Winter 2007

Colored de Bruijn Graphs and the Genome Halving Problem

Max A. Alekseyev
University of South Carolina - Columbia, maxal@gwu.edu

Pavel A. Pevzner

Publication Info

Published in *IEEE ACM Transactions on Computational Biology and Bioinformatics*, Volume 4, Issue 1, Winter 2007, pages 98-107.
http://ieeexplore.ieee.org/servlet/opac?punumber=8857

Colored de Bruijn Graphs and the Genome Halving Problem

Max A. Alekseyev and Pavel A. Pevzner

Abstract—Breakpoint graph analysis is a key algorithmic technique in studies of genome rearrangements. However, breakpoint graphs are defined only for genomes without duplicated genes, thus limiting their applications in rearrangement analysis. We discuss a connection between the breakpoint graphs and de Bruijn graphs that leads to a generalization of the notion of breakpoint graph for genomes with duplicated genes. We further use the generalized breakpoint graphs to study the Genome Halving Problem (first introduced and solved by Nadia El-Mabrouk and David Sankoff). The El-Mabrouk-Sankoff algorithm is rather complex, and, in this paper, we present an alternative approach that is based on generalized breakpoint graphs. The generalized breakpoint graphs make the El-Mabrouk-Sankoff result more transparent and promise to be useful in future studies of genome rearrangements.

Index Terms—Genome duplication, genome halving, genome rearrangement, reversal, breakpoint graph, de Bruijn graph.

1 Introduction

The Genome Halving Problem is motivated by the *whole genome duplication* events in molecular evolution [18], [25], [20], [17], [9]. 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 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 *preduplicated genome* R from the rearranged duplicated genome P (Fig. 1a).

From an 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 a *signed permutation*.

For unichromosomal genomes, the rearrangements are limited to *reversals* (also known as invertions). The reversal (i, j) over genome $x_1 x_2 \dots x_n$ "flips" genes $x_i \dots x_j$ as follows:

$$x_1 \dots x_{i-1} \quad x_i \quad x_{i+1} \dots x_j x_{j+1} \dots x_n$$
$$\longrightarrow$$
$$\downarrow$$
$$x_1 \dots x_{i-1} - x_j - x_{j-1} \dots - x_i x_{j+1} \dots x_n.$$
$$\longleftarrow$$

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

There are two natural ways to represent duplication of the circular genome R resulting in a unichromosomal genome $R \oplus R$ (Fig. 1b, left) and a multichromosomal genome $2R$ (Fig. 1b, right) but only the former is applicable to unichromosomal genomes.

For unichromosomal genomes, the *whole genome duplication* is a concatenation of the genome R with itself resulting in a *perfect duplicated genome* $R \oplus R$ (Fig. 1b, left). 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 an ancestral preduplicated 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 [10], [11], [12] culminating in a rather complex algorithm in [14]. The El-Mabrouk-Sankoff algorithm is one of the most technically challenging results in bioinformatics and its proof spans a few dozen pages in [14] (covering both unichromosomal and multichromosomal genomes). In this paper, we revisit the El-Mabrouk-Sankoff work and present an alternative approach for the case of unichromosomal genomes.[1]

The crux of our approach is a new construction that generalizes the notion of breakpoint graph for any set of

• *The authors are with the Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093-0114. E-mail: {maxal, ppevzner}@cs.ucsd.edu.*

Manuscript received 17 Sept. 2004; revised 2 June 2005; accepted 21 Nov. 2005; published online 9 Jan. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0149-0904.
Digital Object Identifier no. 10.1109/TCBB.2007.1002.

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

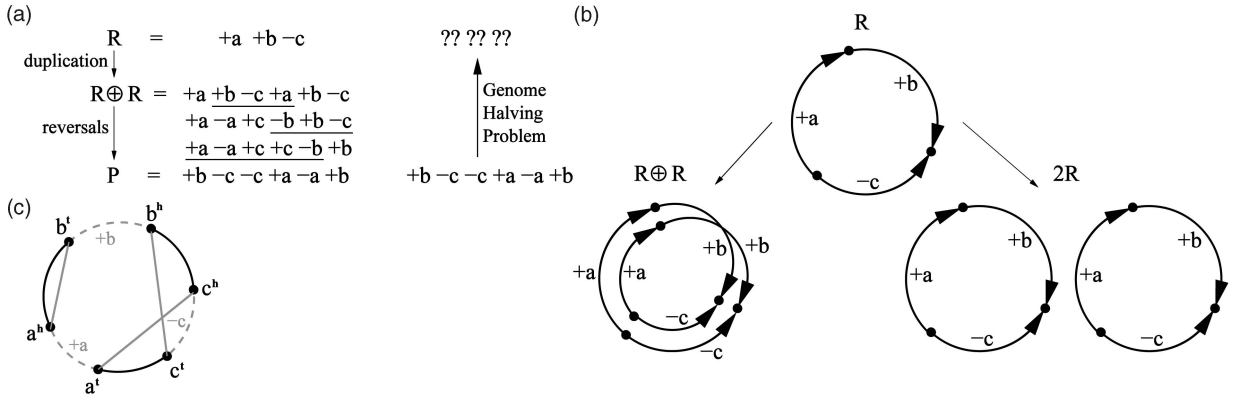

Fig. 1. (a) Whole genome duplication of genome $R = +a + b - c$ into a perfect duplicated genome $R \oplus R = +a + b - c + a + b - c$ followed by three reversals. (b) Whole genome duplication of a circular genome R (center) resulting in $R \oplus R$ (left) or $2R$ (right). (c) Breakpoint graph of genomes $+a + b - c$ and $+a + b + c$.

genomes with duplicated genes (any gene may be present in an arbitrary number of copies). This construction is related to well-known de Bruijn graphs and proved to be useful in studies of the Genome Halving Problem.

