2007

Whole Genome Duplications and Contracted Breakpoint Graphs

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

Pavel A. Pevzner

Publication Info

Published in *SIAM Journal on Computing*, ed. Madhu Sudan, Volume 36, Issue 6, 2007, pages 1748-1763.

SIAM J. COMPUT.
Vol. 36, No. 6, pp. 1748–1763

WHOLE GENOME DUPLICATIONS AND CONTRACTED BREAKPOINT GRAPHS*

MAX A. ALEKSEYEV† AND PAVEL A. PEVZNER†

Abstract. The genome halving problem, motivated by the whole genome duplication events in molecular evolution, was solved by El-Mabrouk and Sankoff in the pioneering paper [*SIAM J. Comput.*, 32 (2003), pp. 754–792]. The El-Mabrouk–Sankoff algorithm is rather complex, inspiring a quest for a simpler solution. An alternative approach to the genome halving problem based on the notion of the contracted breakpoint graph was recently proposed in [M. A. Alekseyev and P. A. Pevzner, *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 4 (2007), pp. 98–107]. This new technique reveals that while the El-Mabrouk–Sankoff result is correct in most cases, it does not hold in the case of unichromosomal genomes. This raises a problem of correcting a flaw in the El-Mabrouk–Sankoff analysis and devising an algorithm that deals adequately with all genomes. In this paper we efficiently classify all genomes into two classes and show that while the El-Mabrouk–Sankoff theorem holds for the first class, it is incorrect for the second class. The crux of our analysis is a new combinatorial invariant defined on duplicated permutations. Using this invariant we were able to come up with a full proof of the genome halving theorem and a polynomial algorithm for the genome halving problem.

Key words. genome duplication, genome halving, genome rearrangement, breakpoint graph, de Bruijn graph

AMS subject classifications. 90C27, 90C35, 90C46, 94C15

DOI. 10.1137/05064727X

1. Introduction.

In 1970 Susumu Ohno came up with two fundamental theories of chromosome evolution that were the subjects of many controversies in the last 35 years [31]. The first, random breakage theory, was embraced by biologists from the very beginning but was refuted by Pevzner and Tesler in 2003 [35] and Murphy et al. in 2005 [30]. The second, whole genome duplication theory, postulated a new type of evolutionary event and had a very different fate. It was subject to controversy in the first 35 years and only recently was proven to be correct [27, 15]. Kellis, Birren, and Lander in 2004 [27] sequenced the yeast *K. waltii* genome, compared it with the yeast *S. cerevisiae* genome, and demonstrated that nearly every region in *K. waltii* corresponds to two regions in *S. cerevisiae*, thus proving that there was a whole genome duplication event in the course of yeast evolution. This discovery was quickly followed by the discovery of whole genome duplications in vertebrates [24, 36, 12] and plants [21]. Finally, in September, 2005 Dehal and Boore [14] found evidence of two rounds of whole genome duplications on the evolutionary path from early vertebrates to humans. Shortly afterwards, Meyer and Van de Peer [28] found evidence of yet another (third) round of whole genome duplications in ray-finned fishes, thus implying that nearly every human gene could have existed in as many as eight copies at different stages of evolution.

These recent studies provided irrefutable evidence that whole genome duplications represent a new type of event that may explain phenomena which classical evolutionary studies have had difficulty explaining (e.g., emergence of new metabolic

*Received by the editors December 12, 2005; accepted for publication (in revised form) October 19, 2006; published electronically March 19, 2007.
http://www.siam.org/journals/sicomp/36-6/64727.html
†Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093 (maxal@cs.ucsd.edu, ppevzner@cs.ucsd.edu).

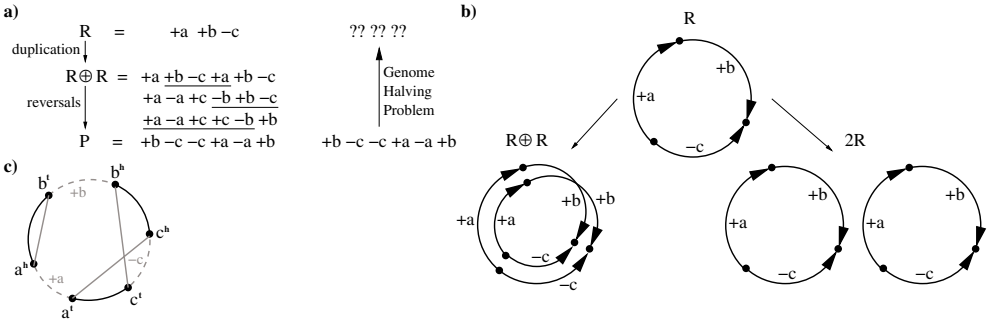

FIG. 1. (a) *Whole genome duplication of genome* $R = +a + b - c$ *into a perfect duplicated genome* $R \oplus R = +a + b - c + a + b - c$ *followed by three reversals.* (b) *Whole genome duplication of a circular genome* R *(center) resulting in* $R \oplus R$ *(left) or* $2R$ *(right).* (c) *Breakpoint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

pathways [27]). At the same time, they raised the problem of reconstructing the genomic architecture of the ancestral preduplicated genomes. Unfortunately, since the El-Mabrouk–Sankoff algorithm for solving this problem [20] has not yet resulted in a software tool, the recent studies of whole genome duplications did not attempt to rigorously reconstruct the architecture of the preduplicated genomes. We revisited the El-Mabrouk–Sankoff result, found a flaw in their approach, reformulated and proved the genome halving theorem, and developed a new algorithm and software tool for studies of genome duplications.

Whole genome duplications 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 (Figure 1(a)).

From an algorithmic perspective, the genome is a collection of chromosomes, and each chromosome is a *signed sequence* over a finite alphabet. 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 "-". In this paper we focus on the basic *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, is referred to as *signed permutation*. For unichromosomal genomes the rearrangements are limited to *reversals* that "flip" genes $x_i \dots x_j$ in a genome $x_1 x_2 \dots x_n$ as follows:

$$x_1 \dots x_{i-1} \underrightarrow{x_i \quad x_{i+1} \dots \quad x_j} x_{j+1} \dots x_n \longrightarrow x_1 \dots x_{i-1} \underleftarrow{-x_j - x_{j-1} \cdots - x_i} x_{j+1} \dots x_n.$$

The *reversal distance* between two genomes is defined as the minimum number of reversals required to transform one genome into the other (see Chapter 10 of [34] for a review of genome rearrangement algorithms).

We represent a circular genome R as a cycle formed by directed edges encoding the genes and their directions (Figure 2(b), center). There are two natural ways to represent duplication of the genome R resulting in a unichromosomal genome $R \oplus R$ (Figure 1(b), left) and a multichromosomal genome $2R$ (Figure 1(b), right) but only the former is applicable to unichromosomal genomes.

