Journal of Logic and Computation Advance Access published December 9, 2011

Combinatorics of chromosomal rearrangements based on synteny blocks and synteny packs

GUÉNOLA DRILLON and ALESSANDRA CARBONE, *Université Pierre et Marie Curie, UMR7238, Génomique Analytique, 15 rue de l'Ecole de Médecine, F-75006 Paris, France; CNRS, UMR7238, Laboratoire de Génomique des Microorganismes, F-75006 Paris, France.*
E-mail: guenoladrillon@gmail.com; alessandra.carbone@lip6.fr

GILLES FISCHER, *CNRS, UMR7238, Laboratoire de Génomique des Microorganismes, F-75006 Paris, France.*
E-mail: gilles.fischer@upmc.fr

Abstract

The reconstruction of the history of rearrangements and the reconstruction of ancestral genomes are some of the challenges of bioinformatics today. Many algorithms already exist, treating one or the other question but none treating both. These reconstructions are interdependent and we argue on the interest of treating both problems in parallel to lead to a richer and more complete output. We also argue on the importance of redefining several steps of these algorithms to improve both reconstructions: the identification of synteny blocks has to be as precise as possible, and the treatment of multiple genomes has to be based on pairwise comparisons to ensure the most detailed reconstructions. In this article, we highlight novel solutions to these points and focus on the need of explicitly treating overlapping, included, duplicated and unsigned synteny blocks. To do so, we introduce the new notion of *synteny pack*, which is a representation of local hypothetical intermediate ancestral genomes. We discuss a number of examples on yeast genomes to illustrate the importance of such a definition.

Keywords: Synteny block, combinatorics of genome rearrangements, ancestral genome, pairwise comparison.

1 Introduction

The availability of full genome sequences has revolutionized genomics and especially our possibilities to understand evolution. The reconstruction of ancestral genomic sequences is one of the challenges of bioinformatics today. Different species share large or small sets of genes depending on phylogenetic proximity. These genes are inherited from their common ancestor and are not necessarily ordered in a similar manner along the chromosomes. This is due to chromosomal rearrangements, such as inversions within a chromosome or translocations between chromosomes, transforming ancestors into new species observable today (Section 2).

The two main questions that are biologically relevant in this field concern (i) the reconstruction of the history of the rearrangements, i.e. the succession of rearrangements that occurred along the branches of a phylogenetic tree, and (ii) the reconstruction of the ancestral genomes from available complete genome sequences, i.e. the genomes located at the internal nodes of the phylogenetic tree (Figure 1). In fact, a better understanding of the rearrangements is behind these questions and is needed for a better estimation of rearrangement rates [6], a precise identification of biological

doi:10.1093/logcom/exr047

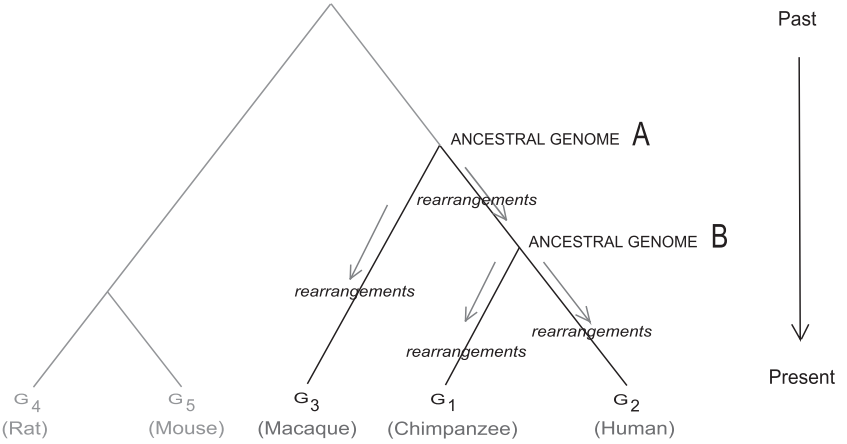

FIGURE 1. A schematic representation of a phylogenetic tree. The leaves of the tree are labelled by the genomes of five extant species (G_1, G_2, G_3, G_4 and G_5). The internal nodes represent *ancestral genomes* and *rearrangements* are associated to each branch. Several outgroup genomes can be used for the reconstruction of the ancestral genome B: G_3, G_4 and G_5. This tree does not take into account a precise estimation of the evolutionary time between species, usually encoded in the length of the branches.

operations allowing for genome evolution (other than translocation and inversion) [25], a more accurate identification of the regions where chromosomal breaks took place (are they randomly distributed or biased towards rearrangement hotspots?) [21] and a more detailed analysis of the genetic elements occurring in these regions [14].

Many algorithms have been proposed to answer these questions. Some of them aim at reconstructing the history of rearrangements [11, 13, 24, 26]; others try to reconstruct the ancestral genomes [2, 7, 8, 12, 17, 18]. They are all based on the principle of parsimony, which expects the transformation requiring the smaller number of rearrangements, between blocks of physically close genes (synteny blocks), to be the one chosen by nature. Among them, two algorithms can be distinguished: *MGRA*, from Alekseyev and Pevzner [2], and *inferCARs*, from Ma and collaborators [17]. They both reconstruct ancestral genomes taking into account several genomes with multiple chromosomes at once (Section 3).

In this article, we argue on the importance of redefining crucial steps of these algorithms such as the construction of the synteny blocks and the multiple genome comparison. The identification of synteny blocks, on which reconstructions directly depend, has to be precise, doable over all genomes to be compared and based on gene sequence only (Section 4). Multiple genomes can share very different levels of conserved synteny, and we shall do pairwise comparison instead of multiple genomes comparison to preserve most of the information they contain (Section 6). Moreover, ancestral genome structures and rearrangements are interdependent and we shall treat them in parallel during reconstruction. As a consequence, genome and history reconstructions will be reliable and coherent. Such join outcome is important to gain insights on the different mechanisms of rearrangement (Section 5).

In addition, our new way to identify synteny blocks demands to deal with overlapping, included, duplicated and unsigned synteny blocks. To do so, we introduce the new notion of *synteny pack*,

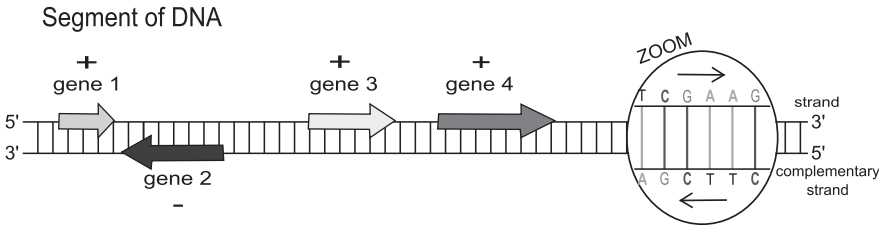

FIGURE 2. Zoom in a double-stranded DNA. This segment contains four genes, each gene has a position on the DNA sequence and an orientation (defined by the reading frame $5' \to 3'$ associated to the strand on which the gene is localized).

which is a representation of local hypothetical intermediate ancestral genomes. This notion asks for an explicit treatment of the combinatorial relationships between synteny blocks that is much more complex than the ones previously studied. We discuss a number of examples on yeast genomes to illustrate the importance of such a definition (Section 7).

2 Some biological notions in genome rearrangements

In this section, we shall list and fix the biological terminology to help the reader in following the text. We assume the reader to be familiar with the biological notions of DNA, RNA and proteins. For further reading, refer to [1, 9, 15, 16, 22]. To understand the challenge of genome reconstruction, it is important to understand genome structure. Genomes can differ from one species to another, and their characteristics might have an important impact on the complexity of the problem. Genomes may contain one or several *chromosomes*, which may be circular or linear. In eukaryotic cells, such as human or yeast cells, the genome is a set of several linear chromosomes (Figure 2). The two extremities of a linear chromosome are called *telomeres*. Chromosomes are made of a double-stranded DNA molecule, in which each *strand* is a sequence of *nucleotides*. Nucleotides can be of four types: A, T, C and G, where A, T and C, G are complements of each other. The genetic information in a genome is held within genes:

DEFINITION 1
A gene is a segment of DNA, a sequence of consecutive nucleotides, that is transcribed into a single-stranded RNA molecule.

Some of these genes encode for proteins (a sequence of *amino acids*) but others generate catalytic, structural or regulatory RNA molecules. Each gene has a position along the chromosome and an orientation (Figure 2).

DEFINITION 2
The orientation of a gene is determined by the strand on which it is encoded. Since double-stranded DNA molecule could be read in two different ways (but always from 5' to 3'), the positive orientation (or positive strand) is arbitrarily fixed while sequencing, and the negative orientation is the remaining one.

Different species can have many genes in common, inherited from their common ancestor. We are able to identify them by the homology of their sequences.

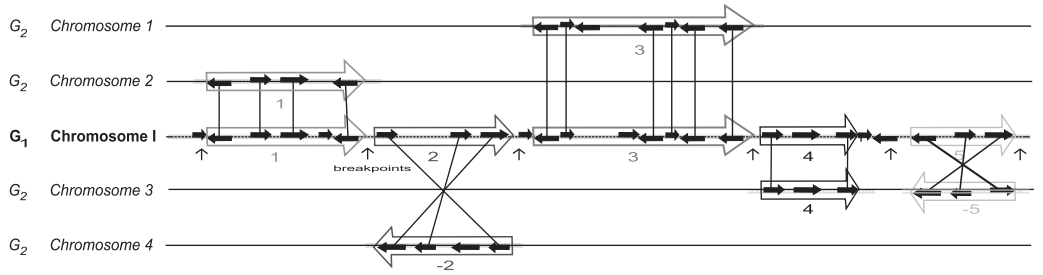

FIGURE 3. Segment of the chromosome I of genome G_1. Genes (small black arrows) are grouped in synteny blocks (long arrows) depending on their adjacency in G_1 and G_2. Synteny blocks are projected on their homologs (homologous genes are linked by a line) on the four chromosomes of G_2. For each block in G_2, as in G_1, the local order of genes is preserved. A block n in G_1 (where $n = 1, 2, 3, 4, 5$) is found in G_2 either on the same orientation (n) or in the opposite orientation ($-n$). Breakpoints are indicated between blocks, along chromosome I, by vertical arrows.