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. We present our Genome Halving Algorithm in Section 6.

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 [15]. The algorithm was further simplified and improved in a series of papers [4], [16], [2], [21] using the breakpoint graph construction introduced in [3]. Recently, Bergeron et al. [5] 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 [3] 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 with two elements x^t and x^h in the unsigned permutation.[2] 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 $\pi(P)$. 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 $\pi(P)$ and $\pi(Q)$ be corresponding unsigned permutations. The *breakpoint graph*[3] $G = G(P, Q)$ 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. 1c). Edges of each color form a matching[4] on V:

- Pairs of obverse elements form an *obverse matching*.
- Adjacent elements in $\pi(P)$, other than obverses, form a *black matching*.
- Adjacent elements in $\pi(Q)$, other than obverses, 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 $\pi(P)$ can be read along a single black-obverse cycle, while the permutation $\pi(Q)$ 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

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

where $|P| = |Q|$ is the size of the permutations P and Q, $c(G)$ is the number of black-gray cycles in the breakpoint graph G, and $\mathbf{h}(G)$ is an 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 labeled genomes P and Q into unsigned permutations $\pi(P)$ and $\pi(Q)$ by

2. Indices "t" and "h" stand for "tail" and "head," respectively.

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

4. In this paper, "matching" always means a *perfect* matching.

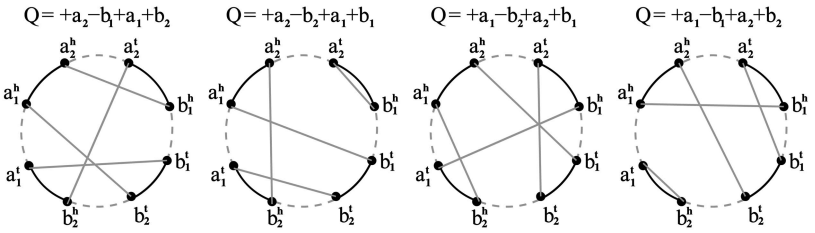

Fig. 2. Breakpoint graphs corresponding to four different labelings of $P = +a - a - b + b$ and $Q = +a - b + a + b$ (without loss of generality, we assume that labeling of $P = +a_1 - a_2 - b_1 + b_2$ is fixed). Two out of four breakpoint graphs have $c(G) = 1$, while two others have $c(G) = 2$.

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. The breakpoint graph $G(P, Q)$ of the labeled 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 $\pi(P)$ and $\pi(Q)$ (and, thus, the original genomes P and Q) as well as an intergenome correspondence between gene copies.

We remark that different labelings may lead to different breakpoint graphs for the same genomes P and Q (Fig. 2) and it is not clear how to choose a labeling that results in the minimum reversal distance between the labeled copies of P and Q.

Recently, there were many attempts to generalize the Hannenhalli-Pevzner theory for genomes with duplicated and deleted genes [6], [8], [13], [22], [23], [24]. However, the only known option for solving the reversal distance problem for duplicated genomes *exactly* is to consider all possible labelings, to compute the reversal distance problem for each labeling, and to choose the labeling with the minimal reversal distance. For duplicated genomes with n genes, this leads to 2^n invocations of the Hannenhalli-Pevzner algorithm, rendering this approach impractical. Moreover, the problem remains open if one of the genomes is perfectly duplicated (i.e., computing the reversal distance $d(P, R \oplus R)$). Surprisingly, 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 (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 labeling of gene copies such that the breakpoint graph $G(P, R \oplus R)$ of the labeled genomes P and $R \oplus R$ attains the minimum value of $|P| - c(G) + \mathbf{h}(G)$. Since $|P|$ is constant and the existing results [7] suggests that $\mathbf{h}(G)$ is typically small, the value of $d(P, Q)$ depends mostly on $c(G)$. El-Mabrouk and Sankoff [14] established that the problems of maximizing $c(G)$ and minimizing $\mathbf{h}(G)$ can be solved separately in a consecutive manner. In this paper, we focus on the former and harder problem:

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

4 Contracted Breakpoint Graph

To introduce breakpoint graphs of genomes with duplicated genes, we first revisit the notion of breakpoint graph and discuss the relationships between breakpoint graphs and de Bruijn graphs. We find it convenient to represent a circular signed permutation as an alternating cycle formed by edges of two colors with one color reserved for directed obverse edges. For example, Fig. 3 shows a black-obverse cycle representation of permutation $P = +a - a - b + b$ (Fig. 3a) and a gray-obverse cycle representation of permutation $Q = +a - b + a + b$ (Fig. 3b; the obverse edges in these cycles are labeled). Given a set of edge-labeled

