "hurdles" (components of a

particular type). As the probability for a component to be a hurdle is low, h and f are usually close to 0. Therefore, the number of cycles c is the dominant parameter in the formula. In other words, the more cycles there are, the less inversions we need to transform O into O'. For example in Figure 3(c), $n = 4$, $c = 3$, $h = 0$ and $f = 0$, which leads to $d_{inv}(O, O') = 2$.

4.2 Enumerating the Compatible Orders

We say that an order O' is *compatible* with a phylogeny T iff (T, O') is a duplication tree. As mentioned in the introduction, the considered duplication trees are equivalent to binary search trees. Therefore, to enumerate all the orders compatible with T, we associate a binary variable b_i to each internal node i of T. Each b_i defines an order relation between the left and right descendant leaves of i. By setting b_i to 0, we make all the left descendants smaller than the right ones. Conversely, by setting b_i to 1, all left descendants are considered larger than the right ones (see Figure 3(a)(b)). Assigning a value to all internal nodes of T defines a total order O' on its leaves: the order between two leaves is determined by the b_i value of their closest common ancestor. Otherwise, the order is partial since some pairs of leaves are incomparable. We will denote such a partial order as O^*. Note that every order admits two transcriptional orientations according to our definition of a duplication tree. Therefore, if n is the number of leaves in T, there are 2^{n-1} possible assignments of the b_i variables, each with two possible transcriptional orientations. This leads to 2^n distinct orders O' compatible with T. Hereafter, for clarity of presentation, we will only consider one of the two orientations. Figure 3

4.3 A Lower Bound for the Inversion Distance

To avoid computing $d_{inv}(O, O')$ for each of the 2^{n-1} orders O' compatible with T, we consider a branch-and-bound strategy similar to the one used by Zheng *et al.* (2003). The idea is to compute a lower bound on $d_{inv}(O, O')$ as we progressively define O^* by

updating the partial breakpoint graph of (O, O^*). In order to progressively construct this graph, it is essential to define the b_i values in a post-order traversal of T: the binary variables of all the descendant nodes of i should be defined before b_i. This insures that the two subtrees of i have a total order on their leaves.

Consequently, if we set b_i to 0, the greatest left descendant leaf l_{max} of node i will immediately precede its smallest right descendant leaf r_{min} in O'. Conversely, if b_i is set to 1, the greatest right descendant r_{max} will immediately precede the smallest left descendant l_{min}. Therefore, the assignment of a b_i value allows us to add a desired edge in the partial breakpoint graph between l_{max} and r_{min} (or r_{max} and l_{min}) (see Figure 3(c)).

Let O^* be the partial order obtained at a given stage of the procedure. Let e be the number of cycles and p the number of paths of the corresponding *partial breakpoint graph*. The remaining desired edges can create at most p cycles, ending with a breakpoint graph with at most $c = e + p$ cycles. Thus, any total order O' that can be obtained from the partial order O^* is such that:

$$d_{inv}(O, O') = n + 1 - c + h + f \geq n + 1 - c \geq n + 1 - p - e = d^*_{inv}(O, O^*).$$

4.4 Algorithm

The branch-and-bound algorithm proceeds as follows (see Algorithm 1). Denote O' the best order obtained so far at a given step and $\min_{inv} = d_{inv}(O, O')$ the corresponding inversion distance. Each following step assigns the values of the binary variables in a post-order traversal of T that progressively defines a partial order O^*. This procedure stops and backtracks when the current partial order O^* is such that $d^*_{inv}(O, O^*) > \min_{inv}$. This is justified by the fact that any total order that can be obtained from O^* cannot lead to a smaller inversion distance. If no bound were used, the assignment procedure would explore all the 2^{n-1} possible configurations of the binary variables. Finally, every time a total order is reached, the inversion distance is computed using the HP algorithm

and $\min_{inv}$ and O' are updated, if necessary.

The efficiency of a branch-and-bound algorithm is usually correlated with its initial solution. Here, we use the initial order O' obtained with the polynomial-time algorithm described in the next section.

Algorithm 1: Branch-and-bound

Data: A signed ordered phylogeny (T, O) with n leaves.

Result: An order O' such that (T, O') is a duplication tree and $d_{inv}(O, O')$ is minimal.

begin

O' is the initial order obtained with the polynomial-time algorithm (c.f. Section 5.2)

$min_{inv} \leftarrow d_{inv}(O, O')$

O^* is an empty partial order, and PBPG(O, O^*) the corresponding partial breakpoint graph

Label the $n - 1$ internal nodes of T according to a post-order traversal ($i < j$ if node i is a descendant of node j)

Associate a binary variable b_i to each internal node i of T

RECURSIVE_EXPLORE(1)

return O'

end

Procedure RECURSIVE_EXPLORE(*integer* i)

begin

if $i = n - 1$ **then**

if $d_{inv}(O, O^*) < min_{inv}$ **then**

$O' \leftarrow O^*$

$min_{inv} \leftarrow d_{inv}(O, O^*)$

end

else

$b_i \leftarrow 0$

Add adjacency (l_{max}, r_{min}) in PBPG(O, O^*)

if $d^*_{inv}(O, O^*) < min_{inv}$ **then**

RECURSIVE_EXPLORE$(i + 1)$

end

Remove adjacency (l_{max}, r_{min}) in PBPG(O, O^*)

$b_i \leftarrow 1$

Add adjacency (r_{max}, l_{min}) in PBPG(O, O^*)

if $d^*_{inv}(O, O^*) < min_{inv}$ **then**

RECURSIVE_EXPLORE$(i + 1)$

end

Remove adjacency (r_{max}, l_{min}) in PBPG(O, O^*)

end

end

Where l_{min} and l_{max} are respectively the smallest and greatest left descendant leaf of node i, and r_{min}, r_{max}, the smallest and greatest right descendant leaf of i.

5 Minimizing the Breakpoint Distance

5.1 The Minimum Breakpoint Duplication Problem

Genome rearrangement mechanisms such as inversions cannot be observed directly from the data and can only be inferred from different theoretical probabilistic, algorithmic or phylogenetic methods. Evidence for the occurrence of such mechanisms during evolution is reflected by the presence of breakpoints, that is inverted genes or genes that are adjacent in one genome but separated in another related genome. In contrast with rearrangement mechanisms, breakpoints can be directly observed from data. The breakpoint distance is the most widely used measure of gene order conservation, and usually considered as a first attempt to solve a given genome rearrangement problem. Moreover it provides an upper bound for the inversion distance.

