Reconstruction of Ancestral Genome subject to Whole Genome Duplication, Speciation, Rearrangement and Loss

Denis Bertrand[1], Yves Gagnon[1], Mathieu Blanchette[2], and Nadia El-Mabrouk[1]

[1] DIRO, Université de Montréal, H3C 3J7, Canada, (dbertran78@gmail.com, y.gagnon@umontreal.ca, mabrouk@iro.umontreal.ca)
[2] McGill Centre for Bioinformatics, McGill University, H3A 2B4, Canada, blanchem@mcb.mcgill.ca.

Abstract. Whole genome duplication (WGD) is a rare evolutionary event that has played a dramatic role in the diversification of most eucaryotic lineages. Given a set of species known to have evolved from a common ancestor through one or many rounds of WGD together with a set of genome rearrangements, and a phylogenetic tree for these species, the goal is to infer the pre-duplicated ancestral genomes. We use a two step approach: (1) Compute a score for each possible ancestral adjacency at each internal node of the phylogeny; (2) Combine adjacencies to form ancestral chromosomes. The main contribution of our method is the computation of a rigorous score for each potential ancestral adjacency (a, b), reflecting the maximum number of times a and b can be adjacent in the whole phylogeny, for any setting of ancestral genomes. We first apply our method on simulated datasets and show a high accuracy for adjacency prediction. We then infer the pre-duplicated ancestor of a set of 11 yeast species and compare it to a manually assembled ancestral genome obtained by Gordon *et al.* (2009).

1 Introduction

Whole genome duplication (WGD) is a spectacular evolutionary event that has the effect of simultaneously doubling all the chromosomes of a genome. Evidence for WGD events has shown up across the whole eukaryote spectrum, from the protist *Giardia* to the yeast species [7], including most plant lineages, several insect, fish, amphibians, and even to mammalian species. For some genomes, recent duplication is easily detected by the presence of a nearly complete set of duplicated chromosomes. However, in most cases, due to a series of intrachromosomal and interchromosomal movements disrupting the initial perfectly doubled structure of the genome, all that we can observe is a set of duplicated blocks (chromosomal segments or genes) representing a high proportion of the genome, scattered throughout the genome.

Studying the evolution of a set of species that have been subject to one or many WGD events during their common evolution is challenging due to the high rates of paralogy in their genomes. Inferring the content and chromosome

organization of ancestral genomes preceding the WGD events is a major step towards solving this difficulty, and also answering numerous biological questions such as the mechanisms of polyploid formation, the variation in rearrangement rates, gene losses and gains through the phylogenetic tree, and the consequence of such variations on the genetic and physiological specificities of species.

In 2003, we have presented the first formal result related to genome duplication, which is an exact linear-time algorithm for solving the *genome halving problem* [5]: Given a present-day genome G represented as a set of strings (chromosomes) with each block present exactly twice, the genome halving problem asks to infer a perfectly duplicated genome H (a genome with exactly two copies of each chromosome) minimizing the rearrangement distance to G (inversions, reciprocal translocations or both). Our results have been reformulated recently by Alekseyev and Pevzner [1] using an alternative representation of the breakpoint graph. Subsequently, Sankoff and colleagues [16, 15], and more recently Gavranović and Tannier [6], used variations of the genome halving strategy (Guided Genome Halving or GGH) to find the preduplicated ancestor of a doubled genome in the presence of a non-duplicated outgroup [16, 15]. As noticed in [7], the GGH algorithms can hardly be generalized to a complete phylogenetic tree, with more than one WGD event on a path from an extant species to the root of the tree, and an arbitrary number of post-WGD genomes and non-WGD outgroups. Moreover, as for genome halving, GGH algorithms can only consider genes that have retained two copies after the WGD. In the case of reconstructing the ancestor of *Saccharomyces cerevisiae*, Gordon *et al.* [7] have noticed that less than 20% of all genes can be taken into account by the GGH strategy. Subsequent work shows that this limitation can be circumvented by grouping genes into double conserved syntenies [12, 6].

In this paper, we consider the general problem of inferring the pre-duplicated genome preceding the first duplication event in a multi-species evolutionary history involving WGDs, rearrangement events, and block losses. The input of our problem is a set of extant genomes, each represented as a set of strings on an alphabet of blocks, each block potentially present more than once in each genome, a phylogenetic tree representing the evolution of the species, with specific branches marked with WGD events. Such data and phylogenetic information is available for a variety of eukaryotic lineages, such as the yeast species [7], grass genomes [14], angiosperms [13] and many other lineages. Our approach for ancestral genome prediction is to maximize the conservation of block adjacencies in the phylogeny. We use a two-step methodology: (1) at each node of the phylogeny, compute the adjacency score of each pair (a, b) of blocks; (2) infer a pre-duplicated ancestral genome by an optimal chaining of adjacencies.

The main contribution of our method is the computation of a rigorous score for each potential ancestral adjacency (a, b), reflecting the maximum number of times a and b can be adjacent in the whole phylogeny, for any setting of ancestral genomes. As it is the case for the other local approaches [4], in the absence of a complete set of reliable syntenies, the output of our algorithm is

a set of *Contiguous Ancestral Regions fragments* (CAR) [11, 4], rather than a completely assembled ancestral genome.

