University of South Carolina
Scholar Commons

Faculty Publications | Computer Science and Engineering, Department of

11-8-2008

Multi-Break Rearrangements and Breakpoint Re-Uses: From Circular to Linear Genomes

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

Follow this and additional works at: https://scholarcommons.sc.edu/csce_facpub

 Part of the Computer Engineering Commons, and the Genetics and Genomics Commons

Publication Info

Published in *Journal of Computational Biology*, Volume 15, Issue 8, 2008, pages 1117-1131.
http://www.liebertpub.com/products/product.aspx?pid=31
© by Mary Ann Liebert, Inc.

This Article is brought to you by the Computer Science and Engineering, Department of at Scholar Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 8, 2008
© Mary Ann Liebert, Inc.
Pp. 1117–1131
DOI: 10.1089/cmb.2008.0080

Multi-Break Rearrangements and Breakpoint Re-Uses: From Circular to Linear Genomes

MAX A. ALEKSEYEV

ABSTRACT

Multi-break rearrangements break a genome into multiple fragments and further glue them together in a new order. While 2-break rearrangements represent standard reversals, fusions, fissions, and translocations, 3-break rearrangements represent a natural generalization of transpositions. Alekseyev and Pevzner (2007a, 2008a) studied multi-break rearrangements in circular genomes and further applied them to the analysis of chromosomal evolution in mammalian genomes. In this paper, we extend these results to the more difficult case of linear genomes. In particular, we give lower bounds for the rearrangement distance between linear genomes and for the breakpoint re-use rate as functions of the number and proportion of transpositions. We further use these results to analyze comparative genomic architecture of mammalian genomes.

Key words: algorithms, combinatorics, computational molecular biology, evolution, genomic rearrangements.

1. INTRODUCTION

REARRANGEMENTS ARE GENOMIC "EARTHQUAKES" that change the chromosomal architecture. Each of the standard rearrangement operations (i.e., reversal, translocation, fusion, or fission) can be viewed as making up to two breaks in a genome and gluing the resulting fragments in a certain order. More complex rearrangement operations such as transpositions[1] require three breaks. Alekseyev and Pevzner (2008a) introduced a generalized *k-break* rearrangement operation that makes k breaks in a genomes and glues the resulting fragments in a new order, and studied such operations in the case of *circular genomes* (i.e., genomes consisting of circular chromosomes).

While 2-breaks correspond to the standard rearrangement operations; 3-breaks add transpositions, 3-way fusions, and 3-way fissions to the set of rearrangement operations. Although transpositions are believed to be rare (as compared to reversals and translocations) and 3-way fusions/fissions were never described before in an evolutionary context, these complex rearrangements may be involved in chromosome aberrations in irradiated genomes (Sachs et al., 2002, 2004; Levy et al., 2004; Vazquez et al., 2002).

Recent studies revealed that the *multi-break* distance (and the DCJ distance [Yancopoulos et al., 2005] in particular) provides a simpler framework for analyzing rearrangements than the *genomic distance*

Department of Computer Science and Engineering, University of California at San Diego, La Jolla, California.
[1]Throughout this paper, transpositions refer to both transpositions and inverted transpositions.

(Hannenhalli and Pevzner, 1995). In particular, switching to multi-break rearrangements in circular genomes recently led to solving difficult combinatorial problems (Bergeron et al., 2006; Alekseyev and Pevzner, 2007b, 2008a) that may be intractable in the case of linear genomes. However, since multi-breaks are defined for circular genomes, their application to linear genomes requires *circularization*[2] of linear chromosomes (Bergeron et al., 2006; Alekseyev and Pevzner, 2007a). This raises a question to what extent such circularization distorts the rearrangement analysis and affects the computation of rearrangement scenarios. In this paper, we address this concern by establishing a link between rearrangements in linear genomes and their circularized versions and proving a lower bound for the genomic distance between linear genomes in the case of difficult to analyze transposition-like operations. To evaluate the obtained lower bounds, we further compute and compare them to genomic distances for a number of mammalian genomes. We demonstrate that circularization has little impact on the genomic distance, providing a rationale for using it in the rearrangement analysis.

Another application of multi-break rearrangements is the *breakpoint re-use* analysis. In particular, recent analysis of breakpoint re-use in Alekseyev and Pevzner (2007a) addressed the *"FBM versus RBM"* controversy in mammalian evolution. The *Random Breakage Model (RBM)* (Nadeau and Taylor, 1984) postulates that rearrangements happen at "random" genomic positions, resulting in low *breakpoint re-use rate*. Pevzner and Tesler (2003b) came up with an alternative *Fragile Breakage Model (FBM)* that postulates existence of *fragile* genomic regions that are more likely to be broken by rearrangements than the rest of the genome, implying (in contrast to the RBM) high breakpoint re-use rate. A variety of further studies argued for existence of fragile regions in mammalian genomes (Murphy et al., 2005; van der Wind et al., 2004; Bailey et al., 2004; Zhao et al., 2004; Webber and Ponting, 2005; Hinsch and Hannenhalli, 2006; Ruiz-Herrera et al., 2006; Mehan et al., 2007; Kikuta et al., 2007; Caceres et al., 2007; Gordon et al., 2007). In the current paper, we extend the results from Alekseyev and Pevzner (2007a) to the case of linear genomes and provide the foundation for further identification and analysis of fragile regions in mammalian genomes (Alekseyev and Pevzner, 2008b).

The paper is organized as follows. In Section 2, we review the recent results from Alekseyev and Pevzner (2007a, 2008a) on the multi-break distances and breakpoint re-use in circular genomes. In Section 3, we extend these results to the case of linear genomes and demonstrate that the circularization of mammalian genomes adopted in Alekseyev and Pevzner (2007a) has little impact on the rearrangement analysis. In Section 4, we apply our lower bounds to analyzing six mammalian genomes and pose some open problems.

2. MULTI-BREAK REARRANGEMENTS IN CIRCULAR GENOMES

We will find it convenient to represent a circular chromosome with *elements* (e.g., genes or synteny blocks) $x_1, \ldots, x_n$ as a cycle (Fig. 1) composed of n directed labelled edges (corresponding to elements) and n undirected unlabeled edges (connecting adjacent elements). The directions of the edges correspond to *signs* (strands) of the elements. We label the *tail* and *head* of a directed edge x_i as x_i^t and x_i^h, respectively. Vertex x_i^t is called the *obverse* of vertex x_i^h, and vice versa. Vertices in a chromosome connected by an undirected edge are called *adjacent*. We represent a genome as a collection of disjoint cycles (chromosomes) with edges of two *alternating* colors: one color (usually black or gray) reserved for undirected edges and the other ("obverse"[3]) color reserved for directed edges. We do not explicitly show the directions of obverse edges since they are defined by superscripts "t" and "h" at the endpoints (Fig. 1).

Let P be a genome represented as a collection of *alternating cycles* with black and obverse edges (a cycle is alternating if colors of its edges alternate). For any two black edges (u, v) and (x, y) in the genome (graph) P we define a *2-break* rearrangement as replacement of these edges with either a pair of edges

[2]While multi-breaks in linear genomes can be defined similarly to circular genomes, the linear case is harder to analyze. In contrast to circular genomes, not every multi-break can be performed over a linear genome: multi-breaks that create circular chromosomes are not allowed.

[3]We have chosen rather unusual name "obverse" for the color to be consistent with previous papers on genome rearrangements.