DEFINITION 3
Two genes showing relatively close[1] nucleotide sequences or coding for proteins with relatively close amino acid sequences are called homologs. (See Section 4 for more details and a formal definition of homology.)

Among homologs, we are specially interested in *orthologs*.

DEFINITION 4
Two genes from two different species that come from a same ancestral gene are called orthologs.

Because of frequent duplication events, some genes can be homologous without being orthologous. Orthologous genes are important for the reconstruction of the ancestral genome but yet, sequence alignment is only able to determine homologs. When we compare the chromosomal location of homologs in two different species, we easily observe *synteny blocks*.

DEFINITION 5
A conserved synteny block corresponds to the collocation of a series of homologs on chromosomes from different species.

A formal definition of synteny block is given in Section 4. For the moment, we can simply keep in mind that genes in these blocks are *syntenic homologs*. They probably share their proximity with their last common ancestor and therefore we consider them as orthologs. While comparing two genomes, we observe that they share synteny blocks distributed throughout their chromosomes in a different order and/or orientation (Figure 3) due to the accumulation of chromosomal rearrangements (Definition 8) during evolution. For us, a chromosome is an ordered list of signed synteny blocks.

DEFINITION 6
Two homologous synteny blocks have the same orientation if their genes are ordered in the same way along the chromosomes in both species. By convention, one of the two genomes is the reference and all its synteny blocks have positive orientation.

[1]To decode whether two genes are *relatively close* or not, we align the gene sequences (either nucleotide or amino acid sequences) and compute a score of sequence similarity depending on the number of insertions, deletions and mismatches, and ask for this score to be bigger than some threshold.

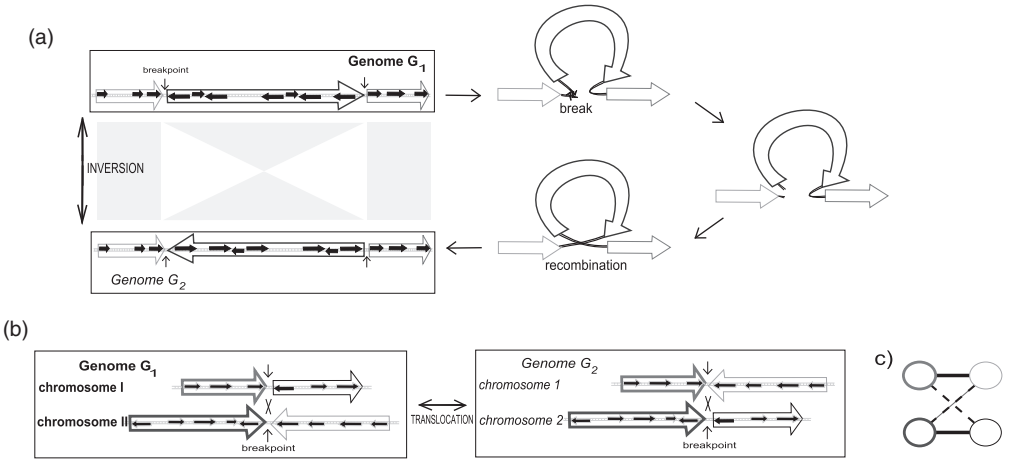

FIGURE 4. (**a**) G_1 and G_2 are unichromosomal and partitioned in three blocks. G_1 differs from G_2 by one inversion (left). An inversion can be explained biologically by a recombination between inverted repeated sequences after at least one DNA double strand break (right). (**b**) G_1 and G_2 are partitioned in two chromosomes and four synteny blocks. G_1 differs from G_2 by one translocation. A translocation is due to a recombination between two different chromosomes and leads to two new chromosomes. (**c**) Breakpoint graph of the genomes G_1 and G_2 of (b). Vertices represent synteny blocks and edges represent physical proximity between blocks (bold lines for G_1, dashed lines for G_2).

Differences in orientation can either be due to a chromosomal rearrangement as seen further, or to a difference in the 'choice' of the positive strand (at the moment of sequencing) for each chromosome in the two genomes.

DEFINITION 7
The region of the chromosome lying between two consecutive synteny blocks, is called a synteny breakpoint. Each breakpoint is characterized by the two ordered and oriented synteny blocks surrounding it.

By convention, we accept as breakpoints, those regions lying between a synteny block and a telomere. (Notice that all telomeres are identified and denoted with the same letter '0'.) Given two consecutive synteny blocks B and D and their associated breakpoint $(B;D)$, by convention we have: $(B;D) \neq (D;B)$ and $(B;D) = (-D;-B)$ with $-B$ being the equivalent of B but oriented in the opposite way. Given a telomeric block A, we also have: $(0;A) = (-A;0)$. Notice that the same notation will be used to represent *block adjacency*: i.e. for instance, the fact that B is next to D along the chromosome.

DEFINITION 8
Chromosomal rearrangements result from chromosome breakage into parts, followed by the chromosome reconstitution based on an abnormal combination of the parts that implies a structural variation of the DNA molecule.

There exist intrachromosomal rearrangements, such as *deletions* (losses), *insertions* or *duplications* (gains) of a gene or a group of genes. For the specific study of rearrangements and genomes reconstruction, one often focuses on *inversions* (also called *reversals*) of chromosomal segments containing from one to few hundreds of genes (Figure 4a). There exist also interchromosomal

rearrangements, such as the end to end *fusion* of two chromosomes into one or the *fission* of a chromosome into two. But the most common interchromosomal rearrangement is the *translocation*, that is the breaking off of the ends of two chromosomes and their reciprocal exchange (Figure 4b). Translocations involving only one fragment are called *non-reciprocal*. (for more details see [1] p. 453–466).

3 Previous approaches

3.1 The principle of parsimony

Most existing models are based on the principle of parsimony, which expects the transformation that requires the minimal number of rearrangements to be the one chosen by nature. Rearrangements are rare events, as illustrated by drosophila evolution, where only about 10 rearrangements are estimated by million years [6]. It is the low frequency of the events that justifies parsimony.

3.2 First attempts to model rearrangements

Unichromosomal genomes: in 1988, the two mitochondrial unichromosomal genomes of the turnip and the cabbage have been observed to be rearranged [20]. A formal framework modelling chromosomal rearrangements, based on the unique operation of inversion, has been introduced to solve the underlying NP-hard problem [11]. This first approximation algorithm of complexity $O(n^4)$ was improved the year after with an algorithm of complexity $O(n^2)$ [13], with n being the number of entities to be rearranged (genes or synteny blocks). Both algorithms work only for pairs of unichromosomal genomes.

Pairwise comparisons: models developed after 1988 are all based on this pioneer approach. Genomes are compared in pairs and rearrangements that explain how to go from one genome to another are identified. But this approach is far from any biological reality as illustrated in Figure 1. Strictly speaking, there is no pathway from G_1 to G_2 that would have existed during evolution, but rather two distinguished paths from the ancestor B towards G_1 and G_2. To reconstruct these two paths, we have to consider outgroup genomes, that is a group of genomes (possibly one) that diverge from the G_1 and G_2 lineage before their last common ancestor. Therefore, for a reconstruction of the ancestor and a localization of the rearrangements along the branches, a multiple genomes comparison is needed. The more genomes are considered, the more precise will be the reconstruction.

The reconstruction of ancestral genomes: the study of chromosomal rearrangements involving several genomes demands for the exploration of a large combinatorial space of chromosomal arrangements and requires rather sophisticated algorithmic approaches. Given two genomes, there is a huge space of feasible configurations representing all potential histories of rearrangements leading to different ancestral genome reconstructions. How to find the good one among them? All proposed algorithms are based on parsimony combined with some additional principle and lead to a specific subset of solutions. The question is to manage to reduce as much as possible this subset around the true solution. We might not be able to identify precisely this latter and this impossibility is due to several reasons: (i) the sequenced genomes can contain mistakes (on assembling or on gene detection) and (ii) available genomes can be so distantly related that all their rearrangements cannot be traced back for a reliable reconstruction. In our work, we prefer to obtain a partially but accurately reconstructed ancestral genome rather than a complete but inexact reconstruction of the ancestor.

3.3 Two different models for multichromosomal genome reconstruction and their limitations

There are two main models dealing with multichromosomal linear genomes for the reconstruction of ancestral genomes that we will discuss: the rearrangement-based model and the cytogenetic-based model.

The rearrangement-based model: this kind of models is the more common and has never stopped to be improved since 1996 [4]. The biological principle guiding this algorithmic approach is the fact that inversion and translocation are two rearrangements that both involve two breakpoints (Figure 4a and b). This means that the existence of a breakpoint implies the existence of another breakpoint elsewhere. For instance, let us take two different breakpoints $(A;B)$ in G_1 and $(A;C)$ in G_2 resulting from the comparison of G_1 with G_2 (in short, the G_1G_2 comparison). They imply the existence of at least another breakpoint: $(X;B)$ in G_2 if $B \neq 0$ or $(Y;C)$ in G_1 if $C \neq 0$, where X and Y are either blocks or telomeres. So breakpoints in a genome are *linked* and therefore, they *cannot* be treated one by one but at least two by two and possibly more in case of re-use of breakpoints (detailed later).