As mentioned in Section 4.1, the orders considered in this paper represent clusters of tandemly duplicated genes, and as such, can always be considered linear. Let O and $\hat{O}$ be two signed orders, not necessarily on the same set of genes. A *breakpoint* of $\hat{O}$ with respect to O is a pair (j, k) of consecutive elements in $\hat{O}$ which is not present in O, neither in the form (j, k) nor in the form $(-k, -j)$. To account for breakpoints at cluster extremities, we add two "artificial genes" s and f so that O becomes (s, O, f) and $\hat{O}$ becomes $(s, \hat{O}, f)$ (see Figure 4). Figure 4

We denote by $d_{bp}(O, \hat{O})$ the number of breakpoints in $(s, \hat{O}, f)$ with respect to (s, O, f). When O and $\hat{O}$ are two permutations on the same set of genes, then $d_{bp}(O, \hat{O}) = d_{bp}(\hat{O}, O)$, and d_{bp} is a distance.

The breakpoint distance is correlated to the inversion distance. Indeed, any sequence of inversions transforming $\hat{O}$ into O will eliminate all the breakpoints of $\hat{O}$ with respect to O. The following is a well known property.

Property 1 *Let O and $\hat{O}$ be two signed orders on the same set of genes. We have:*

$$\frac{d_{bp}(O,\hat{O})}{2} \leq d_{inv}(O,\hat{O}) \leq d_{bp}(O,\hat{O}).$$

In this section, we present an exact polynomial-time algorithm solving the following problem.

Minimum-Breakpoint Duplication problem

Input: A signed ordered phylogeny (T,O),

Output: An order $\hat{O}$ such that $(T,\hat{O})$ is a duplication tree and $d_{bp}(O,\hat{O})$ is minimal.

A solution to this problem is an upper bound for the Minimum-Inversion Duplication problem. Indeed, let (T,O) be a signed ordered phylogeny, and O' and $\hat{O}$ be two orders such that (T,O') and $(T,\hat{O})$ are two duplication trees and $d_{inv}(O,O')$, $d_{bp}(O,\hat{O})$ are minimal. Then, from Property 1 we have:

$$d_{inv}(O,O') \leq d_{inv}(O,\hat{O}) \leq d_{bp}(O,\hat{O}).$$

The bound $d_{bp}(O,\hat{O})$ is not very tight as each inversion could creates two breakpoints. A much better bound is $d_{inv}(O,\hat{O})$, which is obtained by using the HP algorithm with $\hat{O}$ outputted by the polynomial-time algorithm we present in the next section.

5.2 A Dynamic Programming Algorithm

For the purpose of our dynamic programming algorithm, orders extremities must be ignored while computing the number of breakpoints in intermediate sub-orders. Hence, extremities s and f will only be considered at the end of the procedure. We use the notation $d^*_{bp}(O,\hat{O})$ to refer to the number of breakpoints in $\hat{O}$ with respect to O. We denote by $\hat{O}[x,y]$ the sub-permutation of $\hat{O}$ beginning with element x and ending with element y. For example if $\hat{O} = (4,2,3,5,1)$, then $\hat{O}[2,5] = (2,3,5)$.

Let (T, O) be a signed ordered phylogeny, and $\hat{O}$ be an alternative order on the leaves of T such that $(T, \hat{O})$ is a duplication tree. By definition, all genes in $\hat{O}$ must have the same sign. For clarity of presentation and w.l.o.g, we consider that they are positive.

Assume that $\hat{O} = \hat{O}[i, l]$, that is $\hat{O}$ begins with element i, ends with element l. Then, the duplication tree $(T, \hat{O}[i, l])$ can be defined recursively as the combination of two duplication trees $(T_1, \hat{O}[i, j])$ and $(T_2, \hat{O}[k, l])$ (see Figure 5), where j and k are two adjacent elements in $\hat{O}$ such that the least common ancestor of i, j and the least common ancestor of k, l are the two children of the root of T. Consequently, the breakpoint distance between $\hat{O}[i, l]$ and O can be expressed as follows:

$$d_{bp}(O, \hat{O}[i, l]) = d^*_{bp}(O, \hat{O}[i, j]) + d^*_{bp}(O, \hat{O}[k, l]) + bp(s, i) + bp(j, k) + bp(l, f), \quad (1)$$

where $bp(x, y) = \begin{cases} 1 & \text{if the pair } (x, y) \text{ is a breakpoint with respect to } (s, O, f) \\ 0 & \text{otherwise.} \end{cases}$

Figure 5

Now we describe the central recursion of the algorithm. Let denote by $B[i, l]$ the minimum number of breakpoints (with respect to O) we can get among the set of orders compatible with T which start with i and end with l. Consider the subtree labeling of Figure 5 and assume that $i \in T_{11}$ and $l \in T_{22}$.

Then,

$$B[i, l] = \min_{(j \in T_{12}\ ,\ k \in T_{21})} \Big(B[i, j] + B[k, l] + bp(j, k) \Big) \quad (2)$$

with the initial condition $B[i, i] = 0$ for every leaf i.

The $B[i, l]$ values can be computed recursively as follows. We consider every subtree T_x of T in a bottom-up approach (post-order traversal), beginning with the leaves of T and ending with T itself. For each T_x, using Recurrence 2, we compute $B[i, l]$ for every pair of leaves (i, l) whose least common ancestor is the root of T_x. It is easy to see from Figure 5 that this condition on (i, l) is necessary and sufficient for the existence of a duplication tree $(T_x, \hat{O}[i, l])$.

Finally, the breakpoint distance $d_{bp}(O, \hat{O})$ for an optimal order $\hat{O}$ such that $(T, \hat{O})$ is a duplication tree is

$$d_{bp}(O, \hat{O}) = \min_{(i,l)} \Big(B[i, l] + bp(s, i) + bp(l, f) \Big) \tag{3}$$

over the pairs (i, l) whose least common ancestor is the root of T. The order $\hat{O}$ is then simply constructed by backtracking in the dynamic programming table.

