Evolution of Tandemly Repeated Sequences Through Duplication and Inversion

Denis Bertrand, Mathieu Lajoie, Nadia El-Mabrouk, Olivier Gascuel

Evolution of Tandemly Repeated Sequences through Duplication and Inversion

Denis Bertrand[1], Mathieu Lajoie[2⋆], Nadia El-Mabrouk[2], and Olivier Gascuel[1]

[1] LIRMM, UMR 5506, CNRS et Univ. Montpellier 2, 161 rue Ada, 34392 Montpellier Cedex 5 France
{dbertran, gascuel}@lirmm.fr
[2] DIRO, Université de Montréal, CP 6128 succ Centre-Ville, Montréal QC, H3C 3J7, Canada
{lajoimat,mabrouk}@iro.umontreal.ca

Abstract. Given a phylogenetic tree T for a family of tandemly repeated genes and their *signed* order O 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 time-efficient branch-and-bound algorithm and show, using simulated data, that it can be used to detect "wrong" phylogenies among a set of putative ones for a given gene family. An application on a published phylogeny of KRAB zinc finger genes is presented.

Key words: gene family, gene order, inversion, duplication, phylogeny

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. Such large segment duplication is a primary mechanism for generating gene clusters on chromosomes.

Many gene families of the human genome are organized in tandem, including HOX genes [31], immunoglobulin and T-cell receptor genes [21], MHC genes [20] and olfactory receptor genes [11]. 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.

Based on the initial evolutionary model of tandemly repeated sequences introduced by Fitch [9], a number of recent studies have considered the problem of reconstructing a tandem duplication history of a gene family [5–7, 16, 28, 32].

⋆ The two first authors have contributed equally to this work

These are essentially phylogenetic inference methods using the additional constraint that the resulting tree should induce a duplication history concordant with the given gene order. When a phylogeny is already available, a linear-time algorithm can be used to check whether it is a duplication tree [32]. However, even for gene families that have evolved through tandem duplications, it is often impossible to reconstruct a duplication history [7]. 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 [25]. Although genome rearrangement with inversions has received large attention in the last decade [14, 17, 4, 26, 2], beginning with the polynomial-time algorithm of Hannenhalli and Pevzner for computing the reversal distance between two signed gene orders [14], 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 [12, 19, 18, 3]. 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 may be very large, we restrict ourselves to finding the minimum number of inversions required to explain a given ordered phylogeny. As a first attempt, we only considered tandem duplications involving single genes. Though the model described by Fitch [9] allows for simultaneous duplications of several gene copies, single duplications are known to be predominant over multiple duplications [1, 9, 28].

After describing the evolutionary models in section 2 and the optimization problem in section 3, we present our main branch-and-bound algorithm in section 4. Finally, in section 5, we test the algorithm's time-efficiency on simulated data and show its usefulness to detect, among a set of possible phylogenies, the "wrong" ones. An application on KRAB zinc finger genes is presented.

2 The Evolutionary Model

2.1 Duplication Model

This model, first introduced by Fitch [9], is based on unequal recombination during meiosis, which is assumed to be the sole evolutionary mechanism (except point mutations) acting on sequences. Consequently, 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 1.a). Duplications may be *simple* (duplication of a single gene) or multiple (simultaneous duplication of neighboring genes). In this paper, we only consider simple duplications.

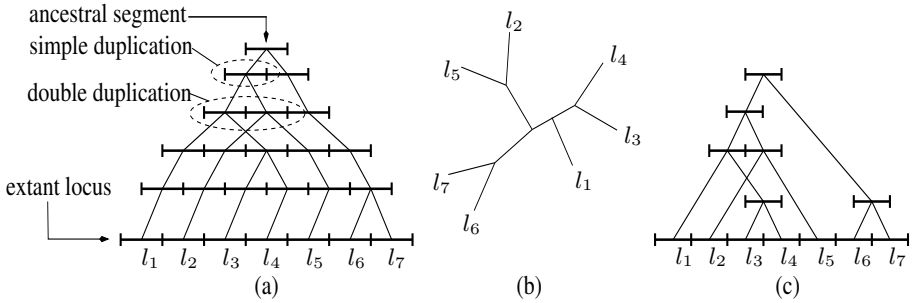

Fig. 1. (a) Duplication history; each segment represents a copy. (b) The unrooted duplication tree corresponding to history (a). (c) The duplication tree corresponding to history (a).