The approaches most comparable to ours are those developed by Ma *et al.* (see the method in [11] for single gene copies, and its generalization to genomes with duplications in [9]), both of which using a Fitch approach for inferring ancestral adjacencies. We show that our approach outperforms the former on simulated data. The latter can only be used if accurate gene trees, with branch lengths, are available, which is often limiting. In contrast, our approach works under stronger assumptions but requires only a species tree and extant genomes as input. Our paper is structured as follows: after introducing basic notations, we introduce the notion of adjacency scores, show how to compute it, and how to use it to assemble putative ancestral genomes by solving an instance of the traveling salesperson problem. We then show, using simulated data, that the predicted pre-WGD genomes are highly accurate, even in the presence of a large number of rearrangements. Finally, we apply our approach to the prediction of the ancestral pre-WGD yeast genome and obtain results very similar to the hand-curated ancestral genome produced by Gordon *et al.* [7].

2 Preliminaries

Notation: Let $\mathcal{B}$ be a set of unsigned blocks (e.g. genes, or any other type of genomic markers). A *string* is a sequence of blocks from $\mathcal{B}$, where each block is signed ($+$ or $-$) to mark its orientation. A *genome* G is a collection of strings $C_1, C_2, \cdots C_N$ called its *chromosomes*, where each element of $\mathcal{B}$ may be present more than once. To represent chromosomal ends, we add an artificial block O, which is also added to our alphabet $\mathcal{B}$, at an extremity of each chromosome, and consider each chromosome as circular. We denote by $\Sigma_G \subseteq \mathcal{B}$ the set of blocks present in G (including O), and by $mult(a, G)$ the multiplicity of block a in G. In particular, the multiplicity of O is the number of chromosomes of G. We denote by $\pm\Sigma_G$ the set obtained from Σ_G by considering each block in its positive and negative directions. By convention, the artificial block O is always considered positively signed. A *multiset* of $\pm\Sigma_G$ is a subset of $\pm\Sigma_G$ with possibly repeated blocks.

Let $a \in \Sigma_G$ and $b \in \pm\Sigma_G$. We say that b is a left-adjacency of a in G iff "$b\ +a$" or "$-a\ -b''$ is a substring of G. Symmetrically, b is a right-adjacency of a in G iff "$+a\ b''$ or "$-b\ -a$" is a subsequence of G. We denote by $LA(G, a)$ and $RA(G, a)$ the *multisets* of left- and right-adjacencies of the one or more copies of a in G.

Evolutionary model: A *Whole Genome Duplication* (or WGD for short) is an event transforming a genome $G = \{C_1, C_2 \cdots C_N\}$ into a genome G^D containing $2N$ chromosomes, i.e. $G^D = \{C_1, C'_1, C_2, C'_2 \cdots C_N, C'_N\}$, where, for each $1 \leq i \leq N$, $C_i = C'_i$. Let $G_1, G_2, \cdots G_n$ be a set of n related species at the leaves of a species tree T, assumed to have evolved from a common ancestor through WGD events, intra-chromosomal (inversions or transpositions of chromosomal segments) and inter-chromosomal (reciprocal translocations between

two chromosomes, fusions of two chromosomes or fissions of one chromosome) rearrangements, and block losses. A phylogeny for $(G_i)_{i=1}^n$ is a tree T with n leaves, where G_i, for $1 \leq i \leq n$, is the label of leaf i, and each internal node (also called *speciation node*) has exactly two children and represents a speciation event.

In our model, WGDs are the only duplication events responsible for block multiplicity (in particular, single-block duplications are not considered). In addition, we suppose that, in each genome, at least one block reflects the doubling status of the genome, i.e. there exists a block that has not lost any copies. As noticed by Zheng *et al.* [16], under this assumption, a history with a minimum number of WGD events can be easily deduced from the number of copies of the most frequent genes found in each genome. In order to account for those duplication events, we create new internal nodes in T, called *WGD nodes*, and position them appropriately on the edges of T. Contrary to speciation nodes, each WGD node has only a single child. Moreover, if all genomes G_i, for $1 \leq i \leq n$, have multiplicity greater than 1, then we have to add one or more WGD nodes above the root r of T. In this case, we create a new root D, that we call *the duplication root of* T.

Assuming a model with no convergent evolution and minimum losses, the multiset of blocks Σ_u present at node u can be obtained as follows. Let $A(a)$ be the node of T representing the least common ancestor of the leaves that contain a given block a. Then, we assign a to each node belonging to a path from $A(a)$ to any leaf containing a. In order to define the multiset Σ_u, we also need to know the multiplicity of each block at u, which can be recursively defined as the maximum of its multiplicities in u's two children. See Figure 1 left, for an example.

3 Problem definition

Given a species tree T for the genomes $(G_i)_{i=1}^n$, augmented with WGD nodes as described in the previous section, we want to infer the pre-duplicated ancestral genomes, i.e. the ancestral genomes just preceding the WGD nodes on the paths from the root D of T to a leaf. We will use a parsimony criteria seeking a solution with a minimum number of adjacency changes along the branches of T, or, equivalently, a maximum of adjacency conservation.