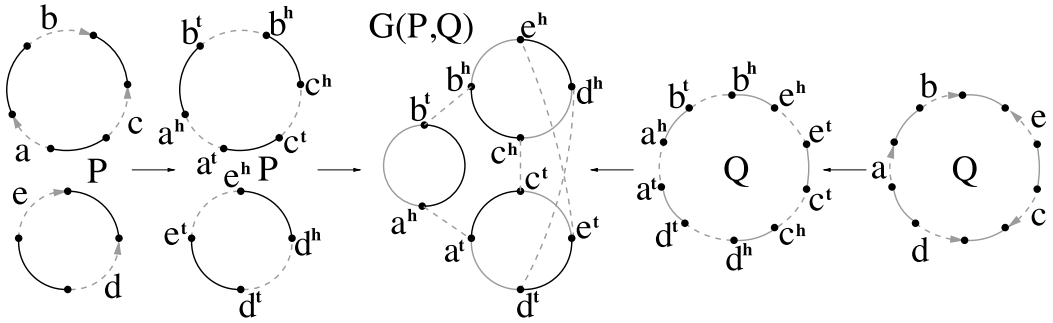

FIG. 1. The breakpoint graph $G(P, Q)$ of a two-chromosomal genome $P = (+a + b - c)(-d + e)$ and a unichromosomal genome $Q = (+a + b - e + c - d)$ represented as two black-obverse cycles and a gray-obverse cycle correspondingly.

(u, x), (v, y), or a pair of edges (u, y), (v, x). 2-Breaks correspond to standard rearrangement operations of reversals (Fig. 2a), fissions (Fig. 2b), or fusions/translocations[4] (Fig. 2c). 2-Break rearrangements can be generalized as follows. Given k black edges forming a matching on $2k$ vertices, define a *k-break* as replacement of these edges with a set of k black edges forming another matching on the same set of $2k$ vertices. Note that a 2-break is a particular case of a 3-break (as well as of a k-break for $k > 3$), in which case only two edges are replaced and the third one remains the same.

Let P and Q be two signed genomes on the same set of elements $\mathcal{G}$. The *breakpoint graph* $G(P, Q)$ 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. 1). Edges of each color form a matching on V: *obverse matching* (pairs of obverse vertices), *black matching* (adjacent vertices in P), and *gray matching* (adjacent vertices in Q). Every pair of matchings forms a collection of alternating cycles in $G(P, Q)$, called *black-gray*, *black-obverse*, and *gray-obverse* cycles respectively. The chromosomes of the genome P (resp. Q) can be read along black-obverse (resp. gray-obverse) cycles. The black-gray cycles in the breakpoint graph play an important role in analyzing rearrangements (Bafna and Pevzner, 1996). (For background information on genome rearrangements, see Chapter 10 in Pevzner [2000]).

2.1. Multi-break distance between circular genomes

The *k-break distance* between two genomes is defined as the minimum number of k-breaks required to transform one genome into the other. In contrast to the genomic distance (for linear multichromosomal genomes) (Hannenhalli and Pevzner, 1995; Tesler, 2002a; Ozery-Flato and Shamir, 2003), computing the 2-break distance for circular multichromosomal genomes is a trivial problem:

Theorem 1 (Yancopoulos et al., 2005; Alekseyev and Pevzner, 2007b, 2008a). *The 2-break distance between circular genomes P and Q is $d_2(P, Q) = |P| - c(P, Q)$ where $c(P, Q)$ is the number of black-gray cycles in $G(P, Q)$.*

While 2-breaks correspond to standard rearrangements, 3-breaks add transposition-like operations as well as 3-way fissions and fusions to the set of rearrangements (Fig. 2d). In contrast to the standard rearrangements (modelled as 2-breaks), transpositions introduce 3 breaks in the genome, making them notoriously difficult to analyze. Computing the minimum number of transpositions transforming one genome into another is called *sorting by transpositions*. A number of researchers considered transpositions in conjunction with other rearrangement operations (Walter et al., 1998; Gu et al., 1999; Lin and Xue, 2001; Meidanis et al., 2002; Hartman and Sharan, 2004; Lin et al., 2005; Radcliffe et al., 2005; Bader

[4]This definition of elementary rearrangement operations follows the standard definitions of reversals, translocations, fissions, and fusions for the case of circular chromosomes. For circular chromosomes fusions and translocations are not distinguishable.

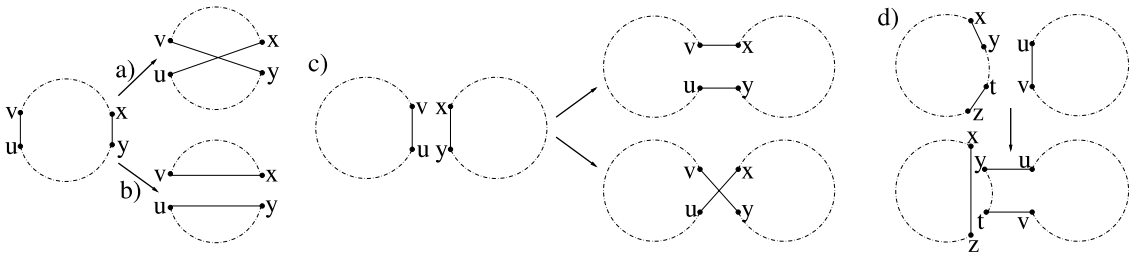

FIG. 2. Two-breaks on edges (u, v) and (x, y) corresponding to the following: **(a)** Reversal: the edges belong to the same black-obverse cycle that is rearranged after 2-break. **(b)** Fission: the edges belong to the same black-obverse cycle that is split by 2-break. **(c)** Translocation/fusion: the edges belong to different black-obverse cycles that are joined by 2-break. **(d)** Three-break on edges (u, v), (x, y) and (z, t) corresponding to transposition of a segment $y \ldots t$ from one chromosome to another.

and Ohlebusch, 2006). Despite many studies, the complexity of sorting by transpositions remains unknown (Bafna and Pevzner, 1998; Christie, 1999; Walter et al., 2003; Hartman, 2003; Elias and Hartman, 2005).

Let $c^{odd}(P, Q)$ be the number of black-gray cycles in the breakpoint graph $G(P, Q)$ with an odd number of black edges (*odd cycles*).

Theorem 2 (Alekseyev and Pevzner, 2007b, 2008a). *The 3-break distance between a black matching P and a gray matching Q is $d_3(P, Q) = \frac{|P|-c^{odd}(P,Q)}{2}$.*

The general formula as well as algorithms for computing the k-break distance can be found in Alekseyev and Pevzner (2008a).

2.2. Breakpoint re-use in circular genomes

If each of $d_3(P, Q)$ 3-breaks on a shortest evolutionary path from a circular genome P to a circular genome Q made 3 breaks (*complete* 3-breaks), it would result in a *breakpoint re-use rate* of $\frac{3 \cdot d_3(P,Q)}{|P|-c^{trivial}(P,Q)}$ where $c^{trivial}(P, Q)$ is the number of *trivial cycles* in $G(P, Q)$ (i.e., cycles consisting of parallel black and gray edges).[5] In reality, some 3-breaks can make two breaks (*incomplete* 3-breaks) as 2-breaks are particular cases of 3-breaks, reducing the estimate for the breakpoint re-use rate. Moreover, the minimum breakpoint re-use rate may be achieved on a suboptimal evolutionary path from P to Q.

We address the question of finding a transformation of one genome into the other by 3-breaks that makes the minimum number of individual breaks. The following theorem shows that there exists a series of $d_3(P, Q)$ 3-breaks that makes the minimum number of breaks while transforming P into Q:

Theorem 3 (Alekseyev and Pevzner, 2007a). *Any series of m k-breaks transforming a circular genome P into a circular genome Q makes at least $m + d_2(P, Q)$ breaks. Moreover, there exists a series of $d_3(P, Q)$ 3-breaks transforming P into Q that makes $d_3(P, Q) + d_2(P, Q)$ breaks.*

The following theorem shows how the minimum number of breaks in a series of 3-breaks transforming P into Q depends on the number of complete 3-breaks.

Theorem 4 (Alekseyev and Pevzner, 2007a). *Any series of 3-breaks with t complete 3-breaks, transforming a circular genome P into a circular genome Q, makes at least $d_2(P, Q) + \max\{d_2(P, Q) - t, d_3(P, Q)\}$ breaks. In particular, any such series of 3-breaks with $t \leq d_2(P, Q) - d_3(P, Q)$ complete 3-breaks makes at least $2d_2(P, Q) - t$ breaks.*

---

[5] Alternatively, the value of $|P| - c^{trivial}(P, Q)$ can be viewed as the number of *breakpoints* between the genomes P and Q. It is often assumed that no two blocks adjacent in P are adjacent in Q (and vice versa), a typical property of the blocks produced by synteny blocks generation algorithms (for two genomes). In this case $c^{trivial}(P, Q) = 0$ and the number of breakpoints is simply equal to the number of blocks $|P| = |Q|$.

