JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 7, 2008
© Mary Ann Liebert, Inc.
Pp. 1–20
DOI: 10.1089/cmb.2007.0136

Sorting Genomes with Centromeres by Translocations

MICHAL OZERY-FLATO and RON SHAMIR

ABSTRACT

A *centromere* is a special region in the chromosome that plays a vital role during cell division. Every new chromosome created by a genome rearrangement event must have a centromere in order to survive. This constraint has been ignored in the computational modeling and analysis of genome rearrangements to date. Unlike genes, the different centromeres are indistinguishable, they have no orientation, and only their location is known. A prevalent rearrangement event in the evolution of multi-chromosomal species is translocation (i.e., the exchange of tails between two chromosomes). A translocation may create a chromosome with no centromere in it. In this paper, we study for the first time centromeres-aware genome rearrangements. We present a polynomial time algorithm for computing a shortest sequence of translocations transforming one genome into the other, where all of the intermediate chromosomes must contain centromeres. We view this as a first step towards analysis of more general genome rearrangement models that take centromeres into consideration.

Key words: sorting by translocations, genome rearrangements, comparative genomics, combinatorics.

1. INTRODUCTION

Genomes of related species tend to have similar genes that are, however, ordered differently. The distinct orderings of the genes are the result of genome rearrangements. Inferring the sequence of genome rearrangements that took place during the course of evolution is an important question in comparative genomics. The genomes of higher organisms, such as plants and animals, are partitioned into continuous units called *chromosomes*. Every chromosome contains a special region called *a centromere*, which plays a vital role during cell division. An *acentric* chromosome (i.e., one that lacks a centromere) is likely to be lost during subsequent cell divisions (Sullivan et al., 2001). Thus, a rearrangement scenario that preserves a centromere in each chromosome is more biologically realistic than one that does not. The computational studies on genome rearrangements to date have ignored the existence and role of centromeres. Hence, the rearrangement scenarios for multi-chromosomal genomes produced by current algorithms may include genomes with non-viable chromosomes. In this study, we begin to address the centromeres in the computational analysis of genome rearrangements.

School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel.

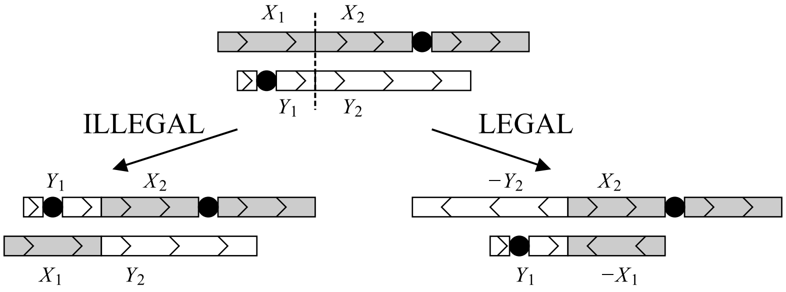


FIG. 1. An example of legal and illegal translocations for a certain cut of two chromosomes. The black circles denote the location of the centromeres; the broken line indicates the positions where the two chromosomes were cut.

Since sequencing a centromere is almost impossible due to the repeated sequences it contains, the only information we have on a centromere is its location in the genome. Therefore, in the model we define, centromeres appear as anonymous and orientation-less elements. We say that a genome is *legal* if each of its chromosomes contains a single centromere. A *legal* rearrangement operation results in a legal genome (Fig. 1). The *legal rearrangement sorting problem* is defined as follows: given two legal genomes A and B, find a shortest sequence of legal rearrangement operations that transforms A into B. The length of this sequence is the *legal distance* between A and B.

A *reciprocal translocation* is a rearrangement in which two chromosomes exchange non-empty ends. A reciprocal translocation results in an illegal genome if exactly one of the exchanged ends contains a centromere. In this paper, we focus on the problem of legal sorting by reciprocal translocations (LSRT). This problem is a refinement of the "sorting by reciprocal translocations" problem (SRT), which ignores centromeres. SRT was studied in Hannenhalli (1996), Bergeron et al. (2006), and Ozery-Flato and Shamir (2006a,b), and is solvable in polynomial time. Clearly, a solution to SRT may not be a solution to LSRT, since 50% of the possible reciprocal translocations are illegal (Fig. 1). Indeed, in many cases, more rearrangements are needed in order to legally sort a genome.

In this study we present a polynomial time algorithm for LSRT. The basic idea is to transform LSRT into SRT, by replacing pairs of centromeres in the two genomes by new unique oriented elements. Our algorithm is based on finding a mapping between the centromeres of the two given genomes such that the solution to the resulting SRT instance is minimum. We show that an optimal mapping can be found in polynomial time. To the best of our knowledge, this is the first rearrangement algorithm that considers centromeres. While a model that permits only reciprocal translocations is admittedly quite remote from the biological reality, we hope that the principles and structure revealed here will be instrumental for analyzing more realistic models in the future. One additional advantage of centromere-aware models is that they restrict drastically the allowed sequences of operations, and therefore are less likely to suffer from high multiplicity of optimal sequences.