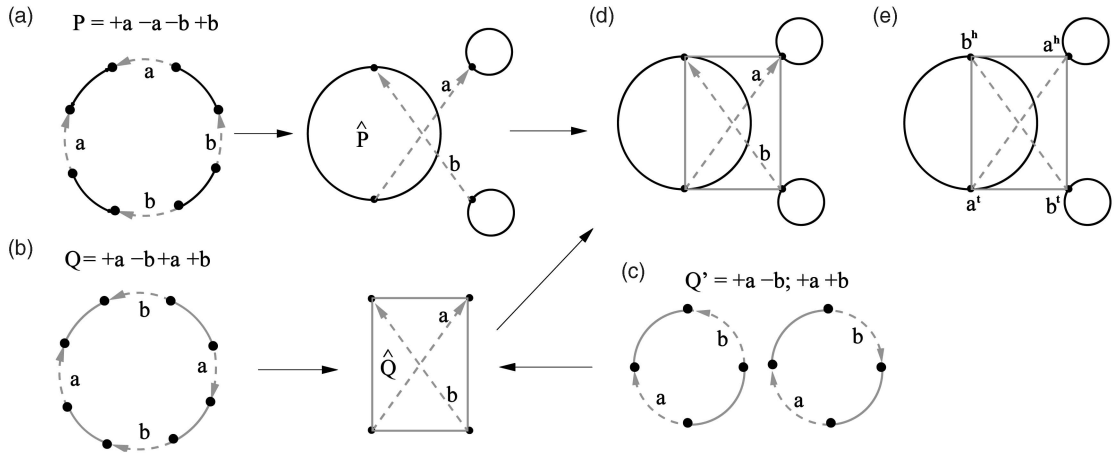

Fig. 3. (a) Genome $P = +a - a - b + b$ as a black-obverse cycle and its transformation into $\hat{P}$ by gluing identically labeled edges. (b) Genome $Q = +a - b + a + b$ as a gray-obverse cycle and its transformation into $\hat{Q}$ by gluing identically labeled edges. (c) Two-chromosomal genome $Q' = (+a - b)(+a + b)$ that is equivalent to the genome Q ($\hat{Q}' = \hat{Q}$). (d) de Bruijn graph for P and Q. (e) Contracted breakpoint graph $G'(P, Q)$.

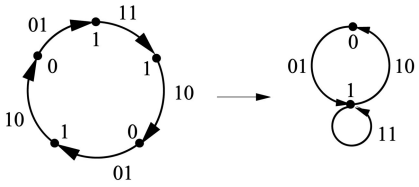

Fig. 4. de Bruijn graph $G_2(01101)$ of the circular sequence 01101.

graphs, the *de Bruijn graph* of this set is defined as the result of "gluing"[5] edges with the same label in all graphs in the set (compare with Pevzner et al. [19]). The de Bruijn graph for the two cycles in Figs. 3a and 3b is shown in Fig. 3d.

For any genome P (represented as a cycle), we define $\hat{P}$ as the graph obtained from P by gluing identically labeled edges. Obviously, the de Bruijn graph of P and Q coincides with the de Bruijn graph of $\hat{P}$ and $\hat{Q}$ (Fig. 3).

While our definition of the de Bruijn graphs is somewhat different from the usual definition, one can see that it produces the same graphs. For example, the classical de Bruijn graph $G_l(x_1 \dots x_n)$ of a circular sequence $x_1 \dots x_n$ parameterized with an integer $l \geq 2$ is defined as a graph with vertices corresponding to all $(l-1)$-tuples and edges corresponding to all l-tuples that occur in $x_1 \dots x_n$ (edge $x_i \dots x_{i+l-1}$ connects vertices $x_i \dots x_{i+l-2}$ and $x_{i+1} \dots x_{i+l-1}$). One can see that $G_l(x_1 \dots x_n)$ is identical to our construction if the circular sequence $x_1 \dots x_n$ is first represented as a cycle passing through all l-tuples in the sequence with further gluing of identically labeled edges of this cycle (Fig. 4).

While our de Bruijn graph construction is merely an equivalent definition of the breakpoint graph, it provides an important new insight. While it was not clear how to generalize the classical notion of breakpoint graph for genomes with duplicated genes, the de Bruijn graphs automatically provide such a generalization. In fact, the de Bruijn graphs are defined as a gluing operation on an *arbitrary* set of graphs and, therefore, are applicable to any set of genomes, including multichromosomal ones (each genome is represented as a set of cycles). The contracted breakpoint graph defined below is simply the de Bruijn graph of duplicated genomes.

The conventional breakpoint graph (Bafna and Pevzner [3]) of signed permutations P and Q on n elements can be defined as the gluing of n *pairs* of obverse edges in the corresponding permutations $\pi(P)$ and $\pi(Q)$ represented as black-obverse and gray-obverse alternating cycles. The *contracted breakpoint graph* of duplicated genomes P and Q on n elements is simply the gluing of n *quartets* of obverse edges. Below, we give an equivalent and a somewhat more formal definition of the 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 labeling of P and Q. The *contracted breakpoint graph* $G'(P,Q)$ is the result of contracting every pair of vertices x_1^j, x_2^j (where $x \in \mathcal{G}, j \in \{t,h\}$) in the breakpoint graph G into a single vertex x^j. 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. 3e). The contracted breakpoint graph $G'(P,Q)$ is uniquely defined by P and Q and does not depend on a particular labeling. 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 the union of black and obverse edges (*black-obverse connected*), and*
- *H is connected with respect to the union of gray and obverse edges (*gray-obverse connected*).*

Proof. Suppose that graph H is a contracted breakpoint graph of the genomes P and Q (represented as black-obverse P-cycle and gray-obverse Q-cycle). The graph H is simply the result of gluing these P-cycle and Q-cycle. Since gluing cycles cannot disconnect them, the graph H is both black-obverse and gray-obverse connected.

Consider a black-obverse and gray-obverse connected graph H where each vertex is incident to two black edges, two gray edges, and a pair of parallel obverse edges. Label endpoints of each obverse edge x in H by x^t and x^h. Since the graph H is black-obverse connected, there exists an alternating Eulerian black-obverse cycle traversing all black and obverse 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 and obverse edges that defines some duplicated genome Q. Then, the graph H is a contracted breakpoint graph for the genomes P and Q. □

In the case when Q is a 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 pairs of parallel obverse edges, the double gray edges form a matching in G' (Fig. 5a).

