

# Genome Halving Problem Revisited

Max A. Alekseyev and Pavel A. Pevzner

Department of Computer Science and Engineering,  
University of California at San Diego,  
La Jolla, CA 92093-0114, U.S.A.  
{maxal, ppevzner}@cs.ucsd.edu

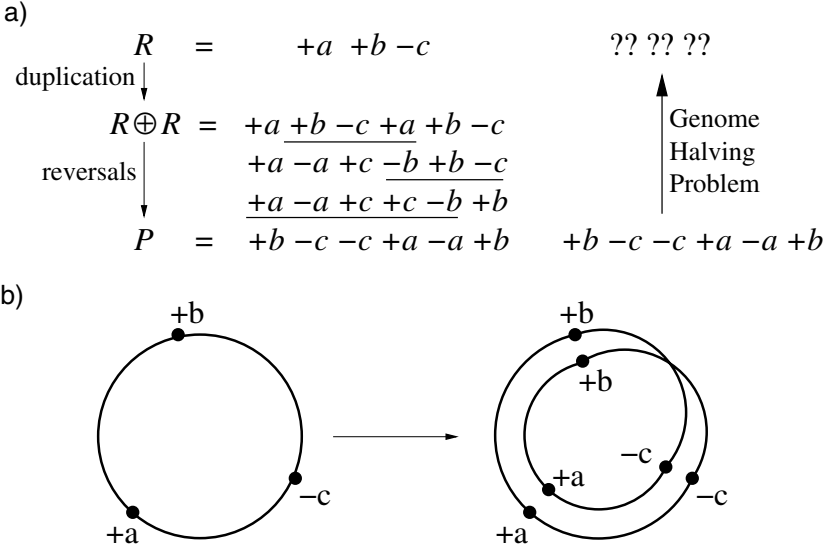
**Abstract.** The Genome Halving Problem is motivated by the whole genome duplication events in molecular evolution that double the gene content of a genome and result in a perfect duplicated genome that contains two identical copies of each chromosome. The genome then becomes a subject to rearrangements resulting in some rearranged duplicated genome. The Genome Halving Problem (first introduced and solved by Nadia El-Mabrouk and David Sankoff) is to reconstruct the ancestral pre-duplicated genome from the rearranged duplicated genome. The El-Mabrouk–Sankoff algorithm is rather complex and in this paper we present a simpler algorithm that is based on a generalization of the notion of the breakpoint graph to the case of duplicated genomes. This generalization makes the El-Mabrouk–Sankoff result more transparent and promises to be useful in future studies of genome duplications.

## 1 Introduction

The Genome Halving Problem is motivated by the *whole genome duplication* events in molecular evolution [13], [17], [15], [12], [5]. These dramatic evolutionary events double the gene content of a genome  $R$  and result in a *perfect duplicated genome*  $R \oplus R$  that contains two identical copies of each chromosome. The genome then becomes a subject to rearrangements that shuffle the genes in  $R \oplus R$  resulting in some *rearranged duplicated genome*  $P$ . The *Genome Halving Problem* is to reconstruct the ancestral *pre-duplicated genome*  $R$  from the rearranged duplicated genome  $P$  (Fig. 1a).

From the algorithmic perspective, the genome is a collection of chromosomes, and each chromosome is a sequence over a finite alphabet (depending on the scale, the alphabet may vary from *genes* to *synteny blocks*). DNA has two strands and genes on a chromosome have directionality that reflects the strand of the genes. We represent the order and directions of the genes on each chromosome as a sequence of *signed elements*, i.e., elements with signs “+” and “-”.

For the sake of simplicity, we focus on the *unichromosomal* case, where the genomes consist of just one chromosome and assume that the genomes are *circular*. A unichromosomal genome where each gene appears in a single copy sometimes is referred to as *signed permutation*.



**Fig. 1.** a) Whole genome duplication of genome  $R = +a+b-c$  into a perfect duplicated genome  $R \oplus R = +a + b - c + a + b - c$  followed by three reversals. b) Whole genome duplication of a circular genome

For unichromosomal genomes the rearrangements are limited to *reversals*. The reversal  $(i, j)$  over genome  $x_1x_2 \dots x_n$  “flips” genes  $x_i \dots x_j$  as follows:

$$\begin{array}{c}
 x_1 \dots x_{i-1} \quad \underline{x_i \quad x_{i+1} \dots x_j x_{j+1} \dots x_n} \\
 \downarrow \\
 x_1 \dots x_{i-1} \quad \underline{-x_j - x_{j-1} \dots -x_i x_{j+1} \dots x_n}
 \end{array}$$

The *reversal distance* between two genomes is defined as the minimal number of reversals required to transform one genome into another.

The *whole genome duplication* is a concatenation of the genome  $R$  with itself resulting in a *perfect duplicated genome*  $R \oplus R$  (Fig. 1b). The genome  $R \oplus R$  becomes a subject to reversals that change the order and signs of the genes and transforms  $R \oplus R$  into a *duplicated genome*  $P$ . The Genome Halving Problem is formulated as follows.

**Genome Halving Problem.** Given a duplicated genome  $P$ , recover the ancestral pre-duplicated genome  $R$  minimizing the reversal distance from the perfect duplicated genome  $R \oplus R$  to the duplicated genome  $P$ .

The Genome Halving Problem was solved in a series of papers by El-Mabrouk and Sankoff [6], [7], [8] culminating in a rather complex algorithm in [9]. The El-Mabrouk-Sankoff algorithm is one of the most technically challenging results in bioinformatics and its proof spans almost 40 pages in [9] (covering

both unichromosomal and multichromosomal genomes). In this paper we revisit the El-Mabrouk–Sankoff work and present a simpler algorithm for the case of unichromosomal genomes.<sup>1</sup>

The paper is organized as follows. Section 2 reviews the Hannenhalli-Pevzner theory and formulates the duality theorem for genomes without duplicated genes. Section 3 discusses the problem of finding rearrangement distance between duplicated genomes and extension of the Hannenhalli-Pevzner theory to this case. Section 4 introduces the concept of the contracted breakpoint graph for duplicated genomes. In section 5 we study cycle decompositions of contracted breakpoint graphs in the case when one of the genomes is perfect duplicated. Finally, in section 6 we present our new Genome Halving Algorithm.