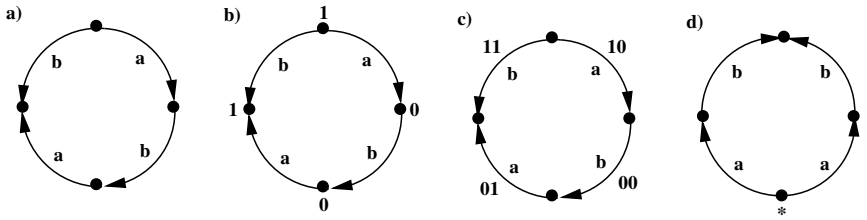

FIG. 2. (a) *Circular genome* $P = +a - b + a + b$ *represented as a cycle with directed edges.* (b) 01*-labeling of the vertices of the cycle defined by* P. (c) *Induced labeling of the genes of* P *that is consistent.* (d) *For some genomes consistent labelings do not exist: for genome* $Q = +a + b - b - a$ *the labels of both copies of gene* a *start with the same digit ("*$*$*") so they cannot be inversions of each other.*

For unichromosomal genomes, *whole genome duplication* is a concatenation of the genome R with itself, resulting in a *perfect duplicated genome* $R \oplus R$. The genome $R \oplus R$ becomes subject to reversals that change the order and signs of the genes and transform $R \oplus R$ 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* $d(P, R \oplus R)$ *from the perfect duplicated genome* $R \oplus R$ *to* P.

The genome halving problem was solved in a series of papers [18, 19, 17] culminating in a rather complex algorithm by El-Mabrouk and Sankoff in [20]. The El-Mabrouk–Sankoff algorithm is one of the most technically challenging results in computational biology and its proof spans over 30 pages in [20]. Recently Alekseyev and Pevzner [1] revisited the El-Mabrouk–Sankoff work and presented an alternative approach based on the notion of a *contracted breakpoint graph*.

After paper [1] was submitted, our studies of the contracted breakpoint graph led us to realize that the El-Mabrouk–Sankoff analysis has a flaw and that the problem of finding $\min_R d(P, R \oplus R)$ remains unsolved in the simplest case when P is a unichromosomal genome. Below we show that this flaw is a rule rather than a pathological case: it affects a large family of duplicated genomes. We further proceed to give a full analysis of the genome halving problem that is based on introducing an invariant that divides the set of all rearranged duplicated genomes into two classes. We show that the El-Mabrouk–Sankoff formula is correct for the first class but is off by 1 for the second class. We remark that our approach is very different from [20] and we do not know whether the technique in [20] can be adjusted to address the described complication.

To introduce a new combinatorial invariant of duplicated genomes, consider labelings of vertices in the cycle defined by the duplicated rearranged genome P with numbers 0 and 1 (Figure 2(b)). Every such labeling induces a two-digit labeling of the genes (edges): a label of each gene is formed by the labels of the incident vertices (Figure 2(c)). A 01-labeling of the vertices is called *consistent* if for every pair of identical genes in P the label of one copy is an inversion of the other. If there exists a consistent labeling of genome P, we define the *parity index* of P as the number of genes labeled "01" modulo 2. Below we prove that the parity index is well defined, i.e., the parity index is the same for all consistent labelings of a genome. It turns out that the El-Mabrouk–Sankoff theorem fails on genomes with the parity index 0.

The paper is organized as follows. Section 2 presents the concept of contracted

breakpoint graph and reviews some results from [1]. Section 3 describes a flaw in the El-Mabrouk–Sankoff analysis. Section 4 presents a solution to the genome halving problem (for unichromosomal circular genomes) and a classification of the genomes for which the original El-Mabrouk–Sankoff theorem is incorrect. Section 5 outlines our genome halving algorithm. Finally, section 6 discusses potential biological applications of the presented algorithm.

2. Reversal distance between duplicated genomes and contracted breakpoint graphs. A duality theorem and a polynomial algorithm for computing reversal distance between two signed permutations was proposed by Hannenhalli and Pevzner [23] and later was generalized for multichromosomal genomes [22]. The algorithm was further simplified and improved in a series of papers [6, 25, 3, 7, 42, 26] and applied in a variety of biological studies [29, 10, 8, 33, 5].

A signed permutation on n elements can be transformed into an unsigned permutation on $2n$ elements (see [4]) by substituting every element x in the signed permutation with two elements x^t and x^h in the unsigned permutation (indices t and h stand for tail and head, respectively). 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 the *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 let $\pi(P)$ and $\pi(Q)$ be corresponding unsigned permutations. The *breakpoint graph* $G = G(P, Q)$ 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 (Figure 1(c)). Edges of each color form a matching on V as follows:

- 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*, 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 play an important role in computing the reversal distance. According to the Hannenhalli–Pevzner theorem, the reversal distance between permutations P and Q is given by the formula

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

where $|P| = |Q|$ is the size of P and Q, $c(P, Q) = c(G(P, Q))$ is the number of black-gray cycles in the breakpoint graph G, and $h(G)$ is an easily computable combinatorial parameter. While this result 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 $\mathcal{G}$. If one labels copies of each gene x as x_1 and x_2, then the genomes P and Q become signed permutations and (1) applies. 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 $\pi(Q)$. We remark that different labelings may lead to different breakpoint graphs (on the same vertex set) for the same genomes P and Q (Figure 3) 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. We also remark that pairs of vertices x_1^j and x_2^j form yet another

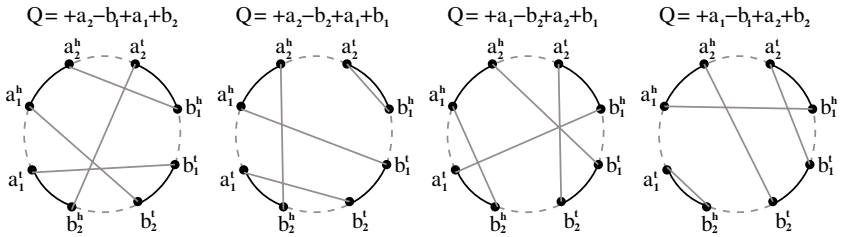

FIG. 3. *Breakpoint graphs for $P = +a - a - b + b$ and four different labelings $Q = +a - b + a + b$ (we assume that the labeling of $P = +a_1 - a_2 - b_1 + b_2$ is fixed). Two out of four breakpoint graphs have $c(G) = 1$, while two others have $c(G) = 2$. The counterpart matching in these graphs is formed by pairs (a_1^t, a_2^t), (a_1^h, a_2^h), (b_1^t, b_2^t), (b_1^h, b_2^h).*

matching in the breakpoint graph G called *counterpart*. Counterpart of a vertex v is denoted $\bar{v}$ so that $\bar{x}_1^j = x_2^j$ and $\bar{x}_2^j = x_1^j$ (see legend for Figure 3).

Recently there were many attempts to generalize the Hannenhalli–Pevzner theory for genomes with duplicated and deleted genes [9, 11, 16, 37, 40, 41]. 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 $d(P, R \oplus R)$). 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 formula (1), the genome halving problem can be posed as follows. For a 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 $|P| - c(P, Q) + h(P, Q)$. Since $|P|$ is constant and $h(G)$ is typically small (see [34]), the value of $d(P, Q)$ depends mainly on $c(P, Q)$. El-Mabrouk and Sankoff [20] established that the problems of maximizing $c(P, Q)$ and minimizing $h(P, Q)$ can be solved separately in a consecutive manner.[1] In this paper we focus on the former and harder problem as follows.

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 $c(P, R \oplus R)$ in the breakpoint graph $G(P, R \oplus R)$ of the labeled genomes P and $R \oplus R$.*

From now on, we will find it convenient to represent a circular signed permutation as an alternating cycle formed by edges of two colors with one color reserved for obverse edges. For example, Figures 4(a),(b) show a black-obverse cycle representation of permutation $P = +a - a - b + b$ and a gray-obverse cycle representation of permutation $Q = +a - b + a + b$ (the obverse edges in these cycles are labeled and directed). Given a set of edge-labeled graphs, the *de Bruijn graph* of this set is defined as the result of "gluing"[2] edges with the same label in all graphs in the set (compare with Pevzner,

[1] In paper [2] we describe an analogue of formula (1) without the "$h(G)$" term and give a short proof of the genome halving theorem (for multichromosomal genomes) that does not rely on the analysis in [20].