Let $G(P,Q)$ be a breakpoint graph for some labeling of P and Q. A set of black-gray cycles in $G(P,Q)$ is 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 labeling induces a black-gray cycle decomposition of the contracted breakpoint graph. We are interested in the reverse problem:

Labeling Problem. *Given a black-gray cycle decomposition of the contracted breakpoint graph $G'(P,Q)$ of duplicated genomes P and Q, find labeling of P and Q that induces this cycle decomposition.*

This problem does not always have a solution (Fig. 6). Below, we show how to address the Labeling Problem by considering *multichromosomal* rather than unichromosomal genomes. A multichromosomal duplicated genome is a set of circular chromosomes with every gene present in two

5. Gluing takes into account the directions of edges, i.e., tails (or heads) of all edges with a given label are glued into a single vertex.

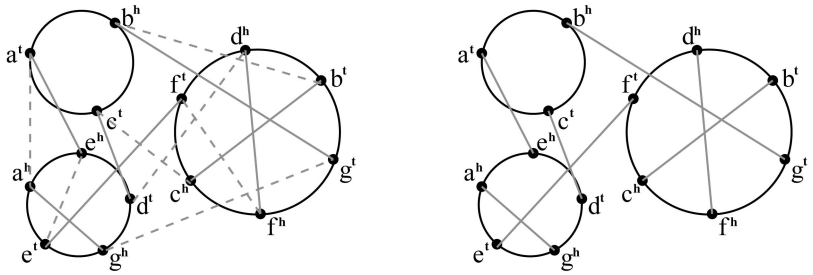

Fig. 5. (a) Contracted breakpoint graph $G'(P, R \oplus R)$ of genomes $P = -a - b + g + d + f + g + e - a + c - f - c - b - d - e$ and $R = +a - g - b - c + d - f + e$. (b) BG-graph corresponding to $G'(P, R \oplus R)$.

copies. For example, Fig. 3c presents a multichromosomal duplicated genome Q' consisting of two chromosomes $+a - b$ and $+a + b$, each of which forms a gray-obverse cycle. We remark that the de Bruijn graph of the genome Q' coincides with the de Bruijn graph of a unichromosomal genome $Q = +a - b + a + b$ (Fig. 3b) and, hence, the contracted breakpoint graphs $G'(P, Q)$ and $G'(P, Q')$ are the same for any genome P (Fig. 3d). We call genomes Q and Q' *equivalent* if their de Bruijn graphs are equal, i.e., $\hat{Q} = \hat{Q}'$.

Theorem 2. *Let P and Q be unichromosomal duplicated genomes and C be a black-gray cycle decomposition of the contracted breakpoint graph $G'(P, Q)$. Then, there exists a multichromosomal genome Q' such that 1) Q' is equivalent to Q and 2) there exists some labeling of P and Q' that induces the cycle decomposition C.*

Proof. Consider a contracted breakpoint graph $G' = G'(P, Q)$ of the genomes P and Q and its black-gray cycle decomposition C (Fig. 7a gives an example of a contracted breakpoint graph while Fig. 7c gives an example of its black-gray cycle decomposition). In order to prove the theorem, it is enough to construct a breakpoint graph $G(P, Q')$ of the genome P and some genome Q' equivalent to Q such that the black-gray cycle decomposition of G is contracted into C. Then, any labeling of the vertices of $G(P, Q')$ imposes a labeling of the genomes P and Q' that induces C.

We find it convenient to represent the cycle decomposition of G' as a graph H where every cycle from C forms its own connected component (Fig. 7c) and assume that every vertex of the graph G' has two copies in H with identical labels (i.e., graph H has twice the number of vertices as compared to G'). To transform the graph H into a breakpoint graph $G(P, Q')$, we will add obverse edges to H as described below.

The genome P defines a black-obverse cycle (Fig. 7d). Traversing black edges in graph H in the order given by this cycle defines a set of obverse edges in H (Fig. 7e). These obverse edges form a matching in H and define a gray-obverse cycle decomposition. We define the genome Q' as a collection of chromosomes corresponding to these gray-obverse cycles so that the graph H, together with the set of obverse edges, represents the breakpoint graph $G(P, Q')$. By the construction, we have $G'(P, Q') = G'(P, Q)$, meaning that the genomes Q' and Q are equivalent. □

Lemma 1. *The perfect duplicated (unichromosomal) genome $R \oplus R$ is equivalent to the two-chromosomal genome $2R$. Moreover, $2R$ is the only genome equivalent to $R \oplus R$.*

Proof. It is easy to see that the gluing of both $R \oplus R$ and $2R$, represented as gray-obverse cycles, results in a single gray-obverse cycle c that traverses R in order (every edge in this cycle has multiplicity 2) (Fig. 8). Any other genome that is glued into c cannot have a cycle shorter than c since such a short cycle would remain short after gluing. This implies that every genome that is glued into c either traverses c twice ($R \oplus R$) or is formed by two cycles each of which traverses c once ($2R$). □