These models are based on *breakpoint graphs*, where nodes are the synteny blocks and edges are defined between neighbouring blocks within a chromosome: for each block B with two, left and right, neighbouring blocks C and D along the genome, there is an edge from B to C and a edge from B to D in the breakpoint graph. (Figure 4c shows a small breakpoint graph of four nodes corresponding to a translocation.) Each genome under comparison has its own edges in the graph, corresponding to its own neighbouring blocks, and therefore each node has exactly n edges incident to it, respectively, associated to the n genomes.

For a long time, breakpoint graphs were only used to compare two genomes: where each cycle, formed by alternating edges from the two genomes, corresponds to one or several rearrangement that occurred between the two genomes. For instance, any cycle of length 4 in the breakpoint graph implies the existence of a rearrangement corresponding to an inversion or a translocation. MGRA published in 2009 [2], is the first model that deals with several multichromosomal genomes (involving *multiple breakpoint graphs* and a new type of cycles, the degree of the nodes being different than 2) and looks for rearrangements between genomes as well as for the reconstruction of a phylogenetic tree and the associated ancestral genomes. Even though the combinatorial structure of multiple breakpoint graphs is much more complicated than one of the breakpoint graphs describing two genomes, the new notion of cycle still allows the identification of rearrangements.

This last model highlights complex cyclic relations between breakpoints across species. A lack of precision in the definition of synteny blocks and the fact that some rearrangements, like insertion, duplication or other, are not taken into account can lead to cyclic structures which have no immediate biological interpretation. Therefore, they cannot be used to reconstruct reliable ancestral adjacencies, even though they could be useful to better understand rearrangement mechanisms. Yet, this model reconstructs only ancestral genomes and not the rearrangement history. Moreover, this model does not consider a phylogenetic tree as input but reconstructs it instead. This may be an advantage when the phylogenetic tree is unknown or uncertain. But, we expect that the reconstruction of an ancestral genome (as B in Figure 1) will depend more on the information contained in certain genomes (as G_1 and G_2 because of their phylogenetic proximity) and less on others (as G_4 and G_5). We expect also that the reconstruction may be more precise for *recent* ancestors (as B) than for *older* ones (as A), which are separated by a greater evolutionary distance to the extant species.

The cytogenetic-based model: the cytogenetic approach draws its inspiration from the experimental technique of 'chromosomal painting', where one is able, by hybridization, to recognize similar chromosomal segments between two species. Different models have been defined [8, 10] but inferCARs from Ma and collaborators [17] can be seen as the gold standard. Its originality comes from the fact that it exploits potential local similarities between genomes. It predicts the ancestral order and orientation of the blocks from their adjacencies observed in modern species, using a phylogenetic tree. The guiding principle is that if two blocks are contiguous in one of the extant species under comparison (for instance, G_1 in Figure 1) and in one outgroup (G_3), then they were probably contiguous in the ancestor of the extant species (the genome B).

However, this model does not consider the information coming from linked breakpoints and treats breakpoints one by one. Each breakpoint in a species is compared with the outgroup. As soon as the pair of consecutive synteny blocks characterizing a breakpoint is found conserved in at least one outgroup genome, the algorithm imposes the ancestor to contain these two synteny blocks in the same order. Since local similarities between genomes are unlikely to happen by chance, the reconstruction of the ancestor based on the similarity principle becomes reliable. A disadvantage of inferCARs is that if a succession of rearrangements occurs in the same region (by breakpoint re-use), it would be impossible to trace back the history of rearrangement and the ancestral adjacencies.

Advantages and disadvantages of both methods: these two algorithms reconstruct ancestral genomes but exploiting radically different biological information. One bases its reconstruction on breakpoint links and the other bases its reconstruction on individual breakpoints and on the local differences that they imply. Both algorithms do not provide any information on which rearrangements took place or on re-use breakpoints, even if MGRA use information on cyclic relations during its ancestral genome reconstructions. Moreover, they compare several genomes by identifying blocks which are shared by all species at the cost of loosing blocks only shared by pairs or by a subgroup of closely related species. This implies also the inability to incorporate distant genomes in the analysis.

We are interested in two main things: (i) to keep pairwise comparison as the basic tool of our analysis instead of multiple genome comparison and (ii) to consider any information on rearrangement that may be useful to understand the evolutionary process underlying the rearrangement. Our algorithm is presented in four sections. In Section 4, the *construction* of the synteny blocks is detailed. In Section 5, breakpoints resulting from synteny blocks are *linked* by using the same approach as rearrangement-based algorithms. As in Alekseyev and Pevzner's algorithm, we aim to reconstruct reliable ancestral genomes through the identification of a list of rearrangements explaining breakpoints. In Section 6, linked breakpoints are *validated* and the corresponding rearrangements are located on a given branch of the phylogenetic tree by using the guiding principle of cytogenetic-based algorithms. As inferCARs algorithm, we want to use a phylogenetic tree and local comparisons with outgroup genomes to validate these rearrangements. In Section 7, we introduce the new notion of *synteny pack*. The construction of synteny blocks closer to biological data, in Section 4, induces the creation of synteny blocks that can overlap, be included, be duplicated or be unsigned (with an undefined orientation). These cases were never explicitly considered before but they are important, since even small approximations can impair a reliable reconstruction, as we will see.

4 Construction of synteny blocks

A pioneering study [19] introduced the notion of 'conserved segments', that is sequences of consecutive nucleotides that are relatively close to each other and preserve gene order (with no disruption induced by rearrangements). We study 'synteny blocks' instead of conserved segments,

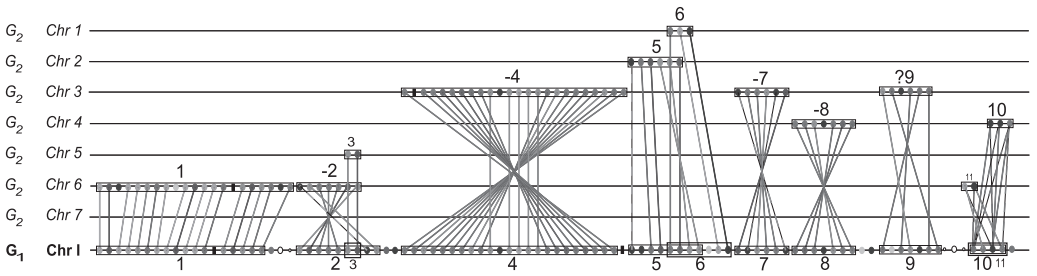

FIGURE 5. Fragments of a chromosome of G_1 mapping on the chromosomes of G_2. Each dot represents a gene. Homologs in both species are linked by lines. The picture shows some basic characteristics of the blocks described in the text: inclusion (Block 3 in Block 2 in G_1), microrearrangement (Block 4), overlapping (Blocks 5 and 6 in G_1), unsigned block (Block 9 in G_2), duplication (Blocks 10 and 11 in G_2).

a concept introduced in [23] together with the GRIMM-Synteny algorithm. A synteny block is a segment that can be converted into conserved segments by microrearrangements as inversions, insertions, deletions and duplications of very small genome segments (whose length is defined by a parameter). We base our synteny blocks construction on genes homology, instead of whole genome alignment (as done for vertebrates with MGRA and inferCARs). This allows us to consider genomes which are so distant that alignment of their nucleotide sequences becomes impossible, as in yeast.

The construction of the synteny blocks between two genomes G_1 and G_2 is done in four steps: (i) identification of *highly* homologous genes, keeping only those sharing synteny; (ii) identification of genes sharing *a weak* condition of homology and being syntenic to the ones identified in (i); (iii) construction of synteny blocks; and (iv) definition of block's signs. Figure 5 illustrates the main characteristics of a genome structure as it appears after the identification of synteny blocks.

4.1 Search of syntenic strong homology

Homology is a continuous trait and two proteins can be similar at various degrees. Therefore, the imposed threshold used to define homology has to be smooth enough not to miss divergent orthologs, having been subjected to an important number of mutations (nucleotides changes), but stringent enough to avoid an excess of false positives. To do so, we use Blast [3], a program that for a gene g of G_1 and a genome G_2 returns all genes g' in G_2, called *matches*, for which the *similarity score* between g and g' is above a certain threshold. This similarity score is computed from the amino acid sequences alignment (not nucleotides sequences) of the two corresponding proteins. It is the sum of the scores on all positions of the sequence alignment, depending on the residues similarity (two amino acids are called *similar* if they share similar physical and chemical characteristics). The gene g' of G_2, for which the score is the highest, is called the *best match*.

DEFINITION 9
Given two proteins p_1 and p_2 encoded by two genes, g_1 and g_2, occurring respectively in G_1 and G_2, we say that g_1 and g_2 are a bidirectional best hit, in short BDBH or BDBH homologs, if the best match of p_1 in G_2 is p_2 and, reciprocally, the best match of p_2 in G_1 is p_1.

As an example, about 80% of the genes in two distantly related genomes such as human and fish, are BDBH.

Syntenic homologs sharing a BDBH relationship are subsequently considered as orthologs. They are useful to identify ancestral synteny still shared by extant species. A parameter Δ is used to formalize the syntenic relationship between pairs of BDBH homologs. Notice that at this step, we consider only genes in G_1 that are found to be BDBH of some genes in G_2 and reciprocally. We shall refer to them as *BDBH-genes*.