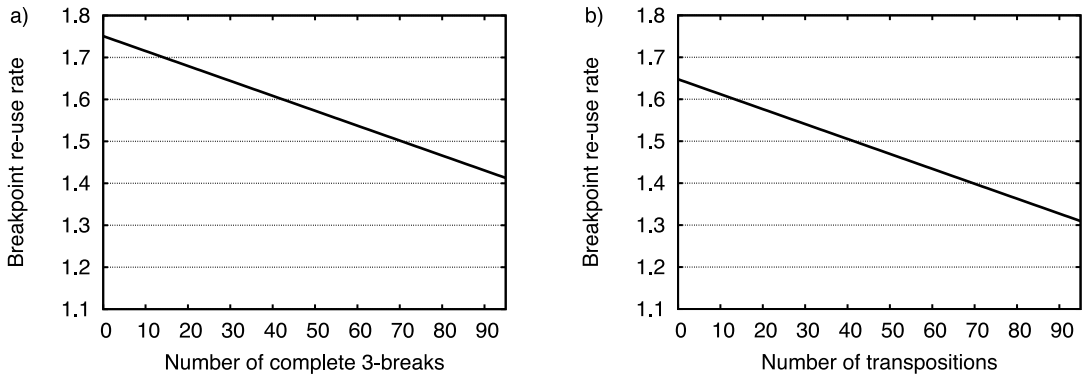

FIG. 3. The lower bound for the breakpoint re-use rate between the human and mouse genomes based on 281 synteny blocks from Pevzner and Tesler (2003a). The lower bound is represented as a function of the number of complete 3-breaks in a series of 3-breaks between the circularized human and mouse genomes (reproduced from Alekseyev and Pevzner, 2007a) **(a)** and transpositions in a series of rearrangements between the linear human and mouse genomes **(b)**.

Alekseyev and Pevzner (2007a) applied this theorem to the *Human* genome H and *Mouse* genome M consisting of the 281 synteny blocks from Pevzner and Tesler (2003a) under the assumption that all chromosomes are circular (*naive circularization*). In the next section, we will provide similar estimates for the more relevant case of linear genomes.

The breakpoint graph $G(H, M)$ contains 35 black-gray cycles, including three odd black-gray cycles, implying that $d_2(H, M) = 281 - 35 = 246$ (Theorem 1) and $d_3(H, M) = \frac{281-3}{2} = 139$ (Theorem 2). Theorem 3 implies that the minimum number of breaks for 2-break sorting H into M is $d_2(H, M) + d_2(H, M) = 246 + 246 = 492$ while for 3-break sorting the minimum number of breaks is $d_3(H, M) + d_2(H, M) = 139 + 246 = 385$.

Figure 3a gives the lower bound for the breakpoint re-use rate as a function of the number of complete 3-breaks in a series of rearrangements transforming H into M. Alekseyev and Pevzner (2007a) argue that in order to achieve a small breakpoint re-use rate between the genomes H and M, the number of complete 3-breaks must be large.

To estimate the breakpoint re-use rate as a function of the *proportion* of complete 3-breaks we need the following slightly stronger variant of Theorem 6 from Alekseyev and Pevzner (2007a) (the original proof of Theorem 6 in Alekseyev and Pevzner [2007a] works for this variant too):

Theorem 5 (Alekseyev and Pevzner, 2007a). *For any series of m 3-breaks with t complete 3-breaks, transforming a genome P into a genome Q, $m \geq \max\{d_2(P, Q) - t, d_3(P, Q)\}$. Moreover, there exists a series of $\max\{d_2(P, Q) - t, d_3(P, Q)\}$ 3-breaks transforming P into Q with $\min\{t, d_2(P, Q) - d_3(P, Q)\}$ complete 3-breaks.*

Theorem 5 immediately implies:

Corollary 6. *Let P and Q be circular genomes and t be a non-negative integer not exceeding $d_2(P, Q) - d_3(P, Q)$. A shortest series of 3-breaks with exactly t complete 3-breaks, transforming P into Q, consists of $d_2(P, Q) - t$ 3-breaks.*

Note that the restriction $t \leq d_2(P, Q) - d_3(P, Q)$ well covers the biologically reasonable cases. Indeed, any shortest series of 3-breaks transforming a genome P into a genome Q consists of $d_3(P, Q)$ 3-breaks out of which exactly $d_2(P, Q) - d_3(P, Q)$ 3-breaks are complete. A larger number of complete 3-breaks in a series can decrease neither the length of this series (i.e., the distance between P and Q) nor the total number of breaks made in it. Hence, we will focus on the case $t \leq d_2(P, Q) - d_3(P, Q)$ below.

Theorem 7. *Let P and Q be circular genomes and δ be a non-negative number not exceeding $\frac{d_2(P,Q)-d_3(P,Q)}{d_3(P,Q)}$. A series of 3-breaks with the proportion δ of complete 3-breaks, transforming P into Q, makes at least $\frac{2+\delta}{1+\delta} d_2(P, Q)$ breaks.*

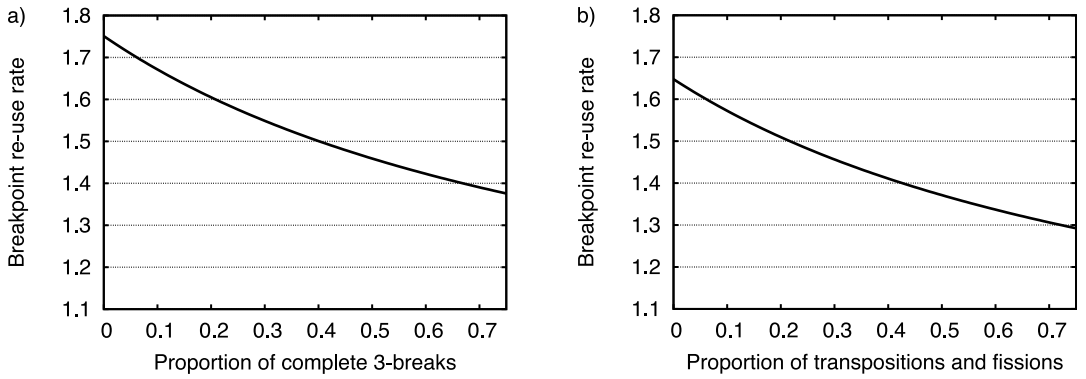

FIG. 4. The lower bound for the breakpoint re-use rate between the human and mouse genomes based on 281 synteny blocks from Pevzner and Tesler (2003a). The lower bound is represented as a function of the proportion of complete 3-breaks in a series of 3-breaks between the circularized human and mouse genomes **(a)** and transpositions and fissions in a series of rearrangements between the linear human and mouse genomes **(b)**.

Proof. Corollary 6 implies that for any shortest series of 3-breaks transforming a genome P into a genome Q with t complete 3-breaks, $t \leq d_2(P, Q) - d_3(P, Q)$, the proportion of complete 3-breaks in this series is $\delta = \frac{t}{d_2(P,Q)-t} \leq \frac{d_2(P,Q)-d_3(P,Q)}{d_3(P,Q)}$. Furthermore, the number of complete 3-breaks in the series can be expressed as $t = \frac{\delta}{1+\delta} d_2(P, Q)$, and Theorem 4 implies that the total number of breaks made is at least $\frac{2+\delta}{1+\delta} d_2(P, Q)$. To complete the proof, it is enough to notice that a series of the length d with the (fixed) proportion δ of complete 3-breaks makes $3\delta d + 2(d - \delta d) = (2 + \delta)d$ breaks in total and this number grows with d (hence, a shortest such series makes the smallest number of breaks). ■

For the human and mouse genomes, Theorem 7 implies the lower bound for the breakpoint re-use rate (as a function of the proportion of complete 3-breaks) shown in Figure 4a.