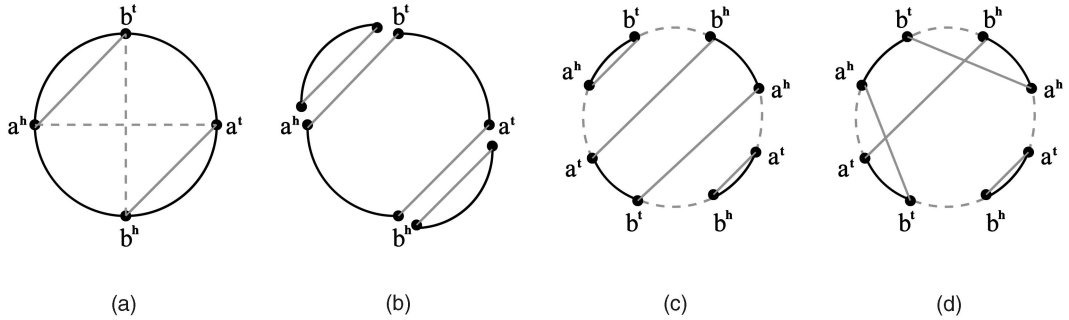

Fig. 6. (a) Contracted breakpoint graph $G'(P, R \oplus R)$ for $P = +a + b - a - b$ and $R = +a + b$. (b) Black-gray cycle decomposition C of G', which is not induced by any labeling of P and $R \oplus R$. (c) Breakpoint graph $G(P, 2R)$ inducing C. (d) Breakpoint graph $G(P, R \oplus R)$ (unique up to relabeling of vertices with $c(G) = 2 < |C| = 3$.

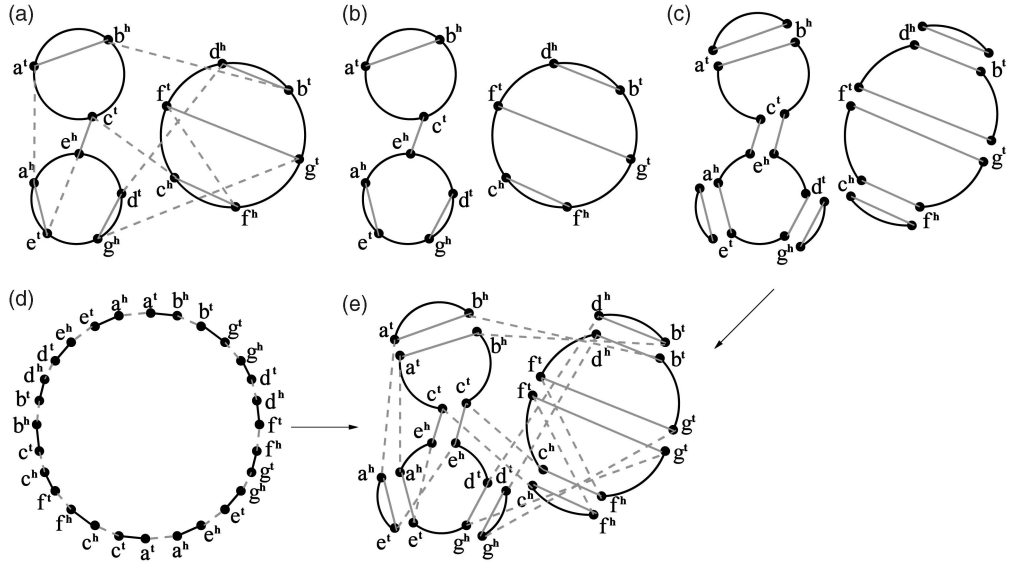

Fig. 7. For genomes $P = -a - b + g + d + f + g + e - a + c - f - c - b - d - e$ and $R = -a - b - d - g + f - c - e$, (a) contracted breakpoint graph $G'(P, R \oplus R)$. (b) BG-graph corresponding to G'. (c) Maximal black-gray cycle decomposition C of G' forming graph H. (d) Genome P as a black-obverse cycle. (e) Breakpoint graph inducing the cycle decomposition C. Note that this figure can be viewed as a superimposition of (c) and (d).

Theorem 2 and Lemma 1 imply:

Theorem 3. *Let P and $R \oplus R$ be unichromosomal duplicated genomes and C be a black-gray cycle decomposition of the contracted breakpoint graph $G'(P, R \oplus R)$. Then, there exists some labeling of either $R \oplus R$ or $2R$ that induces the cycle decomposition C.*

Theorem 3 reveals the connection between the Weak Genome Halving Problem and maximal cycle decomposition and breaks the analysis of the Weak Genome Halving Problem into two cases depending on whether the maximal cycle decomposition is induced by $R \oplus R$ or $2R$. In this paper, we focus on the $R \oplus R$ case while the $2R$ case is analyzed in [1].

Let $c_{max}(G')$ be the number of cycles in a maximal black-gray cycle decomposition of the contracted breakpoint graph $G' = G'(P, R \oplus R)$. The $R \oplus R$ case of Theorem 3 motivates the following reformulation of the Weak Genome Halving Problem.

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 bicolored 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. 5b). 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 a cycle decomposition of the BG-graph G, and we formulate necessary and sufficient conditions for achieving this bound.

A BG-graph is *connected* if it is connected with respect to the union of 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

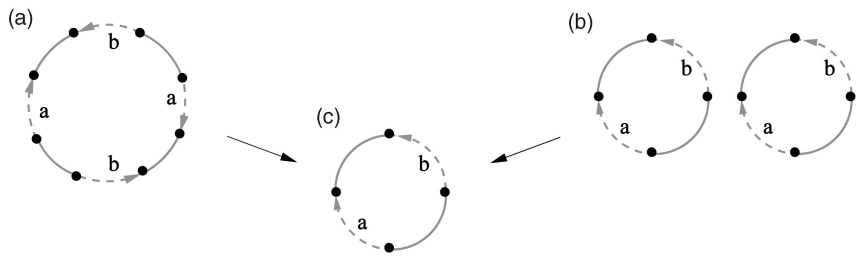

Fig. 8. For genome $R = +a - b$ (a) perfect duplicated genome $Q = R \oplus R$ as a gray-obverse cycle. (b) Two-chromosomal genome $Q' = 2R$ as gray-obverse cycles. (c) de Bruijn graph $\hat{Q} = \hat{Q}'$, where every edge has multiplicity 2.