Ancestral genome assignment. We assume that the multiset Σ_u of blocks present at each internal node of T has already been determined. A *genome assignment* $G(u)$ at u is a genome on $\mathcal{B}$ respecting the content and multiplicity constraints given by Σ_u. If u is a WGD node, an additional constraint is that $G(u)$ must be a duplicated genome.

Let u and v be two nodes of T with u being the parent of v. In the case of genomes with single gene copies, it is easy to define the number of adjacencies preserved along branch (u, v) as the number of common substrings of size 2 between them. This definition is not directly transposable to the case of

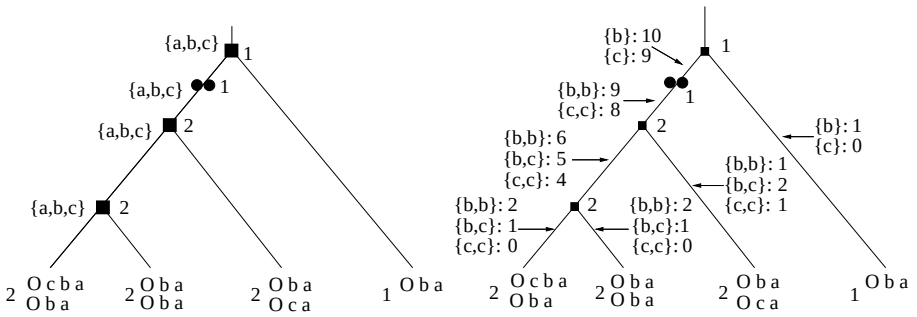

Fig. 1. Left: A species tree with each leaf labeled with its corresponding genome and multiplicity number, and each internal node labeled with the multiplicity and block set of the ancestral genome just preceding the speciation or WGD event. Squares indicate speciation nodes, and the double circle indicates a WGD node. Right: An illustration of Algorithm L^{below} for the adjacencies of gene a. Each edge (u, v) is labeled by its valid multisubsets. For each branch (u, v) and each multisubset X, the indicated number is the value of $L^{below}_{(u,v)}(a, X)$.

genomes with multiple gene copies, as the one to one orthology between genes is not set. Instead, for each block a, we compare its left and right-adjacency multisets in $G(u)$ and $G(v)$. More precisely, we define $adjCons(a, G(u), G(v)) = |LA(G(u), a) \cap LA(G(v), a)| + |RA(G(u), a) \cap RA(G(v), a)|$, as the number of left and right conserved adjacencies of a on the branch (u, v), and

$$adjCons(G(u), G(v)) = \sum_{a \in \Sigma_u \cap \Sigma_v} adjCons(a, G(u), G(v))$$

as the total number of left and right conserved adjacencies on the branch (u, v). In both formulas, intersections and cardinalities are taken over multisets. Notice that $adjCons(G(u), G(v))$ accounts for each adjacency conservation twice.

We then define $adjCons(T)$ as the maximum number of conserved adjacencies in T, over all possible ancestral genome assignments $G(u_1), \cdots G(u_k)$ at all internal (speciation and WGD) nodes $n_1, \cdots n_k$ of T:

$$adjCons(T) = \max_{G(u_1),\ldots,G(u_k)} \sum_{(u,v) \in E(T)} adjCons(G(u), G(v))$$

Finally, for a given ancestral node u_i with genome assignment $H(u_i)$, define

$$adjCons(T|_{G(u_i)=H(u_i)}) = \max_{G(u_1),\ldots,G(u_k)|G(u_i)=H(u_i)} \sum_{(u,v) \in E(T)} adjCons(G(u), G(v)),$$

which is the maximum number of adjacencies that can be preserved along the branches of T, if the genome at node u_i is set to $H(u_i)$. We can now state our optimization problem precisely.

Ancestral Genome Assignment Problem:
INPUT: A species tree T for the genomes $(G_i)_{i=1}^n$ augmented with one or more WGD nodes as described in the previous section; The multiset of blocks at each internal node; A particular WGD node u of interest.
OUTPUT: An ancestral genome assignment $H(u)$ to u such that $adjCons(T|_{G(u)=H(u)})$ is maximized.

Of course, one may formulate the same reconstruction problem for non-WGD nodes. However, those are typically less well constrained by the data at the leaves, yielding a large number of potential optimal solutions. In contrast, optimal assignments at WGD nodes, although not generally unique, have less flexibility. In this paper, we focus on inferring the pre-duplicated genomes preceding a first WGD event on a branch from the root of T to a leaf. In other words, u is the first WGD node on a branch from the root of T to a leaf.