Computing a given $B[i, l]$ value using Recurrence 2 takes $O(n^2)$ time in the worst case when the tree is balanced ($O(n)$ for a caterpillar tree). Since $B[i, l]$ is computed once for every pair (i, l), the worst-time complexity for the whole algorithm is $O(n^4)$.

6 Results with Simulated and Biological Data

To simulate the evolution of a gene family and obtain the corresponding ordered phylogeny (T, O), we first generate T using the Yule (1924) model and define an order O' such that (T, O') is a duplication tree. Then, we obtain O by applying a fixed number of inversions to O'.

6.1 Execution Time

To compare the execution time of the algorithms, we applied them on simulated ordered phylogenies with size varying from 10 to 40 leaves, that underwent 4, 8, 16 and 32 inversions. Results are averaged over 1,000 phylogenies and are given in Figure 6. We observe that the branch-and-bound performance depends significantly on the number of inversions. Nevertheless, it can be used on relatively important phylogenies within reasonable time (30 seconds on average for an ordered phylogeny with 40 leaves and 32 inversions). On the other hand, the execution time of the polynomial-time algorithm depends uniquely on the size of the phylogeny and requires less than a second for all the instances.

Figure 6

6.2 Using the Polynomial-time Algorithm as a Heuristic

The polynomial-time algorithm finds a duplication tree $(T, \hat{O})$ such that the breakpoint distance between $\hat{O}$ and the order O observed on the chromosome is minimal. To see if $\hat{O}$ can be used as an approximation to the Minimum-Inversion Duplication problem, we applied the algorithm on simulated data and compared $d_{inv}(O, \hat{O})$ (computed using the HP algorithm) with the optimal value returned by the branch-and-bound. We used ordered phylogenies with 10 and 20 leaves, which underwent 1 to 16 inversions. The results are averaged over 1,000 phylogenies and are presented in Figure 7. We see that when the number of inversions is low, the inversion distance obtained with the polynomial-time

algorithm is very close to the optimal one. Figure 7

6.3 Improving Phylogenetic Inference

We applied our algorithms on simulated data to verify how they could be used to validate inferred phylogenies of tandemly repeated gene families. The idea is that a wrong phylogeny should require more inversions than the true one. We simulated ordered phylogenies with 10 and 20 leaves, which underwent 1, 2, 4 and 6 inversions. These are the observable states (T_{true}, O) resulting from "true" duplication/inversion histories. For each T_{true}, we then generated four "wrong" (but close) phylogenies T_{wrong}, by applying one to four random Nearest Neighbor Interchange rearrangements (NNI) (e.g Swofford *et al.*, 1996, chap. 7). Those "wrong" phylogenies can be seen as the ones we would obtain from biological data when a few branches have weak statistical support. We then used the branch-and-bound algorithm to compute the minimum number of inversions inv() necessary to explain (T_{true}, O) and its associated (T_{wrong}, O). We did the same procedure with the polynomial-time algorithm which instead compute the minimum number of breakpoints bp() between the order of an inferred duplication tree and the order on the chromosome. The results are averaged over 1,000 phylogenies and are presented in Figure 8 and 9. Surprisingly, the results are very similar although the breakpoint distance is slightly less sensitive to wrong phylogenies. Figure 8

Results can be interpreted as follows. For a wrong 10 leaves phylogeny that differs by one NNI from the true one, roughly 50% of the time on average our algorithms report an excess of inversions/breakpoints, otherwise they report the same number compared to the true phylogeny. Suppose we have a set of putative phylogenies for a given gene family, and one is correct while the others differ by a few NNI. According to Figure 8 and 9, for wrong trees, the algorithms almost always reports the same number of inversions/breakpoints or more as in the true tree. Thus, choosing the phylogeny with the lowest number of inversions/breakpoints is either a winning strategy, or not enough to select a single phylogeny as several ones require the same number of inversions/breakpoints, but is Figure 9

almost never misleading. Of course, this ability to discard wrong phylogenies decreases as the true number of inversions increases, but even with 6 inversions and 4 NNI the number of misleading cases remains low.

6.4 Application on Biological Data

The KRAB-zinc finger gene family encodes for transcription factors. It contains more than 400 active members physically grouped into clusters. In a recent study, Hamilton *et al.* (2006) proposed a phylogeny of the primate specific ZNF91 sub-family based on their tether[1] and flanking sequences. This phylogeny (obtained by Neighbor-Joining (Saitou and Nei, 1987)) contains a monophyletic group of 6 genes clustered at the telomere of HSA4p, which may have been derived from a single ancestor through successive tandem duplications.

We applied the branch-and-bound algorithm on this cluster using the proposed phylogeny, and found that a duplication/inversion history would require at least 4 inversions, which seems relatively high considering that only 6 genes are involved.

To test whether a "better" phylogeny could be proposed, we used the MrBayes software (F. Ronquist, 2003) to obtain a sample from the posterior probability distribution of all possible phylogenies. The tether (+100 flanking bp) sequences were downloaded from the Human KZNF Gene Catalog[2] (Huntley *et al.*, 2006) and aligned using ClustalW (Thompson *et al.*, 1994) with default settings. The ZNF160 tether sequence was used as an outgroup to obtain a rooted tree. We performed 500,000 MCMC generations with MrBayes under the GTR model (Lanave *et al.*, 1984; Tavare, 1986) and a gamma-shaped rate variation with a proportion of invariable sites. Convergence was easily attained and the experiment was repeated three times with similar results. Finally we applied the branch-and-bound on the sampled phylogenies and observed that the best one (p=0.4) is compatible with an optimal duplication/inversion history involving only

[1]The region upstream from the first finger.

[2]`http://znf.llnl.gov/catalog/`

two inversions. This provides strong support for the tandem duplication/inversion model and indicates that our phylogeny is probably the correct one. Results obtained with the polynomial-time algorithm are similar although less discriminative. Phylogenies are presented in Figure 10 with both their associated numbers of inversions / breakpoints. Figure 10

7 Conclusion

This work represents the first attempt to account for inversions in an evolutionary model of tandemly repeated genes. We presented a time-efficient branch-and-bound algorithm for finding the minimal number of inversions in an evolutionary history of a gene family characterized by an ordered phylogeny. We have also developed a polynomial-time algorithm based on the breakpoint distance. We demonstrated, using simulations, that it is a good heuristic for the original problem. Though only simple duplications were considered here, the model has been shown useful to select an appropriate phylogeny among a set of possible ones. These are encouraging results that motivate further extensions.