DEFINITION 10
Two genes g_i and g_j in G_1 having BDBH homologs in G_2 are called syntenic BDBH if one of the two conditions is satisfied:

- they are at most Δ BDBH-genes apart from each other and their homologs in G_2 are also at most Δ BDBH-genes apart;
- there exists a chain of genes in G_1, say $h_1 h_2 \ldots h_n$, not necessarily ordered along the chromosome, where $h_1 = g_i$ and $h_n = g_j$ and where $h_i h_{i+1}$ are at most Δ BDBH-genes apart from each other as their respective BDBH homologs in G_2.

Δ has to be a balance between: being too big and creating *wrong synteny* corresponding to no common ancestral adjacency and being too small with the risk of losing some of them. A value $\Delta = 5$ was used to define synteny blocks considered in Figure 5.

4.2 *Search of weaker syntenic homology*

To better identify breakpoint regions and to increase the number of genes that structure the ancestral blocks, we consider all genes (not only BDBH-genes) and we relax the notion of homology but only at the proximity of syntenic BDBH (in this case, Δ does not represent anymore a number of 'BDBH-genes' but simply a number of 'genes', therefore it represents *smaller* distances and this allows us to search for homologs satisfying less stringent conditions of homology).

DEFINITION 11
Two genes g_i in G_1 and g_j in G_2 are called syntenic 30% homologs if they satisfy the three following conditions:

- they do not both correspond to already defined syntenic BDBH;
- they have at least 30% of similarity (in number of similar residues) over at least 50% of their length; and
- they are both less than Δ genes away from a syntenic BDBH.

Similarity between genes and proximity with syntenic BDBH genes are indicators used to validate orthology.

4.3 *Definition of synteny blocks*

From syntenic homologs, we are able to define *synteny blocks*:

DEFINITION 12
A synteny block, defined between two species, is made of at least two syntenic BDBH. It may contain an unrestricted number of homologous genes: BDBH or 30%. Its homologs do not have to be ordered exactly in the same way in both species. Each homolog (BDBH or 30%) has to be close to at least one syntenic BDBH (with respect to its own Δ distance definition; see Definitions 10 and 11).

Figure 5 shows syntenic homologous genes obtained by the two homology criteria: BDBH are represented by solid lines and 30% homologs by dashed lines. For block 7, the third step of evaluation has allowed to find two more genes in synteny, probably orthologous. We could have chosen to construct also blocks with only syntenic 30% homologs, without BDBH, but in addition to the high computational time, this would have created a lot of *wrong* blocks in the subtelomeric regions, regions that are duplicated a lot and rearranged in a different way from the rest of the genome. The disadvantage is that if the genomes are more distant and do not share many BDBH relationships, only few blocks are found.

Block 4 in Figure 5, is a block that has been subjected to a microrearrangement (a small inversion of six genes in the middle). Depending on the value of Δ, this block could also have been represented as three distinct blocks. Ideally, the reconstruction of the history of rearrangements between two genomes should take into account not only as rearrangements but also microrearrangements and therefore Block 4 should be considered as three distinct blocks. Since the three blocks are close to each other in both species, and *a fortiori* in the ancestral genome, it is reasonable to treat them as a unique block and postpone the resolution of their microrearrangement. This choice has the effect of imposing an order in the process of identification of the rearrangements but not in the order of the rearrangements themselves. In particular, it allows us to keep track of the *position* of close groups of genes, like in Block 9 of Figure 5. In this short chromosomal segment that is preserved in G_1 and in G_2, an important number of microrearrangements occurred. It is important to consider Block 9 as a whole because rearrangements that have occurred between it and its neighbouring blocks (8 and 10 in G_1) might be easier to identify if microrearrangements are treated afterwards.

4.4 Sign of a synteny block

Each block is defined by a number between 1 and the total number N of blocks in the genome, and by a sign. For a reference genome, we associate to each block a different positive integer $1 \ldots N$, going from left to right by convention. For the other genome, we associate to the corresponding blocks the numbers defined in the reference genome. The sign of the blocks are positive when they present the same orientation as in the reference genome, otherwise negative. In Figure 5, Block 1 is positive and Block 7 has a negative sign in G_2. There are situations where the sign is not obvious due to many microrearrangements as in Block 6 or 9.

DEFINITION 13
The sign of a block B is positive if the first (last) gene of B, along the chromosome, in G_1 is homologous to the first (last) gene of B in G_2. Otherwise, if the first (last) gene of B in G_1 is homologous to the last (first) gene of B in G_2, the sign of B is negative. If neither of these two conditions is satisfied, then we say that the sign is undefined.

This is one of the differences between our definition of a synteny block compared to those used in other algorithms, where blocks are either positive or negative. The other differences are illustrated in Figure 5 where blocks might present unusual configurations: (i) blocks might be included one in another (as Block 3 is included in Block 2 in G_1); (ii) blocks can overlap with their neighbours (as Blocks 5 and 6 in G_1); (iii) blocks can be unsigned (as Block 9 in G_2); (iv) blocks can be duplicated (as Blocks 10 and 11 in G_2). We will see how to handle these blocks, after having presented the general approach, in Section 7.

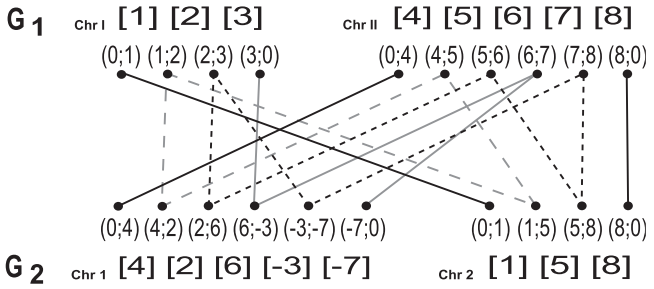

FIGURE 6. The *adjacency graph* of two genomes G_1 and G_2. G_1 and G_2 are both constituted by two chromosomes. The graph represents the links between different adjacencies. There are two cycles in black and grey dashed lines and four paths between two telomeres (represented as 0) in black and grey solid lines.

5 Identification of linked breakpoints

Linked breakpoints are the key combinatorial notion to reconstruct the history of rearrangements, as in the rearrangement-based models. We aim at linking all the breakpoints that are involved in a given rearrangement and for this, we use a partition graph, called *adjacency graph* [5]. We start by defining adjacency graphs and give a biological interpretation of different combinatorial structures appearing in adjacency graphs. *Linking breakpoints/adjacencies* is not a new concept. We shall provide an innovative interpretation of this concept though, that will be important for trusting and validating genome reconstructions.

5.1 Adjacency graphs

Genomes can be represented by the set of their adjacencies where the telomeres are also represented as adjacencies; $(0;N)$ denotes the left telomeres and $(M;0)$ the right ones (Figure 6). The comparison between two genomes is described by a graph, called *adjacency graph* [5] defined as follows.

DEFINITION 14
The adjacency graph of two genomes G_1 and G_2 is a partition graph whose nodes are the adjacencies of G_1 and G_2 and such that for each block B there is an edge between $(A;B)$ in G_1 and $(X;B)$ in G_2 and an edge between $(B;C)$ in G_1 and $(B;Y)$ in G_2. Each adjacency being defined by at most two blocks, adjacency graphs have nodes with degree at most 2.

A concrete example of adjacency graph is illustrated in Figure 6.

5.2 Interpretation of path structures in adjacency graphs

There are two different kinds of path structures in an adjacency graph: either a path links two telomeres or a path is cyclic and involves no telomere. Note that nodes in a adjacency graph can be involved in exactly one path or cycle. Different combinatorial properties of paths and cycles correspond to different rearrangements [5].

An adjacency graph might be seen as a 'projection' of all rearrangements that happened during the evolution of the species. The full reconstruction of all rearrangements from such graphs might not always be possible. Let us consider a pair of two linked breakpoints, for instance $(1;2)$ and

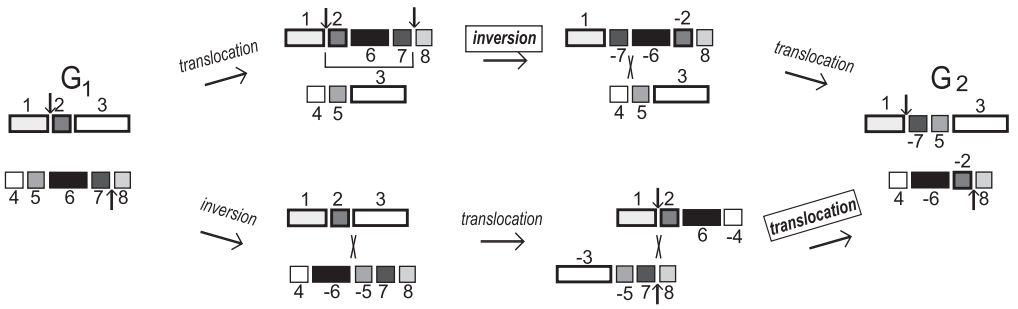

FIGURE 7. Example of scenarios of rearrangements that *transform* G_1 into G_2. G_1 and G_2 are both constituted by two chromosomes and height blocks. This example shows that it cannot be determined whether a pair of breakpoints (for instance $(1;-7)$ and $(-2;8)$ in G_2) is the result of an inversion (top) or of a translocation (bottom).

$(7;8)$ in G_1 and $(1;-7)$ and $(-2;8)$ in G_2 (Figure 7). Each breakpoint is localized on different chromosomes in both genomes, and this suggests that they could result from a translocation. But the fact that each breakpoint in G_2 comprises one negative block suggests that they could result from an inversion. In fact, both rearrangements could be responsible for the observed breakpoints (see the different scenarios in Figure 7). This means that paths and cycles are not sufficient to identify which rearrangement occurred. Even if we can independently find several disjoint chains of linked breakpoints involved in different rearrangements, they are all dependant from each other since they make a part of the same history.