4 Method

We start by defining an upper bound on our objective function, $adjCons(T|_{G(u)=H(u)})$. We define $adjCons(a, T|_{u,\mathcal{Z}})$ as the maximum number of left and right adjacencies of a that can be preserved over the whole tree, for any assignment $G(u_1), ..., G(u_k)$ of ancestral genomes such that $G(u)$ satisfies the set of constraints specified in $\mathcal{Z}$. Then, it is straightforward to show that

$$\begin{aligned} adjCons(T|_{G(u)=H(u)}) &\leq \sum_a adjCons(a, T|_{LA(a,G(u))=LA(a,H(u)),RA(a,G(u))=RA(a,H(u))}) \\ &\leq \sum_a adjCons(a, T|_{LA(a,G(u))=LA(a,H(u))}) + \\ &\qquad adjCons(a, T|_{RA(a,G(u))=RA(a,H(u))}) \end{aligned}$$

Our ancestral reconstruction algorithm thus seeks a genome H such that the above term is maximized. It proceeds in two steps:

1. For each internal node u of the tree (speciation or WGD node), each block $a \in \Sigma_u$, and each multisets X of possible left adjacencies of a at node u, we compute $adjCons(a, T|_{LA(a,G(u))=X})$, reflecting the maximum number of left-adjacencies that can be preserved over the whole tree, for any setting of ancestral genome assignment to internal nodes with the condition that the genome $G(u)$ satisfies $LA(a, G(u)) = X$. These quantities are computed using a dynamic programming algorithm described below. We then proceed similarly for right adjacencies.
2. For the WGD node u for which an ancestral genome is sought, we obtain the desired pre-duplicated genome by chaining the adjacencies at node u in a optimal way.

4.1 Computing adjacency scores

We first describe how to compute $adjCons(a, T|_{LA(a,G(u_i))=X})$, for any node u_i, block $a \in \Sigma_{u_i}$, and candidate left-adjacencies X. The algorithm to compute

right-adjacencies is very similar. Consider an edge (u, v) in T, where u is the parent of v. Let X be a multisubset of $\pm\Sigma_u$. Let $G(u)$ be a genome assignment at node u such that $LA(a, G(u)) = X$. We define $L^{below}_{(u,v)}(a, X)$ as the maximum number (over all possible genome assignments of T's internal nodes) of left-adjacencies involving the copies of a that can be preserved along the branch (u, v) and all the branches of the subtree rooted at node u. Similarly, we define $L^{above}_{(u,v)}(a, X)$ as the maximum number of left-adjacencies involving a that can be preserved, along branch (u, v) and all the branches outside the subtree rooted at node u. Then, for an internal node u with children v and w and parent p, we obtain

$$adjCons(a, T|_{LA(a,G(u))=X}) = L^{below}_{(u,v)}(a, X) + L^{below}_{(u,w)}(a, X) + L^{above}_{(p,u)}(a, X).$$

Notice that, if u is a WGD node, then u has a single child v, and thus the term $L^{below}_{(u,w)}(a, X)$ should be removed from the above formula. Similarly, if u is the root of the tree, then the term $L^{above}_{(p,u)}(a, X)$ should be removed.

We are thus interested in calculating the tables $L^{below}_{(u,v)}$ and $L^{above}_{(u,v)}$ for each edge (u, v) of T. Those are obtained by the dynamic programming algorithms shown in Figures 5 and 6. An illustration of this algorithm is given on the right tree of Figure 1.

Although expressed in a recursive manner for simplicity, both algorithms can be re-written using a dynamic programming approach that proceeds in a bottom-up fashion to obtain L^{above} and in a top-down fashion to obtain L^{below}. The running time to compute $adjCons(a, T|_{LA(a,G(u))=X})$ is thus $\Sigma_{(u,v)\in T}(|\pm\Sigma_u|^{mult(a,G(u))} \times |\pm\Sigma_v|^{mult(a,G(v))})$.

4.2 Assembling an adjacency-preserving pre-duplication ancestral genome

We now seek to build a solution to the Ancestral Genome Assignment Problem, i.e. to infer a pre-duplication genome at a given WGD node u, aiming to maximize the number of conserved adjacencies on T. We achieve this by solving a Traveling Salesperson Problem (TSP) on a complete undirected graph where vertices correspond to blocks. We initially weighted edges according to our upper bound L^{all}_u. However, this weighting gives too much importance to adjacencies implying blocks with high multiplicity. Thus, we decided to weight the edges according to the ratio $rL^{all}_u(a, X) = L^{all}_u(a, X)/adjConsMax(a, T)$, where $adjConsMax(a, T) = \Sigma_{(u,v)\in E(T)} \min(mult(a, G_u), mult(a, G_v))$. Notice that $adjConsMax(a, T)$ represents the number of conserved adjacencies for the block a in T if a is always adjacent to the same gene in all leaves of T. This ratio allows us to evaluate the confidence of an inferred adjacency (see Figures 3 and 4 top right).

More precisely, we build an undirected graph Q that contains a pair of vertices a^t, a^h for each block $a \in \Sigma_u$, as well as a set of vertices O_1, O_2, O_k marking chromosome ends, where k is chosen to be at least as large as (but possibly larger

than) the maximum number of chromosomes in the ancestral genome we seek to infer. Edges weights are chosen as follows, for $a \neq b \in \Sigma_u$, $i \neq j \in \{1, ..., k\}$, and M some large number:

$w(a^h, b^t) = rR_u^{all}(a, \{+b\}) + rL_u^{all}(b, \{+a\})$ $\quad$ $w(a^t, a^h) = M$
$w(a^h, b^h) = rR_u^{all}(a, \{-b\}) + rR_u^{all}(b, \{-a\})$ $\quad$ $w(O_i, a^t) = 2 \times rL_u^{all}(a, \{O\})$
$w(a^t, b^t) = rL_u^{all}(b, \{-a\}) + rL_u^{all}(a, \{-b\})$ $\quad$ $w(O_i, a^h) = 2 \times rR_u^{all}(a, \{O\})$
$w(a^t, b^h) = rL_u^{all}(a, \{+b\}) + rR_u^{all}(b, \{+a\})$ $\quad$ $w(O_i, O_j) = 0$

Because a^t and a^h are connected by heavy edges, any maximum weight hamiltonian path must include all of them. A hamiltonian cycle through Q thus defines a set of strings (chromosomes; delimited by O vertices), with some possibly empty (two consecutive O vertices). Starting from O_1, the cycle visits pairs (a^t, a^h) (corresponding to $+a$) or (a^h, a^t) (corresponding to $-a$). The heaviest hamiltonian cycle through Q thus corresponds to an hypothetical ancestral genome H at u that preserves a large number of adjacencies.

The instance of the TSP we need to solve here is a symmetrical weighted graph with $2 \cdot |\Sigma_u| + k$ vertices. In the case of the application to the reconstruction of the ancestral pre-duplication yeast genome, $|\Sigma_u| = 4705$, so the graph is quite large. Although an NP-Complete problem, TSP is one of the best studied algorithmic problems and excellent heuristics exist. We considered two of them. The first is a simple greedy approach that repeatedly selects the heaviest edge remaining unless this results in the premature closing of a cycle. The second is the Chained Lin-Kernighan heuristic [10] implemented in the Concord package [2]. Although not guaranteed to produce an optimal solution, this heuristic has proved highly accurate in other contexts [3].

5 Results

Studying the evolution of genomes through whole genome duplication is only possible on species exhibiting clear traces of genome duplication. Moreover, a strong prerequisite for reconstructing accurate ancestral genomes is to have enough data on extant species, and sufficient colinearity of gene order among a reasonably large number of related species. Yeast genomes are a perfect example of a data set satisfying all these conditions. Following the extensive work of Wolfe and colleagues during the last decade, it is now almost universally accepted that *Saccharomyces cerevisiae* is the descendant of an ancient whole-genome duplication event. Moreover, the availability of a large number of completely sequenced yeast genomes spanning a large evolutionary time-depth (comparable to that of the vertebrates), as well as the Yeast Gene Order Browser [7], provides the material for an accurate ancestral genome reconstruction. We therefore focus, on this paper, on the study of the yeast species data sets.

To be able to evaluate our reconstructed pre-duplicated ancestor of *Saccharomyces cerevisiae*, we first test our method on a data set obtained through a simulated evolution that is as close as possible to the one observed for yeast species. This is explained in the following section.

5.1 Simulated data sets

The phylogenetic tree given in Figure 2 reflects the evolution of the 11 yeast species recorded in the Yeast Gene Order Browser, as given by [8]. Gene sets at leaves are those provided in [7], and gene sets at internal nodes, as well as the number of gene losses on each branch, are directly inferred from those at the leaves. To simplify the study, we only consider the 4705 genes present at the pre-duplicated ancestral node (i.e. we remove from the gene content of each leaf those genes that are not in the ancestor).

Based on this tree, we simulate the evolution of 11 genomes, starting from an ancestor with 4705 genes distributed among 8 chromosomes (the number of chromosomes of the pre-duplicated ancestor as predicted by Gordon *et al.*), and performing a certain number of rearrangements (inversion, translocation, fusion, fission) and gene losses on each branch of the tree. The number of gene losses is simply the one observed in the phylogenetic tree of Figure 2 (genes to be removed are chosen randomly). The number of rearrangements is selected randomly from an interval $[\mu/2, \mu]$, where μ is a parameter chosen prior to the generation, and the size of each rearrangement is random. As for the rate of rearrangement operations it is chosen to be similar to that reported for *S. cerevisiae* in [7]. More precisely we choose the rates (Inv : Trans : Fus+Fiss) = (5 : 4 : 1).

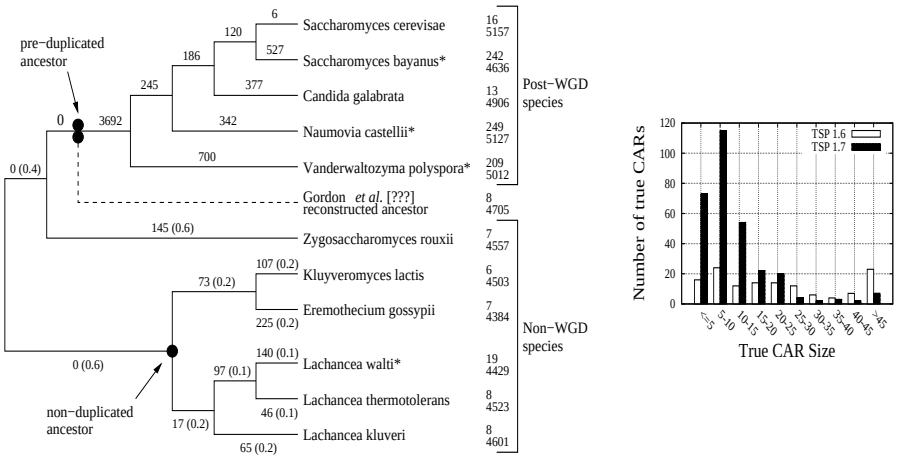

Fig. 2. Left:Yeast phylogenetic tree used for the simulations. The ancestors (pre-duplicated and non-duplicated) are represented by black dots. The branch lengths represents the number of gene losses. The ones in parentheses are the distances used for the Ma *et al.* method. * indicates partially sequenced organisms. On leaves, the top number is the number of chromosomes, contigs or scaffolds. The bottom number is the number of genes. Right: Size distribution of true CARs in our inferred pre-duplicated yeast ancestor, considering the Gordon *et al.* [7] ancestor as the "true" ancestor.