## 2 Hannenhalli-Pevzner Theory

A duality theorem and a polynomial algorithm for computing reversal distance between two signed permutations was first proposed by Hannenhalli and Pevzner [10]. The algorithm was further simplified and improved in a series of papers [3], [11], [1], [16] using the *breakpoint graph* construction introduced in [2]. Recently, Bergeron et al., [4] proposed yet another simplification of the Hannenhalli-Pevzner proof that does not use the breakpoint graph construction.

Let  $P$  be a circular signed permutation. Bafna and Pevzner [2] described a transformation of a signed permutation on  $n$  elements into an unsigned permutation on  $2n$  elements by substituting every element  $x$  in the signed permutation by two elements  $x^t$  and  $x^h$  in the unsigned permutation.<sup>2</sup> Each element  $+x$  in the permutation  $P$  is replaced with  $x^t x^h$ , and each element  $-x$  is replaced with  $x^h x^t$  resulting in an unsigned permutation  $P'$ . For example, a permutation  $+a + b - c$  will be transformed into  $a^t a^h b^t b^h c^h c^t$ . Element  $x^t$  is called an *obverse* of element  $x^h$ , and vice versa.

Let  $P$  and  $Q$  be two circular signed permutations on the same set of elements  $\mathcal{G}$ , and  $P'$  and  $Q'$  be corresponding unsigned permutations. The breakpoint graph<sup>3</sup>  $G = G(P, Q)$  is defined on the set of vertices  $V = \{x^t, x^h \mid x \in \mathcal{G}\}$  with edges of three colors: “obverse”, black, and gray (Fig. 3). Edges of each color form a matching on  $V$ :

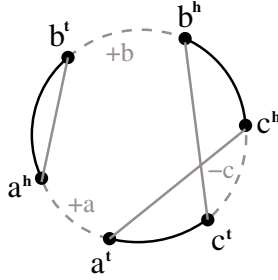
- pairs of obverse elements form an *obverse matching*;
- adjacent elements in  $P'$ , other than obverse, form a *black matching*;
- adjacent elements in  $Q'$ , other than obverse, form a *gray matching*.

Every pair of matchings forms a collection of *alternating* cycles in  $G$ , called *black-gray*, *black-obverse*, and *gray-obverse* cycles respectively (a cycle is alternating if colors of its edges alternate). The permutation  $P'$  can be read along a

<sup>1</sup> A generalization of our results to *multichromosomal* and *linear* genomes will be discussed elsewhere.

<sup>2</sup> Indices “t” and “h” stand for “tail” and “head” respectively.

<sup>3</sup> Our definition of the breakpoint graph is slightly different from the original definition from [2] and is more suitable for analysis of duplicated genomes.



**Fig. 2.** The breakpoint graph  $G(P, Q)$  for  $P = +a + b - c$  and  $Q = +a + b + c$

single black-obverse cycle while the permutation  $Q'$  can be read along a single gray-obverse cycle in  $G$ . The black-gray cycles in the breakpoint graph  $G$  play an important role in computing the reversal distance between the permutations  $P$  and  $Q$ . According to the Hannenhalli-Pevzner theory, the reversal distance between permutations  $P$  and  $Q$  is given by the formula:

$$d(P, Q) = b(G) - c(G) + h(G) \tag{1}$$

where  $b(G)$  is the number of black edges in the breakpoint graph  $G$ ,  $c(G)$  is the number of black-gray cycles in the breakpoint graph  $G$ , and  $h(G)$  is a small easily computable combinatorial parameter.

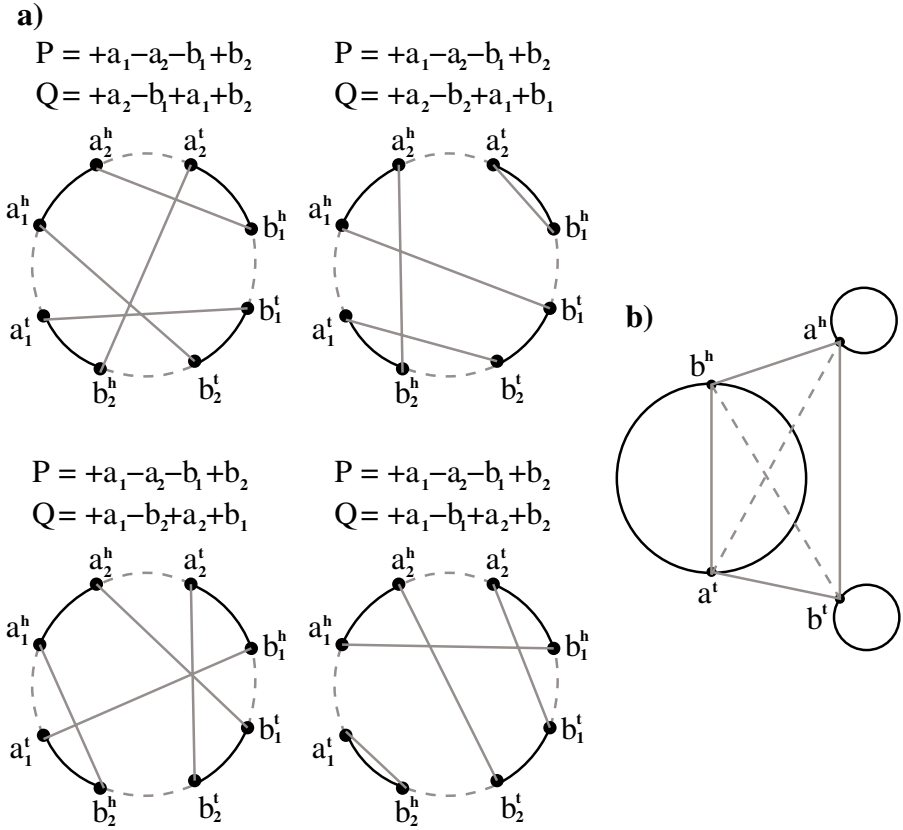
### 3 Reversal Distance Between Duplicated Genomes

While the Hannenhalli-Pevzner theory leads to a fast algorithm for computing reversal distance between two signed permutations, the problem of computing reversal distance between two genomes with duplicated genes remains unsolved.