There are four different types of paths and cycles that can be differentiated depending on their length.

Short cycles of length 4: they correspond to either an inversion or a translocation, which are the only rearrangements involving two breakpoints in a genome. Figure 6 illustrates a cycle of length 4 in grey dashed line.

Short paths of length 2 or 3 that include telomeres: a path of length 2 starting and finishing in two telomeres of G_1 (G_2) represents the *fusion* of two chromosomes in a single chromosome in G_2 (G_1), or equivalently, the *fission* of a chromosome in G_2 (G_1) into two chromosomes in G_1 (G_2).

A path of length 3 represents either a telomeric inversion or a non-reciprocal translocation. In Figure 6, the grey solid line path represents a telomeric inversion of the blocks [7][3] into [−3][−7].

Cycles of length >4: long cycles imply the use of the same breakpoints several times, leading to the notion of *breakpoint re-use*. For a cycle of length $2l$, one needs at least $l-1$ translocations or inversions to explain it [4]. Without re-use, each translocation and inversion generates two breakpoints in a genome, therefore we should identify $2*(l-1)$ breakpoints in G_1 (G_2). In reality, we only observe l breakpoints in G_1 (G_2) and the difference $l-2$ corresponds to the number of *re-used* breakpoints. The black dashed line cycle in Figure 6 corresponds to a cycle of length 6. Either two translocations or two inversions or one translocation and one inversion with 1 re-used breakpoint $(1=3-2)$ could have generated this cycle.

Paths of length >3 that include telomeres: as for long cycles, long paths imply the re-use of breakpoints. Rearrangements contain several possible inversions and translocations and at least one rearrangement involving telomeres: fusion, fission, telomeric inversion or non-reciprocal translocation.

The principle of parsimony favours small cycles (paths) since small cycles (paths) imply fewer rearrangements. For l *observed* breakpoints in G_1, one long cycle would imply $l-1$ rearrangements, while many distinct small cycles of length 4 would imply only $l/2$ rearrangements. Therefore, for large values of l, the biological significance of a cycle becomes suspicious. The presence of a long cycle in a adjacency graph could be: (i) either an artefact due to a bad reconstruction of synteny blocks or to rearrangements other than inversions, translocations, fusion or fission that have not been explicitly considered in our analysis; and (ii) or a sign of multiple re-use of the same breakpoints.

6 Comparison with outgroup genomes

The latter step of the algorithm returns a list of linked adjacencies corresponding to the rearrangements between the genomes G_1 and G_2 under comparison. We are interested in distinguishing which rearrangements occurred during the evolution of G_1 and which during the evolution of G_2 as well as in determining the local ancestral order of their corresponding blocks. Given a pair of adjacencies in G_1 and the corresponding pair in G_2, we know that either one or the other has been obtained by a rearrangement, say the pair in G_1. This means that the ancestral genome should share the adjacencies present in G_2. In Figure 7, for instance, which one of the pairs of adjacencies $(1;2)$ and $(7;8)$ in G_1 or $(1;-7)$ and $(-2;8)$ in G_2 was present in the ancestral genome?

To be able to answer to this question, we need of at least one outgroup genome G_3. The idea of the algorithm is simple: each time that synteny blocks in G_1 and G_2 do not share the same adjacencies, we look at a genome G_3 in the outgroup. If G_3 shares the same adjacencies as G_1, we deduce that the difference come from a rearrangement specific to G_2 and that the ancestor of G_1 and G_2 was locally like G_3 and G_1. By the parsimony principle, in fact, if this was not the case and if the ancestral genome was like G_2, two distinct but similar rearrangements should have happened during the evolution of G_1 and during the evolution of G_3 to make them locally similar. If G_3 is neither like G_1 nor G_2, then either we look to another genome in the outgroup, if any, or we are not able to trace back the origin of these breakpoints. Figure 1 highlights several genomes as potential candidates for forming the outgroup useful to reconstruct the ancestral genome B. The closest a genome is to G_1 and G_2, the better is for the comparison. This is the case of G_3, since genomes G_4 and G_5 are further away in the tree. The integration of the information coming from G_4 and G_5 could improve the reconstruction based on G_3 though.

6.1 From a pairwise comparison to a n by n comparison

The need to compare two genomes G_1 and G_2 with several genomes $G_3 \ldots G_n$ in the outgroup requires either to carefully define the combinatorial structure *shared* by all genomes or to determine a suitable algorithmic strategy for a pairwise comparison. The first option is followed by all known algorithms and demands all genomes to share the same blocks. Usually this induces an important loss of information on the ancestor to reconstruct, since the only genes and the only blocks that could be described in the ancestral genome are exactly those that are common to all genomes. We prefer to choose multiple pairwise comparisons between G_1 and G_2 and all the outgroup genomes, by only asking to both genomes G_i and G_j, of the G_iG_j comparison, where $i \in \{1,2\}$ and $j \in \{3,..n\}$, to share the same blocks. Then we work with the G_1G_2 comparison and all the G_iG_j comparison at the gene level as described below.

6.2 Algorithm for the reconstruction of translocations and inversions

For each pair of linked breakpoints in G_1, resulting from the G_1G_2 comparison, we are interested in checking whether the associated adjacencies exist also in G_i, where $i \in \{3,..n\}$. If G_1 and G_i share

a same adjacency, genes around the breakpoint in G_1 must belong to the same block in the G_1G_i comparison. More explicitly, to check the existence of the adjacency $(B;D)$ of G_1 in G_i, we take all genes g of B and define the set S_B collecting all blocks, in the G_1G_i comparison, containing the g's. The same is done for D and S_D is defined. Three different cases can arise. (i) $S_B \cap S_D \neq \emptyset$: B's genes and D's genes belong at least to one same block in the G_1G_i comparison, the adjacency represented by the blocks B and D in the G_1G_2 comparison, is shared by the genome G_i; we assign a score: $score_i((B;D)) = 1$. (ii) $S_B \cap S_D = \emptyset$: we infer that the breakpoint $(B;D)$ of the G_1G_2 comparison appears also in the G_1G_i comparison (G_1 and G_i are locally different); we fix a score: $score_i((B;D)) = 0$. (iii) $S_B = \emptyset$ or $S_D = \emptyset$: nothing can be inferred because we do not have the orthologs of B or D in G_i; we fix a null score: $score_i((B;D)) = 0$. In practice, we refine the 0,1 scores, considering values between 0 and 1 depending on a number of conditions, such as, the number of intervening genes between orthologs of surrounding genes of a given breakpoint in G_1, in the outgroup genome G_i. We will not detail all conditions here, but we just present the idea. Let us consider the breakpoint pair $(A;B)$ and $(C;D)$ in G_1. We compute a score $Score_1(A,B,C,D)$ as defined below:

$$\text{Score}_1(A,B,C,D) = \max_{i\in\{3,\ldots n\}} \text{score}_i((A;B)) + \max_{i\in\{3,\ldots n\}} \text{score}_i((C;D))$$

The same computation is done for the pair of breakpoints in G_2, which corresponds to the one already considered for G_1, for instance $(A;D)$ and $(C;B)$. A final score $Score_2(A,B,C,D)$ is computed as before.

We define a *confidence score* associated to the pairs of breakpoints in G_1G_2 as:

$$CS(A,B,C,D) = \frac{|\text{Score}_1(A,B,C,D) - \text{Score}_2(A,B,C,D)|}{2}$$

These two pairs of breakpoints correspond to a unique rearrangement. If $Score_1 > Score_2$, the rearrangement occurred along the branch from the ancestor of G_1 and G_2 to G_2. If $Score_1 < Score_2$ then the rearrangement occurred along the branch from the ancestor of G_1 and G_2 to G_1. If $Score_1 = Score_2$ ($CS = 0$), we cannot conclude. (References to blocks A, B, C, D are missing to simplify the notation.) The respective reconstructed ancestral adjacencies have confidence score CS.

Roughly speaking, this means that even if only one adjacency (over the two ancestral ones) is conserved in exactly one distant genome, the algorithm will validate both as ancestral.

6.3 The cases of re-used breakpoints

Cycles of length bigger than four represent several rearrangements involving breakpoint re-use. Some rearrangements could have happened along the branch from the ancestor of G_1 and G_2 to G_1 and others along the branch from the ancestor to G_2. Therefore, some adjacencies of G_1 and others from G_2 will appear in the outgroup. For two linked adjacencies, at least one adjacency has to be found to validate both adjacencies as ancestral. In the same way, in the general case, at least $n-1$ adjacencies have to be found to validate n adjacencies as ancestral. The longer is the cycle, the harder is to find $n-1$ adjacencies conserved in the outgroup genomes. For most long cycles, all adjacencies are not found, and breakpoints are treated independently as in the inferCARs algorithm: each adjacency found in an outgroup genome is validated as ancestral. Notice that in this latter case, the links between breakpoints become useless for ancestral genome reconstruction. However, knowing about their existence might be useful for a biologist searching for new insights linked to

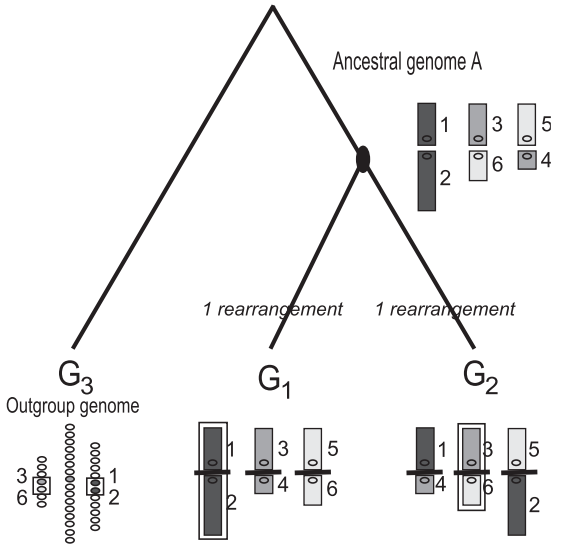

FIGURE 8. Difference with the inferCARs algorithm. Representation of six linked breakpoints in G_1 and G_2. The six grey rectangle represent synteny blocks. Small circles represent genes. Orthologous genes found in the outgroup have the same colour and number as the block of their ortholog. The outgroup genome shares one adjacency with G_1 (1/2) and another with G_2 (3/6). The adjacencies 1/2 and 3/6 are ancestral, as well as the adjacency 5/4 because the breakpoints are linked. Two of the rearrangements involved a re-use of the breakpoint 5/4. Notice that inferCARs cannot create an adjacency in the ancestral genome that is not shared with any extant species.

the difficulty of the reconstruction. This highlights that we are far from finding every rearrangements responsible for the observed breakpoints involved in long cycles.