[2] 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.

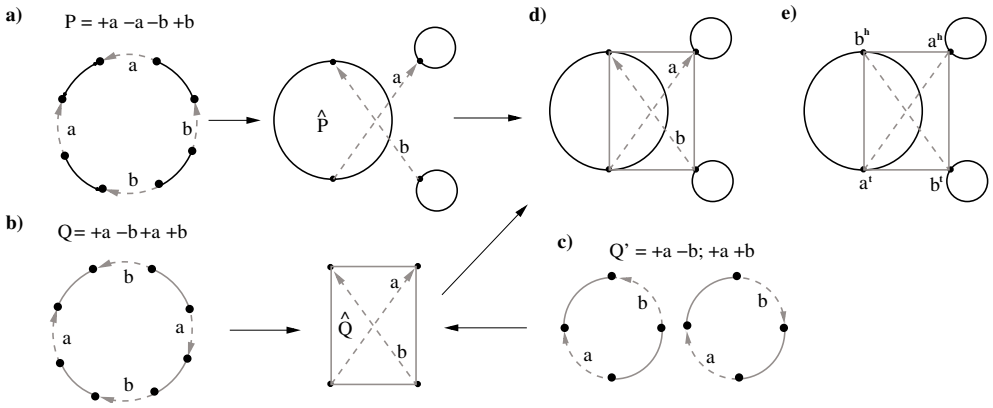

FIG. 4. (a) *Genome $P = +a - a - b + b$ as a black-obverse cycle and its transformation into $\hat{P}$ by gluing identically labeled edges.* (b) *Genome $Q = +a - b + a + b$ as a gray-obverse cycle and its transformation into $\hat{Q}$ by gluing identically labeled edges.* (c) *Two-chromosomal genome $Q' = (+a - b)(+a + b)$ that is equivalent to the genome Q (i.e., $\hat{Q}' = \hat{Q}$).* (d) *The de Bruijn graph of genomes P and Q.* (e) *The contracted breakpoint graph $G'(P, Q)$.*

Tang, and Tesler [32]). The de Bruijn graph for two cycles in Figures 4(a),(b) is shown in Figure 4(d).

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 $\hat{Q}$ (Figure 4). For a duplicated genome P, the black edges of $\hat{P}$ form a set of vertex-disjoint black cycles. We denote by $b_e(P)$ the number of even black cycles in $\hat{P}$.

The conventional breakpoint graph [4] of signed permutations P and Q on n elements can be defined as the gluing of n *pairs* of identically labeled obverse edges in the corresponding permutations (represented as black-obverse and gray-obverse 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 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 let 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 (Figure 4(e)). The contracted breakpoint graph $G'(P, Q)$ does not depend on a particular labeling of P and Q. The following theorem gives a characterization of the contracted breakpoint graphs (for unichromosomal genomes).

THEOREM 1 (see [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 two 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).*

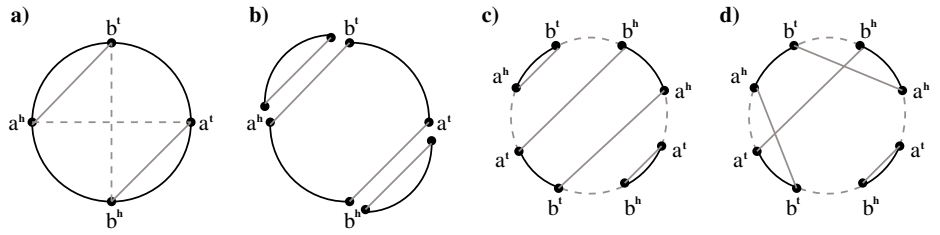

FIG. 5. (a) *Contracted breakpoint graph* $G'(P, R \oplus R)$ *for* $P = +a + b - a - b$ *and* $R = +a + b$. (b) *Black-gray cycle decomposition* C *of* G' *which is not induced by any labeling of* P *and* $R \oplus R$. (c) *Breakpoint graph* $G(P, 2R)$ *inducing* C. (d) *Breakpoint graph* $G(P, R \oplus R)$ *(unique up to relabeling of vertices) with* $c(G) = 2 < |C| = 3$.

In the case when $Q = R \oplus R$ is a perfect duplicated genome, 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. Similar to the double obverse edges, the double gray edges form a matching in G' (Figure 5(a)).

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 $G'(P, Q)$. We are interested in the reverse problem as follows.

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

This problem may not always have a solution for unichromosomal genomes (Figure 5) and this is exactly the factor that leads to a counterexample in section 3. This complication will be addressed in section 4 using the following three theorems proved in [1].

THEOREM 2 (see [1]). *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* P *and either* $R \oplus R$ *or* $2R$ *that induces the cycle decomposition* C.

Let $c_{max}(P, R \oplus R) = c_{max}(G'(P, R \oplus R))$ be the number of cycles in a maximal black-gray cycle decompositions of the contracted breakpoint graph $G'(P, R \oplus R)$. Theorem 2 motivates the following problem that will later help us to solve the weak genome halving problem.

CYCLE DECOMPOSITION PROBLEM. *For a given duplicated genome* P*, find a perfect duplicated genome* $R \oplus R$ *maximizing* $c_{max}(P, R \oplus R)$.

Although the maximal black-gray cycle decomposition of $G'(P, R \oplus R)$ may correspond to a breakpoint graph $G(P, 2R)$ (Figure 5), we will prove below that there exists a breakpoint graph $G(P, R \oplus R)$ having "almost" the same number of black-gray cycle as $G(P, 2R)$ (Figure 5(d)). Later we will classify all the cases in which there exists a labeled genome $R' \oplus R'$ such that $c(P, R' \oplus R') = c(P, 2R)$.

The solution to the cycle decomposition problem is given by the following two theorems.

THEOREM 3 (see [1]). *For a given duplicated genome* P *and any perfect duplicated genome* $R \oplus R$,

$$c_{max}(P, R \oplus R) \leq |P|/2 + b_e(P),$$

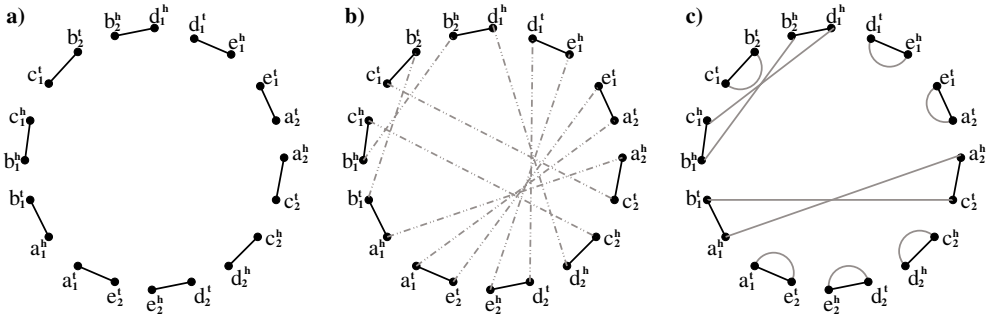

FIG. 6. (a) *A set of black edges forming the partial graph* $\mathcal{G}(\mathbf{V}, A)$ *corresponding to the genome* $P = +a+b-c+b-d-e+a+c-d-e$. (b) *Natural graphs as connected components in the partial graph with counterpart edges.* (c) *A completed graph* $\mathcal{G}(\mathbf{V}, A, \Gamma)$ *with maximum number of cycles* $c(G) = 8$. $\mathcal{G}(\mathbf{V}, A, \Gamma)$ *is a breakpoint graph of the circular genome* $P = +a_1+b_1-c_1+b_2-d_1-e_1+a_2+c_2-d_2-e_2$ *and a perfect duplicated genome* $(-a_1 + e_2 + d_2 - c_2 + b_1)(-b_2 + c_1 - d_1 - e_1 + a_2)$ *(of the form* $R \ominus R$*).*

where $|P|/2$ *represents the number of unique genes in* P *and* $b_e(P)$ *is the number of even black cycles in* $\hat{P}$*. Moreover, if* $c_{max}(P, R \oplus R) = |P|/2 + b_e(P)$*, then each black-gray connected component of* $G'(P, R \oplus R)$ *contains either a single even black cycle (simple component) or a pair of odd black cycles (paired component).*