3. REARRANGEMENTS IN LINEAR GENOMES

A *linear genome* is a collection of linear chromosomes represented as sequences of signed elements (genes or synteny blocks). Similarly to circular genomes, we represent each linear chromosome on n elements as a sequence of n directed obverse edges (encoding elements and their direction) and $n - 1$ undirected black edges (connecting adjacent elements). So, each linear chromosome is an alternating *path* of obverse and black edges (starting and ending with obverse edges), and a linear genome is a collection of such paths.

Every linear genome P with m chromosomes has $2m$ vertices representing endpoints of the chromosomes. If we introduce an arbitrary perfect matching on these $2m$ vertices, consisting of black *closing edges*, the resulting graph will represent some circular genome that contains P as a subgraph. We call the resulting genome a *closure* of P and note that in general it is not uniquely defined. Black edges that belong to P are called *non-closing*.

Throughout this section, we assume that P and Q are linear genomes on the same set of elements.

3.1. Rearrangement distance between linear genomes

Let $d_2^l(P, Q)$ be the genomic distance between the genomes P and Q, i.e., the minimum number of reversals, translocations, fissions, and fusions required to transform P into Q. Also, let $d_3^l(P, Q)$ be the minimum number of reversals, translocations, fissions, and fusions as well as transpositions[6] required to transform P into Q.

[6]We do not consider 3-way fusions and 3-way fissions since such operations were never reported in biological literature.

Theorem 8. *Let P and Q be linear genomes. For any closure P' of the genome P, there exists a closure Q' of the genome Q such that $d_2^l(P, Q) \geq d_2(P', Q')$. Similarly, for any closure P' of the genome P, there exists a closure Q' of the genome Q such that $d_3^l(P, Q) \geq d_3(P', Q')$.*

Proof. Let S' be a closure of a linear genome S. We note that any reversal, translocation, fission, or fusions transforming the genome S into a linear genome T corresponds to a 2-break transforming the closure S' into some closure T' of the genome T (Fig. 5a–d). Similarly, any transposition transforming the genome S into a linear genome T corresponds to a 3-break transforming the closure S' into some closure T' of the genome T (Fig. 5e).

For the genomes P and Q, consider a series of $d_k^l(P, Q)$ ($k = 2$ or $k = 3$) rearrangements transforming P into Q. This series corresponds to a series of k-breaks transforming P' into some circular genome Q' that is a closure of the genome Q. To complete the proof it is sufficient to notice that the distance $d_k(P', Q')$ between the genomes P' and Q' does not exceed $d_k^l(P, Q)$, i.e., $d_k(P', Q') \leq d_k^l(P, Q)$. ■

Theorem 8 immediately implies:

Corollary 9. *For any linear genomes P and Q, $k = 2$ or $k = 3$,*

$$d_k^l(P, Q) \geq \max_{P'} \min_{Q'} d_k(P', Q')$$

$$d_k^l(P, Q) = d_k^l(Q, P) \geq \max_{Q'} \min_{P'} d_k(P', Q').$$

where P' and Q' vary over all possible closures of the genomes P and Q respectively.

Since the k-break distance $d_k(P', Q')$ ($k = 2$ or $k = 3$) gives a lower bound for the linear distance $d_k^l(P, Q)$, our goal is to make this bound as tight as possible by choosing appropriate closures P' and Q'. We start with defining the breakpoint graph of linear genomes and a number of its characteristics that we will find useful.

Let P' and Q' be closures of linear genomes P and Q. The breakpoint graph $G(P, Q)$ is defined as a result of removal of all closing edges from the breakpoint graph $G(P', Q')$ (of circular genomes P' and Q'). It is easy to see that $G(P, Q)$ is well-defined by the genomes P and Q and does not depend on a particular choice of closures P' and Q'. Every cycle in $G(P', Q')$ with m closing edges will be split into m paths in $G(P, Q)$. Therefore, the black-gray connected components of $G(P, Q)$ are formed by $c(P, Q)$ black-gray cycles and a number of black-gray paths. We distinguish between black-gray paths with both terminal edges of black color (*bb-paths*), with both terminal edges of gray color (*gg-paths*), and with terminal edges of different colors (*bg-paths*), including isolated vertices viewed as bg-paths with zero black and zero gray edges. We denote the number of such paths by $l_{bb}(P, Q)$, $l_{gg}(P, Q)$, and $l_{bg}(P, Q)$

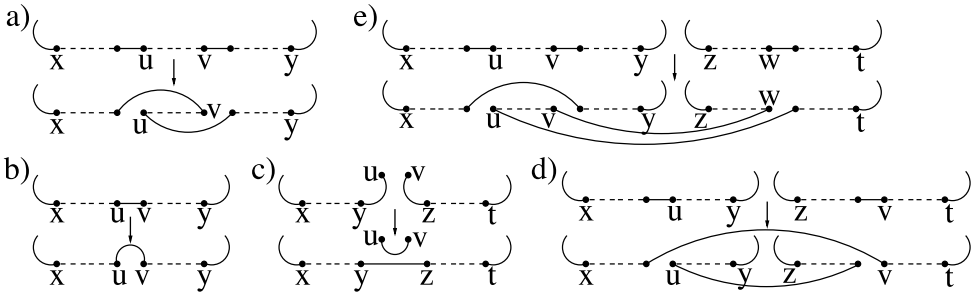

FIG. 5. Rearrangements of linear genomes correspond to k-breaks over closures: **(a)** Reversal of the region (u, v) is a 2-break over non-closing black edges. **(b)** Fission at the black edge (u, v) is the identity multi-break over the edge (u, v), re-claiming this edge as closing. **(c)** Fusion of the chromosomes endpoints y and z is a 2-break replacing closing edges (y, u) and (z, v) with a non-closing edge (y, z) and a closing edge (u, v). **(d)** Translocation exchanging chromosomes parts (u, y) and (v, t) is a 2-break operating over non-closing edges. **(e)** Transposition is a 3-break operating over non-closing edges.

respectively (note that the number $l_{bg}(P, Q)$ is always even). The total number of black-gray connected components in $G(P, Q)$ is

$$cc(P, Q) = c(P, Q) + l_{bb}(P, Q) + l_{gg}(P, Q) + l_{bg}(P, Q).$$

We also distinguish between black-gray connected components with odd/even number of black/gray edges and call them *b-odd*, *b-even*, *g-odd*, *g-even* respectively. To refer to the number of such components we will use these aliases as superscripts. Similarly to cycles, bg-paths have the same number of black and gray edges, so we call bg-paths simply *odd* and *even*, depending on the oddness of the number of black edges. We denote the number of such bg-paths by $l_{bg}^{odd}(P, Q)$ and $l_{bg}^{even}(P, Q)$ respectively. We rely on the following identities:

$$\forall j \in \{bb, bg, gg\}, \qquad l_j(P, Q) = l_j^{b\text{-}odd}(P, Q) + l_j^{b\text{-}even}(P, Q),$$

$$l_j(P, Q) = l_j^{g\text{-}odd}(P, Q) + l_j^{g\text{-}even}(P, Q);$$

$$\forall j \in \{bb, gg\}, \qquad l_j^{b\text{-}odd}(P, Q) = l_j^{g\text{-}even}(P, Q),$$

$$l_j^{b\text{-}even}(P, Q) = l_j^{g\text{-}odd}(P, Q).$$

These identities allow computation of all parameters in terms of $c(P, Q)$, $c^{odd}(P, Q)$, $l_{bb}(P, Q)$, $l_{bb}^{b\text{-}odd}(P, Q)$, $l_{gg}(P, Q)$, $l_{gg}^{b\text{-}odd}(P, Q)$, $l_{bg}(P, Q)$, $l_{bg}^{odd}(P, Q)$.

Similarly to the breakpoints graphs for circular and linear genomes, we can define breakpoint graphs and associated characteristics in the case when one genome is circular while the other is linear (in such a graph all paths are either bb-paths or gg-paths).