Figure 8 illustrates a cycle of length six, a case of re-use, where two over three adjacencies are found in the outgroup genome. The third ancestral adjacency is deduced from the other two, even if it is not present in any extant species. This is an example where inferCARs [17] would not have been able to reconstruct the third ancestral adjacency.

7 Treatment of ambiguous cases based on synteny packs

The algorithm described until here is made for well-defined and non-overlapping synteny blocks, as were inferCARs and MGRA on which it is based. As seen in Section 4, on real data, blocks are not always well defined, some may be included, duplicated, overlapping or unsigned. These situations are not taken into account by current algorithms that only consider simpler block configurations as in Figure 3. They may be due to: (i) weak homology not corresponding to orthology (false positive); (ii) transpositions or telomeric rearrangements; or (iii) micro-inversions. In our goal to find as many as possible translocations and inversions, we delete ambiguous blocks in the first two cases, and in the third case, we explicitly undo the micro-inversions to be able to recover rearrangements involving ambiguous blocks. Because we do not know the origin of the ambiguity, the idea is to test different arrangements and look at the cycles length to see if it brings more noise than information.

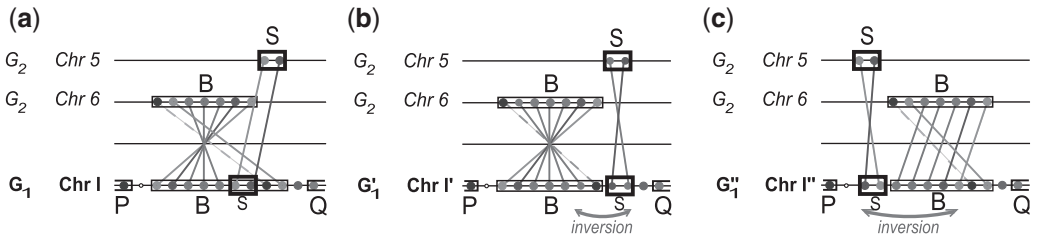

FIGURE 9. Example of the inclusion of a block. (**a**) Pairwise comparison of two genomes G_1 and G_2. Homologs constituting the synteny block S are included in G_1 among genes that belong to block B. (**b**) Pairwise comparison of G_2 with a *virtual* genome G'_1, resulting from an inversion in G_1 (last four genes) and identification of breakpoints between P and B, B and S, S and Q in G'_1. (**c**) Pairwise comparison of G_2 with a *virtual* genome G''_1, resulting from an inversion in G_1 (first seven genes) that leads to the identification of breakpoints between P and S, S and B, B and Q. The breakpoint between B and S and the one between S and B are different since they lie, respectively, on the right (left) and left (right) of B (S). Notice that genes order is different between G''_1 and G_2; configuration (b), which involves less microrearrangements, is more likely to be *real* than (c).

An arrangement will be preferred if it implies minimal cycles. To do that, we introduce a new notion: the *synteny pack*.

DEFINITION 15
A synteny pack corresponds to several alternatives for a local arrangement of several synteny blocks. It is a collection of ordered sequences of blocks $B_1, B_2, \ldots B_n$, where a B_i can have either a positive or a negative sign or be missing, and where consecutive blocks do not overlap, nor are included one in the other, nor are duplicated. (See Section 7.1 for each situation details.)

We start by illustrating the idea with an example.

Example with an included block: in Figure 9a, if we apply a small inversion of four genes on chromosome I of G_1, we obtain the genome G'_1 of Figure 9b. The block S in G'_1 is not included in block B anymore. The rearrangements (maybe translocations) involving the block S and the block B (on each side), which have occurred either during the evolution of G_1 and/or during the evolution of G_2, look easier to find from the comparison of G'_1 with G_2 than from the comparison of G_1 with G_2. To be able to reconstruct them, we need to consider a special order of B and S in the analysis: B being on the left of S and both being between P and Q along the chromosome. As shown in Figure 9c, another inversion could be responsible for the inclusion of S into B. We should try both (even if the first one is the most likely, by parsimony; see legend). These two block orders allow for the identification of different breakpoints: respectively $(P;B)$, $(B;S)$, $(S;Q)$ and $(P;S)$, $(S;B)$, $(B;Q)$ which were not identifiable in G_1. They will be useful to find rearrangements and local ancestral adjacencies.

Existing algorithms, when applied to the example illustrated in Figure 9, would propose a simplified approach which would drop blocks like S. As a consequence, they would have missed the reconstruction of the corresponding rearrangements (in Section 7.3, we shall discuss the different advantages to explicitly treat included blocks).

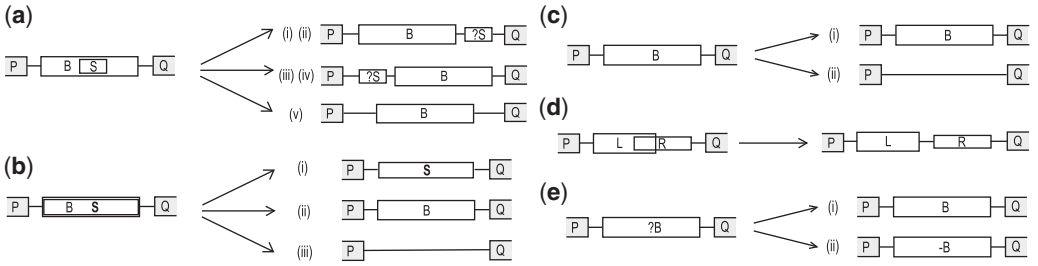

FIGURE 10. Five different cases where blocks are not well defined and where synteny packs are needed. See the text for a detailed definition of each cases: (**a**) inclusion, (**b**) duplication, (**c**) poorly defined region, (**d**) overlap and (**e**) unsigned.

7.1 Different cases involving synteny packs

To solve ambiguous cases in genome reconstruction, we introduce *synteny packs* representing different possible blocks arrangements. We shall then validate the soundness of the rearrangements that occurred on each side of these ambiguous blocks. (As we will see, testing the soundness of the rearrangements allows us to identify the *correct* order and to validate it as an *ancestral intermediate*.) This will be done in the same spirit as the treatment of inclusion discussed above. In the following, we detail the five different configurations involving a *synteny pack*. We shall analyse a genome as being composed by two entities, synteny blocks and synteny packs, which do not overlap each other. A block is represented with [ ] and a synteny pack with {}. In Figure 9a, we can locally describe G_1 as ...$[P]\{BS|SB\}[Q]$... where BS and SB represent the two possible situations associated to the overlapping of S with B.

The case of blocks inclusion: if S is included in B in G_1, we consider five different blocks arrangements (Figure 10a): $\{BS|B-S|SB|-SB|B\}$. The first four are justified by the fact that the inclusion can come from one or more inversions, and the last one, by the fact that S may have been inserted or be a *wrong* orthologous block (inclusions are sometimes due to homology and not orthology). For several included blocks $S_1 \ldots S_n$, we test: (i) all combinations involving each included block S_i individually ($\{BS_i|B-S_i|S_iB|-S_iB\}$); (ii) all combinations involving two included blocks S_i and S_j with $i<j$ at a time, considering only arrangements resulting from two inversions ($\{-S_iB-S_j|BS_j-S_i|B-S_i-S_j|-S_jS_iB|-S_i-S_jB\}$). We do not consider triplet or even more blocks together because of the high computational time.

The case of blocks duplication: if S and B are the same block in G_1, we consider three different blocks arrangements (Figure 10b): $\{S|B|\emptyset\}$. There is only one block in G_1 for two in G_2: S and B. We are interested to find which block in G_2 is ortholog to the one in G_1 and we consider three different cases: S, B or none. The *none* case is tested when no signal of orthology is found for S and B. (This possibility corresponds to the biologically sound situation of loss of the ortholog.)

The case of blocks localized in a poorly defined region: for B in G_1, if B and its homolog in G_2 are *telomeric* (i.e. all genes in the block lie in the first or the last 30 genes of a chromosome) or the homolog of B in G_2 is included in a block, then we consider two different blocks arrangements (Figure 10c): $\{B|\emptyset\}$. Notice that telomeric regions are known to be highly rearranged and, as a consequence, rearrangements involving B might interfere with other ones. For this reason, it might be wise to avoid their reconstruction (this corresponds to the option $\emptyset$). Similarly, if B is an included

block in G_2, it might have been a *wrong* block as explained above for the inclusion case, and not considering this block would be the best option in this case too.