THEOREM 4 (see [1]). *For any duplicated genome* P*, there exists a perfect duplicated genome* $R \oplus R$ *such that*

$$c_{max}(P, R \oplus R) = |P|/2 + b_e(P)$$

and each paired component of $G'(P, R \oplus R)$ *contains a single interedge (a double gray edge connecting distinct black cycles).*

3. A flaw in the El-Mabrouk–Sankoff analysis. El-Mabrouk and Sankoff came up with a theorem describing the minimum distance from the given rearranged duplicated genome to a perfect duplicated genome. Given a rearranged duplicated genome P, the crux of their approach is an algorithm for computing $c(G)$—the number of cycles of a so-called *maximal completed graph*, i.e., a breakpoint graph[3] with the maximum number of black-gray cycles. In [20] they demonstrate that $c(G)$ equals the number of genes plus $\gamma(G)$, where $\gamma(G)$ is the parameter defined below. We illustrate the concepts from [20] using the genome $P = +a + b - c + b - d - e + a + c - d - e$ on the set of genes $\mathcal{B} = \{a, b, c, d, e\}$ (page 757 in [20]). El-Mabrouk and Sankoff first arbitrarily label two copies of each gene x as x_1 and x_2 for each $x \in \mathcal{B}$ and further transform the signed permutation G into an unsigned permutation $a_1^t a_1^h b_1^t b_1^h c_1^h c_1^t b_2^t b_2^h d_1^h d_1^t e_1^h e_1^t a_2^t a_2^h c_2^t c_2^h d_2^h d_2^t e_2^h e_2^t$.

Let $\mathbf{V} = \{x_1^t, x_1^h, x_2^t, x_2^h \mid x \in \mathcal{B}\}$. The *partial graph* $\mathcal{G}(\mathbf{V}, A)$ associated with P has the edge set A of black edges linking adjacent terms (other than obverses x_i^t and x_i^h) in the corresponding unsigned permutation (Figure 6(a)).

Black edges together with counterpart edges (i.e., edges between x_1^t and x_2^t or between x_1^h and x_2^h) form a graph shown in Figure 6(b). The connected components of this graph are called *natural graphs* in [20]. There are four connected components (natural graphs) in the graph in Figure 6(b), two of them have three black edges (odd

[3]Following El-Mabrouk and Sankoff [20] we ignore obverse edges in breakpoint graphs throughout section 3.

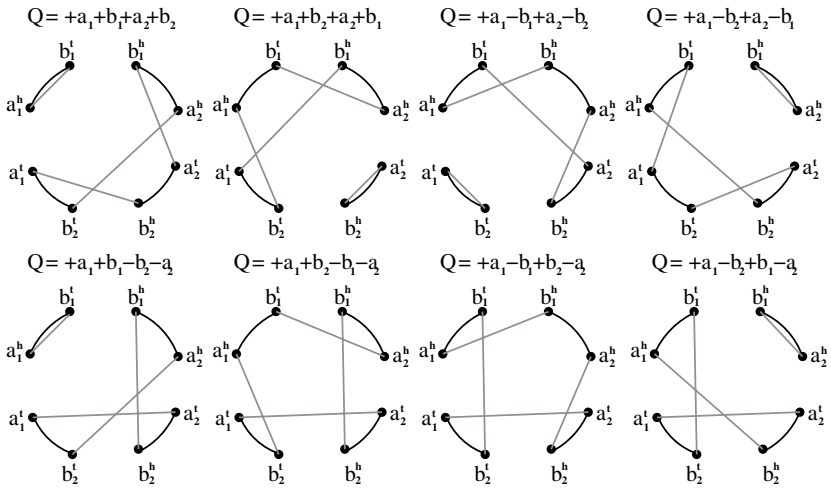

FIG. 7. *Breakpoint graphs of the circular genome $P = +a + b - a - b$ and all possible labelings of all possible perfect duplicated genomes Q (without loss of generality, we assume that labeling of $P = +a_1 + b_1 - a_2 - b_2$ is fixed). In terms of* [20]*, the top four graphs correspond to an $R \oplus R$ duplication pattern, while the bottom four graphs correspond to an $R \ominus R$ duplication pattern.*

natural graphs) and two of them have two black edges (even natural graphs). Let NE be the number of even natural graphs ($NE = 2$ in Figure 6(b)).

El-Mabrouk and Sankoff define the parameter

$$\gamma(G) = \begin{cases} NE & \text{if all natural graphs are even,} \\ NE + 1 & \text{otherwise.} \end{cases}$$

A graph $\mathcal{G}(\mathbf{V}, A, \Gamma)$ obtained from the partial graph $\mathcal{G}(\mathbf{V}, A)$ by introducing a set of gray edges Γ is called a *completed graph* if $\mathcal{G}(\mathbf{V}, A, \Gamma)$ is a breakpoint graph for some genomes on the set of genes $\{x_1, x_2 \mid x \in \mathcal{B}\}$. The following theorem (Theorem 7.7 in [20]) characterizes the maximum number of cycles in the completed graph $\mathcal{G}(\mathbf{V}, A, \Gamma)$.

THEOREM. *The maximal number of cycles in a completed graph of $\mathcal{G}(\mathbf{V}, A)$ is $c(G) = \frac{|A|}{2} + \gamma(G)$.*

For the genome in Figure 6 we have $\gamma(G) = NE + 1 = 3$ and $c(G) = \frac{|A|}{2} + \gamma(G) = \frac{10}{2} + 3 = 8$. A completed graph $\mathcal{G}(\mathbf{V}, A, \Gamma)$ with eight cycles is shown at Figure 6(c).[4] Below we provide a counterexample to Theorem 7.7 from [20].

Consider a circular genome $P = +a + b - a - b$ labeled as $+a_1 + b_1 - a_2 - b_2$. The genome P defines a partial graph $\mathcal{G}(\mathbf{V}, A)$ with a single natural graph of even size implying $\gamma(G) = 1$. It follows from Theorem 7.7 in [20] that there exists a perfect duplicated genome Q such that the breakpoint graph $G = G(P, Q)$ consists of $\frac{|A|}{2} + \gamma(G) = 3$ cycles. However, the direct enumeration of all possible perfect duplicated genomes Q shows that there is no breakpoint graph $G(P, Q)$ with three cycles. There exist eight distinct labeled perfect duplicated genomes Q giving rise to eight breakpoint graphs $G(P, Q)$ shown in Figure 7. All of them have less than three cycles. In the next section we explain what particular property of the genome $+a + b - a - b$ was not addressed properly in the El-Mabrouk–Sankoff analysis.