One of the next steps of this work will be to account for multiple duplications in the evolutionary model, although generalizing the Minimum-Inversion Duplication problem to this model is far from being straightforward. From this perspective, it seems reasonable to begin with a simpler rearrangement distance such as the breakpoint distance. Another important generalization will be to consider a family of tandemly duplicated genes with orthologs in two or more genomes. For example, Shannon *et al.* (2003) identified homologous ZNF gene family regions in human and mouse. A phylogenetic tree involving such tandemly repeated genes in human and mouse clusters was established. It would be of major interest to develop an algorithm allowing one to explain such a phylogeny based on an evolutionary model involving tandem duplication, inversion and speciation events.

The biological assumption used in this paper is that inversion is the only rearrangement mechanism leading to the presence of different transcriptional orientations in a tandemly duplicated gene family. Another possible rearrangement mechanism that has been documented and that gives an alternative explanation for the presence of inverted genes is *inverted tandem duplication* as a single event (Stankiewicz and Lupski, 2002). However, "simple" inverted duplication (inverted duplication of a single gene) is not always sufficient to explain the fact that a tree representing a tandemly duplicated gene

family is not a duplication tree, thus multiple duplication and deletion cannot be ignored in this case. Using such a biological assumption is thus as difficult as using the assumption of multiple duplications and inversions, and will be considered in future developments.

Acknowledgments

The authors wish to thank M. Aubry and H. Tadepally for their help on zinc finger genes and the anonymous referees for their valuable comments. This work was supported by grants from the Fonds québécois de la recherche sur la nature et les technologies (FQRNT), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR) and the French ACI IMPBIO: REPEVOL project.

References

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

Benson, G. and Dong, L., 1999. Reconstructing the duplication history of a tandem repeat. In *Proceedings of Intelligent Systems in Molecular Biology (ISMB1999), Heidelberg, Germany*, 44–53. AAAI.

Bergeron, A., Mixtacki, J., and Stoye, J., 2004. Reversal distance without hurdles and fortresses. *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.

Chen, K., Durand, D., and Farach-Colton, M., 2000. Notung: Dating gene duplications using gene family trees. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*. ACM, New York.

El-Mabrouk, N., 2000. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *CPM 2000*, *LNCS*, volume 1848, 222- 234.

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

Elemento, O. and Gascuel, O., 2005. An exact and polynomial distance-based algorithm to reconstruct single copy tandem duplication trees. *Journal of Discrete Algorithms* 2, 362–374.

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

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

Fitch, W., 1977. 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.

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. Oxford University Press.

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.

Glusman, G., Yanai, I., Rubin, I., and Lancet, D., 2001. The complete human olfactory subgenome. *Genome Research* 11, 685–702.

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

Hamilton, A., Huntley, S., Tran-Gyamfi, M., Baggott, D., L.Gordon, and Stubbs, L., 2006. Evolutionary expansion and divergence in the znf91 subfamily of primate-specific zinc finger genes. *Genome Research* 16, 584–594.

Hannenhalli, S. and Pevzner, P. A., 1999. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *J. ACM* 48, 1–27.

Huntley, S., Baggot, D., Hamilton, A., M. TranGyamfi, S. Y., 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.

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.

Lanave, C., Preparata, G., Saccone, C., and Serio, G., 1984. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20, 86–93.

Ma, B., Li, M., and Zhang, L., 1998. On reconstructing species trees from gene trees in term of duplications and losses. In Istrail, S., Pevzner, P., and Waterman, M., eds., *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, 182–191. ACM, New York.

Page, R. and Charleston, M., 1997. Reconciled trees and incongruent gene and species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 37, 57–70.

Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425.

Setubal, J. and Meidanis, J., 1997. *Introduction to Computational Molecular Biology*, chapter 7. PWS Pub. Co. SET j 97:1 1.Ex.

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

Siepel, A., 2002. Algorithm to find all sorting reversals. In *Proceedings of the second conference on computational molecular biology (RECOMB'02)*, 281 - 290. ACM Press.

Stankiewicz, P. and Lupski, J., 2002. Genome architecture, rearrangements and genomic disorders. *TRENDS in Genetic* 18, 74 - 82.

Swofford, D., Olsen, P., Waddell, P., and Hillis, D., 1996. *Molecular Systematics*, chapter Phylogenetic Inference, 407–514. Sinauer Associates, Sunderland, Massachusetts.

Tang, M., Waterman, M., and Yooseph, S., 2001. Zinc finger gene clusters and tandem gene duplication. In *Proceedings of International Conference on Research in Molecular Biology (RECOMB2001)*, 297–304.

Tavare, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17.

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.

Yule, G., 1924. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character Immunology Review* 213, 21–87.

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

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

Zheng, C., Lenert, A., and Sankoff, D., 2003. Reversal distance for partially ordered genomes. *Bioinformatics* 21, i502 - i508.

List of Figures

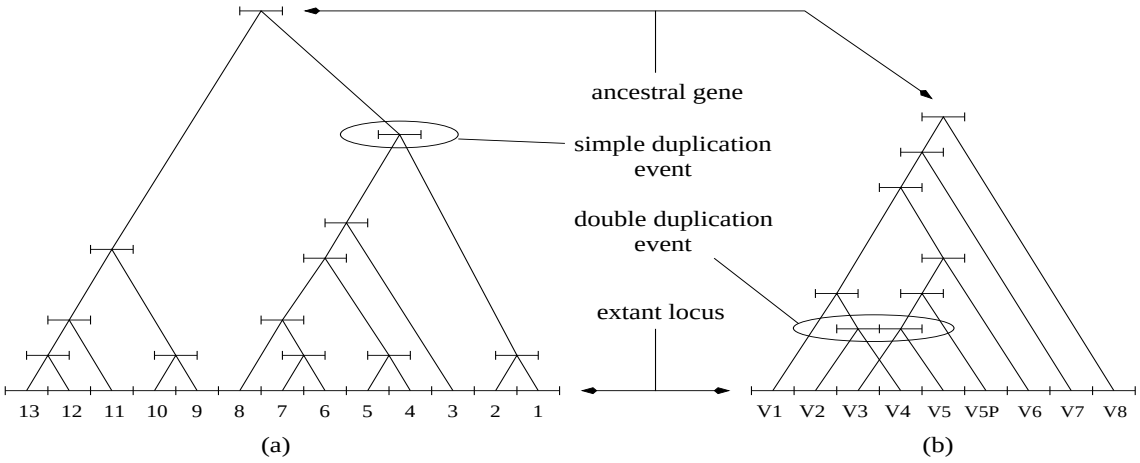

Figure 1: (a) Simple rooted duplication tree of the 13 Antennapedia-class homeobox genes from the cognate group (Zhang and Nei, 1996). (b) Rooted duplication tree of the 9 variable genes of the human T cell receptor Gamma (TRGV) locus (Elemento *et al.*, 2002). In both examples, the contemporary genes are adjacent and linearly ordered along the extant locus.