Lemma 10. *For a circular genome P' and a linear genome Q,*

$$\min_{Q'} d_2(P', Q') = |P'| - cc(P', Q) \quad \textit{and} \quad \min_{Q'} d_3(P', Q') = \frac{|P'| - cc^{b\text{-}odd}(P', Q)}{2}$$

where Q' varies over all possible closures of the genome Q.

Proof. Theorem 1 implies that $\min_{Q'} d_2(P', Q') = |P'| - \max_{Q'} c(P', Q')$. To maximize $c(P', Q')$, the closure Q' needs to be chosen in such a way that it closes each path in the breakpoint graph $G(P', Q)$ into a separate black-gray cycle. Therefore, $\max_{Q'} c(P', Q') = cc(P', Q)$.

Similarly, Theorem 2 implies that

$$\min_{Q'} d_3(P', Q') = \frac{|P'| - \max_{Q'} c^{odd}(P', Q')}{2}.$$

To maximize $c^{odd}(P', Q')$, the closure Q' needs to be chosen in such a way that it closes each b-odd path in the breakpoint graph $G(P', Q)$ into a separate black-gray cycle. Therefore, $\max_{Q'} c(P', Q') = cc^{b\text{-}odd}(P', Q)$. ■

Theorem 11. *For linear genomes P and Q, $\max_{P'} \min_{Q'} d_2(P', Q') = B_2(P, Q)$ where*

$$B_2(P, Q) = |P| - c(P, Q) - \max\{1, \frac{l_{bg}(P, Q)}{2}\} - l_{bb}(P, Q),$$

implying that $d_2^l(P, Q) \geq \max\{B_2(P, Q), B_2(Q, P)\}$.

Proof. By Lemma 10 we have $\max_{P'} \min_{Q'} d_2(P', Q') = |P| - \min_{P'} cc(P', Q)$. In order to minimize $cc(P', Q)$, the closure P' needs to be chosen in such a way that it minimizes the number of black-gray connected components in $G(P, Q)$. This can be done as follows. If $l_{bg}(P, Q) = 0$, then we will connect (using closing black edges) all the gg-paths into a single cycle. If $l_{bg}(P, Q) > 0$, we will

first connect a pair of bg-paths and all the gg-paths into a single bb-path, and then form pairs of the remaining bg-paths and connect bg-paths in each pair into a bb-path. As a result, $\min_{P'} cc(P', Q) = c(P, Q) + \max\{1, \frac{l_{bg}(P,Q)}{2}\} + l_{bb}(P, Q)$. Therefore, $\max_{P'} \min_{Q'} d_2(P', Q') = B_2(P, Q)$ and by Corollary 9, $d_2^l(P, Q) \geq B_2(P, Q)$. Moreover, since $d_2^l(P, Q) = d_2^l(Q, P) \geq B_2(Q, P)$, we have $d_2^l(P, Q) \geq \max\{B_2(P, Q), B_2(Q, P)\}$. ■

Lemma 12. *For linear genomes P and Q, $\min_{P'} cc^{b\text{-}odd}(P', Q) = L_3(P, Q)$ where*

$$L_3(P, Q) = c^{odd}(P, Q) + l_{bb}^{b\text{-}odd}(P, Q) + \delta(P, Q) + \max\left\{0, \frac{|l_{bg}^{odd}(P, Q) - l_{bg}^{even}(P, Q)|}{2} - l_{gg}^{b\text{-}even}(P, Q)\right\}$$

and

$$\delta(P, Q) = \max\left\{0, l_{gg}^{b\text{-}even}(P, Q) - \frac{|l_{bg}^{odd}(P, Q) - l_{bg}^{even}(P, Q)|}{2}\right\} \bmod 2.$$

Proof. Note that in any closure of P, the closing (black) edges connect gg-paths and bg-paths from $G(P, Q)$ into $m_1 = \frac{l_{bg}(P,Q)}{2}$ bb-paths and a number of cycles. Note that if $l_{bg}(P, Q) = 0$ then connecting all gg-paths into a single cycle (which will be odd iff $l_{gg}^{b\text{-}even}(P, Q)$ is odd) gives an optimal closure P'' (i.e., for which $\min_{P'} cc^{b\text{-}odd}(P', Q) = cc^{b\text{-}odd}(P'', Q)$). It is easy to check that in this case $cc^{b\text{-}odd}(P'', Q) = L_3(P, Q)$. For the rest of the proof, we assume that $l_{bg}(P, Q) > 0$.

We will show that there exists an optimal closure where the closing edges do not connect any gg-paths into a cycle. Such an optimal closure can be obtained from an arbitrary optimal closure P'' as explained below. Since $l_{bg}(P, Q) > 0$, the closing edges in $G(P'', Q)$ create at least one bb-path formed by two bg-paths at the ends and possibly gg-paths in the middle. We re-connect (modifying the set of closing edges) all the gg-paths from $G(P, Q)$, that are connected into cycles in $G(P'', Q)$, into the middle of this bb-path. Note that such modification of the closure may change the b-oddness of the affected bb-path but only if at least one of the destroyed cycles was odd. In any case the number of b-odd connected components is not increased. Therefore, the modified closure is optimal and satisfies the required property by construction. Without loss of generality we will assume that the closing edges create no cycles.

Bringing black closing edges into $G(P, Q)$ can be viewed as a two-step procedure: first, connecting gg-paths into longer gg-paths; and second, connecting pairs of bg-paths and maybe single gg-paths into bb-paths. Our goal is to minimize the number of b-odd bb-paths or, equivalently, to maximize the number of b-even bb-paths.

Consider an outcome of the first step. It is clear that connection of two b-odd gg-paths or two b-even gg-paths results in a b-odd gg-path, while connection of b-odd and b-even gg-paths results in a b-even gg-path. As we will see b-even gg-paths are more preferable than b-odd gg-paths. After the first step we can have up to $m_2 = l_{gg}^{b\text{-}even}(P, Q)$ b-even gg-paths.

Now, consider the second step. Connection of an odd bg-path and an even bg-path with an optional b-odd gg-path in between create a b-even bb-path. At the same time connection of a pair of odd bg-paths or a pair of even bg-paths requires a b-even gg-path in between in order to produce a b-even bb-path. All other combinations of bg-paths and gg-paths result in b-odd bb-paths.

We can create $m_3 = \min\{l_{bg}^{odd}(P, Q), l_{bg}^{even}(P, Q)\}$ b-even bb-paths without any use of gg-paths, and up to $m_4 = \frac{|l_{bg}^{odd}(P,Q) - l_{bg}^{even}(P,Q)|}{2}$ b-even bb-paths (note that $m_3 + m_4 = m_1$), each of which requires a b-even gg-path in the middle. Hence, we can create $m_5 = m_3 + \min\{m_4, m_2\} = \min\{m_1, m_2 + m_3\}$ b-even bb-paths. The other $m_6 = m_1 - m_5 = \max\{0, m_4 - m_2\}$ bb-paths (formed by pairs of bg-paths of the same oddness) will be b-odd. So far we have used $\min\{m_4, m_2\}$ b-even gg-paths. The other gg-paths (if any) can be connected (at the first step) into a single gg-path that is b-odd iff $m_2 - \min\{m_4, m_2\} = \max\{m_2 - m_4, 0\}$ is odd (i.e., $\delta(P, Q) = 1$). The b-odd gg-path can be easily incorporated into any of created bb-paths without changing its b-oddness. The b-even gg-path we have to incorporate into some of created b-even bb-paths and turn it into a b-odd bb-path. Hence, for an optimal closure P', there

are $l_{bb}^{b\text{-}odd}(P, Q) + m_6 + \delta(P, Q)$ b-odd bb-paths and $c^{odd}(P, Q)$ odd cycles in $G(P', Q)$, implying that $cc^{b\text{-}odd}(P', Q) = c^{odd}(P, Q) + l_{bb}^{b\text{-}odd}(P, Q) + m_6 + \delta(P, Q)$. ■

Theorem 13. *For linear genomes P and Q, $d_3^l(P, Q) \geq \max\{B_3(P, Q), B_3(Q, P)\}$ where $B_3(P, Q) = \frac{|P|-L_3(P,Q)}{2}$.*

Proof. Since $d_3^l(P, Q) = d_3^l(Q, P)$ it is sufficient to show that $d_3^l(P, Q) \geq B_3(P, Q)$. Corollary 9 and Lemma 10 imply

$$d_3^l(P, Q) \geq \max_{P'} \min_{Q'} d_3(P', Q') = \frac{|P| - \min_{P'} cc^{b\text{-}odd}(P', Q)}{2}.$$

Now, applying Lemma 12 completes the proof. ■

Since there is a particular interest in sorting unichromosomal genomes with reversals and transpositions (Walter et al., 1998; Meidanis et al., 2002), below we give a simplified version of Theorem 13 for unichromosomal genomes.