[4] While we do not explicitly consider $R \ominus R$ duplications shown in this figure (see [20] for details), our counterexample works for both $R \oplus R$ and $R \ominus R$ duplications.

4. Classification of duplicated genomes. The labeling problem can be addressed by considering *multichromosomal* genomes.[5] A multichromosomal duplicated genome is a set of circular chromosomes with every gene present in two copies. For example, Figure 4(c) presents a multichromosomal duplicated genome Q' consisting of two circular 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 $Q = +a-b+a+b$ (Figure 4(b)) and, hence, the contracted breakpoint graphs $G'(P,Q)$ and $G'(P,Q')$ are the same for any genome P (Figure 4(d)). We call genomes Q and Q' *equivalent* if their de Bruijn graphs are the same, i.e., $\hat{Q} = \hat{Q}'$.

It is easy to see that $R \oplus R$ is equivalent to $2R$ and, thus, $G'(P, R \oplus R) = G'(P, 2R)$ for any duplicated genome P. But in contrast to the breakpoint graph $G(P, R \oplus R)$ (for any labeling of P and $R \oplus R$) that contains a single gray-obverse cycle, the breakpoint graph $G(P, 2R)$ contains two gray-obverse cycles. The following theorem reveals the relationship between $G(P, R \oplus R)$ and $G(P, 2R)$.

THEOREM 5. *For any labeling of the genomes P and $2R$, there exists a labeling of the genome $R \oplus R$ such that $|c(P, R \oplus R) - c(P, 2R)| \leq 1$. Moreover, if there are two gray edges (x, y) and $(\bar{x}, \bar{y})$ belonging to the same black-gray cycle in $G(P, 2R)$, then there exists a labeling of $R \oplus R$ with $c(P, R \oplus R) \geq c(P, 2R)$.*

Proof. Let (x, y) be a gray edge in the breakpoint graph $G(P, 2R)$. Since the genome $2R$ is perfect duplicated there exists a gray edge $(\bar{x}, \bar{y})$ connecting counterparts of x and y. Define a graph H having the same vertices and edges as $G(P, 2R)$ except the gray edges (x, y) and $(\bar{x}, \bar{y})$ that are replaced with the gray edges $(x, \bar{y})$ and $(\bar{x}, y)$. Since the graph $G(P, 2R)$ consists of two gray-obverse cycles, the gray edges (x, y) and $(\bar{x}, \bar{y})$ belong to different gray-obverse cycles. Therefore, the graph H contains a single gray-obverse cycle (as well as a single black-obverse cycle inherited from $G(P, 2R)$). This implies that H is a breakpoint of the genomes P and $R \oplus R$ (i.e., $H = G(P, R \oplus R)$), where the labeling of P is the same as in $G(P, 2R)$.

If the gray edges (x, y) and $(\bar{x}, \bar{y})$ belong to the same black-gray cycle in $G(P, 2R)$, then this cycle may be split into two in H, while the other black-gray cycles are not affected. Conversely, if the gray edges (x, y) and $(\bar{x}, \bar{y})$ belong to different black-gray cycles in $G(P, 2R)$, then these cycles may be joined into a single cycle in H. In either case the difference $|c(P, R \oplus R) - c(P, 2R)|$ does not exceed 1. ∎

We redefine the notion of parity of a genome P in terms of the de Bruijn graph $\hat{P}$. A genome P is called *singular* if all black cycles in $\hat{P}$ are even. For a nonsingular genome P, define parity$(P) = \infty$. For a singular genome P, we clockwise label edges of each black cycle in $\hat{P}$ with alternating numbers $\{0, 1\}$ so that every two adjacent edges are labeled differently (Figure 8(a)). Labels of black edges in cycle P classify obverse edges in P into two classes: *even* if its flanking black edges have the same labels, and *odd* if its flanking black edges have different labels (Figure 8(b)). Let m_{even} and m_{odd} be the number of even/odd obverse edges in P correspondingly. Obviously, both m_{even} and m_{odd} are even numbers. We define parity$(P) = m_{odd}/2 \bmod 2$.

This definition of the parity index coincides with the one given in the introduction. To establish a correspondence between them one can consider a genome P as a black-obverse cycle and contract each black edge into a single vertex that inherits the label from the black edge. Since every pair of adjacent black edges of $\hat{P}$ is labeled differently, every pair of counterpart vertices is labeled differently as well. This implies that two-

[5] We emphasize that in this paper we consider only the genome halving problem for unichromosomal genomes and use multichromosomal genomes only to prove some auxiliary results.

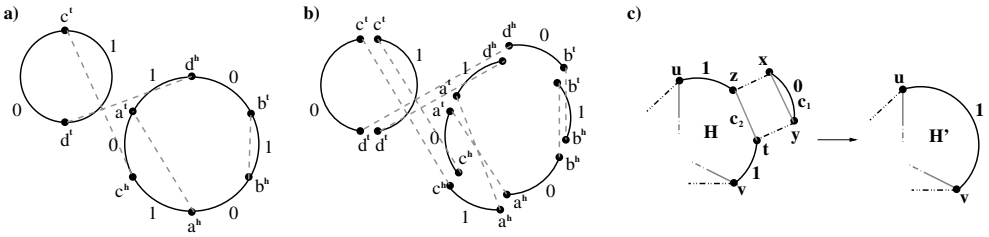

FIG. 8. *For the genome* $P = +a-b-b-d+c-a-d+c$, (a) 01*-labeling of the de Bruijn graph* $\hat{P}$; (b) *induced labeling of black-obverse cycle* P *with* $m_{odd} = 4$ *and* $m_{even} = 4$; (c) *transformation of the graph* H *into* H' *by removing vertices* x, y, z, t *and incident edges and adding a black edge* (u, v) *labeled the same as* (u, x) *and* (v, t).

digit labels of every pair of obverse edges are inversions of each other.

THEOREM 6. *The parity index of a singular genome is well defined.*

Proof. Let P be a singular genome. If $\hat{P}$ has k black cycles, then there are 2^k different 01-labelings of its black edges (two possible labelings per cycle). Therefore, it is sufficient to show that a change of 01-labeling of a particular black cycle c does not affect parity(P).

Let m^c_{even} and m^c_{odd} be the number of even/odd obverse edges in cycle P connecting black edges of c with black edges outside c. Since double obverse edges form a matching in the de Bruijn graph $\hat{P}$, the total number of double obverse edges connecting c with other black cycles is even and, thus, $m^c_{even} + m^c_{odd}$ is a multiple of 4.

Changing the 01-labeling of the black cycle c reverses the labels $0 \leftrightarrow 1$ in c. Reversed labeling of c does not change parity of obverse edges connecting two black edges in c (since both endpoint labels change) or two black edges outside of c (since neither of the endpoint labels changes). At the same time, each of the $m^c_{even} + m^c_{odd}$ obverse edges connecting black edges in c with black edges outside of c changes its parity (i.e., even edges become odd and vice versa). Then m_{odd} changes as follows:

$$m'_{odd} = m_{odd} - m^c_{odd} + m^c_{even} = m_{odd} - (m^c_{odd} + m^c_{even}) + 2m^c_{even}.$$

Since both $m^c_{odd} + m^c_{even}$ and $2m^c_{even}$ are multiples of 4, the parity of $m'_{odd}/2$ and $m_{odd}/2$ is the same, implying that parity(P) is well defined. □

Our goal is to prove the following theorem.