The paper is organized as follows. Section 2 gives the necessary preliminaries. In Section 3, we model LSRT and present some elementary properties of it. Section 4 describes an exponential algorithm for LSRT, which searches for an optimal mapping between the centromeres of A and B (i.e., one that leads to a minimum SRT solution). In Section 5, we take a first step towards a polynomial time algorithm for LSRT by proving a bound that is at most two translocations away from the legal translocation distance. In Section 6, we present a theorem leading to a polynomial time algorithm for computing the legal translocation distance and solving LSRT.

A preliminary version of this study appeared in the proceedings of RECOMB 2007 (Ozery-Flato and Shamir, 2007).

2. PRELIMINARIES

This section provides the needed background for SRT. The definitions follow previous literature on translocations (Hannenhalli, 1996; Bergeron et al., 2006; Ozery-Flato and Shamir, 2006a, 2006b). In the model we consider, a *genome* is a set of chromosomes. A *chromosome* is a sequence of genes. A *gene* is identified by a positive integer. All genes in the genome are distinct. When it appears in a genome, a gene is assigned a sign of plus or minus. The following is an example of a genome with two chromosomes and six genes: $\{(1, -5), (-4, -3, -2, 6)\}$.

The *reverse* of a sequence of genes $I = (x_1, \ldots, x_l)$ is $-I = (-x_l, \ldots, -x_1)$. Two chromosomes, X and Y, are called *identical* if either $X = Y$ or $X = -Y$. Therefore, *flipping* chromosome X into $-X$ does not affect the chromosome it represents.

Let $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ be two chromosomes, where X_1, X_2, Y_1, Y_2 are sequences of genes. A *translocation* cuts X into X_1 and X_2 and Y into Y_1 and Y_2 and exchanges segments between the chromosomes. It is called *reciprocal* if X_1,X_2, Y_1 and Y_2 are all non-empty. There are two types of translocations on X and Y. A *prefix-suffix* translocation switches X_1 with Y_2:

$$(\underline{X_1}, X_2), (Y_1, \underline{Y_2}) \Rightarrow (-Y_2, X_2), (Y_1, -X_1).$$

A *prefix-prefix* translocation switches X_1 with Y_1:

$$(\underline{X_1}, X_2), (\underline{Y_1}, Y_2) \Rightarrow (Y_1, X_2), (X_1, Y_2).$$

Note that we can mimic one type of translocation by a flip of one of the chromosomes followed by a translocation of the other type.

For a chromosome $X = (x_1, \ldots, x_k)$, define $\mathit{Tails}(X) = \{x_1, -x_k\}$. Note that flipping X does not change $\mathit{Tails}(X)$. For a genome A, define $\mathit{Tails}(A) = \bigcup_{X \in A} \mathit{Tails}(X)$. For example:

$$\mathit{Tails}(\{(1, -3, -2, 4, -7, 8), (6, 5)\}) = \{1, -8, 6, -5\}.$$

Two genomes A_1 and A_2 are *co-tailed* if $\mathit{Tails}(A_1) = \mathit{Tails}(A_2)$. In particular, two co-tailed genomes have the same number of chromosomes. Note that if A_2 was obtained from A_1 by performing a reciprocal translocation, then $\mathit{Tails}(A_2) = \mathit{Tails}(A_1)$. Therefore, SRT is solvable only for genomes that are co-tailed. For the rest of this paper, the word “translocation” refers to a reciprocal translocation, and we assume that the given genomes, A and B, are co-tailed. Denote the set of tails of A and B by *Tails*.

2.1. Cycle graph

Let n and N be the number of genes and chromosomes in A (equivalently, B), respectively. We shall always assume that both A and B consist of the genes $\{1, \ldots, n\}$. The *cycle graph* of A and B, denoted $G(A, B)$, is defined as follows. The set of vertices is $\bigcup_{i=1}^{n} \{i^0, i^1\}$. The vertices i^0 and i^1 are called the two *ends* of gene i (think of them as ends of a small arrow directed from i^0 to i^1). For every two genes, i and j, where j immediately follows i in some chromosome of A (respectively, B) add a black (respectively, gray) edge $(i, j) \equiv (out(i), in(j))$, where $out(i) = i^1$ if i has a positive sign in A (respectively, B) and otherwise $out(i) = i^0$, and $in(j) = j^0$ if j has a positive sign in A (respectively, B) and otherwise $in(j) = j^1$. An example is given in Figure 2a. There are $n - N$ black edges and $n - N$ gray edges in $G(A, B)$. A gray edge (i, j) is *external* if the genes i and j belong to different chromosomes of A, otherwise it is *internal*. A cycle is *external* if it contains an external edge, otherwise it is *internal*.