Corollary 14. *For unichromosomal linear genomes P and Q, $d_3^l(P, Q) \geq \frac{|P|-c^{odd}(P,Q)}{2} - 1$.*

Proof. Note that for unichromosomal genomes P and Q, the breakpoint graph $G(P, Q)$ contains exactly two black-gray paths with equal number of black and gray edges in total and two black terminal edges and two gray terminal edges. The two possible cases are considered below.

If $l_{bb}(P, Q) = l_{gg}(P, Q) = 1$ and $l_{bg}(P, Q) = 0$, then $L_3(P, Q) \leq c^{odd}(P, Q) + 2$ where the upper bound is attained for $l_{bb}^{b\text{-}odd}(P, Q) = l_{gg}^{b\text{-}even}(P, Q) = 1$.

If $l_{bb}(P, Q) = l_{gg}(P, Q) = 0$ and $l_{bg}(P, Q) = 2$, then $L_3(P, Q) \leq c^{odd}(P, Q) + 1$ where the upper bound is attained when both bg-paths are of the same oddness. ■

3.2. Breakpoint re-use in linear genomes

Pevzner and Tesler (2003b) emphasized the importance of proper analysis of rearrangements affecting the chromosome ends for computing the breakpoint re-use rates. If one ignores the specifics of breakpoint re-uses at the chromosomes ends, that is, external breakpoints (Pevzner and Tesler, 2003b), then the human genome can be transformed into the mouse genome without any breakpoint re-uses thus immediately "invalidating" all arguments in Pevzner and Tesler (2003b). Indeed, one can simply break the human genome into individual synteny blocks by fissions and further assemble these blocks into the mouse genome by fusions. This paradox emphasizes the potential pitfalls of ignoring external breakpoints and fissions/fusions in computing breakpoint re-use rates since in reality the described rearrangement scenario has a very large breakpoint re-use rate, thus confirming the conclusions in Pevzner and Tesler (2003b). Below we describe how to compute the breakpoint re-use rates for linear genomes.

Similarly to the case of circular genomes, we are interested in estimating the total number of breaks required to transform a linear genome P into a linear genome Q with reversals, fusions, fissions, translocations, and transpositions. According to Theorem 8, any series of such rearrangements corresponds to a series of 3-breaks transforming some closure P' of the genome P into some closure Q' of the genome Q. Let $b^c(P, Q)$ be the minimum number of breaks made in such a series of 3-breaks (over all possible closures P' and Q'). Theorems 8 and 3 imply:

Corollary 15. *For linear genomes P and Q,*

$$b^c(P, Q) \geq \max_{P'} \min_{Q'} d_3(P', Q') + d_2(P', Q')$$

$$b^c(P, Q) = b^c(Q, P) \geq \max_{Q'} \min_{P'} d_3(P', Q') + d_2(P', Q')$$

where P' and Q' vary over all possible closures of the genomes P and Q respectively.

To find out the exact value of $\max_{P'} \min_{Q'} d_3(P', Q') + d_2(P', Q')$, we need the following lemma:

Lemma 16. *For a circular genome P' and a linear genome Q,*

$$\min_{Q'} d_2(P', Q') + d_3(P', Q') = \frac{3}{2}|P'| - \frac{3cc^{b\text{-}odd}(P', Q) + 2cc^{b\text{-}even}(P', Q)}{2}.$$

Proof. Theorems 1 and 2 imply that

$$\min_{Q'} d_3(P', Q') + d_2(P', Q') = \frac{3}{2}|P'| - \frac{\max_{Q'} 3c^{odd}(P', Q') + 2c^{even}(P', Q')}{2}.$$

To maximize $3c^{odd}(P', Q') + 2c^{even}(P', Q')$, a closure Q' has to be chosen in such a way that it closes each path in the breakpoint graph $G(P', Q)$, into a separate black-gray cycle. Indeed, having $m > 1$ paths connected into a single cycle is always worse than connecting each of these paths into a separate cycle as $3 < 2m$. Therefore, for an optimal closure Q', we have $c^{odd}(P', Q') = cc^{b\text{-}odd}(P', Q)$ and $c^{even}(P', Q') = cc^{b\text{-}even}(P', Q)$. ■

Theorem 17. *For linear genomes P and Q, $\max_{P'} \min_{Q'} d_3(P', Q') + d_2(P', Q') = B_{23}(P, Q)$ where*

$$B_{23}(P, Q) = \frac{3}{2}|P| - c(P, Q) - l_{bb}(P, Q) - \frac{l_{bg}(P, Q) + L_3(P, Q)}{2},$$

implying that $b^c(P, Q) \geq \max\{B_{23}(P, Q), B_{23}(Q, P)\}$.

Proof. By Lemma 16 we have

$$\max_{P'} \min_{Q'} d_3(P', Q') + d_2(P', Q') = \frac{3}{2}|P| - \frac{\min_{P'} 3cc^{b\text{-}odd}(P', Q) + 2cc^{b\text{-}even}(P', Q)}{2}.$$

Note that if $l_{bg}(P, Q) = 0$ then connecting all gg-paths into a single cycle (which will be odd if $l_{gg}^{b\text{-}even}(P, Q)$ is odd) gives an optimal closure P'. It is easy to check that in this case $\max_{P'} \min_{Q'} d_3(P', Q') + d_2(P', Q') = B_{23}(P, Q)$. For the rest of the proof, we assume that $l_{bg}(P, Q) > 0$.

Note that in any closure of P, the closing (black) edges connect gg-paths and bg-paths from $G(P, Q)$ into bb-paths and cycles. We will show that in an optimal closure the closing edges do not connect any gg-paths into a cycle. Indeed, since $l_{bg}(P, Q) > 0$, the closing edges create at least one bb-path formed by two bg-paths at the ends and possibly gg-paths in the middle. It is easy to see that it is always better to include more gg-paths in the middle of this bb-path (maybe letting the objective function increase by one) rather than to create a separate cycle out of these gg-paths (in which case the objective function would increase by at least 2). Therefore, closing edges in an optimal closure P' connect gg-paths and bg-paths from $G(P, Q)$ into $\frac{l_{bg}(P,Q)}{2}$ bb-paths in $G(P', Q)$. As the total number of new bb-paths is fixed, the problem of minimizing $3cc^{b\text{-}odd}(P', Q) + 2cc^{b\text{-}even}(P', Q)$ is equivalent to minimizing $cc^{b\text{-}odd}(P', Q)$. For an optimal closure P', Lemma 12 gives $cc^{b\text{-}odd}(P', Q) = L_3(P, Q)$, implying that

$$\begin{aligned} 3cc^{b\text{-}odd}(P', Q) + 2cc^{b\text{-}even}(P', Q) &= 2cc(P', Q) + cc^{b\text{-}odd}(P', Q) \\ &= 2(c(P, Q) + l_{bb}(P, Q) + \frac{l_{bg}(P, Q)}{2}) + L_3(P, Q) \end{aligned}$$

and thus $\max_{P'} \min_{Q'} d_3(P', Q') + d_2(P', Q') = B_{23}(P, Q)$.