THEOREM 7. *For a duplicated genome* P,

$$\max_R c(P, R \oplus R) = \begin{cases} |P|/2 + b_e(P) & \text{if parity}(P) \neq 0, \\ |P|/2 + b_e(P) - 1 & \text{otherwise.} \end{cases}$$

The proof of Theorem 7 is split into two cases depending on whether P is singular or nonsingular.

THEOREM 8. *For a nonsingular genome* P, $\max_R c(P, R \oplus R) = |P|/2 + b_e(P)$.

Proof. If P is a nonsingular genome, then $\hat{P}$ has an odd black cycle. According to Theorem 4 there exists a perfect duplicated genome $R \oplus R$ such that $c_{max}(P, R \oplus R) = |P|/2 + b_e(P)$. Theorem 2 ensures that the maximum cycle decomposition of the contracted breakpoint graph $G'(P, R \oplus R)$ is induced by a labeling of either $R \oplus R$ or $2R$. If it is $R \oplus R$, then the theorem holds. Otherwise, consider a paired component in $G'(P, R \oplus R)$ (which exists since $\hat{P}$ has an odd black cycle) and an interedge e in it. Let (x, y) and $(\bar{x}, \bar{y})$ be gray edges in $G(P, 2R)$ corresponding to the interedge e in $G'(P, R \oplus R)$. Since e is the only bridge between two different black cycles (Theorem 4)

in $G'(P, R \oplus R)$, the gray edges (x, y) and $(\bar{x}, \bar{y})$ must belong to the same black-gray cycle in $G(P, 2R)$. Applying Theorem 5 to these gray edges, we obtain a labeled genome $R \oplus R$ with $c(P, R \oplus R) = c(P, 2R) = |P|/2 + b_e(P)$. □

For a singular genome P, we first fix some alternating 01-labeling of black edges in every black cycle of $\hat{P}$. The labeling of edges imposes labeling of vertices of any breakpoint graph $G(P, Q)$ (for any genome Q) so that each vertex inherits a label from an incident black edge. Note that every pair of counterpart vertices get different labels, as their incident black edges are adjacent in $\hat{P}$. A labeling of vertices of $G(P, Q)$ is called *uniform* if endpoints of every gray edge have identical labels (i.e., every gray edge is even). We will need the following theorem.

THEOREM 9. *Let P be a singular genome and Q be a perfect duplicated genome with $c(P, Q) = |P|/2 + b_e(P)$. Then every alternating* 01*-labeling of $\hat{P}$ imposes a uniform labeling on vertices of $G(P, Q)$.*

While the definition of the breakpoint graph does not explicitly specify the counterpart edges, one can derive them for $G(P, Q)$ in Theorem 9 from the vertex labels. Also, it is easy to see that gray and counterpart edges in $G(P, Q)$ form cycles of length 4 as soon as Q is a perfect duplicated genome. We take the liberty of restating the condition $c(P, Q) = |P|/2 + b_e(P)$ as $c_{bg}(G) = n + c_{bc}(G)$, where $c_{bg}(G)$ is the number of black-gray edges in G, n is the number of unique genes in P, and $c_{bc}(G)$ is the number of black-counterpart cycles in G. Also, every alternating 01-labeling of $\hat{P}$ corresponds to an alternating labeling of black edges within black-counterpart cycles. This leads to the following reformulation of Theorem 9.

THEOREM 10. *Let H be a graph on $4n$ vertices consisting of three perfect matchings black, gray, and counterpart such that* (i) *gray and counterpart matchings form cycles of length* 4*, and* (ii) *$c_{bg}(H) = n + c_{bc}(H)$. Then every alternating* 01*-labeling of black edges within black-counterpart cycles imposes a uniform labeling on vertices of H.*

Proof. The proof is done by induction on n. If $n = 1$, then the graph H consists of a gray-counterpart cycle with two black edges parallel to the gray edges, and the theorem holds. Assume that the theorem holds for graphs with less than $4n$ vertices.

Since H has $2n$ black edges and $c_{bg}(H) = n + c_{bc}(H) > n$, the pigeonhole principle implies that there exists a black-gray cycle c_1 of length 2 in H. Let $e_1 = (x, y)$ be a gray edge in the cycle c_1 (thus, e_1 is even) and let (x, z) and (y, t) be adjacent counterpart edges. Then there is a gray edge $e_2 = (z, t)$ belonging to the same gray-counterpart cycle as e_1. Let c_2 be a black-gray cycle c_2 containing the gray edge e_2.

If the cycle c_2 has length 2, then the endpoints of e_2 have identical labels. In this case we define a new graph H' as the graph H without vertices x, y, z, t and all incident edges. It is easy to see that H' is a graph on $4(n-1)$ vertices satisfying the conditions of the theorem. Indeed, the number of black-gray cycles in H' is reduced by 2 and the number of black-counterpart cycles is reduced by 1 (as compared to H), i.e., $c_{bg}(H') = c_{bg}(H) - 2$ and $c_{bc}(H') = c_{bc}(H) - 1$. Therefore, $c_{bg}(H') = (n-1) + c_{bc}(H')$. By the induction assumption, every alternating 01-labeling of H' imposes uniform labeling on vertices of H'. It implies that every alternating 01-labeling of H imposes uniform labeling on vertices of H.

If the cycle c_2 has length greater than 2, let (u, z) and (t, v) be black edges adjacent to e_2. These black edges are neighbors of the black edge (x, y) on a black-counterpart cycle (passing through the vertices u, z, x, y, t, v), so they have the same label l which is different from the label of (x, y). Therefore, the endpoints of the gray edge e_2 have identical labels. We define a new graph H' as the graph H with vertices x, y, z, t and all incident edges removed but with a single black edge (u, v) labeled l added

(Figure 8(c)). The graph H' has $4(n-1)$ vertices, $c_{bc}(H') = c_{bc}(H)$ black-counterpart cycles, and $c_{bg}(H') = c_{bg}(H) - 1$ black-gray cycles; thus, $c_{bg}(H') = n - 1 + c_{bc}(H')$ and the induction applies. □

To complete the proof of Theorem 7 we need one more theorem.

THEOREM 11. *For a singular genome P and a perfect duplicated genome Q with $c(P,Q) = |P|/2 + b_e(P)$,*

- *$Q = R \oplus R$ if and only if parity$(P) = 1$;*
- *$Q = 2R$ if and only if parity$(P) = 0$.*

Proof. According to Theorem 2, the graph $G(P,Q)$ has either a single gray-obverse cycle (i.e., $Q = R \oplus R$) or two symmetric gray-obverse cycles (i.e., $Q = 2R$). Theorem 9 implies that all gray edges in G are even (i.e., have identically labeled endpoints) for every alternating 01-labeling of black edges of P.

Case 1. Graph G has a single gray-obverse cycle c. Consider an arbitrary vertex v in G and its counterpart $\overline{v}$. Vertices v and $\overline{v}$ break c into two paths: c' (from v to $\overline{v}$) and c'' (from $\overline{v}$ to v). For every path (cycle) c denote c_{odd} as the number of odd obverse edges in c. Note that obverse edges are evenly divided between c' and c'', i.e., for every pair of obverse edges connecting counterpart vertices, one edge belongs to c' and the other edge belongs to c''. Therefore, $c'_{odd} = c''_{odd}$. Note that the start (vertex v) and the end (vertex $\overline{v}$) vertices of path c' are labeled differently. Since the total number of odd edges is odd for every path with differently labeled ends, and since all gray edges are even (Theorem 9), the total number of odd obverse edges in the path c' is odd. Therefore, $c_{odd}/2 = c'_{odd}$ is odd, implying that parity$(P) = 1$.