Every vertex in $G(A, B)$ has degree 2 or 0, where vertices of degree 0 (isolated vertices) belong to *Tails*. Therefore, $G(A, B)$ is uniquely decomposed into cycles with alternating gray and black edges. An *adjacency* is a cycle with two edges. A *breakpoint* is a black edge that is not part of an adjacency.

2.2. Overlap graph with chromosomes

A *signed permutation* $\pi = (\pi_1, \ldots, \pi_n)$ is a permutation on the integers $\{1, \ldots, n\}$, where a sign of plus or minus is assigned to each number. If A is a genome with the set of genes $\{1, \ldots, n\}$ then any concatenation π_A of the chromosomes of A is a signed permutation of size n.

Place the vertices of $G(A, B)$ along a straight line according to their order in π_A. Now, every gray edge and every chromosome is associated with an interval of vertices in $G(A, B)$. Two intervals *overlap* if their intersection is not empty but none contains the other. The *overlap graph with chromosomes* of A and B w.r.t. π_A, denoted $\mathit{OVCH}(A, B, \pi_A)$, is defined as follows. The set of nodes is the set of gray

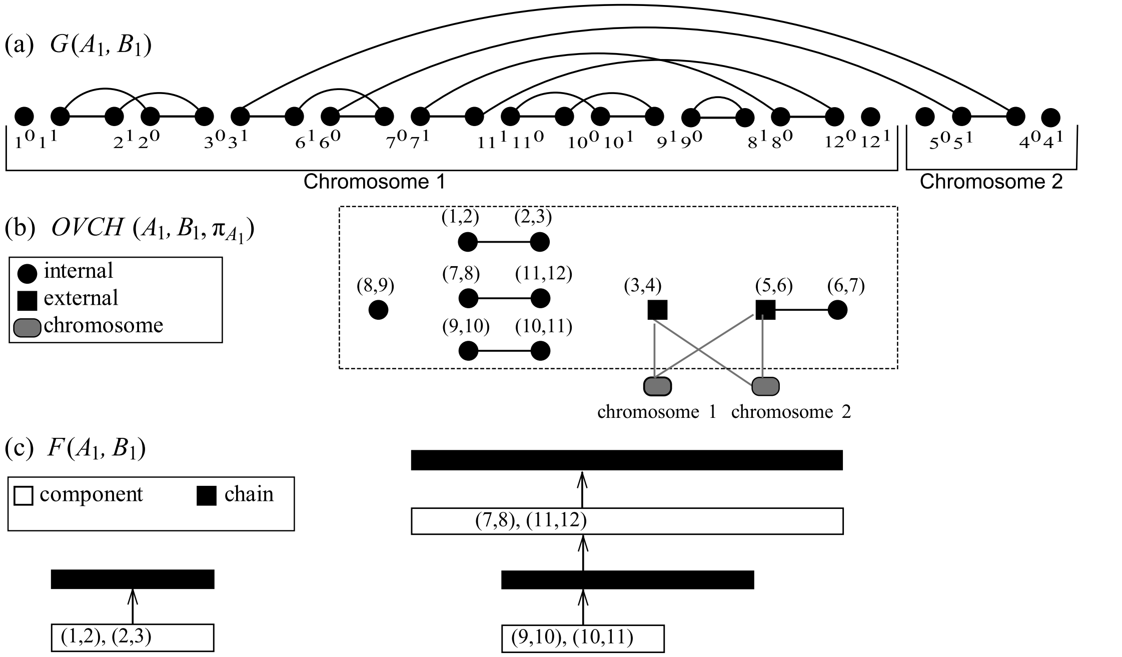


FIG. 2. Auxiliary graphs for $A_1 = \{(1, -2, 3, -6, 7, -11, 10, -9, -8, 12), (5, 4)\}$, $B_1 = \{(1, \ldots, 4), (5, \ldots, 12)\}$ ($\pi_{A_1} = (1, -2, 3, -6, 7, -11, 10, -9, -8, 12, 5, 4)$). **(a)** The cycle graph. Black edges are horizontal; gray edges are curved. **(b)** The overlap graph with chromosomes. The graph induced by the vertices within the dashed rectangle is $OV(A_1, B_1, \pi_{A_1})$. **(c)** The forest of internal components.

edges and chromosomes in $G(A, B)$. Two nodes are connected if their corresponding intervals overlap. An example is given in Figure 2b. This graph is an extension of the overlap graph of a signed permutation defined in (Kaplan et al., 2000). Let $OV(A, B, \pi_A)$ be the subgraph of $OVCH(A, B, \pi_A)$ induced by the set of nodes that correspond to gray edges (i.e., excluding the chromosomes' nodes). We shall use the word "component" for a connected component of $OV(A, B, \pi_A)$.