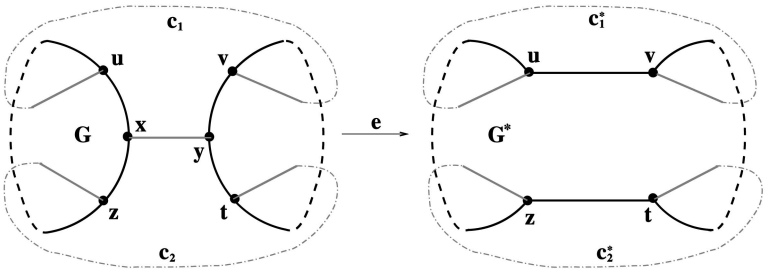

Fig. 9. *e*-transformation of a graph G into a graph $G^\star$. Black-gray cycles c_1, c_2 in G passing through interedge $e = (x, y)$ are transformed into black-gray cycles $c_1^\star, c_2^\star$ in $G^\star$. The black-gray cycle c_1 traverses edges $(u, x), (x, y), (y, v)$ that are replaced by a single edge (u, v), while black-gray cycle c_2 traverses edges $(z, x), (x, y), (y, t)$ that are replaced by a single edge (z, t). As a result, the black cycles connected by e in G are merged into a single black cycle in $G^\star$.

$e = (x, y)$ be an interedge in G. We define an *e-transformation* $(G, C) \xrightarrow{e} (G^\star, C^\star)$ of the graph G and its black-gray cycle decomposition C into a new BG-graph $G^\star$ on $2(n-1)$ vertices with $m-1$ black cycles and a black-gray cycle decomposition $C^\star$ of $G^\star$ of the same size as C (Fig. 9). 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 $G^\star$ from G, we replace these triples of edges with single black edges (u, v) and (z, t), respectively, and delete vertices x and y. This operation transforms the cycles c_1 and c_2 in G into cycles $c_1^\star$ and $c_2^\star$ in $G^\star$. We define the black-gray cycle decomposition $C^\star$ as C with the cycles c_1 and c_2 replaced with $c_1^\star$ and $c_2^\star$.

Lemma 2. *Let C be a maximal black-gray cycle decomposition of a BG-graph G and $(G, C) \xrightarrow{e} (G^\star, C^\star)$ be the e-transformation for some interedge $e = (x, y)$ in G. Then, $c_{max}(G) = c_{max}(G^\star)$.*

Proof. It follows from the definition of *e*-transformation that $c_{max}(G) = |C| = |C^\star| \le c_{max}(G^\star)$. On the other hand, every black-gray cycle decomposition $D^\star$ of the graph $G^\star$ 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 $D^\star$ by black-gray-black triples $(u, x), (x, y), (y, v)$ and $(z, x), (x, y), (y, t)$). Therefore, $c_{max}(G^\star) \le c_{max}(G)$. □

Theorem 4. *If G is a connected BG-graph with $2n$ vertices and m black cycles, then*

$$c_{max}(G) \le n + 2 - m = |P|/2 + 2 - m.$$

Proof. Suppose that $c_{max}(G) = k$, i.e., a maximal cycle decomposition of G contains k black-gray cycles. We find it convenient to view these cycles as k disconnected cycles (i.e., every cycle forms its own connected component) that are later contracted in the BG-graph G 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, $n \ge k - 1$. 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 $(G, C) \xrightarrow{e} (G^\star, C^\star)$. Lemma 2 implies $c_{max}(G) = c_{max}(G^\star)$. Note that $G^\star$ 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, $c_{max}(G) = c_{max}(G^+) \le n + 2 - m$. □

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 equals $|P|$, Theorem 4 implies

$$\begin{aligned} c_{max}(G) &= \sum_H c_{max}(H) \le \sum_H |H|/2 + 2 - m_H \\ &= |P|/2 + \textstyle\sum_{m=1}^{\infty} (2-m) \cdot s_m \le |P|/2 + s_1, \end{aligned}$$

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. Note that since gray edges form a matching in BG-graph, a single odd cycle cannot form a connected component. Therefore, s_1 does not exceed $b_e(G)$:

$$c_{max}(G) \le |P|/2 + b_e(G). \tag{2}$$

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. 7b 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 *noncrossing* if its intraedges (as chords within each black circle) do not cross. The BG-graph in Fig. 7b is noncrossing, while the BG-graph in Fig. 5b is not.

Theorem 5. *For a simple BG-graph G on $2n$ vertices, $c_{max}(G) = n + 1$ if and only if G is noncrossing.*

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 $2n$.

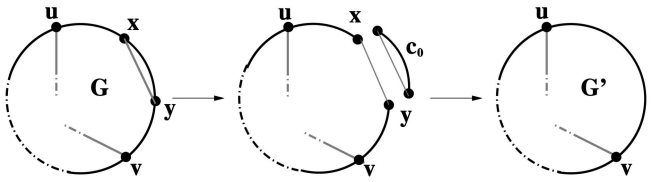

Fig. 10. Transformation of a BG-graph G into a BG-graph G' by splitting a black-gray cycle c_0 consisting of parallel black and gray edges (x, y).

We first prove (the 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 maximal 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 noncrossing, consider a double gray edge e spanning (as a chord) the minimum number of black edges. 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 the 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. 10). Note that $c_{max}(G') = c_{max}(G) - 1$ and G' is noncrossing if and only if G is noncrossing.