The case of blocks overlapping: if two blocks L and R overlap in G_1, where the starting point of L is at the left of the starting point of R, we consider a single block arrangement (Figure 10d): $\{LR\}$. The hypothesis that R precedes L is discharged by the parsimony principle (since it would require at least one more inversion).

The case of unsigned blocks: if a block B is unsigned in G_1, we consider two different blocks arrangements (Figure 10e): $\{B|-B\}$. The block B may have been positive or negative before having been subjected to several microrearrangements and both possibilities are treated.

A synteny pack might be the result of a combination of these five cases. For instance, if blocks S_1 and S_2 (S_1 being before S_2) are included in a block B in G_1 and B is included in another block B' in G_2, the synteny pack in G_1 corresponds to: $\{BS_1|B-S_1|S_1B|-S_1B|BS_2|B-S_2|S_2B|-S_2B|-S_1B-S_2|BS_2-S_1|B-S_1-S_2|-S_2S_1B|-S_1-S_2B|B|\emptyset\}$. It is important to observe that the definition of a synteny pack in G_1 might depend on the definition of a synteny pack in G_2 and vice versa.

7.2 *Synteny packs are solved by using adjacency graphs*

The idea underlying the 'synteny pack' notion is that some blocks order will allow us to find rearrangements and other blocks order do not. An adjacency graph will be constructed for each combination of a synteny pack, hoping that one graph would be formed by short cycles, while the rest by longer ones (being less parsimonious).

Figure 11 shows some details of the comparison between two yeast species, *Lachancea kluyveri* (*LAKL*) and *Lachancea thermotolerans* (*LATH*). Among the few blocks represented, Block [104] (constituted by two genes) is included in Block [100] in *LATH* and this configuration generates a synteny pack in *LATH* and a synteny pack in *LAKL* (refer to *the case of blocks localized in a poorly defined region*). Part of chromosome B of *LATH* can be written as $\{100,-104|100,104|-104,100|104,100|100\}$, this notation representing the five local potential ancestral orders of *LATH*. This implies that the left neighbour of [23] could have been [100] or [104] or [−104] and in the same way, the right neighbour of [104] could have been [100] or [23] or [75]. Cycles and paths are computed for all possible combinations by the algorithm. The principle of parsimony, favouring smallest cycles and smallest paths, plays a crucial role in the identification of the optimal rearrangements and therefore of the expected local intermediate order. In Figure 11, the five pairs of potential intermediate orders of genomes *LATH* and *LAKL* involve different cycles. They are all described separately in the figure, and the principle of parsimony guides the algorithm to choose the three smallest cycles as optimal ones (Figure 11a). This solution provides evidence for a local intermediate order where Block [104] is on the left of Block [100].

If all combinations involve cycles of the same length, the algorithms chooses one of them randomly. In fact, they link the same breakpoints but in a different order.

7.3 *Precision recovered by explicitly treating included blocks*

Based on the yeast example discussed before, we illustrate the advantages of unraveling synteny packs (Figure 12c) instead of (i) splitting including blocks into three blocks: left block, right block and included block (Figure 12a) or (ii) ignoring the existence of the included block (Figure 12b).

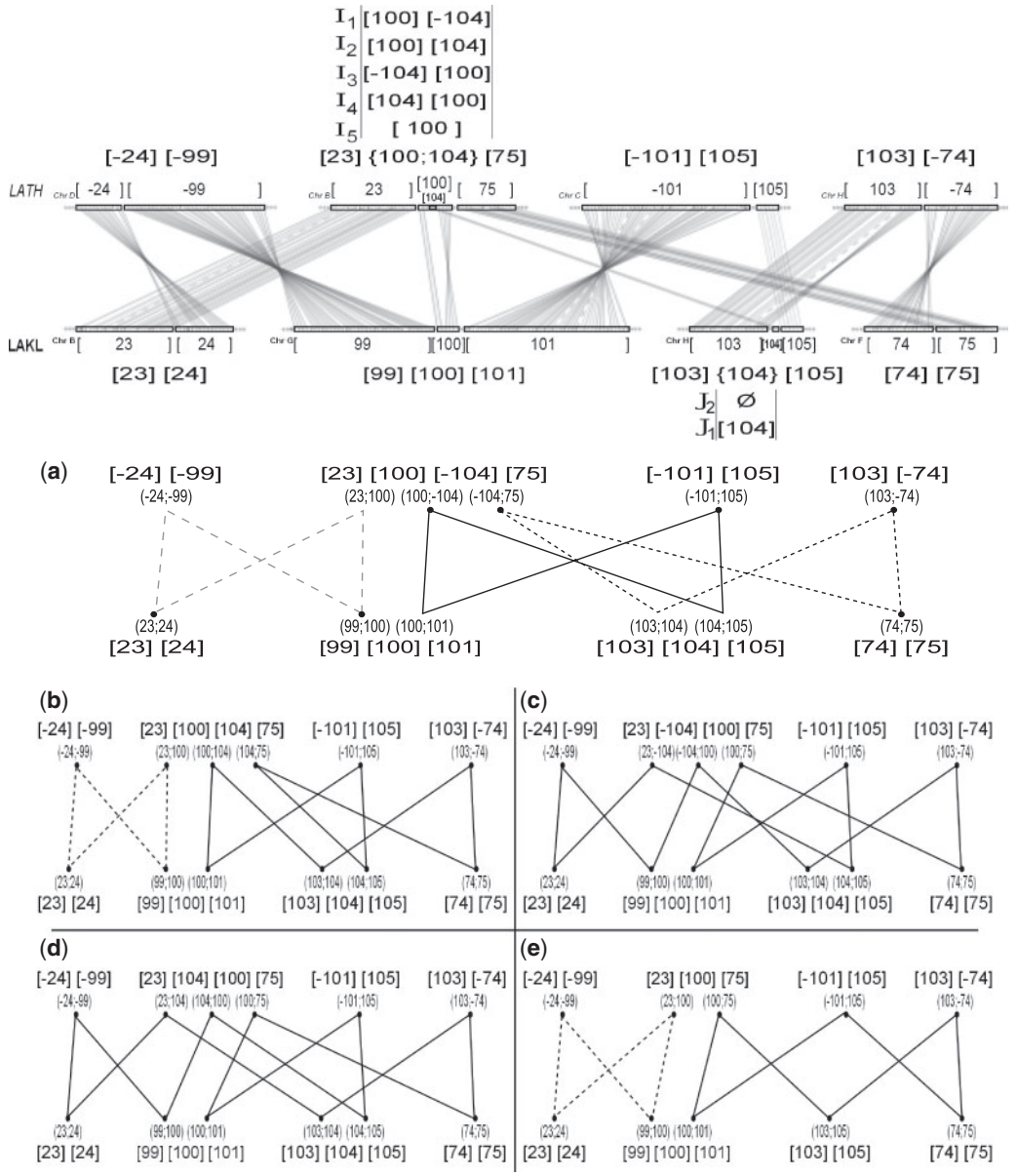

FIGURE 11. Different cycles resulting from different combinations of a synteny pack. Representation of few blocks of the comparison between two genomes: *Lachancea kluyveri* (*LAKL*) and *Lachancea thermotolerans* (*LATH*). Blocks are indicated in bracket. Block [104] is included in Block [100] in *LATH*. The top picture represents some homology relationships between genes localized in regions that are involved in mutual rearrangements. The different block orders and respective breakpoints of the five local potential intermediate orders of *LATH* and the two orders of *LAKL* are represented. These different orders generate five possible combinations of linking breakpoints. No other adjacency graphs are possible. (**a**) I_1 and J_1 involve three small cycles of length 4. (**b**) I_2 and J_1 involve two cycles of length 4 and 8, respectively. (**c**) I_3 and J_1 involve a unique long path. (**d**) I_4 and J_1 involve a unique long path. (**e**) I_5 and J_1 involve two cycles of length 4 and 6, respectively.