Let  $P$  and  $Q$  be duplicated genomes on the same set of genes  $\mathcal{G}$  (i.e., each gene appears in two copies). If one labels copies of each gene  $x$  as  $x_1$  and  $x_2$  then genomes  $P$  and  $Q$  become signed permutations and the Hannenhalli-Pevzner theory applies. As before we turn the labelled genomes  $P$  and  $Q$  into unsigned permutations  $P'$  and  $Q'$  by replacing each element  $x_i$  with a pair of obverses  $x_i^t x_i^h$  in the order defined by the sign of  $x_i$ . Breakpoint graph  $G(P, Q)$  of the labelled genomes  $P$  and  $Q$  has a vertex set  $V = \{x_1^t, x_1^h, x_2^t, x_2^h \mid x \in \mathcal{G}\}$  and uniquely defines permutations  $P'$  and  $Q'$  (and, thus, the original genomes  $P$  and  $Q$ ) as well as an inter-genome correspondence between gene copies.

We remark that different labellings may lead to different breakpoint graphs for the same genomes  $P$  and  $Q$  (Fig. 3a) and it is not clear how to choose a labelling that results in the minimum reversal distance between the labelled copies of  $P$  and  $Q$ .

Currently, the only known option for solving the reversal distance problem for duplicated genomes is to consider all possible labellings for each duplicated gene, to solve the reversal distance problem for each labelling, and to



**Fig. 3.** a) The breakpoint graphs corresponding to four different labellings of  $P = +a - a - b + b$  and  $Q = +a - b + a + b$ . b) The contracted breakpoint graph  $G'(P, Q)$

choose the labelling with the minimal reversal distance. For genomes with  $n$  genes, each present in  $k$  copies, it leads to  $(k!)^n$  invocations of the Hannenhalli-Pevzner algorithm rendering this approach impractical. In particular, for duplicated genomes with  $n$  genes (each gene present in 2 copies) it results in  $2^n$  calls to the Hannenhalli-Pevzner algorithm. Moreover, the problem remains open if one of the genomes is perfectly duplicated (i.e., computing  $d(P, R \oplus R)$ ). Surprisingly enough the problem of computing  $\min_R d(P, R \oplus R)$  that we address in this paper is solvable in polynomial time.

Using the concept of the breakpoint graph and formula (1) the Genome Halving Problem can be posed as follows. For a given duplicated genome  $P$ , find a perfect duplicated genome  $R \oplus R$  and a labelling of gene copies such that the breakpoint graph  $G(P, R \oplus R)$  of the labelled genomes  $P$  and  $R \oplus R$  attains the minimum value of  $b(G) - c(G) + h(G)$ . Since  $b(G)$  is constant and  $h(G)$  is typically small, the value of  $d(P, Q)$  mostly depends on  $c(G)$ . El-Mabrouk and

Sankoff [9] showed that the problems of maximizing  $c(G)$  and minimizing  $h(G)$  can be solved separately in a consecutive manner. In this paper we focus on the former, much harder, problem.

**Weak Genome Halving Problem.** *For a given duplicated genome  $P$ , find a perfect duplicated genome  $R \oplus R$  and a labelling of gene copies that maximizes the number of black-gray cycles  $c(G)$  in the breakpoint graph  $G$  of labelled genomes  $P$  and  $R \oplus R$ .*

## 4 Contracted Breakpoint Graph

Let  $P$  and  $Q$  be duplicated genomes on the same set of genes  $\mathcal{G}$  and  $G$  be a breakpoint graph defined by some labelling of  $P$  and  $Q$ . For a vertex  $u = x_i^j$  in  $G$  (where  $x \in \mathcal{G}$ ,  $i \in \{1, 2\}$ ,  $j \in \{t, h\}$ ), we denote its *counterpart* by  $\bar{u} = x_{3-i}^j$ . Counterpart vertices form yet another matching in  $G$ .

A *contracted breakpoint graph*  $G'(P, Q)$  is a result of contracting every pair of counterpart vertices in the breakpoint graph  $G$  into a single vertex (e.g.,  $x_1^t$  and  $x_2^t$  are contracted into a single vertex  $x^t$ ). So the contracted breakpoint graph  $G' = G'(P, Q)$  is a graph on the set of vertices  $V' = \{x^t, x^h \mid x \in \mathcal{G}\}$  with each vertex incident to two black, two gray, and a pair of parallel obverse edges (Fig. 3b). The contracted breakpoint graph  $G'(P, Q)$  is uniquely defined by  $P$  and  $Q$  and does not depend on a particular labelling.<sup>4</sup> The following theorem gives a characterization of the contracted breakpoint graphs.

**Theorem 1.** *A graph  $H$  with black, gray, and obverse edges is a contracted breakpoint graph for some duplicated genomes if and only if*

- each vertex in  $H$  is incident to two black edges, two gray edges, and a pair of parallel obverse edges;
- $H$  is connected with respect to black and obverse edges (black-obverse connected);
- $H$  is connected with respect to gray and obverse edges (gray-obverse connected).

*Proof.* Suppose that graph  $H$  is a contracted breakpoint graph of the genomes  $P$  and  $Q$ . Consider the graph  $H$  as a contraction of a breakpoint graph  $G(P, Q)$  for some labelling of the genomes  $P$  and  $Q$ . Since there is a single black-obverse cycle in  $G$  that cannot be split by contraction, the graph  $H$  is black-obverse connected. Similar reasoning implies that the graph  $H$  is gray-obverse connected.

---

<sup>4</sup> The contracted breakpoint graph is a natural generalization of the notion of breakpoint graph for genomes with duplicated genes. The conventional breakpoint graph (Bafna and Pevzner [2]) of signed permutations  $P$  and  $Q$  on  $n$  elements can be defined as gluing of  $n$  pairs of obverse edges in the corresponded unsigned permutations  $P'$  and  $Q'$  (assuming  $P'$  and  $Q'$  are represented as black-obverse and gray-obverse alternating cycles). The breakpoint graph of duplicated genomes  $P$  and  $Q$  on  $n$  elements is simply gluing of  $n$  quartets of obverse edges.