In order to prevent confusion, we will refer to nodes that correspond to chromosomes as "chromosomes" and reserve the word "vertex" for nodes that correspond to gray edges. A vertex is external (resp. internal) if it corresponds to an external (resp. internal) gray edge. Obviously a vertex is external iff it is connected to a chromosome. A component is *external* if it contains an external vertex, otherwise it is *internal*. A component is *trivial* if it is composed of one (internal) vertex. A trivial component corresponds to an adjacency. Note that the internal/external state of a vertex in $OVCH(A, B, \pi_A)$ does not depend on π_A. Therefore, the set of internal components in $OVCH(A, B, \pi_A)$ is independent of π_A. The *span* of a component M is the minimal interval of genes $I(M) = [i, j] \subset \pi_A$ that contains the interval of every vertex in M. Clearly, $I(M)$ is independent of π_A iff M is internal. The following lemma follows from A and B being co-tailed and (Corollary 2.2 in Kaplan et al., 2000):

Lemma 1. *Every internal component corresponds to the set of gray edges of a union of cycles in* $G(A, B)$.

The set of internal components can be computed in linear time using an algorithm in Bader et al. (2001).

2.3. Forest of internal components

$(M_1, \ldots, M_t)$ is a *chain* of components if $I(M_j)$ and $I(M_{j+1})$ overlap in exactly one gene for $j = 1, \ldots, t-1$. The *forest of internal components* (Bergeron et al., 2006), denoted $F(A, B)$, is defined as follows. The vertices of $F(A, B)$ are *(i)* the non-trivial internal components and *(ii)* every maximal chain of internal components that contains at least one non-trivial component. Let M and C be two vertices in $F(A, B)$ where M corresponds to a component and C to a chain. $M \to C$ is an edge of $F(A, B)$ if $M \in C$. $C \to M$ is an edge of $F(A, B)$ if $I(C) \subset I(M)$ and $I(M)$ is minimal (Fig. 2c). We will refer to a component that is a leaf in $F(A, B)$ as simply a *leaf*.

2.4. Reciprocal translocation distance

The *reciprocal translocation distance* between A and B is the length of a shortest sequence of reciprocal translocations that transforms A into B. Let $c(A, B)$ denote the number of cycles in $G(A, B)$. Let $|F(A, B)|$ and $l(A, B)$ denote the number of trees and leaves in $F(A, B)$, respectively. Obviously $|F(A, B)| \leq l(A, B)$. Define

$$\delta(A, B) \equiv \delta(F(A, B)) = \begin{cases} 2 & \text{if } |F(A, B)| = 1 \text{ and } l(A, B) \text{ is even} \\ 1 & \text{if } l(A, B) \text{ is odd} \\ 0 & \text{otherwise } (|F(A, B)| \neq 1 \text{ and } l(A, B) \text{ is even}) \end{cases}$$

Theorem 1 (Bergeron et al., 2006; Hannenhalli, 1996). *The reciprocal translocation distance between A and B is $n - N - c(A, B) + l(A, B) + \delta(A, B)$.*

Let Δc denote the change in the number of cycles after performing a translocation on A. Then $\Delta c \in \{-1, 0, 1\}$ (Hannenhalli, 1996). A translocation is *proper* if $\Delta c = 1$, *improper* if $\Delta c = 0$ and *bad* if $\Delta c = -1$.

Corollary 1. *Every translocation in a shortest sequence of translocations transforming A into B is either proper or bad.*

Proof. An improper translocation cannot decrease the translocation distance since it does not affect any parameter in its formula. ■

3. INCORPORATING CENTROMERES INTO A GENOME

We extend the model described above by adding the requirement that every genome is legal (i.e., every chromosome contains exactly one centromere). We denote the location of a centromere in a chromosome by the element $\bullet$. The element $\bullet$ is unsigned and thus does not change under chromosome flips. The following is an example of a legal genome: $\{(1, 2, 3, \bullet, 4), (\bullet, 5, 6)\}$. The set of tails is defined for regular elements, thus $Tails(\bullet, 5, 6) = \{5, -6\}$. We assume that a cut of a chromosome does not split a centromere. Clearly, for every cut of two chromosomes one translocation is legal while the other is not (Fig. 1).

3.1. A new precondition

We present here a simple condition for the solvability of LSRT. If this condition is not satisfied then A cannot be transformed into B by legal translocations. For chromosome $X = (x_1, \ldots, x_i, \bullet, x_{i+1}, \ldots, x_k)$ define $Elements(X) = \{x_1, \ldots, x_i, -x_{i+1}, \ldots, -x_k\}$. Note that $Elements(X) = Elements(-X)$. For genome A we define $Elements(A) = \bigcup_{X \in A} Elements(X)$. For example:

$$Elements(\{(1, 2, \bullet, 3, 4), (\bullet, 5, 6)\}) = \{1, 2, -3, -4, -5, -6\}.$$

Observation 1. *Let A and B be two legal genomes. If A can be transformed into B by a sequence of legal translocations then $Elements(A) = Elements(B)$.*

We will see later that this condition is also sufficient. Thus, for the rest of this paper we assume that the input to LSRT is co-tailed genomes A and B satisfying $Elements(A) = Elements(B) = Elements$. The cycle graph of A and B, $G(A, B)$, ignores the $\bullet$ elements.