Case 2. Graph G has two gray-obverse cycles c' and c''. Note that obverse edges are evenly divided between c' and c''; i.e., for every pair of obverse edges connecting counterpart vertices, one edge belongs to c' and the other edge belongs to c''. Therefore, $c'_{odd} = c''_{odd}$. Since the total number of odd edges in every cycle is even, and since all gray edges are even (Theorem 9), the total number of odd obverse edges in every cycle is even. Since c'_{odd} is even, the overall number of odd obverse edges is a multiple of 4, implying that parity$(P) = 0$. □

For a singular genome P with parity$(P) = 1$ Theorem 11 implies Theorem 7, while for a singular genome P with parity$(P) = 0$ it implies that there is no genome R for which $c(P, R \oplus R) = |P|/2 + b_e(P)$. In the latter case, there exists a genome R and a labeling of P and $2R$ for which $c(P, 2R) = |P|/2 + b_e(P)$ (Theorem 4). The genome $2R$ can be transformed into a labeled genome $R \oplus R$ with $c(P, R \oplus R) = c(P, 2R) - 1 = |P|/2 + b_e(P) - 1$ (Theorem 5). This completes the proof of Theorem 7.

5. Genome halving algorithm. The classification of circular genomes leads to the following algorithm for the weak genome halving problem.[6]

1. For a given duplicated genome P, find a perfect duplicated genome $R \oplus R$ such that $c_{max}(P, R \oplus R) = |P|/2 + b_e(P)$ (Theorem 4) and decompose $G'(P, R \oplus R)$ into the maximum number of black-gray cycles [1].
2. Find a labeling of the genomes P and Q ($Q = R \oplus R$ or $Q = 2R$) and a breakpoint graph $G(P,Q)$ inducing the maximum black-gray cycle decomposition of $G'(P, R \oplus R)$ (Theorem 2).
3. If $Q = R \oplus R$, then output the breakpoint graph $G(P, R \oplus R)$.
4. If $Q = 2R$ and P is nonsingular, then there is a paired component in $G'(P, R \oplus R)$ with a single interedge (Theorem 4) that corresponds to two gray

[6] The algorithm outputs the breakpoint graph $G(P, R \oplus R)$ (in addition to the preduplicated genome R). This allows one to reconstruct a sequence of reversals transforming $R \oplus R$ into P with the reversal distance algorithm.

edges (x, y) and $(\bar{x}, \bar{y})$ in $G(P, 2R)$. Find a labeling of the genome $R \oplus R$ for which $c(P, R \oplus R) = c(P, 2R)$ (Theorems 5 and 8) and output $G(P, R \oplus R)$.

5. If $Q = 2R$ and P is singular, then parity$(P) = 0$ (Theorem 11). Find a labeling of the genome $R \oplus R$ for which $c(P, R \oplus R) = c(P, 2R) - 1$ (Theorem 5) and output $G(P, R \oplus R)$.

We illustrate the algorithm for a genome $P = +a + b - a - b$ (assuming a fixed labeling of P as $(+a_1 + b_1 - a_2 - b_2)$) using Figure 5. At step 1, we use the algorithm from [1] to construct $R = +a + b$ such that the contracted breakpoint graph $G'(P, R \oplus R)$ (Figure 5(a)) has a black-gray cycle decomposition with $|P|/2 + b_e(P) = 3$ cycles (Figure 5(b)). At step 2, we find a breakpoint graph $G(P, Q)$ (Figure 5(c)) inducing this cycle decomposition (see [1]). The breakpoint graph $G(P, Q)$ has two gray-obverse cycles (Figure 5(c)), thus implying that $Q = 2R$. Since the genome P is singular, we proceed to step 5 and perform a transformation of $Q = 2R$ into a new genome, as described in Theorem 5. We first find a pair of gray edges from two different cycles in $2R$ as described in Theorem 5, for example, a pair of edges labeled (a^h, b^t) in Figure 5(c). Afterwards, we replace these edges with a new pair of gray edges also labeled (a^h, b^t), as shown in Figure 5(d). This operation transforms two cycles in the genome Q into a single cycle (unichromosomal genome) that we represent as $R \oplus R$. According to Theorems 5 and 7, this genome represents a solution to the weak genome halving problem.

The first two steps of the genome halving algorithm can be implemented in $O(|P|^2)$ time (see [1]) while the remaining steps fit this time bound as well. In practice, our genome halving software takes less than a second to halve a "random" duplicated genome with 1000 unique genes on a standard Intel PIII-900MHz CPU.

6. Conclusion. While whole genome duplications in multichromosomal genomes are well established, there are relatively few examples of whole genome duplications in unichromosomal genomes. Undoubtedly, bacterial genomes have undergone a large number of duplications, but it is difficult to distinguish between partial and whole genome duplication scenarios in the case of these genomes (Coissac, Maillier, and Netter [13]). Matters are further complicated by the fact that even a few rearrangements can quickly "randomize" gene orders in bacterial genomes (due to a relatively small number of genes).

Recently, Sugaya et al. [38] studied *Cyanobacterium Anabaena* and came to the conclusion that it has undergone whole genome duplications rather than a series of (tandem) segmental duplications. Indeed, the arrangement of genes in *Cyanobacterium Anabaena* points to whole genome duplications as the most likely scenario (it also has an unusually large genome as compared to other organisms in the *Cyanobacteria* phylum). Also, the recent discovery and sequencing of *Acanthamoeba polyphaga* (the largest known virus to date) revealed an unusually large number of duplicated regions that point to a large duplication event (Suhre [39]). While it remains unclear whether this large segmental duplication represents whole genome duplications or extremely large partial duplication(s), one can argue that it is only a matter of time until ongoing sequencing efforts will reveal traces of whole genome duplications in many unichromosomal (bacterial and viral) genomes.

These recent discoveries raise a new algorithmic challenge that we refer to as the partial genome duplication problem. Let R be a genome with n unique genes (represented as a signed permutation) and R' be a set of m consecutive genes in this genome. We define $R \oplus R'$ as a *perfect partially duplicated genome*. Let P be a genome with $n + m$ genes in which m genes from R' are duplicated and the remaining

$n - m$ genes are unique. Given a genome P, the partial genome duplication problem is to find a perfect partially duplicated genome $R \oplus R'$ such that the reversal distance between $R \oplus R'$ and P is minimal.

Acknowledgments. We are grateful to Mohan Paturi and Dekel Tsur for many insightful comments.

REFERENCES

[1] M. A. Alekseyev and P. A. Pevzner, *Colored de Bruijn graphs and the genome halving problem*, IEEE/ACM Trans. Comput. Biol. Bioinformatics, 4 (2007), pp. 98–107.

[2] M. A. Alekseyev and P. A. Pevzner, *Whole genome duplications, multi-break rearrangements, and genome halving problem*, in Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, Philadelphia, ACM, New York, 2007, pp. 665–679.

[3] 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.

[4] V. Bafna and P. A. Pevzner, *Genome rearrangements and sorting by reversals*, SIAM J. Comput., 25 (1996), pp. 272–289.

[5] E. Belda, A. Moya, and F. J. Silva, *Genome rearrangement distances and gene order phylogeny in γ-proteobacteria*, Mol. Biol. Evol., 22 (2005), pp. 1456–1467.

[6] A. Bergeron, *A very elementary presentation of the Hannenhalli–Pevzner theory*, in Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 2089, Springer, Berlin, 2001, pp. 106–117.