Consider a black-obverse and gray-obverse connected graph  $H$  and label endpoint of each obverse edge  $x$  by  $x^t$  and  $x^h$ . Since the graph  $H$  is black-obverse connected, there exists an alternating Eulerian black-obverse cycle traversing all the black edges in this graph. The order of vertices in this cycle defines some duplicated genome  $P$ . Similarly, since the graph  $H$  is gray-obverse connected, there exists an alternating Eulerian gray-obverse cycle traversing all gray edges that defines some duplicated genome  $Q$ . Then the graph  $H$  is a contracted breakpoint graph for the genomes  $P$  and  $Q$ .  $\square$

In the case when  $Q$  is perfect duplicated genome (i.e.,  $Q = R \oplus R$ ) the gray edges in the contracted breakpoint graph  $G'(P, Q)$  form pairs of parallel gray edges that we refer to as *double* gray edges. Similarly to the obverse edges, the double gray edges form a matching in the contracted breakpoint graph  $G'$ .

Let  $G(P, Q)$  be a breakpoint graph for some labelling of  $P$  and  $Q$ . A set of black-gray cycles in  $G(P, Q)$  is being contracted into a set of black-gray cycles in the contracted breakpoint graph  $G'(P, Q)$  thus forming a black-gray cycle decomposition of  $G'(P, Q)$ . Therefore, each labelling induces a black-gray cycle decomposition of the contracted breakpoint graph. We are interested in a reverse problem: given a black-gray cycle decomposition of the contracted breakpoint graph  $G'(P, Q)$ , find labelling of  $P$  and  $Q$  that induces this cycle decomposition.

**Theorem 2.** *Any black-gray cycle decomposition of the contracted breakpoint graph  $G'(P, R \oplus R)$  is induced by some labelling of  $P$  and  $R \oplus R$ .*

*Proof.* Consider a contracted breakpoint graph  $G' = G'(P, R \oplus R)$  of the genomes  $P$  and  $R \oplus R$  and suppose that labelling of  $P$  is fixed. We will show how a particular black-gray cycle decomposition  $C$  of the graph  $G'$  defines a labelling on  $R \oplus R$  that induces this cycle decomposition.

We will label elements of  $R \oplus R$  in one-by-one fashion. First we represent  $R \oplus R$  as two copies of a linear sequence of unlabeled elements from  $\mathcal{G}$  corresponding to  $R$ . We label the first element  $x$  in the copies of  $R$  by  $x_1$  and  $x_2$ .

Suppose that first  $m$  elements are labelled in both copies of  $R$ . Let  $x$  be the  $m$ -th element in  $R$ . Without loss of generality we assume that  $x$  is labelled as  $x_1$  in the first copy of  $R$  and as  $x_2$  in the second copy. Let  $y$  be the  $(m + 1)$ -th (yet unlabeled) element in  $R$ . Since  $x$  and  $y$  are adjacent elements in  $R \oplus R$ , there exists a double gray edge  $(x, y)$  in the graph  $G'$ . In the black-gray cycle decomposition this edge appears two times. Consider its adjacent black edges for each appearance. Let  $(u, x), (x, y), (y, v)$  and  $(z, x), (x, y), (y, t)$  be two triples of edges consecutively appearing in some black-gray cycles from  $C$ . For the black edges  $(u, x), (y, v), (z, x), (y, t)$  we consider adjacencies in the labelled genome  $P$  that these edges originated from. Without loss of generality we assume that the originating adjacencies are  $(u_i, x_1), (y_j, v_k), (z_l, x_2), (\bar{y}_j, t_m)$  for some indices  $i, j, k, l, m \in \{1, 2\}$ . We label the element  $y$  as  $y_j$  in the first copy of  $R$  and as  $\bar{y}_j$  in the second copy so that elements  $x_1, y_j$  and  $x_2, \bar{y}_j$  will be adjacent in the labelled genome  $R \oplus R$ . We continue labelling in a similar manner until the whole genome  $R \oplus R$  is labelled.

Consider a breakpoint graph  $G(P, R \oplus R)$  for the labelled genomes  $P$  and  $R \oplus R$ . The labelling procedure implies that any black edge and gray edge adjacent in the breakpoint graph  $G(P, R \oplus R)$  are contracted into a pair of adjacent edges in the cycle decomposition  $C$  of the graph  $G'(P, R \oplus R)$ . Hence, the constructed labelling induces the cycle decomposition  $C$ .  $\square$

Let  $c_{max}(G')$  be the number of cycles in a maximal black-gray cycle decomposition of the contracted breakpoint graph  $G' = G'(P, R \oplus R)$ . Theorem 2 implies that the Weak Genome Halving Problem is equivalent to the following.

**Cycle Decomposition Problem.** *For a given duplicated genome  $P$ , find a perfect duplicated genome  $R \oplus R$  maximizing  $c_{max}(G'(P, R \oplus R))$ .*

Black and gray edges of the contracted breakpoint graph  $G'(P, R \oplus R)$  form a bi-colored graph that we study in the next section.