Mathieu Lajoie

Figure 1 (of 10)

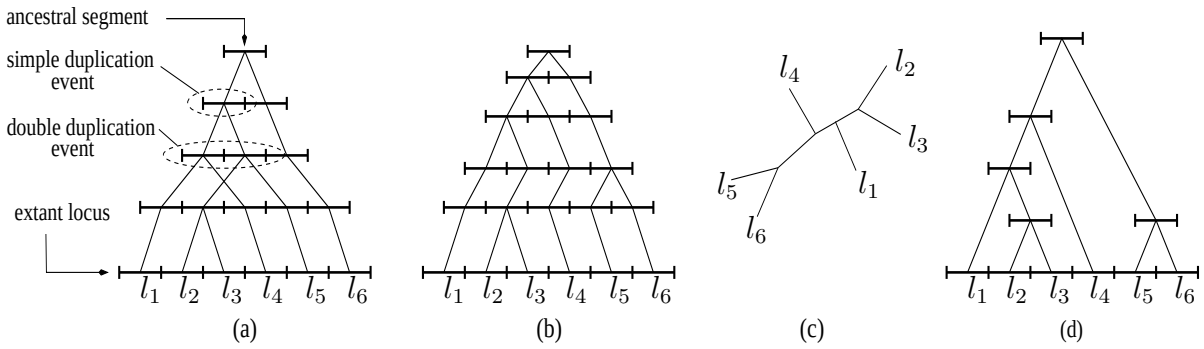

Figure 2: (a) Duplication history; each segment represents a copy. (b) Simple duplication history. (c) The unrooted simple duplication tree corresponding to history (b). (c) A simple rooted duplication tree corresponding to history (b).

Mathieu Lajoie

Figure 2 (of 10)

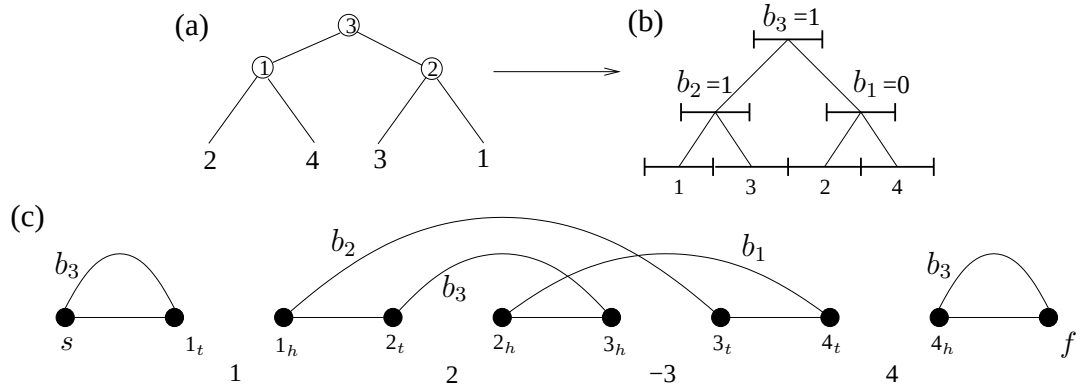

Figure 3: (a) A phylogeny with an appropriate post-order labeling of its internal nodes; (b) The duplication tree corresponding to an assignment of the b_i variables of (a); (c) The breakpoint graph illustrating the difference between the gene order $O' = (1, 3, 2, 4)$ obtained from the duplication tree (b) and the gene order $O = (1, 2, -3, 4)$ observed in the genome. *Desired edges* (curved edges) are added in the same order as the corresponding b_i values (b_1 then b_2 then b_3).

Mathieu Lajoie

Figure 3 (of 10)

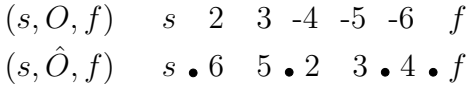

Figure 4: Breakpoints (black dots) in $(s, \hat{O}, f)$ with respect to (s, O, f).

⇑

Mathieu Lajoie

Figure 4 (of 10)

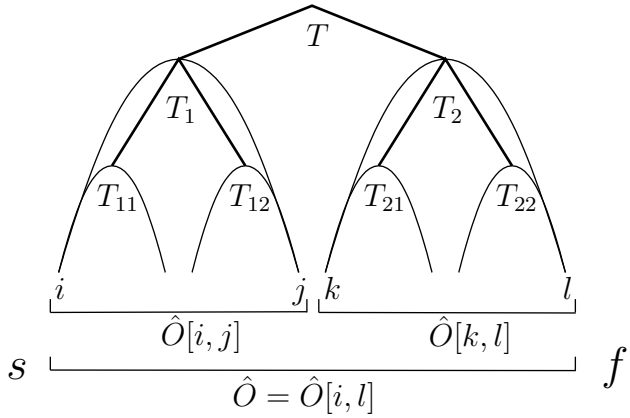

Figure 5: The duplication tree $(T, \hat{O}[i,l])$ can be obtained by combining two duplication trees $(T_1, \hat{O}[i,j])$ and $(T_2, \hat{O}[k,l])$. The "artificial genes" s and f allow to consider breakpoints at cluster's extremities.