By Corollary 15 we have $b^c(P, Q) \geq B_{23}(P, Q)$ and $b^c(P, Q) \geq B_{23}(Q, P)$, implying that $b^c(P, Q) \geq \max\{B_{23}(P, Q), B_{23}(Q, P)\}$. ■

We will now prove the following analogs of Theorems 4 and 7:

Theorem 18. *Any series of rearrangements with t transpositions, transforming a linear genome P into a linear genome Q, makes at least* $\max\{2B_2(P,Q) - t, B_{23}(P,Q)\} - \mathrm{chr}(P) + \mathrm{chr}(Q)$ *breaks, where* $\mathrm{chr}(\cdot)$ *denotes the number of chromosomes. In particular, any such series of rearrangements with* $t \le 2B_2(P,Q) - B_{23}(P,Q)$ *transpositions makes at least* $2B_2(P,Q) - \mathrm{chr}(P) + \mathrm{chr}(Q) - t$ *breaks.*

Proof. Any series of rearrangements transforming the genome P into the genome Q corresponds to series of 3-breaks transforming any particular closure P' into a closure Q' (depending on P'). We note that every rearrangement makes the same number of breaks as the corresponding 3-break in the closures;[7] except for fusions that make smaller number of breaks than the corresponding 2-breaks in the closure (Fig. 5b), and for fissions that make breaks in linear genomes but correspond to identity multi-breaks (making no breaks) over the closures[8] (Fig. 5c).

Let u, v, t be respectively the number of fusions, fissions, and transpositions in a series of m rearrangements transforming the genome P into the genome Q and making b breaks in total. Then there is a series of 3-breaks, transforming a closure P' into a closure Q', that makes $b + u - v$ breaks in total. Since every fusion decreases the number of chromosomes by one, while every fission increases the number of chromosomes by one, $u - v = \mathrm{chr}(P) - \mathrm{chr}(Q)$. By Theorem 4,

$$b + u - v = b + \mathrm{chr}(P) - \mathrm{chr}(Q) \ge d_2(P', Q') + \max\{d_2(P', Q') - t, d_3(P', Q')\},$$

implying that $b \ge \max\{2d_2(P', Q') - t, d_2(P', Q') + d_3(P', Q')\} - \mathrm{chr}(P) + \mathrm{chr}(Q)$. Taking $\max_{P'} \min_{Q'}$ of the right hand side of this inequality, we have $b \ge \max\{2B_2(P,Q) - t, B_{23}(P,Q)\} - \mathrm{chr}(P) + \mathrm{chr}(Q)$. ■

Theorem 19. *Let P and Q be linear genomes and δ be a small enough non-negative number. Any series of rearrangements with the proportion δ of transpositions and fissions, transforming P into Q, makes at least* $\frac{2+\delta}{1+\delta} B_2(P,Q) - \mathrm{chr}(P) + \mathrm{chr}(Q)$ *breaks.*

Proof. Similarly to the proof of Theorem 18, any series of m rearrangements transforming the genome P into the genome Q we can map to a series of 3-breaks transforming any particular closure P' into a closure Q' (depending on P'). We note that under such a mapping transpositions correspond to complete 3-breaks, and vice versa.

Let u, v, t be respectively the number of fusions, fissions, and transpositions in a series of m rearrangements transforming the genome P into the genome Q and making b breaks in total. Since fissions (and only they) correspond to trivial 3-breaks, the proportion of complete 3-breaks in the corresponding series of 3-breaks transforming some closure P' into a closure Q' is

$$\delta' = \frac{t}{m - v} \le \frac{t + v}{m} = \delta.$$

By Theorem 7 and similarly to the proof of Theorem 18, we have

$$b + u - v = b + \mathrm{chr}(P) - \mathrm{chr}(Q) \ge \frac{2 + \delta'}{1 + \delta'} d_2(P', Q') \ge \frac{2 + \delta}{1 + \delta} d_2(P', Q'),$$

implying that $b \ge \frac{2+\delta}{1+\delta} d_2(P', Q') - \mathrm{chr}(P) + \mathrm{chr}(Q)$. Taking $\max_{P'} \min_{Q'}$ of the right hand side of this inequality, we have

$$b \ge \frac{2 + \delta}{1 + \delta} B_2(P, Q) - \mathrm{chr}(P) + \mathrm{chr}(Q).$$ ■

---

[7] We assume that a transposition always makes 3 breaks even if it transposes a part of chromosome starting with one of its ends, a translocation always makes 2 breaks even if it exchanges an entire chromosome with a part of another chromosome, and a reversal always makes 2 breaks even if it involves an end of a chromosome. The biological rationale for this assumption is that chromosomes are flanked by telomeres that, while remaining "invisible" in genomic sequences, can account for breakpoint re-use in the same way as any other genomic position.

[8] Sometimes it is convenient to assume that both fusion and fission make two breaks. The statements of Theorems 18 and 19 can be easily adjusted to address this case by removing the "$-\mathrm{chr}(P) + \mathrm{chr}(Q)$" term.

TABLE 1. PAIRWISE DISTANCES BETWEEN SIX MAMMALIAN GENOMES BASED ON 1360 SYNTENY BLOCKS FROM MA ET AL. (2006)

	Mouse	Rat	Dog	Chimp	Macaque
Human	411 : 410 : 397	756 : 718 : 699	295 : 304 : 284	23 : 22 : 22	110 : 106 : 104
Mouse		503 : 458 : 442	452 : 452 : 438	423 : 422 : 408	409 : 409 : 394
Rat			803 : 760 : 742	768 : 728 : 710	752 : 710 : 692
Dog				304 : 312 : 292	291 : 301 : 283
Chimp					117 : 115 : 113

Each cell contains three numbers $x : y : z$ where x is the genomic distance computed by GRIMM (Tesler, 2002b), y is the 2-break distance (Theorem 1) between the naive circularizations, and z is the lower bound for the genomic distance (Theorem 11).

TABLE 2. PAIRWISE REARRANGEMENT DISTANCES BETWEEN SIX MAMMALIAN GENOMES BASED ON 1360 SYNTENY BLOCKS FROM MA ET AL. (2006), WHERE REARRANGEMENT OPERATIONS INCLUDE REVERSALS, FISSIONS, FUSIONS, TRANSLOCATIONS, AND TRANSPOSITIONS

	Mouse	Rat	Dog	Chimp	Macaque
Human	255	461	201	19	77
Mouse		311	277	263	257
Rat			484	469	459
Dog				207	205
Chimp					86

Using the 281 synteny blocks between the linear human genome H and mouse genome M from Pevzner and Tesler (2003a), we estimate the breakpoint re-use rate across these (linear) genomes. The breakpoint graph $G(H, M)$ have the following parameters: $(c, l_{bb}, l_{gg}, l_{bg}) = (28, 12, 15, 16)$, $(c^{odd}, l_{bb}^{b\text{-}odd}, l_{gg}^{b\text{-}odd}, l_{bg}^{odd}) = (2, 5, 4, 3)$, $\mathrm{chr}(H) = 23$, $\mathrm{chr}(M) = 20$, $B_2(H, M) = 233$, $B_2(M, H) = 230$, $B_3(H, M) = 137$, $B_2(M, H) = 134$, $B_{23}(H, M) = 370$, and $B_{23}(M, H) = 364$. Theorems 11 and 13 imply that

$$d_2^l(H, M) \geq \max\{B_2(H, M), B_2(M, H)\} = 233,$$

$$d_3^l(H, M) \geq \max\{B_3(H, M), B_3(M, H)\} = 137.$$

For the human and mouse genomes, Theorem 18 gives the lower bound for the breakpoint re-use rate as a function of the number of transpositions shown in Figure 3b, while Theorem 19 gives the similar lower bound as a function of the proportion of transpositions and fissions shown in Figure 4b. This illustrates that a very large number of transpositions would be necessary to bring the breakpoint re-use rate below 1.25 rate expected for RBM (Pevzner and Tesler, 2003b; Alekseyev and Pevzner, 2007a).