Notice that four among the species represented in the phylogenetic tree of Figure 2 Left (those indicated by $*$), are partially sequenced species for which only scaffolds are available. To account for this specificity of the data, we perform random fissions on four of our simulated genomes. Moreover, as scaffolds just represent parts of chromosomes, adjacencies at the extremities are not relevant to our study and are not taken into account.

Simulations without WGD.

In the absence of WGD events, the method that is most comparable to ours is the ancestral genome reconstruction methods of Ma *et al.* [11, 9]. The method in [11] has been successfully applied to mammalian genomes, as the predicted Boreoeutherian ancestor appears to approach the results obtained by cytogeneticists. The software availability of [11] and the fact that it is directly applicable to our data sets (DUPCAR [9] requires, in addition to the species tree, a set of reliable gene trees with branch length) make it a natural software to compare with ours. We refer to this software as the Ma method.

We simulate data sets based on the subtree of the yeast phylogeny containing only the five non-duplicated yeast species. Moreover, as the Ma *et al.* algorithm does not support losses, we only consider the set of genes present in all five species. We performed our simulations with 10 different values of μ (the maximum number of rearrangements per branch), varying from 100 to 1000. For each of those 10 μ values, 50 different data sets are obtained (50 different simulated histories), an ancestor is inferred for each dataset and compared to the "true" known ancestor.

Results (error rates and number of preserved CARs) are averaged over all data sets showing a comparable ancestral genome divergence, where the genome divergence of a data set is the fraction of adjacencies in the ancestor that are preserved in at least one leaf of the tree. In Figure 3, the *error rate* is the rate of inferred adjacencies that are not present in the true ancestor, and the proportion of true CARs is the proportion of inferred CARs that are present in the true ancestor. Recall that a CAR is chromosomal segment inferred by the algorithm, and it is "true" if it is a subsequence of the ancestral genome. We arbitrarily imposed a minimum size of 5 adjacencies to consider a CAR a true CAR.

Comparing the error rates of the Ma method, and our methods using the greedy or TSP approach (Figure 3 top left), we first notice that the three methods have a good performance (less than 10% errors for genome divergence of 40%), but with the TSP method outperforming the two others. However, a drawback of the TSP approach is the fact that it outputs very few CARs, typically one or two for a genome divergence above 40%. In all cases, our approaches (greedy and TSP) infer fewer CARs than the Ma method (Figure 3 bottom left), and fewer that the actual number of chromosomes of the true ancestor.

In order to improve the proportion of true CARs inferred, we "force" the production of more CARs by defining "TSP τ" which is the TSP method augmented with the procedure of cutting, from the inferred ancestor, all adjacencies with weight less than a certain threshold τ. Figure 3 top right gives the error rate associated to the set of adjacencies of a given weight (rate of such adjacencies in

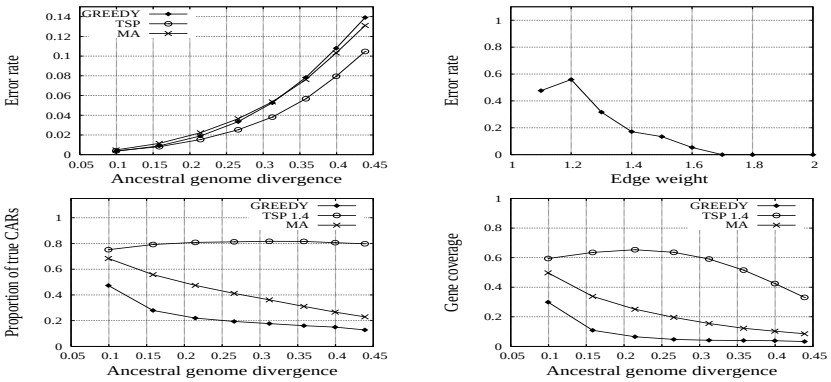

Fig. 3. Simulations for a tree without WGD. (1) Top left: Error rate of the inferred ancestral genomes. (2) Top right: Error rate of adjacencies depending on their weight. (3) Bottom left: Proportion of true CARs inferred. (4) Bottom right: Proportion of genes in the ancestral genome that are covered by the inferred CARs that are true.. See the text for explanation about "Ancestral genome divergence", "Error rate", "TSP", "TSP τ" and "MA'.

our results that are false predictions). Based on this figure, we choose $\tau = 1.4$. As observed in Figure 3 bottom left, the proportion of true CARs inferred is greatly improved compared to the TSP approach without edge cut, but more interestingly compared to the greedy approach and the Ma method. However whereas the "true adjacencies" of the TSP approach where covering more that 90% of the genome, the number of genes covered by the "true CARs" is less than 40% for a genome divergence of more than 40%. However, this gene coverage remains higher that that of the Ma method (Figure 3 bottom right).