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 1.b). Formally, an *ordered phylogeny* is a pair (T, O) where T is a phylogeny and O is the ordered sequence of its leaves. According to this model, all the genes have the same transcriptional orientation.

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 [10] (Figure 1.c).

In the rest of this paper, a *duplication tree* will refer to a *simple rooted duplication tree*, that is a rooted duplication tree that is compatible with at least one history involving only simple duplications. 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 simple duplication model is thus inadequate to describe their evolution. To circumvent this limitation, we propose an extended model of duplication which includes inversions. Thereafter, the transcriptional orientations

of the genes in a *signed* ordered phylogeny (T, O) are specified by signs $(+/-)$ in O. Thus O is formally a signed permutation of the leaves of T. We denote by $d_{inv}(O_i, O_j)$ the inversion distance between the two signed permutations 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 $O_1 = (\pm v)$ one of the two trivial orders.*
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 leaf v. In this case O_{i+1} is obtained from O_i by replacing $\pm v$ by $(\pm u, \pm w)$. 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. 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 $\pm v$ of (T_k, O_k) is replaced by two consecutive leaves $(\pm u, \pm w)$ 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 we have $+v$ in O_k, we obtain O'_{k+1} by replacing $+v$ by $(+u, +w)$ in O'_k. Otherwise we have $-v$ in O_k and we obtain O'_{k+1} by replacing $+v$ by $(+w, +u)$ in O'_k. 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 [14], as it will be used in our approach.

4.1 Hannenhalli-Pevzner (HP) Algorithm

Given two signed permutations 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). The algorithm is based on a bicolored graph, called the *breakpoint graph*, constructed from the two permutations 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. The graph contains two classes of edges: the real and desired edges (as named in [24]). 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*. 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 [14]:

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

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'.

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. 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 2.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 such order admits two transcriptional orientations according to our definition of a duplication tree.

Lemma 2. *An order O' is compatible with T iff it is defined by an assignment of all the binary variables b_i in T and all the genes have the same sign.*

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.

4.3 A Lower Bound for the Inversion Distance

To avoid computing $d_{inv}(O, O')$ for each of the 2^n orders O' compatible with T, we consider a branch-and-bound strategy similar to the one used in [33]. The

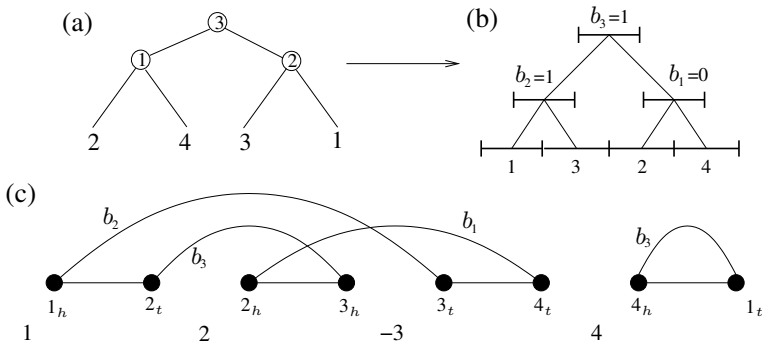

Fig. 2. (a) A phylogeny with an appropriate depth-first labeling of the internal nodes; (b) The duplication tree corresponding to an assignment of the b_i variables of (a); (c) The breakpoint graph illustrating the difference between the gene order $O' = (1, 3, 2, 4)$ obtained from the duplication tree (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). For simplicity, the genome is assumed to be circular (gene 1 next to gene 4).

idea is to compute a lower bound on $d_{inv}(O, O')$ as we progressively define O^* by updating the breakpoint graph of (O, O^*). In order to progressively construct this graph, it is essential to define the b_i values in a depth-first manner according to 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 precedes 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 breakpoint graph between l_{max} and r_{min} (or r_{max} and l_{min}) (see Figure 2.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 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:

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

The branch-and-bound algorithm proceeds as follows. An initial assignment of all binary variables is considered and the corresponding inversion distance is computed. Each following step re-assigns the binary variables in a depth-first manner. At each step, we backtrack to the last node (closest to the root) that has not been re-assigned yet. The re-assignment procedure continues as long as the partial order O^* obtained is such that $d^*_{inv}(O, O^*) < \min_{inv}$, where $\min_{inv}$ is the lowest inversion distance obtained from the previous steps. This is justified

by the fact that any total order that can be obtained from O^* cannot be smaller than the current best value. Every time we reach a leaf, we use the HP algorithm to compute the exact reversal distance $d_{inv}(O, O')$.