⇑

Mathieu Lajoie

Figure 5 (of 10)

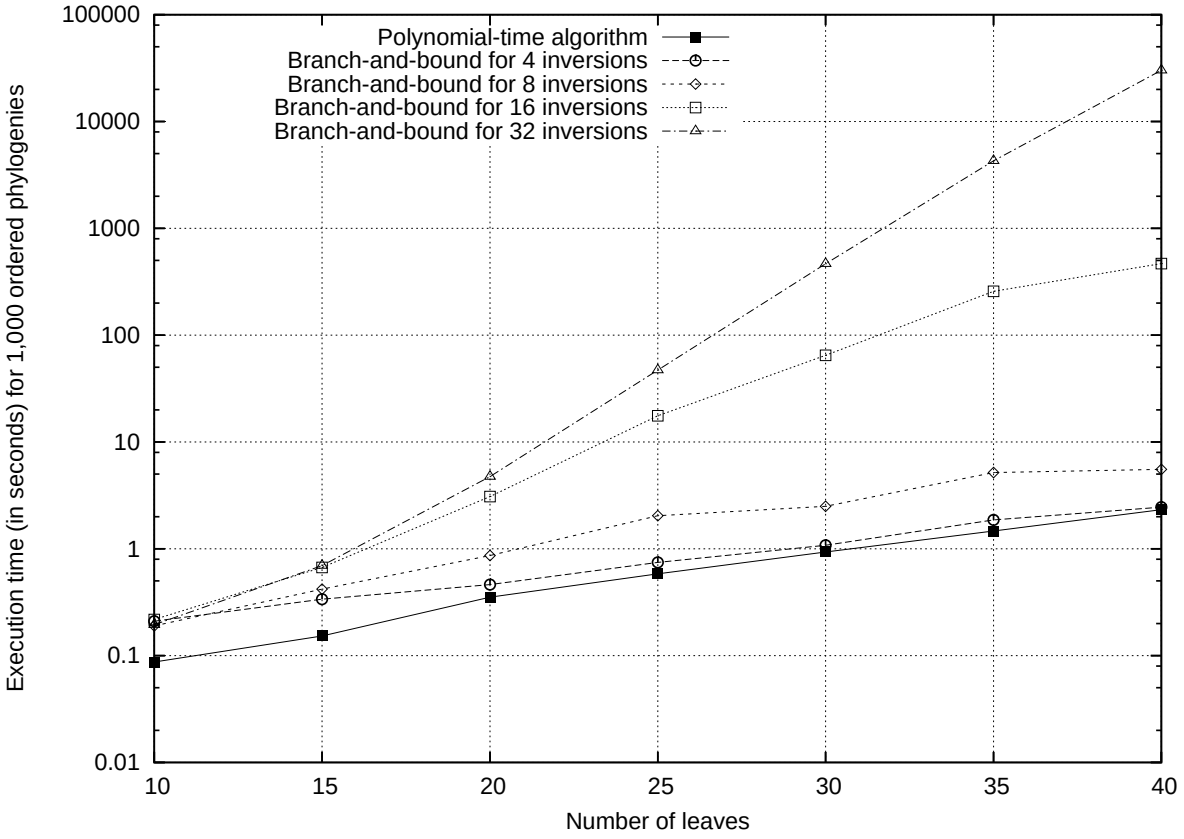

Figure 6: Execution times (in seconds) for 1,000 signed ordered phylogenies with 4, 8, 16 and 32 inversions. The execution time of the polynomial-time algorithm is not affected by the number of inversions.

Mathieu Lajoie

Figure 6 (of 10)