4. CONCLUSION

To estimate the tightness of the obtained lower bounds, we computed pairwise rearrangement distances between six mammalian genomes (Table 1). The results suggest that the 2-break distance between naive circularizations is usually close to the genomic distance, providing a rationale for using circularized genomes in the rearrangement analysis. The lower bounds for the transposition-like rearrangements (Theorems 13, 17, 18, and 19), are particularly important since the exact algorithms for transposition analysis are unknown (Table 2). The lower bounds for the number of breakpoint re-uses between the *human* and *mouse* genomes are illustrated in Figures 3 and 4. The difference in the number of breakpoint re-uses between the linear genomes and their naive circularizations appears to be small.

We remark that our key algorithmic technique for analyzing linear genomes is a transformation of rearrangement scenarios in linear genomes into scenarios (of the same length) in their circularizations (Theorem 8). While a reverse transformation would allow one to prove upper bounds in a similar way, it is not clear how to transform a rearrangement scenario in circularized genomes into a scenario in the original linear genomes. In particular, it is not clear how to avoid intermediate circular chromosomes that are not allowed in a linear scenario. Finding a transformation of a circular scenario into a linear scenario with the minimum increase in the number of rearrangements remains an open problem.

ACKNOWLEDGMENTS

The author thanks Jian Ma for providing the synteny blocks for mammalian genomes and Pavel Pevzner for a number of inspiring discussions.

DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Alekseyev, M.A., and Pevzner, P.A. 2007a. Are there rearrangement hotspots in the human genome? *PLoS Comput. Biol.* 3, e209.

Alekseyev, M.A., and Pevzner, P.A. 2007b. Whole genome duplications, multi-break rearrangements, and genome halving theorem. *SODA* 665–679.

Alekseyev, M.A., and Pevzner, P.A. 2008a. Multi-break rearrangements and chromosomal evolution. *Theoret. Comput. Sci.* 395, 193–202.

Alekseyev, M.A., and Pevzner, P.A. 2008b. Where are the rearrangement hotspots in the human genome? (in preparation).

Bader, M., and Ohlebusch, E. 2006. Sorting by weighted reversals, transpositions, and inverted transpositions. *RECOMB 2006* 563–577.

Bafna, V., and Pevzner, P.A. 1996. Genome rearrangement and sorting by reversals. *SIAM J. Comput.* 25, 272–289.

Bafna, V., and Pevzner, P.A, 1998. Sorting permutations by transpositions. *SIAM J. Discrete Math.* 11, 224–240.

Bailey, J., Baertsch, R., Kent, W., et al. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol.* 5, R23.

Bergeron, A., Mixtacki, J., and Stoye, J. 2006. A unifying view of genome rearrangements. *Lect. Notes Comput. Sci.* 4175, 163–173.

Caceres, M., Sullivan, R.T., and Thomas, J.W. 2007. A recurrent inversion on the eutherian X chromosome. *Proc. Natl. Acad. Sci. USA* 104, 18571–18576.

Christie, D.A. 1999. Genome rearrangement problems [Ph.D. dissertation]. University of Glasgow, Glasgow.

Elias, I., and Hartman, T. 2005. A 1.375-approximation algorithm for sorting by transpositions. *Lect. Notes Comput. Sci.* 3692, 204–214.

Gordon, L., Yang, S., Tran-Gyamfi, M., et al. 2007. Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Res.* 17, 1603–1613.

Gu, Q.P., Peng, S., and Sudborough, H. 1999. A 2-approximation algorithm for genome rearrangements by reversals and transpositions. *Theoret. Comput. Sci.* 210, 327–339.

Hannenhalli, S., and Pevzner, P. 1995. Transforming men into mouse (polynomial algorithm for genomic distance problem). *Proc. 36th Annu. Symp. Found. Comput. Sci.* 581–592.

Hartman, T. 2003. A simpler 1.5-approximation algorithm for sorting by transpositions. *Lect. Notes Comput. Sci.* 2676, 156–169.

Hartman, T., and Sharan, R. 2004. A 1.5-approximation algorithm for sorting by transpositions and transreversals. *Lect. Notes Comput. Sci.* 3240, 50–61.

Hinsch, H., and Hannenhalli, S. 2006. Recurring genomic breaks in independent lineages support genomic fragility. *BMC Evol. Biol.* 6, 90.

Kikuta, H., Laplante, M., Navratilova, P., et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* 17, 545–555.

Levy, D., Vazquez, M., Cornforth, M., et al. 2004. Comparing DNA damage-processing pathways by computer analysis of chromosome painting data. *J. Comput. Biol.* 11, 626–641.

Lin, G.H., and Xue, G. 2001. Signed genome rearrangements by reversals and transpositions: models and approximations. *Theoret. Comput. Sci.* 259, 513–531.

Lin, Y.C., Lu, C.L., Chang, H.-Y., et al. 2005. An efficient algorithm for sorting by block-interchanges and its application to the evolution of Vibrio species. *J. Comput. Biol.* 12, 102–112.

Ma, J., Zhang, L., Suh, B.B., et al. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16, 1557–1565.

Mehan, M.R., Almonte, M., Slaten, E., et al. 2007. Analysis of segmental duplications reveals a distinct pattern of continuation-of-synteny between human and mouse genomes. *Hum. Genet.* 121, 93–100.

Meidanis, J., Walter, M.M., and Dias, Z. 2002. A lower bound on the reversal and transposition diameter. *J. Comput. Biol.* 9, 743–745.

Murphy, W.J., Larkin, D.M., van der Wind, A.E., et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative map. *Science* 309, 613–617.

Nadeau, J.H., and Taylor, B.A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* 81, 814–818.

Ozery-Flato, M., and Shamir, R. 2003. Two notes on genome rearrangement. *J. Bioinform. Comput. Biol.* 1, 71–94.

Pevzner, P., and Tesler, G. 2003a. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13, 37–45.

Pevzner, P.A. 2000. *Computational Molecular Biology: An Algorithmic Approach*. The MIT Press, Cambridge, MA.

Pevzner, P.A., and Tesler, G. 2003b. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 100, 7672–7677.

Radcliffe, A.J., Scott, A.D., and Wilmer, E.L. 2005. Reversals and transpositions over finite alphabets. *SIAM J. Discrete Math.* 19, 224–244.

Ruiz-Herrera, A., Castresana, J., and Robinson, T.J. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol.* 7, R115.

Sachs, R.K., Arsuaga, J., Vazquez, M., et al. 2002. Using graph theory to describe and model chromosome aberrations. *Radiat. Res.* 158, 556–567.

Sachs, R.K., Levy, D., Hahnfeldt, P., et al. 2004. Quantitative analysis of radiation-induced chromosome aberrations. *Cytogenet. Genome Res.* 104, 142–148.

Tesler, G. 2002a. Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.* 65, 587–609.

Tesler, G. 2002b. GRIMM: genome rearrangements web server. *Bioinformatics* 18, 492–493.

van der Wind, A.E., Kata, S.R., Band, M.R., et al. 2004. A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome Res.* 14, 1424–1437.

Vazquez, M., Greulich-Bode, K.M., Arsuaga, J., et al. 2002. Computer analysis of mFISH chromosome aberration data uncovers an excess of very complicated metaphases. *Int. J. Radiat. Biol.* 78, 1103–1115.

Walter, M.E., Dias, Z., and Meidanis, J. 1998. Reversal and transposition distance of linear chromosomes. *SPIRE* 96–102.

Walter, M.E., Reginaldo, L., Curado, A.F., et al. 2003. Working on the problem of sorting by transpositions on genome rearrangements. *Lect. Notes Comput. Sci.* 2676, 372–383.

Webber, C., and Ponting, C.P. 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res.* 15, 1787–1797.

Yancopoulos, S., Attie, O., and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21, 3340–3346.

Zhao, S., Shetty, J., Hou, L., et al. 2004. Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res.* 14, 1851–1860.

Address reprint requests to:
Dr. Max A. Alekseyev
Department of Computer Science and Engineering
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093

E-mail: maxal@cs.ucsd.edu