Simulations with WGD.
We now simulate datasets based on the whole yeast tree (Figure 2). Sets of genomes have been generated, with μ (maximum number of rearrangements per branch) varying from 0 to 500, with increments of 50. For each of those 11 μ values, 50 data sets have been generated. Notice first that the addition, in our simulations, of gene loss, increases the ancestral genome divergence (range x-axis in Figure 3 compared to Figure 4). In this case, the TSP approach clearly infers fewer false adjacencies than the greedy approach, regardless of the ancestral genome divergence. Its error rate remains under 10% for ancestors with divergence under 50%.

Based on Figure 4 top right, we choose two thresholds for edge-cut $\tau = 1.6$ and $\tau = 1.7$. We observe from Figure 4 (bottom left and bottom right) that the proportion of true CARs inferred by TSP1.7 is over 80% for a gene divergence under 0.3 with a gene coverage over 60%, and over 40% for a genome divergence under 0.5, but with a significantly lower gene coverage (only over 20%).

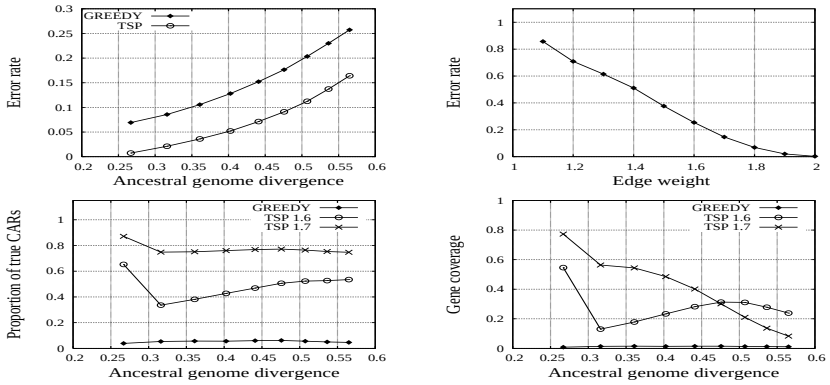

Fig. 4. Simulations for a tree with WGD. See Figure 3.

5.2 Comparison with the Gordon *et al.* ancestor

We applied our method to the yeast species tree (Figure 2) with the gene datasets of the *Yeast Gene Order Browser* [7] described above, to infer the pre-duplicated ancestral genome of *Sccharomyces cerevisiae*. Compared with the ancestral genome manually inferred by Gordon *et al.* [7], about 98% of the adjacencies inferred by our method are also present in the Gordon *et al.* ancestor. However, our TSP approach without edge-cut leads to only 4 CARs compared to the 8 likely ancestral chromosomes.

We tried two cutoff values for edge weight to decrease the number of incorrect adjacencies. With a cutoff value of 1.6, we obtain smaller CARs (average length 26), 84% of them (covering 79% of the genes) being "true" CARs of the Gordon *et al.* ancestor. With a cutoff value of 1.7, CARs are even smaller (12 in average) with 95% true CARs, covering 75% of the genes. Figure 2 Right illustrates the size distribution of true CARs with the different TSP strategies.

6 Conclusion

We have developed a general method for inferring the ancestral pre-duplicated genomes of a set of species known to have evolved through one or many rounds of whole genome duplication, interspersed with genome rearrangements and gene losses. The input to our method is a phylogenetic tree representing the evolution of the species, with positions of the WGD events, and genomes represented as ordered sets of oriented blocks (genes or any other kind of markers), each block appearing in one or many copies in each genome. As WGD is assumed to be the only mechanism giving rise to gene duplicates, gene content and multiplicity at each internal node of the tree can be inferred without resorting to reconciliation. We developed a local approach consisting in inferring ancestral adjacencies and then chaining them in an optimal way. The originality of this method compared

to all other local approaches is the computation of a rigorous score for each ancestral set of adjacencies, reflecting the maximum number of conservation of this set of adjacencies among the whole tree. This is done by a rigorous dynamic programming algorithm running, which is sufficiently fast to run on large data sets (e.g. complete yeast genomes). Chaining adjacencies is then performed using a traveling salesman strategy on a graph representation of all possible adjacencies.

Applying our method, first on simulated datasets and then on the yeast genomes, reveals a high accuracy for adjacency prediction. However, the number of inferred CARs strongly depends on the cutoff value used to separate good adjacencies from noise. Although the TSP strategy seems appropriate, other chaining strategies may be considered and may improve the quality of our results.

In this paper, we focused on inferring the ancestral genomes preceding a first WGD event on the tree. In other words, the inferred genome has a single copy of each chromosome. This restriction is only required for the chaining part of the method, as the first step that consists in computing the score of each set of adjacencies at each internal node of the tree is general. However, if the ancestor of interest has more than one copy of each gene, then it is not clear how to assemble a set of relevant adjacencies to form CARs as the TSP representation breaks down. This is one of the future directions to our project that we aim to pursue.