## 5 Cycle Decomposition of BG-Graphs

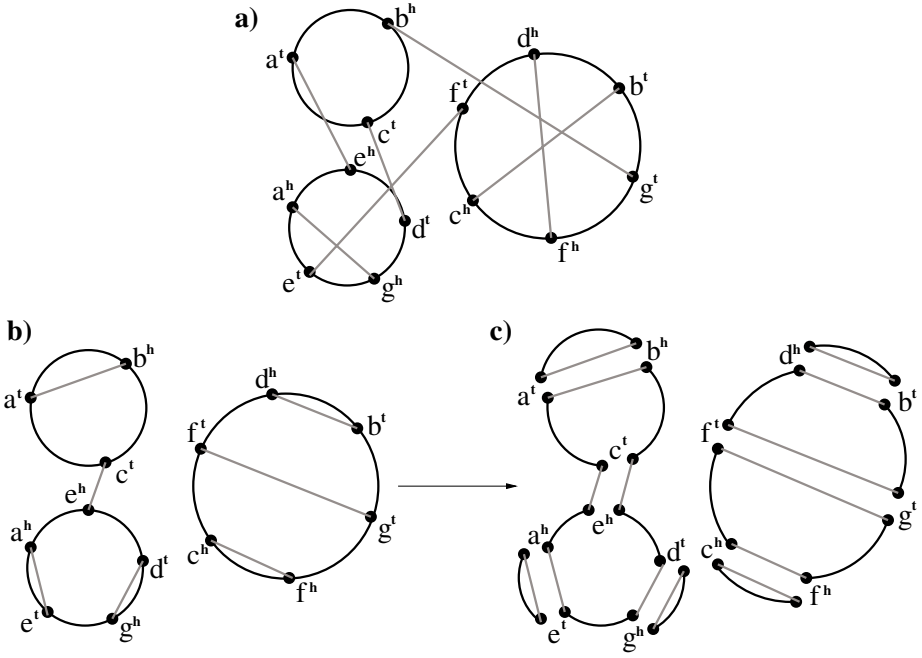
A *BG-graph*  $G$  is a graph with black and gray edges such that the black edges form *black cycles* and the gray edges form gray matching in  $G$  (Fig. 5a). We refer to gray edges in  $G$  as *double* gray edges and assume that every double gray edge is a pair of parallel gray edges. This assumption implies that every BG-graph can be decomposed into edge-disjoint black-gray alternating cycles.

Below we prove an upper bound on the maximal number of black-gray cycles  $c_{max}(G)$  in cycle decomposition of the BG-graph  $G$ , and formulate necessary and sufficient conditions for achieving this bound.

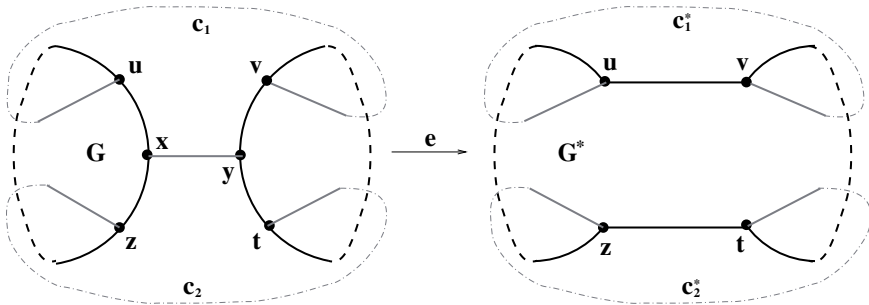
A BG-graph is *connected* if it is connected with respect to both black and gray edges. A double gray edge in the BG-graph connecting vertices of distinct black cycles is called *interedge*. A double gray edge connecting vertices of the same black cycle is called *intraedge*. Note that a connected BG-graph with  $m$  black cycles has at least  $m - 1$  interedges.

Let  $G$  be a BG-graph on  $2n$  vertices with  $m > 1$  black cycles,  $C$  be a black-gray cycle decomposition of  $G$ , and  $e = (x, y)$  be an interedge in  $G$ . We define an *e-transformation*  $(G, C) \xrightarrow{e} (G^*, C^*)$  of the graph  $G$  and its black-gray cycle decomposition  $C$  into a new BG-graph  $G^*$  on  $2(n - 1)$  vertices with  $m - 1$  black cycles and a black-gray cycle decomposition  $C^*$  of  $G^*$  of the same size as  $C$ . In the cycle decomposition  $C$  there are two black-gray cycles  $c_1$  and  $c_2$  passing through edge  $e$  (it may happen that  $c_1 = c_2$  when the same cycle passes through  $e$  two times). Suppose that  $c_1$  traverses edges  $(u, x), (x, y), (y, v)$  while  $c_2$  traverses edges  $(z, x), (x, y), (y, t)$ . To obtain graph  $G^*$  from  $G$  we replace these edges with black edges  $(u, v)$  and  $(z, t)$  respectively and delete vertices  $x$  and  $y$  (Fig. 5). This operation transforms the cycles  $c_1$  and  $c_2$  in  $G$  into into cycles  $c_1^*$  and  $c_2^*$  in  $G^*$ . We define the black-gray cycle decomposition  $C^*$  as cycles  $c_1^*, c_2^*$  and all cycles from  $C$ , except  $c_1$  and  $c_2$ .

**Lemma 1.** *Let  $C$  be a maximal black-gray cycle decomposition of a BG-graph  $G$  and  $(G, C) \xrightarrow{e} (G^*, C^*)$  be the e-transformation for some interedge  $e = (x, y)$  in  $G$ . Then  $c_{max}(G) = c_{max}(G^*)$ .*



**Fig. 4.** For a genome  $P = -a - b + g + d + f + g + e - a + c - f - c - b - d - e$ , a) a BG-graph corresponding to the contracted breakpoint graph  $G'(P, R \oplus R)$  for  $R = +a - g - b - c + d - f + e$ ; b) a BG-graph corresponding to the contracted breakpoint graph  $G'(P, R \oplus R)$  for  $R = -a - b - d - g + f - c - e$ ; c) a maximal cycle decomposition of the BG-graph in b)



**Fig. 5.**  $e$ -transformation of a graph  $G$  into a graph  $G^*$ . Black-gray cycles  $c_1, c_2$  in  $G$  passing through interedge  $e = (x, y)$  are transformed into black-gray cycles  $c_1^*, c_2^*$  in  $G^*$ . The black cycles connected by  $e$  in  $G$  are merged into a single black cycle in  $G^*$

*Proof.* By the definition of  $e$ -transformation,  $c_{max}(G) = |C^*| \leq c_{max}(G^*)$ . On the other hand, every black-gray cycle decomposition  $D^*$  of the graph  $G^*$  can be transformed into a black-gray cycle decomposition  $D$  of  $G$  of the same size (by simply substituting the black edges  $(u, v)$  and  $(z, t)$  in some black-gray cycles in  $D^*$  by black-gray-black triples  $(u, x), (x, y), (y, v)$  and  $(z, x), (x, y), (y, t)$ ). Therefore,  $c_{max}(G^*) \leq c_{max}(G)$ .  $\square$

**Theorem 3.** *If  $G$  is a connected BG-graph with  $2n$  vertices and  $m$  black cycles, then*

$$c_{max}(G) \leq n + 2 - m.$$

*Proof.* Suppose that  $c_{max}(G) = k$ , i.e., a maximal cycle decomposition of  $G$  contains  $k$  black-gray cycles. Consider the BG-graph  $G$  as a result of contracting these  $k$  black-gray cycles by a series of  $n$  gluings of pairs of gray edges into double gray edges. Since one needs at least  $k - 1$  such gluings to contract  $k$  disconnected black-gray cycles into a connected BG-graph,  $k - 1 \leq n$ . It implies the theorem for  $m = 1$ .