5 Results

5.1 Branch-and-Bound Efficiency

To test the efficiency of the branch-and-bound algorithm, we generated 3 sets of 500 phylogenies each with respectively 10, 20 and 40 leaves using r8s [23]. We then defined arbitrarily compatible orders to obtain a total of 1,500 duplication trees. For each of them, we performed 1, 2, 4 and 8 inversions to obtain 12 datasets containing a total of 6,000 signed ordered phylogenies which are no longer duplication trees.

We applied our algorithm on each dataset and measured the execution time (on a Pentium 4) and the average fraction of nodes explored in the search space. Results are given in Table 1. We observe that the algorithm is very efficient and can be used on relatively important phylogenies within reasonable time.

Table 1. Average fraction of nodes explored in the search tree during the branch-and-bound / Execution time (in seconds) for the 500 signed ordered phylogenies.

	1 inversion	2 inversions	4 inversions	8 inversions
10 leaves	1×10^{-2} / 13	2×10^{-2} / 20	6×10^{-2} / 35	0.1/51
20 leaves	3×10^{-5} / 15	8×10^{-5} / 20	2×10^{-4} / 37	2×10^{-3} / 90
40 leaves	7×10^{-11} / 17	2×10^{-10} / 24	1×10^{-9} / 39	2×10^{-8} / 112

5.2 Application on simulated data

We applied our algorithm on simulated data to verify how it could be used to validate inferred phylogenies on tandemly repeated gene families. Using the simulation protocol described in the previous section, we randomly generated 500 duplication trees with 15 leaves. For each one of them we performed 0, 2, 4 and 6 inversions to obtain 4 datasets containing a total of 2,000 signed ordered phylogenies. These are the observable states (T_{true}, O) resulting from "true" duplication/inversion histories. For each T_{true}, we then randomly generated two "wrong" (but close) phylogenies T_{wrong}, that can be obtained by applying respectively one or two Nearest Neighbor Interchange rearrangements (NNI) [27]. Those "wrong" phylogenies can be seen as the ones we could obtain from biological data when a few nodes have weak statistical support. Finally, we used our algorithm to compute the minimum number of inversions inv() necessary in a simple duplication/inversion history to explain each (T_{true}, O) and all its corresponding T_{wrong}. The averaged results are presented in Table 2.

Results can be interpreted as follows. For a wrong phylogeny T_{wrong}, 50% of the time on average our algorithm reports an excess of inversions, otherwise

Table 2. Percentage of times $\text{inv}(T_{wrong}, O)$ is less, equal or greater than $\text{inv}(T_{true}, O)$. Averaged over all possible neighbors for each of the 500 phylogenies.

	Trees distant from one NNI				Trees distant from two NNI			
$\text{inv}(T_{wrong}, O)$	0 *inv.*	2 *inv.*	4 *inv.*	6 *inv.*	0 *inv.*	2 *inv.*	4 *inv.*	6 *inv.*
$< \text{inv}(T_{true}, O)$	0.0	0.2	0.8	1.6	0.0	0.2	0.08	1.6
$= \text{inv}(T_{true}, O)$	49.0	50.4	57.6	68.0	36.6	35.0	41.2	54.4
$> \text{inv}(T_{true}, O)$	51.0	49.4	41.6	30.4	66.4	64.8	58.0	44.0

it reports the same number of inversions compared to the true phylogeny T_{true}. Suppose that we are presented some ordered phylogenies. One is correct and the others differ by a few NNIs. According to Table 2, for wrong trees, the algorithm almost always reports the same number of inversions or more as in the true tree. Thus, choosing the phylogeny with the lowest number of inversions is either a winning strategy (roughly 50% of the time) or useless, but is almost never misleading. Of course, this ability to discard wrong phylogenies decreases as the true number of inversions increases.