By induction, the graph G' is noncrossing if and only if $c_{max}(G') = n$. Therefore, G is noncrossing if and only if $c_{max}(G') = n + 1$. □

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 $(G, C) \xrightarrow{e} (G^\star, C^\star)$ be an e-transformation of G. Note that the graph $G^\star$ is a simple BG-graph on $2(n - 1)$ vertices. Lemma 2 and Theorem 4 imply $c_{max}(G) = c_{max}(G^\star) \leq n$. Therefore, according to Theorem 5, $c_{max}(G) = n$ if and only if the BG-graph $G^\star$ is noncrossing. We are interested in a particular case of this statement.

Theorem 6. *For a paired BG-graph G of size $2n$ with a single interedge, $c_{max}(G) = n$ if and only if G is noncrossing.*

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 noncrossing BG-graph if and only if G is noncrossing. □

We call a BG-graph *optimal* if its connected components are either simple BG-graphs or paired BG-graphs with single interedges. Theorems 5 and 6 imply

Theorem 7. *For an optimal BG-graph G,*

$$c_{max}(G) = |P|/2 + b_e(G).$$

An optimal BG-graph and its maximal cycle decomposition are shown in Figs. 7b and 7c.

6 Genome Halving Algorithm

In order to solve the Cycle Decomposition Problem for a given genome P, we will construct a contracted breakpoint graph $G'(P, R \oplus R)$ which achieves the upper bound (2). The de Bruijn graph $\hat{P}$, being a subgraph of $G'(P, R \oplus R)$ (for any preduplicated genome R), completely defines a vertex set, an obverse matching, and a set of black cycles in G' (Figs. 3a and 3d). We will show how to complete the graph $\hat{P}$ with a set of double gray edges to obtain a contracted breakpoint $G'(P, R \oplus R)$ with the maximum value of $c_{max}(G')$.

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 $\hat{P}$). A *BOG-graph* is a graph with black, obverse, and gray edges where 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 noncrossing BG-graph or a paired noncrossing BG-graph with a single interedge.

The arguments above suggest that the Cycle Decomposition Problem for a genome P can be reformulated as follows: For a given BO-graph G (defined as $G = \hat{P}$), find a gray-obverse connected BOG-graph G' having G as a BO-subgraph. Theorems 1 and 7 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 3. *For a BOG-graph with more than one gray-obverse cycle, there exists a black edge connecting two different gray-obverse cycles.*

Proof. Let H be a BOG-graph with two or more gray-obverse cycles. Since H is black-obverse connected, there exists a black-obverse cycle in H traversing all obverse edges of H. Therefore, there exists a black edge connecting obverse edges from different gray-obverse cycles in H. □

Theorem 8. *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 arbitrarily) and introduce an arbitrary interedge connecting cycles in each pair. Then, we complete each black cycle with an arbitrary noncrossing 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

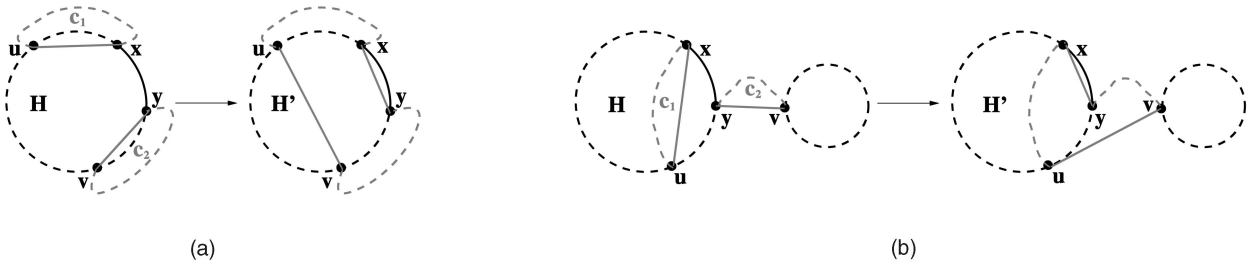

Fig. 11. Merging gray-obverse cycles c_1, c_2 connected by a black edge (x, y) passing through (a) intraedges (x, u) and (y, v) and (b) an intraedge (x, u) and an interedge (y, v).

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 3, 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. 11 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 necessarily cross (x, u) or (y, v) (as chords), a contradiction of the fact that H has a noncrossing BG-subgraph. Hence, the BG-subgraph of H' is noncrossing. 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. 11a), and a paired black-gray connected component with a single interedge into a paired black-gray connected component with a single interedge (Fig. 11b). 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.

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. □

We outline the Genome Halving Algorithm for a duplicated genome P as follows (the $R \oplus R$ case):

1. Construct a BO-graph $G = \hat{P}$.
2. Find a BOG-graph G' with a single gray-obverse cycle having G as a BO-subgraph (Theorem 8).
3. Read a preduplicated genome R along the gray-obverse cycle in G'.
4. Find a maximal black-gray cycle decomposition of G' (Theorems 5 and 6) and a labeling of the genomes P and Q ($Q = R \oplus R$ or $Q = 2R$) inducing this cycle decomposition (Theorem 3).
5. If $Q = R \oplus R$, then we are done.

The case $Q = 2R$ is addressed in [1].

To estimate the complexity of the Genome Halving Algorithm, we assume that every graph is implemented as a collection of sets: a set of vertices, sets of edges of each color, and an array of sets of incident edges indexed by vertices and colors. Note that for a given genome P with n genes, all graphs appearing in the algorithm have vertex and edge sets of order $O(n)$, while every set of incident edges contains at most two elements. Therefore, even with a straightforward data structure, each set operation (such as an insertion/deletion of an element or a membership query) takes $O(n)$ time.