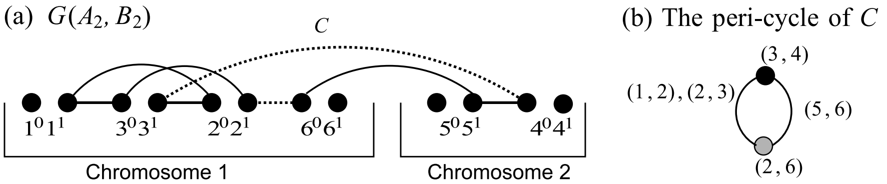


FIG. 3. Pericentric edges and peri-cycles. $A_2 = \{(1, 3, 2, \bullet, 6), (\bullet, 5, 4)\}$, $B_2 = \{(1, 2, 3, \bullet, 4), (\bullet, 5, 6)\}$. **(a)** The cycle graph $G(A_2, B_2)$. Pericentric edges are denoted by dotted lines. **(b)** The peri-cycle of the single cycle in $G(A_2, B_2)$. The labels of the edges denote the set of gray edges in the corresponding paths.

3.2. On the gap between the legal distance and the "old" distance

Let $d(A, B)$ denote the legal translocation distance between A and B. Let $d_{\text{old}}(A, B)$ denote the translocation distance between A and B when the $\bullet$ elements are ignored. Obviously $d(A, B) \geq d_{\text{old}}(A, B)$. Consider the genomes A_2 and B_2 in Figure 3. It can be easily verified that $d_{\text{old}}(A_2, B_2) = 3$ and $d(A_2, B_2) = 4$. This example is easily extendable to two genomes A_{2k} and B_{2k}, with $2k$ chromosomes each, such that $d_{\text{old}}(A_{2k}, B_{2k}) = 3k$ and $d(A_{2k}, B_{2k}) = 4k$.

3.3. Telocentric chromosomes

A chromosome is *telocentric* if its centromere is located at one of its endpoints. For example the chromosome $(\bullet, 5, 6)$ is telocentric.

Lemma 2. *Let A and B be co-tailed genomes satisfying Elements(A) = Elements(B). Then A and B have the same number of telocentric chromosomes. Moreover, the set of genes adjacent to the centromeres in the telocentric chromosomes is the same.*

Proof. Let i be a gene adjacent to the centromere in a telocentric chromosome in A. Thus i is a tail of A and hence a tail of B (since A and B are co-tailed). Suppose w.l.o.g. that i is the leftmost gene in its chromosome both in A and in B and that the centromere is located to the left of i in A. In this case, since genomes A and B are co-tailed, i has the same sign in A and B. Since $Elements(A) = Elements(B)$ it follows that the centromere is located to the left of i also in B. Thus, i is adjacent to the centromere in B and its chromosome is telocentric. ■

Let η denote the number of non-telocentric chromosomes in A and B. We shall show later how mapping between centromeres in non-telocentric chromosomes in A and B can help us to solve LSRT.

3.4. Pericentric and paracentric edges

A gray (respectively, black) edge in $G(A, B)$ is said to be *pericentric* if the two genes it connects flank a centromere in genome B (respectively, A). Otherwise it is called *paracentric* (Fig. 3a). For a gene i we define:

$$cent(i^0) = \begin{cases} -1 & \text{if } i \text{ has a positive sign in } Elements, \\ 1 & \text{otherwise.} \end{cases} \qquad cent(i^1) = -cent(i^0)$$

In other words, the sign of the end closer to the centromere (in both A and B) is positive, and the sign of the remote end is negative. The legality precondition (Section 3.1) implies the following key property:

Lemma 3. *Let (u, v) be an edge in $G(A, B)$. If (u, v) is pericentric then $cent(u) = cent(v) = 1$. Otherwise $cent(u)cent(v) = -1$.*

Proof. The nodes u and v are the ends of two adjacent genes i and j, respectively, in one of the genomes. Suppose (u, v) is pericentric. Then i and j flank a centromere in one of the genomes. Thus u is the end of i closer to j and hence closer to the centromere (i.e., $cent(u) = 1$). Using similar arguments, $cent(v) = 1$.

Suppose (u, v) is paracentric. Then there is no centromere between i and j. W.l.o.g. assume that i is closer to the centromere than j. Then u is the end of i distant from the centromere and v is the end of j closer to the centromere. Therefore, $cent(u)cent(v) = -1$. ■

3.5. *Peri-cycles*

Let C be a cycle in $G(A, B)$. The *peri-cycle* of C, C^P, is defined as follows. The vertices of C^P are the pericentric edges in C. A vertex in C^P is colored gray (respectively, black) if the corresponding edge in C is gray (respectively, black). A path between two consecutive pericentric edges in C is translated to an edge between the two corresponding vertices in C^P (Fig. 3). Note that if C contains no pericentric edges then its peri-cycle is a null cycle (i.e., a cycle with no vertices).