5.3 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 [13], Hamilton *et al.* proposed a phylogeny of the primate specific ZNF91 sub-family based on their tether[3] and flanking sequences. This phylogeny (obtained by Neighbor-Joining [22]) 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 our 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 [8] 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[4] [15] and aligned using ClustalW [29] 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 [30] 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 our algorithm on the sampled phylogenies and observed that the best one (p=0.4) is compatible with an optimal duplication/inversion history involving only two inversions. Phylogenies are presented in Figure 3.

[3] The region upstream from the first finger.
[4] http://znf.llnl.gov/catalog/

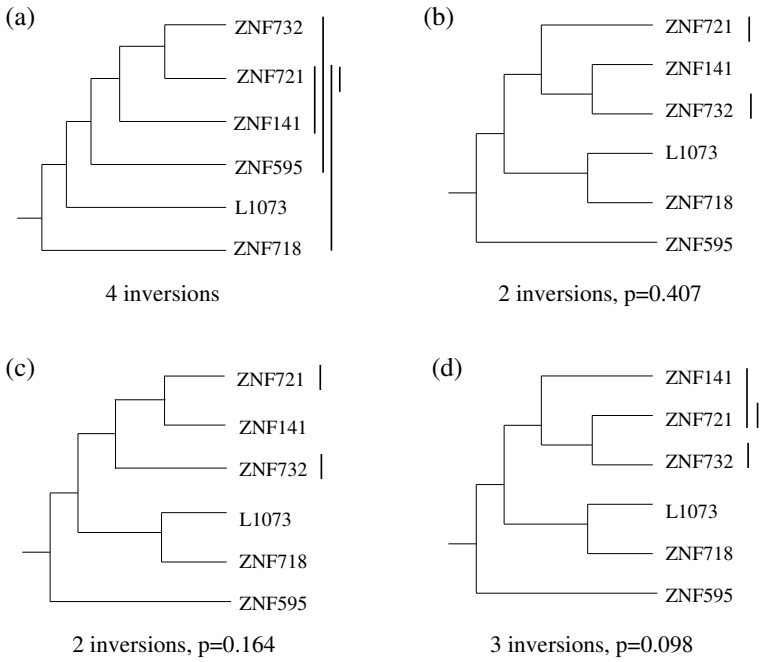

Fig. 3. Different phylogenies for the ZNF141 clade on human chromosome 4, with the associated minimal number of inversions in a dup/inv history. The black vertical lines represent an optimal sequence of inversions leading to the *signed* gene order observed on the chromosome: (+ZNF595,+ZNF718,+L1073,−ZNF732,+ZNF141,−ZNF721). (a) The phylogeny published in [13] 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.

6 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. 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.

The next step of this work will be to account for multiple duplications in the evolutionary model. 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.* [25] 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 to explain such a phylogeny based on an evolutionary model involving tandem duplication, inversion and speciation events.

Acknowledgments. The authors wish to thank M. Aubry and H. Tadepally for their help on zinc finger genes. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (N.E.M.) and the Canadian Institutes of Health Research (M.L.).

References

1. G. Benson and L. Dong. Reconstructing the duplication history of a tandem repeat. In *Proceedings of Intelligent Systems in Molecular Biology (ISMB1999), Heidelberg, Germany*, pages 44–53. AAAI, 1999.
2. A. Bergeron, J. Mixtacki, and J. Stoye. Reversal distance without hurdles and fortresses. volume 3109 of *LNCS*, pages 388 - 399. Springer-Verlag, 2004.
3. K. Chen, D. Durand, and M. Farach-Colton. Notung: Dating gene duplications using gene family trees. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, New York, 2000. ACM.
4. N. El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *CPM 2000*, volume 1848 of *LNCS*, pages 222- 234, 2000.
5. O. Elemento and O. Gascuel. A fast and accurate distance-based algorithm to reconstruct tandem duplication trees. *Bioinformatics*, 18:92–99, 2002.
6. O. Elemento and O. Gascuel. An exact and polynomial distance-based algorithm to reconstruct single copy tandem duplication trees. *Journal of Discrete Algorithms*, 2(2-4):362–374, 2005.
7. O. Elemento, O. Gascuel, and M-P. Lefranc. Reconstructing the duplication history of tandemly repeated genes. *Molecular Biology and Evolution*, 19:278–288, 2002.
8. J.P. Huelsenbeck F. Ronquist. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–4, 2003.