One can demonstrate that every step of the Genome Halving Algorithm can be done in $O(n)$ set operations. Therefore, the overall time complexity of the Genome Halving Algorithm can be estimated as $O(n^2)$. In practice, our implementation of the Genome Halving Algorithm takes less than a second to halve a "random" duplicated genome with 1,000 unique genes with a standard Intel PIII 900 MHz CPU.

REFERENCES

[1] M.A. Alekseyev and P.A. Pevzner, "Whole Genome Duplications and Contracted Breakpoint Graphs," *SIAM J. Computing*, 2007 (in press).

[2] 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. Computer Biology*, vol. 8, pp. 483-491, 2001.

[3] V. Bafna and P.A. Pevzner, "Genome Rearrangement and Sorting by Reversals," *SIAM J. Computing*, vol. 25, pp. 272-289, 1996.

[4] A. Bergeron, "A Very Elementary Presentation of the Hannenhalli-Pevzner Theory," *Lecture Notes in Computer Science*, vol. 2089, pp. 106-117, 2001.

[5] A. Bergeron, J. Mixtacki, and J. Stoye, "Reversal Distance without Hurdles and Fortresses," *Lecture Notes in Computer Science*, vol. 3109, pp. 388-399, 2004.

[6] G. Bourque, Y. Yacef, and N. El-Mabrouk, "Maximizing Synteny Blocks to Identify Ancestral Homologs," *Lecture Notes in Bioinformatics*, vol. 3678, pp. 21-34, 2005.

[7] A. Caprara, "On the Tightness of the Alternating-Cycle Lower Bound for Sorting by Reversals," *J. Combinatorial Optimization*, vol. 3, pp. 149-182, 1999.

[8] X. Chen et al., "Computing the Assignment of Orthologous Genes via Genome Rearrangement," *Proc. Third Asia-Pacific Bioinformatics Conf.*, pp. 363-378, 2005.
[9] F.S. Dietrich et al., "The *Ashbya gossypii* Genome as a Tool for Mapping the Ancient *Saccharomyces cerevisiae* Genome," *Science*, vol. 304, pp. 304-307, 2004.
[10] N. El-Mabrouk, J.H. Nadeau, and D. Sankoff, "Genome Halving," *Lecture Notes in Computer Science*, vol. 1448, pp. 235-250, 1998.
[11] N. El-Mabrouk and D. Sankoff, "On the Reconstruction of Ancient Doubled Circular Genomes Using Minimum Reversal," *Genome Informatics*, vol. 10, pp. 83-93, 1999.
[12] N. El-Mabrouk, B. Bryant, and D. Sankoff, "Reconstructing the Pre-Doubling Genome," *Proc. Third Ann. Int'l Conf. Computational Molecular Biology (RECOMB)*, pp. 154-163, 1999.
[13] N. El-Mabrouk, "Genome Rearrangement by Reversals and Insertions/Deletions of Contiguous Segments," *Lecture Notes in Computer Science*, vol. 1848, pp. 222-234, 2000.
[14] N. El-Mabrouk and D. Sankoff, "The Reconstruction of Doubled Genomes," *SIAM J. Computing*, vol. 32, pp. 754-792, 2003.
[15] S. Hannenhalli and P. Pevzner, "Transforming Cabbage into Turnip (Polynomial Algorithm for Sorting Signed Permutations by Reversals)," *J. ACM*, vol. 46, pp. 1-27 1999
[16] H. Kaplan, R. Shamir, and R. Tarjan, "Faster and Simpler Algorithm for Sorting Signed Permutations by Reversals," *SIAM J. Computing*, vol. 29, pp. 880-892, 1999.
[17] M. Kellis et al., "Proof and Evolutionary Analysis of Ancient Genome Duplication in the Yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, pp. 617-624, 2004.
[18] S. Ohno, U. Wolf, and N. Atkin, "Evolution from Fish to Mammals by Gene Duplication," *Hereditas*, vol. 59, pp. 169-187, 1968.
[19] P. Pevzner, H. Tang, and G. Tesler, "De Novo Repeat Classification and Fragment Assembly," *Genome Research*, vol. 14, 1786-1796, 2004.
[20] L. Skrabanek and K.H. Wolfe, "Eukaryote Genome Duplicatiion—Where's the Evidence?" *Current Opinions in Genetic Development*, vol. 8, pp. 694-700, 1998.
[21] E. Tannier and M.-F. Sagot, "Sorting by Reversals in Subquadratic Time," *Lecture Notes in Computer Science*, vol. 3109, 2004.
[22] D. Sankoff, "Genome Rearrangement with Gene Families," *Bioinformatics*, vol. 15, pp. 909-917, 1999.
[23] K.M. Swenson, M. Marron, J.V. Earnest-DeYoung, and B.M.E. Moret, "Approximating the True Evolutionary Distance between Two Genomes," *Proc. Seventh Workshop Algorithm Eng. and Experiments*, pp. 121-129, 2005.
[24] K.M. Swenson, N.D. Pattengale, and B.M. E. Moret, "A Framework for Orthology Assignment from Gene Rearrangement Data," *Lecture Notes in Bioinformatics*, vol. 3678, pp. 153-166, 2005.
[25] K.H. Wolfe and D.C. Shields, "Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome," *Nature*, vol. 387, pp. 708-713, 1997.

Max A. Alekseyev received the MS degree in mathematics from Nizhni Novgorod State University in 1999. He is pursuing the PhD degree in computer science at University of California San Diego. His current research interests include genome rearrangements and gene duplications analysis.

Pavel A. Pevzner received the PhD degree in 1988 from the Moscow Institute of Physics and Technology. He holds the Ronald R. Taylor Chair in Computer Science at University of California San Diego. His current research interests include genome rearrangements, repeat analysis, and computational proteomics.