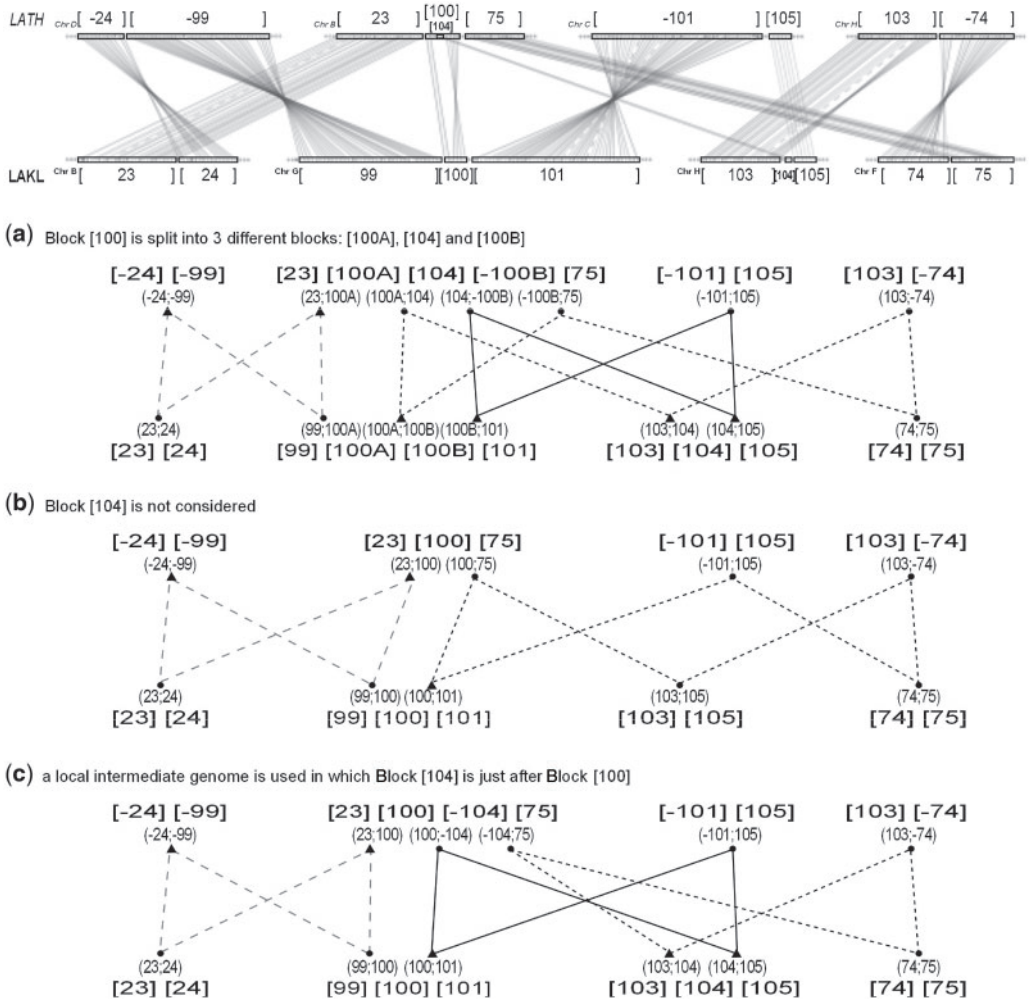

FIGURE 12. Consequences of the different ways to treat an included block. Representation of few blocks of the comparison of two genomes: *Lachancea kluyveri* (*LAKL*) and *Lachancea thermotolerans* (*LATH*). Blocks are indicated in bracket. Block [104] is included in Block [100] in *LATH*. The top picture represents some homology relationships between genes localized in regions that are involved in mutual rearrangements. The three adjacency graphs represent different possibilities to deal with included blocks. (**a**) The block including the other can be split into three blocks: left, right and the included block in question. (**b**) The included block can to not be considered at all. (**c**) A local intermediate genome can be used to solve the microrearrangement responsible for the inclusion. The breakpoints represented by a *triangle* are ancestral adjacencies (found in the outgroup genome *Zygosaccharomyces rouxii*), the *dot* are the one not found in the outgroup genome.

To distinguish and identify each rearrangement: in Figures 12a and b, we observe cycles of length 6, implying a re-use. In Figure 12a, the re-use corresponds to the fact that the inversion responsible for the inclusion of Block [104] into Block [100] took place in the same place (case of re-use) of the rearrangement involving Blocks [−104] and [75] with Blocks [103] and [−74] (Figure 12c). In Figure 12b, the situation is more complicated, since the loss of the Block [104] has merged the rearrangements occurring on its left and on its right (Figure 12c) and they cannot be distinguished anymore. Only in Figure 12c, we are able to identify precisely the four rearrangements that occurred, three being represented on the graph and the fourth one being the small inversion assumed to be responsible for the inclusion treated in the synteny pack.

To validate a maximum of ancestral adjacencies: in Figure 12c, the adjacency (74; 75) is validated as ancestral because (103; 104) is ancestral and is linked to it. This is not the case anymore in Figure 12b, where two adjacencies over three are not found in the outgroup. The disadvantages to merge cycles, in addition to merge rearrangements, are that we cannot use the property 'one breakpoint found, two breakpoints validated' anymore. This will drive to a partial reconstruction of the ancestor.

The impact of using as much information as possible: in the particular example that we discussed, the Block [104], supported by two genes only, was the only proof of the ancestral adjacency of the Blocks [103], [104] and [105], and we would lose this information if the Block [104] would not be considered explicitly (Figure 12b). As a consequence, the reconstructed ancestor would be partial. On the other hand, by splitting blocks, one would risk to create new ones that are very small (as [100*A*]) and with no homologs shared with outgroup. This would imply a loss of information on adjacencies around the small block and lead to the reconstruction of a partial ancestor. (Notice that it is not the case for Block [100A] in Figures 12a though, where the adjacency (23; 100*A*) is anyhow detected as ancestral, its two genes having orthologs in the outgroup.)

To relativize constructed synteny blocks: in Figures 12a, the Block [104] was inserted by an inversion into Block [100]. By splitting the Block [100], we can still recover most of the rearrangements and adjacencies. Included blocks do not always result from an inversion, often they are either just homologs (and not orthologs) or directly inserted. In all these cases, a long cycle is created and it might be preferable to avoid the reconstruction by not considering it or by splitting blocks. Whenever a block is included, it is interesting to see if its analysis would bring information or rather would add some noise to the reconstruction.

8 Conclusions

To help biologists to understand the evolutionary process of chromosomal rearrangements, one needs to reach an accurate definition of breakpoint regions and a precise reconstruction of the events determining these breakpoints. This means that one needs to keep as close as possible to biological data. Our algorithm takes as input a representation of the biological data which is closer to the reality than ever before. Thanks to the introduction of a parameter Δ measuring the conservation of pairwise gene distances in compared genomes and a notion of gene orthology that can be relaxed to low sequence similarity (30% homologs), our synteny blocks are more precise and their combinatorics more realistic. Synteny blocks can overlap, be included in each other, be duplicated and be unsigned. The complexity of synteny blocks leads to several solutions among which the algorithm selects the most parsimonious one.

Our algorithm exploits the unique features of the two previous studied models, the rearrangement-based models and the cytogenetic-based models, to reconstruct both the ancestral genome and the

rearrangements in once. The notion of a pair of linked breakpoints, present in Alekseyev and Pevzner's [2], is revisited using Ma's approach [17] and phylogenetic trees. In addition, a confidence level on the existence of a rearrangement explaining a pair of breakpoints and its position along a given branch of the phylogenetic tree is provided.

With the purpose in mind to be as close as possible to biological data, our algorithm compares genomes in a novel manner. It compares at least three genomes, and possibly many more, and ensures the preservation of all the information coming from pairwise comparisons. The multiple pairwise comparisons with genomes from the outgroup are independent and this has a deep methodological impact. With the incoming of new sequenced genomes, the reconstruction can be incrementally refined and upgraded, to explain breakpoints already introduced but left unresolved, without ever loosing pre-existing information and without re-computing.

Funding

This work was supported by a grant from the Agence Nationale de la Recherche ('GB-3G', ANR-10-BLAN-1606-01).

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, 4th edn., Garland Science, 2002.

[2] M. A. Alekseyev and P. A. Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Research*, **19**, 943–957, 2009.

[3] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402, 1997.

[4] V. Bafna and P. A. Pevzner. Genome rearrangements and sorting by reversals. *SIAM Journal on Computing*, **25**, 272–289, 1996.

[5] A. Bergeron, J. Mixtacki, and J. Stoye. On computing the breakpoint reuse rate in rearrangement scenarios. In *Comparative Genomics*, Vol. 5267 of *Lecture Notes in Computer Science*, C. Nelson and S. Vialette, eds, pp. 226–240. Springer, 2008.

[6] A. Bhutkar, S. W. Schaeffer, S. M. Russo, M. Xu, T. F. Smith, and W. M. Gelbart. Chromosomal rearrangement inferred from comparisons of 12 drosophila genomes. *Genetics*, **179**, 1657–1680, 2008.

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

[8] 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**, e1000234, 2008.

[9] G. Fertin, A. Labarre, I. Rusu, E. Tannier, and S. Vialette. *Combinatorics of Genome Rearrangements*. MIT Press, 2009.

[10] 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**, e1000485, 2009.

[11] S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of Association for Computing Machinery*, **46**, 1–27, 1999.

[12] G. Jean, D. J. Sherman, and M. Nikolski. Mining the semantics of genome super-blocks to infer ancestral architectures. *Journal of Computational Biology*, **16**, 1267–1284, 2009.
[13] H. Kaplan, R. Shamir, and R. E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal of Computing*, **29**, 880–892, 2000.
[14] C. Lemaitre, E. Tannier, C. Gautier, and M.-F. Sagot. Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC Bioinformatics*, **9**, 286, 2008.
[15] W.-H. Li. *Molecular Evolution*. Sinauer Associates, 1997.
[16] M. Lynch. *The Origins of Genome Architecture*, Vol. 98. Sinauer Associates, 2007.
[17] 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, 2006.
[18] B. M. E. Moret, A. C. Siepel, J. Tang, and T. Liu. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *WABI '02: Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pp. 521–536. Springer, 2002.
[19] J. H. Nadeau and B. A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 814–818, 1984.
[20] J. D. Palmer and L. A. Herbon. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, **28**, 87–97, 1988.
[21] Q. Peng, P. A. Pevzner, and G. Tesler. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Computational Biology*, **2**, e14, 2006.
[22] P. A. Pevzner. *Computational Molecular Biology : an Algorithmic Approach*. MIT Press, 2000.
[23] G. Tesler. GRIMM: genome rearrangements web server . *Bioinformatics*, **18**, 492–493, 2002.
[24] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, **21**, 3340–3346, 2005.
[25] F. Zhang, C. M. Carvalho, and J. R. Lupski. Complex human chromosomal and genomic rearrangements. *Trends in Genetics*, **25**, 298–307, 2009.
[26] H. Zhao and G. Bourque. Recovering genome rearrangements in the mammalian phylogeny. *Genome Research*, **19**, 934–942, 2009.

Received 19 November 2009