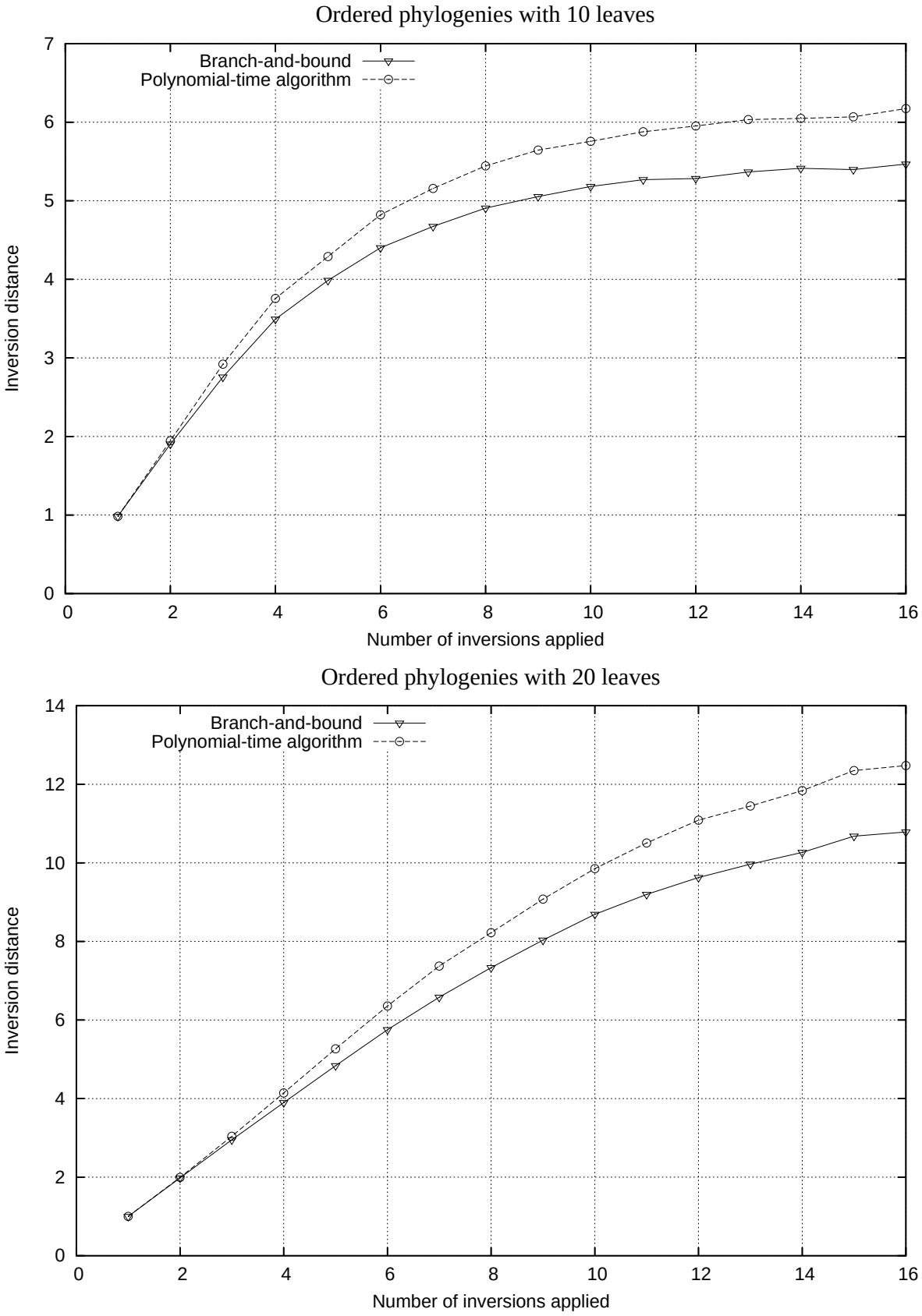

Figure 7: Number of inversions inferred by the polynomial-time algorithm compared to the optimal value obtained by the branch-and-bound. Results are averaged over 1,000 phylogenies.

⇑

Mathieu Lajoie

Figure 7 (of 10)

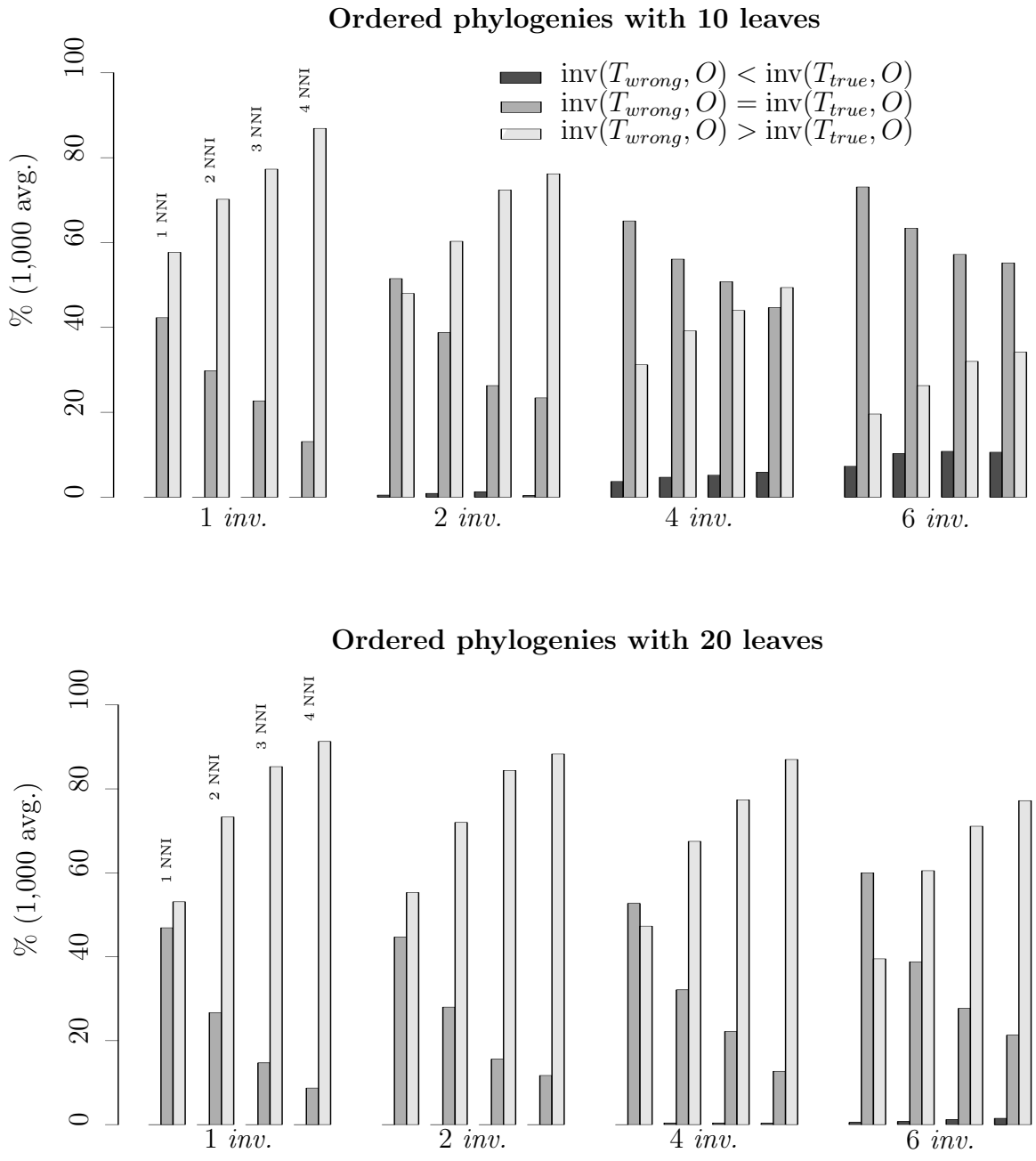

Figure 8: Fraction of times $\text{inv}(T_{wrong}, O)$ is less, equal or greater than $\text{inv}(T_{true}, O)$.

⇑

Mathieu Lajoie

Figure 8 (of 10)