Lemma 4. *Every peri-cycle has an even length and its node colors alternate along the cycle.*

Proof. Let C be a cycle that contains a black pericentric edge (u_1, v_1). Suppose $u_1, v_1, \ldots, u_k, v_k$ is a path between two consecutive black pericentric edges in C. In other words, (u_k, v_k) is a black pericentric edge (possibly $u_1 = u_k$ and $v_1 = v_k$) and there are no other black pericentric edges in this path. Then according to Lemma 3 $cent(v_1) = cent(u_k) = 1$. There is an odd number of edges in the path between v_1 and u_k and thus there must be an odd number of pericentric edges between v_1 and u_k (Lemma 3). It follows that there must exist at least one gray pericentric edge between any two consecutive black pericentric edges. The same argument for a pair of consecutive gray pericentric edges implies that between two such edges there must be at least one black pericentric edge. ■

It follows that every vertex/edge in a peri-cycle has an *opposite* vertex/edge. Removing two opposite vertices/edges from a peri-cycle results in two paths of equal length. We define the *degree* of a cycle as the number of gray (equivalently, black) vertices in its peri-cycle. For example, the single cycle in Figure 3 is of degree 1.

4. MAPPING THE CENTROMERES

This section demonstrates how mapping between the centromeres of A and B can be used to solve LSRT. We shall first see that trying all possible mappings and then solving the resulting SRT gives an exact exponential algorithm for LSRT. Later we shall show how to get an optimal mapping in polynomial time. Let $CEN = \{n + 1, \ldots, n + \eta\}$. For a genome A, let $\dot{\mathbb{A}}$ be the set of all possible genomes obtained by the replacement of each • element in the non-telocentric chromosomes by a distinct element from CEN. Each $i \in CEN$ can be added with either positive or negative sign. Thus $|\dot{\mathbb{A}}| = \eta!2^\eta$. For example, if $A_1 = \{(1, 2, \bullet, 3, 4), (\bullet, 5, 6)\}$ then $\dot{\mathbb{A}}_1$ consists of the genomes $\{(1, 2, 7, 3, 4), (\bullet, 5, 6)\}$ and $\{(1, 2, -7, 3, 4), (\bullet, 5, 6)\}$. Note that every $\dot{A} \in \dot{\mathbb{A}}$ satisfies $Tails(\dot{A}) = Tails$. For each $i \in CEN$ we define $cent(i^0) = cent(i^1) = -1$. A pair $\dot{A} \in \dot{\mathbb{A}}$ and $\dot{B} \in \dot{\mathbb{B}}$ defines a mapping between the centromeres in non-telocentric chromosomes of A and B.

Observation 2. *Let $\dot{A} \in \dot{\mathbb{A}}$ and $\dot{B} \in \dot{\mathbb{B}}$. Then every edge (u, v) in $G(\dot{A}, \dot{B})$ is paracentric and satisfies $cent(u)cent(v) = -1$.*

The notion of legality is easily generalized to partially mapped genomes: a genome is *legal* if each of its chromosomes contains either a single • element or a single, distinct element from CEN (but not both). Since A and $\dot{A} \in \dot{\mathbb{A}}$ differ only in their centromeres, there is a trivial bijection between the set of translocations on $\dot{A}$ and the set of translocations on A. This bijection also preserves legality: a legal translocation on $\dot{A}$ is bijected to a legal translocation on A.

Lemma 5. *Let $\dot{A} \in \dot{\mathbb{A}}$ and $\dot{B} \in \dot{\mathbb{B}}$. Then every proper translocation on $\dot{A}$ is legal and $d(\dot{A}, \dot{B}) = d_{old}(\dot{A}, \dot{B})$.*

Proof. Let $k = d_{\text{old}}(\dot{A}, \dot{B})$. If $k = 0$ then $\dot{A} = \dot{B}$ and hence $d(\dot{A}, \dot{B}) = 0$. Suppose $k > 0$. Let ρ be a translocation on $\dot{A}$ satisfying $d_{\text{old}}(\dot{A} \cdot \rho, \dot{B}) = k - 1$. According to Corollary 1, ρ is either proper or bad. Suppose ρ is bad. Then there is another bad translocation ρ' that cuts the exact positions as ρ, thus satisfying $d_{\text{old}}(\dot{A} \cdot \rho', \dot{B}) = k - 1$, and either ρ or ρ' is legal. Suppose ρ is proper. We shall prove that each of the new chromosomes contains a centromere and hence ρ is legal. Let X be a new chromosome resulting from the translocation ρ and let (u, v) be the new black edge in it. Since ρ is proper, $G(\dot{A} \cdot \rho, \dot{B})$ contains a path between u and v where all the edges existed in $G(\dot{A}, \dot{B})$. This path contains an odd number of edges. Following Observation 2 for $G(\dot{A}, \dot{B})$, $cent(u)cent(v) = -1$. X is composed of two old segments, X_u and X_v, that contain u and v respectively. If $cent(u) = -1$ then X_u contains an element from *CEN*, otherwise X_v contains one. In either case X contains an element from *CEN*. ■