Assume  $m > 1$ . Since the BG-graph  $G$  is connected and contains  $m$  black cycles, there exists an interedge  $e$  in  $G$ . For a maximal cycle decomposition  $C$  of the BG-graph  $G$ , consider an  $e$ -transformation  $(G, C) \xrightarrow{e} (G^*, C^*)$ . Lemma 1 implies  $c_{max}(G) = c_{max}(G^*)$ . Note that  $G^*$  is a connected BG-graph on  $2(n - 1)$  vertices with  $m - 1$  black cycles. Iteratively applying similar  $e$ -transformations  $m - 1$  times we will end up with a BG-graph  $G^+$  of size  $2(n - (m - 1))$  that contains a single black cycle. Hence,  $c_{max}(G) = c_{max}(G^+) \leq n + 2 - m$ .  $\square$

Note that for a BG-graph  $G$ ,  $c_{max}(G)$  equals the sum of  $c_{max}(H)$  over all connected components  $H$  of  $G$ . Since the total size of all connected components is  $b(G)$ , Theorem 3 implies

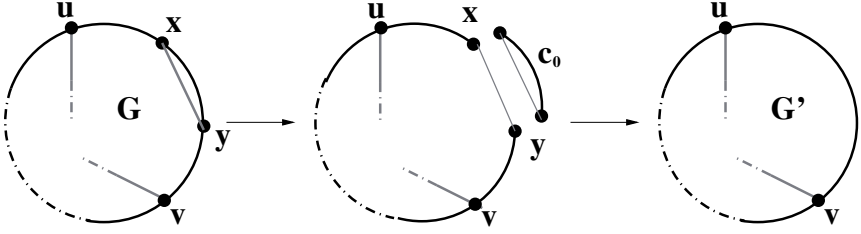
$$c_{max}(G) \leq b(G)/2 + 1 \cdot s_1 + 0 \cdot s_2 + (-1) \cdot s_3 + (-2) \cdot s_4 + \dots,$$

where  $s_m$  is the number of connected components with  $m$  black cycles. Let  $b_e(G)$  be the number of even black cycles (i.e., black cycles of even size) in  $G$ . Since  $s_1$  does not exceed  $b_e(G)$ ,

$$c_{max}(G) \leq b(G)/2 + b_e(G). \tag{2}$$

To achieve the upper bound (2), each connected component of  $G$  must contain either a single even black cycle (a *simple BG-graph*), or a pair of odd black cycles (a *paired BG-graph*). Fig. 5b shows a BG-graph containing an even black cycle forming a simple BG-graph, and a pair of odd black cycles forming a paired BG-graph.

We represent each black cycle of a BG-graph as points on a circle such that the arcs between adjacent points represent the black edges, and intraedges are drawn as straight chords within these circles. A BG-graph is *non-crossing* if its intraedges (as chords within each black circle) do not cross. A BG-graph in Fig. 5b is non-crossing while a BG-graph on in Fig. 5a is not.



**Fig. 6.** Transformation of a BG-graph  $G$  into a BG-graph  $G'$  by splitting a black-gray cycle  $c_0$  consisting of parallel black and gray edges  $(x, y)$

**Theorem 4.** For a simple BG-graph  $G$  on  $2n$  vertices,  $c_{max}(G) = n + 1$  if and only if  $G$  is non-crossing.

*Proof.* We prove the theorem in both directions by induction on  $n$ . The statement is trivial for  $n = 1$ . Assume that the statement is true for any simple BG-graph of size  $2(n - 1)$  and prove it for a simple BG-graph  $G$  of size  $2n$ .

We first prove (reasoning depends on the proof direction) that there exists a double gray edge  $e$  in  $G$  parallel to a black edge (i.e., connecting two adjacent points on a black circle) forming a black-gray cycle  $c_0$  of length 2.

If  $c_{max}(G) = n + 1$ , then a maximum cycle decomposition of the BG-graph  $G$  consists of  $n + 1$  black-gray cycles. Since these cycles contain  $2n$  gray edges in total, the pigeonhole principle implies that there exists a cycle  $c_0$  with a single gray edge  $e$ .

If the BG-graph  $G$  is non-crossing, consider a double gray edge  $e$  with the minimal span. If  $e$  spanned more than one black edge then there would exist a double gray edge with endpoints within the span of  $e$ , i.e., an edge with an even smaller span, a contradiction.

For a found edge  $e = (x, y)$ , let  $u$  and  $v$  be vertices adjacent to  $x$  and  $y$  on the black cycle. Transform  $G$  into a simple BG-graph  $G'$  on  $2(n - 1)$  vertices by removing the vertices  $x$  and  $y$  and all the incident edges, and by adding the black edge  $(u, v)$  (Fig. 5). Note that  $c_{max}(G') = c_{max}(G) - 1$  and  $G'$  is non-crossing if and only if  $G$  is non-crossing.