9. W.M. Fitch. Phylogenies constrained by cross-over process as illustrated by human hemoglobins in a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I. *Genetics*, 86:623–644, 1977.
10. O. Gascuel, D. Bertrand, and O. Elemento. Reconstructing the duplication history of tandemly repeated sequences. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 205–235. Oxford University Press, 2005.
11. G. Glusman, I. Yanai, I. Rubin, and D. Lancet. The complete human olfactory subgenome. *Genome Research*, 11(5):685–702, 2001.
12. R. Guigó, I. Muchnik, and T.F. Smith. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189–213, 1996.
13. A.T. Hamilton, S. Huntley, M. Tran-Gyamfi, D.M. Baggott, L.Gordon, and L. Stubbs. Evolutionary expansion and divergence in the znf91 subfamily of primate-specific zinc finger genes. *Genome Research*, 16(5):584–594, 2006.
14. S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *J. ACM*, 48:1–27, 1999.
15. S. Huntley, D.M. Baggot, A.T. Hamilton, S. Yang M. TranGyamfi, J. Kim, L. Gordon, E. Branscomb, and L. Stubbs. A comprehensive catalogue of human krab-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Research*, 16:669–677, 2006.
16. D. Jaitly, P. Kearney, G. Lin, and B. Ma. Methods for reconstructing the history of tandem repeats and their application to the human genome. *Journal of Computer and System Sciences*, 65:494–507, 2002.
17. H. Kaplan, R. Shamir, and R. E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29:880–892, 2000.
18. Bin Ma, Ming Li, and L. Zhang. On reconstructing species trees from gene trees in term of duplications and losses. In S. Istrail, P.A. Pevzner, and M.S. Waterman, editors, *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pages 182–191, New York, 1998. ACM.
19. R.D.M Page and M.A. Charleston. Reconciled trees and incongruent gene and species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:57–70, 1997.
20. J. Robinson, M.J. Waller, P. Parham, N. de Groot, R. Bontrop, L.J. Kennedy, P. Stoehr, and S.G. Marsh. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Research*, 31(1):311–4, 2003.
21. M. Ruiz, V. Giudicelli, C. Ginestoux, P. Stoehr, J. Robinson, J. Bodmer, S.G. Marsh, R. Bontrop, M. Lemaitre, G. Lefranc, D. Chaume, and M-P. Lefranc. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research*, 28:219–221, 2000.
22. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.
23. M.J. Sanderson. r8s; inferring absolute rates of evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19:301 - 302, 2003.
24. J.C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*, chapter 7. PWS Pub. Co., 1997. SET j 97:1 1.Ex.
25. M. Shannon, A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. Differential expension of Zinc- Finger transcription factor loci in homologous human and mouse gene clusters. *Genome Research*, 13:1097 - 1110, 2003.
26. A. Siepel. Algorithm to find all sorting reversals. In *Proceedings of the second conference on computational molecular biology (RECOMB'02)*, pages 281 - 290. ACM Press, 2002.

27. D.L. Swofford, P.J. Olsen, P.J. Waddell, and D.M. Hillis. *Molecular Systematics*, chapter Phylogenetic Inference, pages 407–514. Sinauer Associates, Sunderland, Massachusetts, 1996.
28. M. Tang, M.S. Waterman, and S. Yooseph. Zinc finger gene clusters and tandem gene duplication. In *Proceedings of International Conference on Research in Molecular Biology (RECOMB2001)*, pages 297–304, 2001.
29. J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673 - 4680, 1994.
30. P. Waddel and M. Steel. General time reversible distances with unequal rates across sites: Mixing t and inverse gaussian distributions with invariant sites. *Molecular Phylogeny and Evolution*, 8:398–314, 1997.
31. J. Zhang and M. Nei. Evolution of antennapedia-class homeobox genes. *Genetics*, 142(1):295–303, 1996.
32. L. Zhang, B. Ma, L. Wang, and Y. Xu. Greedy method for inferring tandem duplication history. *Bioinformatics*, 19:1497–1504, 2003.
33. C. Zheng, A. Lenert, and D. Sankoff. Reversal distance for partially ordered genomes. *Bioinformatics*, 21:i502 - i508, 2003.