Theorem 2. *Let $\dot{A} \in \dot{\mathbb{A}}$. Then $d(A, B) = \min\{d_{old}(\dot{A}, \dot{B}) | \dot{B} \in \dot{\mathbb{B}}\}$.*

Proof. By Lemma 5, $d(\dot{A}, \dot{B}) = d_{\text{old}}(\dot{A}, \dot{B})$ for every $\dot{A} \in \dot{\mathbb{A}}$ and $\dot{B} \in \dot{\mathbb{B}}$. Obviously a legal sorting of $\dot{A}$ into any $\dot{B} \in \dot{\mathbb{B}}$ induces a legal sorting sequence of the same length, of A to B. Thus, $\min\{d_{\text{old}}(\dot{A}, \dot{B}) | \dot{B} \in \dot{\mathbb{B}}\} \geq d(A, B)$. On the other hand, every sequence of legal translocations that sorts A into B induces a legal sorting of $\dot{A}$ into some $\dot{B} \in \dot{\mathbb{B}}$, thus $\min\{d_{\text{old}}(\dot{A}, \dot{B}) | \dot{B} \in \dot{\mathbb{B}}\} \leq d(A, B)$. ■

A pair of genomes, $\dot{A} \in \dot{\mathbb{A}}$ and $\dot{B} \in \dot{\mathbb{B}}$, define an *optimal* mapping between the centromeres of A and B if $d(A, B) = d_{\text{old}}(\dot{A}, \dot{B})$. Theorem 2 and Lemma 5 imply the following algorithm for LSRT:

Algorithm 1. *Sorting by legal translocations*

```
1: Choose $\dot{A} \in \dot{\mathbb{A}}$ arbitrarily.
2: Compute $\dot{B} = arg\min\{d_{\text{old}}(\dot{A}, \ddot{B}) | \ddot{B} \in \dot{\mathbb{B}}\}$.
3: Solve SRT on $\dot{A}$ and $\dot{B}$—making sure that every bad translocation in the sorting sequence is legal.
```

It can be shown, by a minor modification of the algorithm in (Ozery-Flato and Shamir, 2006a), that solving SRT with the additional condition that every bad translocation is legal can be done in $O(n^{3/2}\sqrt{\log(n)})$. Step 2 can be performed by enumerating all possible mappings and computing the SRT distance for each. This implies:

Lemma 6. *LSRT can be solved in $O(\eta!2^{\eta}n + n^{3/2}\sqrt{\log(n)})$.*

Our goal in the rest of this paper is to improve this result by speeding up Step 2 (i.e., finding efficiently an optimal mapping between the centromeres of A and B).

5. CENT-MAPPINGS

Our general strategy will be to iteratively map between two centromeres in A and B and replace them with a regular element until all centromeres in non-telocentric chromosomes are mapped. The resulting instance can be solved using SRT, but the increase in the number of elements may have also increased the solution value. The main effort henceforth will be to guarantee that the overall increase is minimal. For this, we need to study in detail the effect of each mapping step on the the cycle graph $G(A, B)$. Our analysis uses the SRT distance formula (Theorem 1). We shall ignore for now the parameter δ, and focus on the change in the simplified formula $n - c + l$ (N is not changed by mapping operations).

A mapping between two centromeres affects their corresponding black and gray pericentric edges. Let (i, i') and (j, j') be pericentric black and gray edges in $G(A, B)$ respectively. Suppose $cen \in CEN$ is added between i and i' in $\dot{A}$ and between j and j' in $\dot{B}$. In this case, (i, i') and (j, j') in $G(A, B)$ are replaced by the four (paracentric) edges (i, cen), (cen, i'), (j, cen) and (cen, j') in $G(\dot{A}, \dot{B})$. (The first two edges are black, the latter are gray.) We refer to the addition of $cen \in CEN$ between (i, i') and (j, j') as a *cent-mapping* since it maps between two centromeres. Note that for each pair of centromeres in A and B