By induction the graph  $G'$  is non-crossing if and only if  $c_{max}(G') = n$ . Therefore,  $G$  is non-crossing if and only if  $c_{max}(G') = n + 1$ .  $\square$

Let  $G$  be a paired BG-graph  $G$  of size  $2n$  (consisting of two odd black cycles) and  $e$  be an interedge in  $G$ . For a maximal black-gray cycle decomposition  $C$  of  $G$ , let  $(G, C) \xrightarrow{e} (G^*, C^*)$  be an  $e$ -transformation of  $G$ . Note that the graph  $G^*$  is a simple BG-graph on  $2(n - 1)$  vertices. Lemmas 1 and 3 imply  $c_{max}(G) = c_{max}(G^*) \leq n$ . Therefore, according to Theorem 4,  $c_{max}(G) = n$  if and only if the BG-graph  $G^*$  is non-crossing. We are interested in a particular case of this statement.

**Theorem 5.** For a paired BG-graph  $G$  of size  $2n$  with a single interedge,  $c_{max}(G) = n$  if and only if  $G$  is non-crossing.

*Proof.* It is easy to see that for a single interedge  $e$  in a paired BG-graph  $G$ , the  $e$ -transformation turns  $G$  into a non-crossing BG-graph if and only if  $G$  is non-crossing.  $\square$

We call a BG-graph *optimal* if its connected components are either simple BG-graphs, or paired BG-graphs with single interedges. Theorems 4 and 5 imply

**Theorem 6.** *For an optimal BG-graph  $G$ ,  $c_{max}(G) = b(G)/2 + b_e(G)$ .*

An optimal BG-graph and its maximal cycle decomposition are shown at Fig. 5b,c.

## 6 Genome Halving Algorithm

In order to solve the Cycle Decomposition Problem for a genome  $P$ , we will construct a contracted breakpoint graph  $G'(P, \cdot)$  which achieves the upper bound (2). The genome  $P$  alone defines a vertex set of the graph  $G'$ , an obverse matching, and black cycles in  $G'$  so that  $G'$  is black-obverse connected.

A *BO-graph* is a connected graph with black and obverse edges such that the black edges form black cycles and the obverse edges form an obverse matching (every duplicated genome  $P$  corresponds to a BO-graph). A *BOG-graph* is a graph with black, obverse, and gray edges such that black and obverse edges form a BO-graph (a *BO-subgraph*), and black and gray edges form an optimal BG-graph (a *BG-subgraph*). Note that each black-gray connected component of a BOG-graph is a simple non-crossing BG-graph or a paired non-crossing BG-graph with a single interedge.

We now pose the Cycle Decomposition Problem for a genome  $P$  as follows. For a given BO-graph  $G$  (defined by the genome  $P$ ), find a gray-obverse connected BOG-graph  $G'$  having  $G$  as a BO-subgraph. Theorems 1 and 6 imply that such a BOG-graph graph is a contracted breakpoint graph  $G'(P, R \oplus R)$  for some genome  $R$  for which  $c_{max}(G')$  achieves the upper bound (2).

We remark that gray-obverse connected components of a BOG-graph form gray-obverse cycles (alternating double gray and obverse edges). Hence, a BOG-graph is gray-obverse connected if and only if it has a single gray-obverse cycle.

**Lemma 2.** *For a BOG-graph with more than one gray-obverse cycle, there exists a black edge connecting two distinct gray-obverse cycles.*

*Proof.* Let  $H$  be a BOG-graph with more than one gray-obverse cycle. First we will show that there exists a black-gray connected component of the graph  $H$  containing two double gray edges from distinct gray-obverse cycles. Assume that all the double gray edges within each black-gray connected component belong to the same gray-obverse cycle. Then each gray-obverse cycle contains vertices of one or more black cycles. Let  $V_1$  and  $V_2$  be vertex sets of two distinct gray-obverse cycles. Since black and obverse edges connect vertices within the same set, the sets  $V_1$  and  $V_2$  are black-obverse disconnected, a contradiction to black-obverse connectivity of the graph  $H$ .

Let  $C$  be a black-gray connected component of the BOG-graph  $H$  containing two double gray edges from distinct gray-obverse cycles. We represent double gray edges of the component  $C$  as vertices of a graph  $E$  with edges induced by black edges of the component  $C$ . Black-gray connectivity of the component  $C$  implies that the graph  $E$  is connected. If every two double gray edges in  $C$  connected by a black edge belong to the same gray-obverse cycle in  $H$ , then connectivity of the graph  $E$  would imply that all the double gray edges in  $C$  belong to the same gray-obverse cycle.  $\square$

**Theorem 7.** *For a given BO-graph  $G$ , there exists a BOG-graph  $G'$  with a single gray-obverse cycle having  $G$  as a BO-subgraph.*

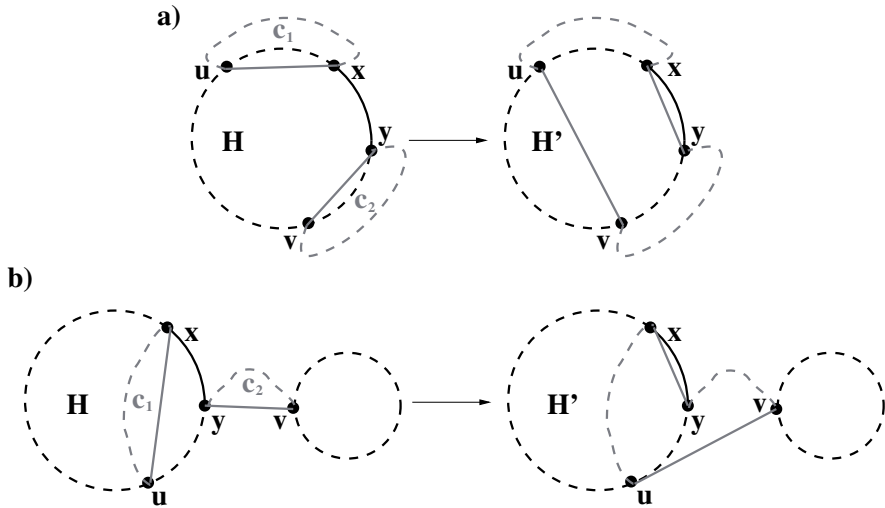
*Proof.* First we group odd black cycles in  $G$  into pairs (formed arbitrary), and introduce an arbitrary interedge connecting cycles in each pair. Then we complete each black cycle with an arbitrary non-crossing gray matching so that each vertex of  $G$  becomes incident to exactly one double gray edge. Denote the resulting graph by  $H$ . Note that  $H$  is a BOG-graph having  $G$  as a BO-subgraph.