References

1. M.A. Alekseyev and P.A. Pevzner. Colored de bruijn graphs and the genome halving problem. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 4(1):98–107, 2007.
2. D. Applegate, R. Bixby, V. Chvatal, and W. Cook. Concorde tsp solver. http://www.tsp.gatech.edu/concorde/, 2006.
3. D. Applegate, W. Cook, and A. Rohe. Chained lin-kernighan for large traveling salesman problems. *INFORMS Journal on Computing*, 15:82 - 92, 2003.
4. C. Chauve and E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *Plos Computational Biology*, 4(11):e1000234, 2008.
5. N. El-Mabrouk and D. Sankoff. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 32(1):754-792, 2003.
6. H. Gavranović and E. Tannier. Guided genome halving: probably optimal solutions provide good insights into the preduplication ancestral genome of Saccharomyces cerevisiae. volume 15 of *Pacific Symposium on Biocomputing*, pages 21-30, 2010.
7. J.L. Gordon, K.P. Byrne, and K.H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PloS Genetics*, 5(5), 2009.
8. S.M. Hedtke, T.M. Townsend, and D.M. Hillis. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology*, 55:522- 529, 2006.
9. B.J. Raney B.B. Suh L. Zhang W. Miller J. Ma, A. Ratan and D. Haussler. Dupcar: Reconstructing contiguous ancestral regions with duplications. *Journal of Computational Biology*, 15(8):1- 21, 2008.

10. S. Lin and B.W. Kernighan. An effective heuristic algorithm for the traveling salesman problem. *Operations Research*, 21:498- 516, 1973.
11. J. Ma, L. Zhang, B.B. Suh, B.J. Raney, R.C. Burhans, W.J. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16:1557- 1565, 2007.
12. E. Tannier. Yeast ancestral genome reconstructions: the possibilities of computational methods. In *Lecture Notes in Computer Science*, volume 5817 of *RECOMB-CG*, pages 1- 12, 2009.
13. D.E. Soltis *et al.* Polyploidy and angiosperm diversification. *American Journal of Botany*, 96:336 - 348, 2009.
14. J. Salse *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution[w]. *The Plant Cell*, 20:11- 24, 2008.
15. C. Zheng, Q. Zhu, Z. Adam, and D. Sankoff. Guided genome halving: hardness, heuristics and the history of the hemiascomycetes. ISMB, pages 96 - 104, 2008.
16. C. Zheng, Q. Zhu, and D. Sankoff. Descendants of whole genome duplication within gene order phylogeny. *Journal of Computational Biology*, 15(8):947-964, 2008.

Annex 1

Algorithm : $L^{below}_{(u,v)}(a, X)$

If $X \not\subset \pm\Sigma_u$ **Or** $|X| \neq mult(a, \Sigma_u)$,
    $L^{below}_{(u,v)}(a, X) = -\infty$ (a solution is impossible);

Otherwise
    If v is a leaf,
        If u is a speciation node,
            $L^{below}_{(u,v)}(a, X) = |X \ \cap \ LA(G(v), a)|$;
        If u is duplication node,
            $L^{below}_{(u,v)}(a, X) = |(X \cup X) \ \cap \ LA(G(v), a)|$;
    Otherwise v is an internal node
        If v is a speciation node with children x and y,
            If u is a speciation node,
                $L^{below}_{(u,v)}(a, X) = max_{X'}\{L^{below}_{(v,x)}(a, X') + L^{below}_{(v,y)}(a, X') + |X \cap X'|\}$;
            If u is a duplication node,
                $L^{below}_{(u,v)}(a, X) = max_{X'}\{L^{below}_{(v,x)}(a, X') + L^{below}_{(v,y)}(a, X')$
                $+|(X \cup X) \cap X'|\}$;
        Otherwise v is a duplication node with one child w,
            $L^{below}_{(u,v)}(a, X) = max_{X'}\{L^{below}_{(v,x)}(a, X') + |X \cap X'|\}$;
        End If
    End If
End If

Fig. 5. Computing $L^{below}_{(u,v)}(a, X)$

Annex 2

Algorithm : $L^{above}_{(p,u)}(a, X)$

If $X \not\subset \pm\Sigma_u$ **Or** $|X| \neq mult(z, \Sigma_u)$,
$L^{above}_{(p,u)}(z, X) = -\infty$ (a solution is impossible);

Otherwise
If u is the root r of T, then $p = D$ and
$L^{above}_{(p,u)}(z, X) = 0$ if X is an eligible set of left adjacencies
of a, and $-\infty$ otherwise;
Otherwise let p' be the parent of p,
If p is a speciation node and s is the sibling of u,
$L^{above}_{(p,u)}(a, X) = max_{X'}\{L^{above}_{(p',p)}(a, X') + L^{below}_{(p,s)}(a, X') + |X \cap X'|\}$,
If p is a duplication node (its only child is u),
$L^{above}_{(p,u)}(a, X) = max_{X'}\{L^{above}_{(p',p)}(a, X') + |X \cap (X' \cup X')|\}$,
End If
End If
End If

Fig. 6. Computing $L^{above}_{(p,u)}(a, X)$