[7] A. Bergeron, J. Mixtacki, and J. Stoye, *Reversal distance without hurdles and fortresses*, in Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 3109, Springer, Berlin, 2004, pp. 388–399.

[8] G. Bourque, P. A. Pevzner, and G. Tesler, *Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes*, Genome Res., 14 (2004), pp. 507–516.

[9] G. Bourque, Y. Yacef, and N. El-Mabrouk, *Maximizing synteny blocks to identify ancestral homologs*, in Comparative Genomics, Lecture Notes in Comput. Sci. 3678, Springer, Berlin, 2005, pp. 21–34.

[10] G. Bourque, E. M. Zdobnov, P. Bork, P. A. Pevzner, and G. Tesler, *Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages*, Genome Res., 15 (2005), pp. 98–110.

[11] X. Chen, J. Zheng, P. Nan, Z. Fu, Y. Zhong, S. Lonardi, and T. Jiang, *Computing the assignment of orthologous genes via genome rearrangement*, in Proceedings of the 3rd Asia Pacific Bioinformatics Conference, Imperial College Press, London, 2005, pp. 363–378.

[12] A. Christoffels, E. G. L. Koh, J. Chia, S. Brenner, S. Aparicio, and B. Venkatesh, *Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes*, Mol. Biol. Evol., 21 (2004), pp. 1146–1151.

[13] E. Coissac, E. Maillier, and P. Netter, *A comparative study of duplications in bacteria and eukaryotes: The importance of telomeres*, Mol. Biol. Evol., 14 (1997), pp. 1062–1074.

[14] P. Dehal and J. L. Boore, *Two rounds of genome duplication in the ancestral vertebrate genome*, PLoS Biol., 3 (2005), e314.

[15] F. S. Dietrich, S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pöhlmann, P. Luedi, S. Choi, R. A. Wing, A. Flavier, T. D. Gaffney, and P. Philippsen, *The* Ashbya gossypii *genome as a tool for mapping the ancient* Saccharomyces cerevisiae *genome*, Science, 304 (2004), pp. 304–307.

[16] N. El-Mabrouk, *Genome rearrangement by reversals and insertions/deletions of contiguous segments*, in Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 1848, Springer, Berlin, 2000, pp. 222–234.

[17] N. El-Mabrouk, D. Bryant, and D. Sankoff, *Reconstructing the pre-doubling genome*, in Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB), ACM, New York, 1999, pp. 154–163.

[18] N. El-Mabrouk, J. H. Nadeau, and D. Sankoff, *Genome halving*, in Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 1448, Springer, Berlin, 1998, pp. 235–250.

[19] N. El-Mabrouk and D. Sankoff, *On the reconstruction of ancient doubled circular genomes*, Genome Informatics, 10 (1999), pp. 83–93.

[20] N. El-Mabrouk and D. Sankoff, *The reconstruction of doubled genomes*, SIAM J. Comput., 32 (2003), pp. 754–792.
[21] R. Guyot and B. Keller, *Ancestral genome duplication in rice*, Genome, 47 (2004), pp. 610–614.
[22] S. Hannenhalli and P. Pevzner, *Transforming men into mouse (polynomial algorithm for genomic distance problem)*, in Proceedings of the 36th Annual Symposium on Foundations of Computer Science, 1995, IEEE, Piscataway, NJ, pp. 581–592.
[23] S. Hannenhalli and P. Pevzner, *Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)*, J. ACM, 46 (1999), pp. 1–27.
[24] O. Jaillon et al., *Genome duplication in the teleost fish* Tetraodon nigroviridis *reveals the early vertebrate proto-karyotype*, Nature, 431 (2004), pp. 946–957.
[25] H. Kaplan, R. Shamir, and R. E. Tarjan, *A faster and simpler algorithm for sorting signed permutations by reversals*, SIAM J. Comput., 29 (1999), pp. 880–892.
[26] H. Kaplan and E. Verbin, *Sorting signed permutations by reversals, revisited*, J. Comput. System Sci., 70 (2005), pp. 321–341.
[27] M. Kellis, B. W. Birren, and E. S. Lander, *Proof and evolutionary analysis of ancient genome duplication in the yeast* Saccharomyces cerevisiae, Nature, 428 (2004), pp. 617–624.
[28] A. Meyer and Y. Van de Peer, *From* $2R$ *to* $3R$*: Evidence for a fish-specific genome duplication (FSGD)*, BioEssays, 27 (2005), pp. 937–945.
[29] W. J. Murphy, G. Bourque, G. Tesler, P. Pevzner, and S. J. O'Brien, *Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps*, Human Genomics, 1 (2003), pp. 30–40.
[30] W. J. Murphy, D. M. Larkin, A. Everts van der Wind, G. Bourque, G. Tesler, L. Auvil, J. E. Beever, B. P. Chowdhary, F. Galibert, L. Gatzke, C. Hitte, C. N. Meyers, D. Milan, E. A. Ostrander, G. Pape, H. G. Parker, T. Raudsepp, M. B. Rogatcheva, L. B. Schook, L. C. Skow, M. Welge, J. E. Womack, S. J. O'Brien, P. A. Pevzner, and H. A. Lewin, *Dynamics of mammalian chromosome evolution inferred from multispecies comparative map*, Science, 309 (2005), pp. 613–617.
[31] S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin, 1970.
[32] P. Pevzner, H. Tang, and G. Tesler, *De novo repeat classification and fragment assembly*, Genome Res., 14 (2004), pp. 1786–1796.
[33] P. Pevzner and G. Tesler, *Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes*, Genome Res., 13 (2003), pp. 37–45.
[34] P. A. Pevzner, *Computational Molecular Biology: An Algorithmic Approach*, The MIT Press, Cambridge, MA, 2000.
[35] P. A. Pevzner and G. Tesler, *Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution*, Proc. Nat. Acad. Sci., 100 (2003), pp. 7672–7677.
[36] M. Robinson-Rechavi, O. Marchand, H. Escriva, and V. Laudet, *An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes*, Curr. Biol., 11 (2001), pp. 458–459.
[37] D. Sankoff, *Genome rearrangement with gene families*, Bioinformatics, 15 (1999), pp. 909–917.
[38] N. Sugaya, M. Sato, H. Murakami, A. Imaizumi, S. Aburatani, and K. Horimoto, *Causes for the large genome size in a Cyanobacterium* Anabaena *sp. PCC*7120, Genome Informatics, 15 (2004), pp. 229–238.
[39] K. Suhre, *Gene and genome duplication in Acanthamoeba polyphaga Mimivirus*, J. Virology, 79 (2005), pp. 14095–14101.
[40] K. M. Swenson, M. Marron, J. V. Earnest-DeYoung, and B. M. E. Moret, *Approximating the true evolutionary distance between two genomes*, in Proceedings of the Seventh Workshop on Algorithm Engineering and Experiments and the Second Workshop on Analytic Algorithmics and Combinatorics, C. Demetrescu, R. Sedgewick, and R. Tamassia, eds., SIAM, Philadelphia, 2005, pp. 121–129.
[41] K. M. Swenson, N. D. Pattengale, and B. M. E. Moret, *A framework for orthology assignment from gene rearrangement data*, in Comparative Genomics, Lecture Notes in Comput. Sci. 3678, Springer, Berlin, 2005, pp. 153–166.
[42] E. Tannier and M. F. Sagot, *Sorting by reversals in subquadratic time*, in Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 3109, Springer, Berlin, 2004, pp. 1–13.