If  $H$  has a single gray-obverse cycle, then the theorem holds for  $G' = H$ . Otherwise, we show how to modify the set of double gray edges in  $H$  to reduce the number of gray-obverse cycles.

Assume that there is more than one gray-obverse cycle in  $H$ . By Lemma 2 there is a black edge  $(x, y)$  connecting distinct gray-obverse cycles  $c_1$  and  $c_2$ . Let  $(x, u)$  and  $(y, v)$  be double gray edges incident to the vertices  $x$  and  $y$  respectively. We replace the edges  $(x, u)$  and  $(y, v)$  in  $H$  with double gray edges  $(x, y)$  and  $(u, v)$  resulting in a graph  $H'$ . Fig. 6 illustrates two cases depending on whether the edge  $(y, v)$  is an interedge (since  $(x, u)$  and  $(y, v)$  belong to the same black-gray connected component, at most one of them can be an interedge).

We will show that the BG-subgraph of  $H'$  is optimal. There are two new double gray edges in the BG-subgraph of  $H'$  compared to  $H$ . Since the introduced double gray edge  $(x, y)$  is parallel to a black edge, it does not cross any other intraedge (as chords). The introduced double gray edge  $(u, v)$  is either an intraedge, or an interedge. In the former case any intraedge crossing the intraedge  $(u, v)$  would necessary cross  $(x, u)$  or  $(y, v)$  (as chords), a contradiction to the fact that  $H$  has a non-crossing BG-subgraph. Hence, the BG-subgraph of  $H'$  is non-crossing. On the other hand, it is easy to see that the transformation  $H \rightarrow H'$  turns a simple black-gray connected component of the graph  $H$  into a simple black-gray connected component of  $H'$  (Fig. 6a), and a paired black-gray connected component with a single interedge into a paired black-gray connected component with a single interedge (Fig. 6b). Hence, the BG-subgraph of  $H'$  is optimal and  $H'$  is a BOG-graph.

Note that the BOG-graph  $H'$  has  $G$  as a BO-subgraph (since black and obverse edges were not affected by the transformation). The graph  $H'$  has the same gray-obverse cycles as  $H$ , except for the gray-obverse cycles  $c_1$  and  $c_2$  which are joined into a single cycle in  $H'$ . Hence, the number of gray-obverse cycles in  $H'$  is reduced as compared to  $H$ .



**Fig. 7.** Merging gray-obverse cycles  $c_1, c_2$  connected by a black edge  $(x, y)$  passing through a) intraedges  $(x, u)$  and  $(y, v)$ ; b) an intraedge  $(x, u)$  and an interedge  $(y, v)$

Iteratively reducing the number of gray-obverse cycles we will eventually come up with a BOG-graph  $G'$  having  $G$  as a BO-subgraph with a single gray-obverse cycle.  $\square$

We outline the Genome Halving Algorithm for a duplicated genome  $P$  as follows.

1. Construct a BO-graph  $G$  defined by the genome  $P$ .
2. Find a BOG-graph  $G'$  with a single gray-obverse cycle having  $G$  as a BO-subgraph (Theorem 7).
3. Read a pre-duplicated genome  $R$  along the gray-obverse cycle in  $G'$ .

## References

1. D.A.Bader, B.M.E.Moret, and M.Yan “A linear-time algorithm for computing inversion distances between signed permutations with an experimental study”. *J. Comput. Biol.*, 8 (2001), pp.483-491.
2. V.Bafna and P.A.Pevzner “Genome rearrangement and sorting by reversals”. *SIAM Journal on Computing*, 25 (1996), pp. 272-289.
3. A.Bergeron. “A very elementary presentation of the Hannenhalli-Pevzner theory”. In *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, 2089 (2001), pp. 106-117.
4. A. Bergeron, J. Mixtacki, and J. Stoye. “Reversal distance without hurdles and fortresses”. In *Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, 3109 (2004), pp. 388-399

5. F.S.Dietrich et al. "The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient *Saccharomyces cerevisiae* Genome". *Science*, 304 (2004), pp. 304-307.
6. N.El-Mabrouk, J.H.Nadeau, and D.Sankoff "Genome halving". In *Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, 1448 (1998), pp. 235-250.
7. N.El-Mabrouk and D.Sankoff "On the reconstruction of ancient doubled circular genomes using minimum reversal". *Genome Informatics*, 10 (1999), pp. 83-93.
8. N.El-Mabrouk, B.Bryant, and D.Sankoff "Reconstructing the pre-doubling genome". In *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB)* (1999), pp. 154-163.
9. N.El-Mabrouk and D.Sankoff "The Reconstruction of Doubled Genomes". *SIAM Journal on Computing*, 32 (2003), pp. 754-792.
10. S.Hannenhalli and P.Pevzner "Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)". In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing* (1995), pp. 178-189. *Journal of the ACM*, 46 (1999), pp. 1-27.
11. H.Kaplan, R.Shamir, and R.Tarjan. "Faster and simpler algorithm for sorting signed permutations by reversals". *SIAM Journal on Computing*, 29 (1999), pp. 880-892
12. M.Kellis et al. "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*". *Nature*, 428 (2004), pp. 617-624.
13. S.Ohno, U.Wolf, and N.Atkin "Evolution from fish to mammals by gene duplication". *Hereditas*, 59 (1968), pp. 169-187.
14. P.Pevzner and G.Tesler "Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes". *Genome Research*, 13 (2003), 37-45.
15. L.Skrabaneck and K.H.Wolfe "Eukaryote genome duplication - where's the evidence?". *Curr. Opin. Genet. Devel.*, 8 (1998), pp. 694-700.
16. E. Tannier and M.-F. Sagot "Sorting by reversals in subquadratic time". In *Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, 3109 (2004).
17. K.H.Wolfe and D.C.Shields "Molecular evidence for an ancient duplication of the entire yeast genome". *Nature*, 387 (1997), pp. 708-713.