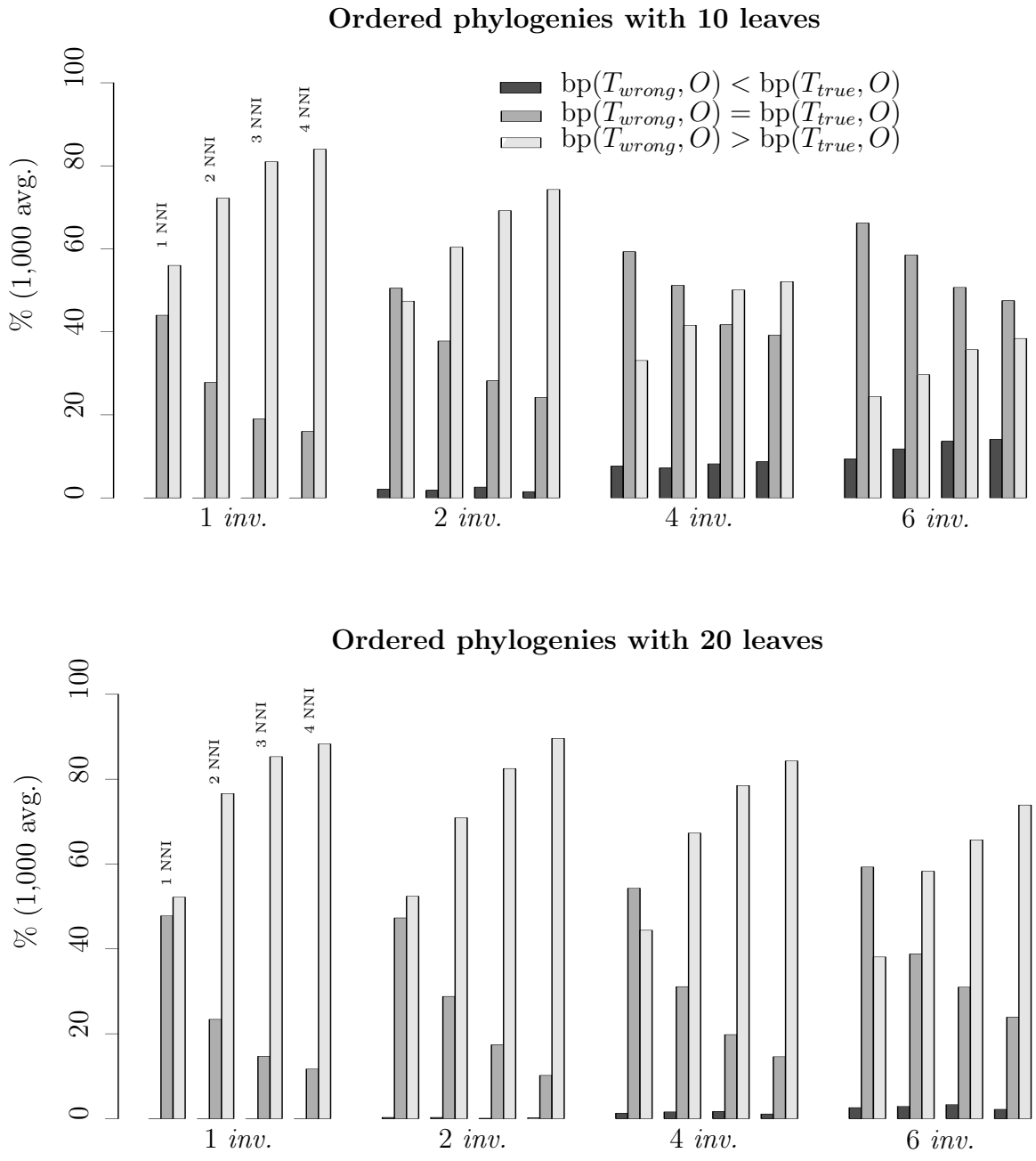

Figure 9: Fraction of time $\mathrm{bp}(T_{wrong}, O)$ is less, equal or greater than $\mathrm{bp}(T_{true}, O)$.

⇑

Mathieu Lajoie

Figure 9 (of 10)

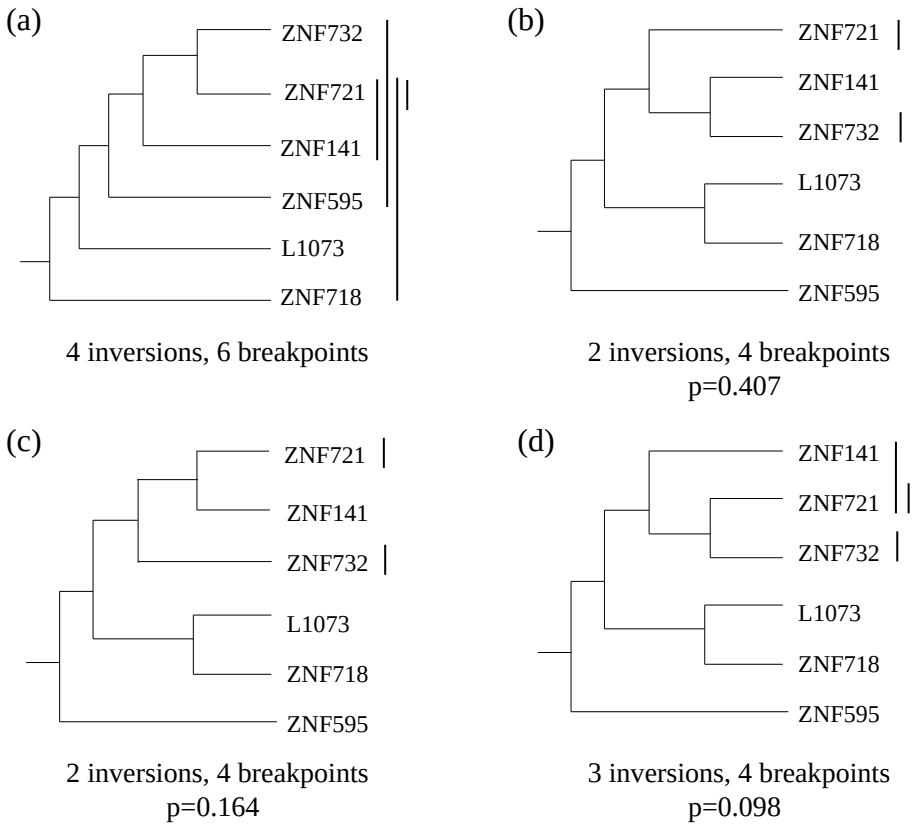

Figure 10: Different phylogenies for the ZNF141 clade on human chromosome 4, with the associated minimal number of inversions/breakpoints. The black vertical lines represent an optimal sequence of inversions leading to the *signed* gene order observed on the chromosome: (+ZNF595, +ZNF718, +L1073, −ZNF732, +ZNF141, −ZNF721). (a) The phylogeny published by Hamilton *et al.* (2006) requires 4 inversions, which is relatively high for 6 genes; (b,c,d) The 3 best phylogenies we obtained with MrBayes, and their associated probabilities. The first two ones require only 2 inversions, which is optimal for this order. The position of the root was determined using ZNF160 as an outgroup.

⇑

Mathieu Lajoie

Figure 10 (of 10)