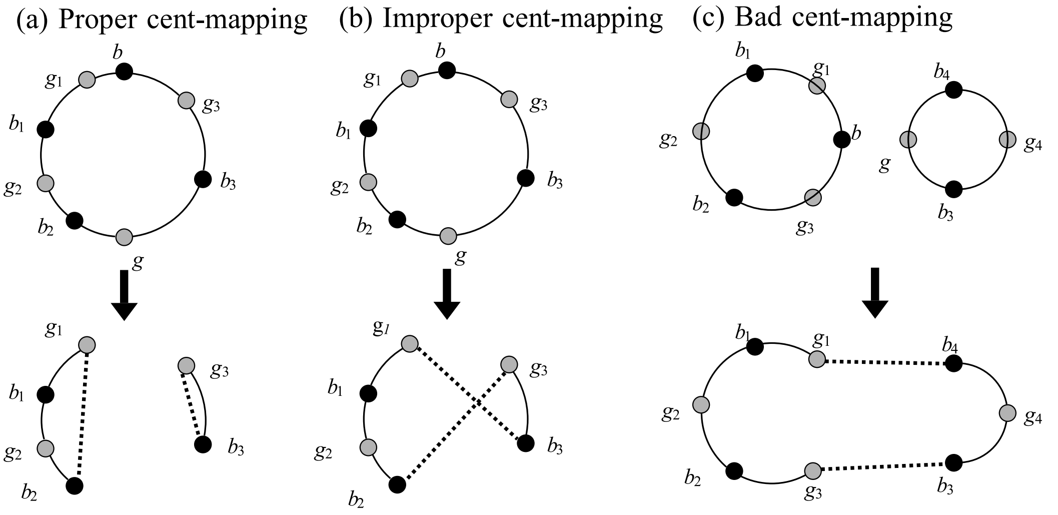


FIG. 4. The effect of a cent-mapping on peri-cycles. Each of the cycles is a peri-cycle with black and gray nodes corresponding to centromeres (pericentric edges) in A and B, respectively. In all cases, a cent-mapping on b and g in the top peri-cycles is performed, and the bottom peri-cycles are the result. Dotted lines denote new edges. **(a,b)** Two alternative cent-mappings of a pair of pericentric edges in the same cycle. **(c)** Each of the two alternatives generates a single cycle.

there are two possible cent-mappings (corresponding to the relative signs of the added elements). Given $\dot{A} \in \dot{\mathbb{A}}$, every $\dot{B} \in \dot{\mathbb{B}}$ defines η disjoint cent-mappings and vice versa. Obviously, every cent-mapping increases the number of genes by one ($\Delta n = +1$).

Lemma 7. *Every cent-mapping satisfies* $\Delta c \in \{-1, 0, 1\}$.

Proof. Let (i, i') and (j, j') be black and gray pericentric edges in $G(A, B)$, respectively. Let $cen \in CEN$ be the element between i and i' in $\dot{A}$. If (i, i') and (j, j') belong to the same cycle before the cent-mapping then $\Delta c \in \{0, 1\}$. If (i, i') and (j, j') belong to different cycles before the cent-mappings then $\Delta c = -1$. ■

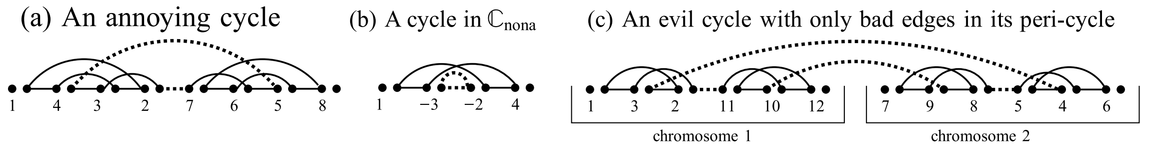
In the rest of the paper, we will analyze the effect of a cent-mapping using peri-cycles. A peri-cycle can be viewed as a compact representation of a cycle focused on pericentric edges, which are the only edges affected by cent-mappings. A cent-mapping is called *proper*, *improper*, *bad* if $\Delta c = 1, 0, -1$ respectively. For illustrations of the three types of cent-mappings, see Figure 4. We say that a cent-mapping *operates* on a cycle C if C contains at least one of the mapped pericentric edges. Proper and improper cent-mappings always operate on one cycle in $G(A, B)$; a bad cent-mapping always operates on two different cycles in $G(A, B)$.

Observation 3. *Every proper cent-mapping satisfies* $\Delta l \in \{0, 1\}$. *An improper cent-mapping satisfies* $\Delta l = 0$. *A bad cent-mapping satisfies* $\Delta l \in \{0, -1, -2\}$.

It follows that a proper cent-mapping satisfies $\Delta(n - c + l) = 0$ iff $\Delta l = 0$; An improper cent-mapping satisfies $\Delta(n - c + l) = 1$; a bad cent-mapping satisfies $\Delta(n - c + l) = 0$ iff $\Delta l = -2$. A proper cent-mapping is *safe* if it satisfies $\Delta l = 0$. In the following sections we present two classes of cycles, "annoying" and "evil" for which any set of proper cent-mappings that eliminates all their pericentric edges is unsafe.

5.1. Annoying cycles

In this section we focus on cycles in leaves. The degree of every cycle in a leaf is at most 1 (otherwise it must be external). Moreover, a leaf can contain at most one cycle of degree 1 (for the same reason).


(a) An annoying cycle

(b) A cycle in $\mathbb{C}_{\text{nona}}$

(c) An evil cycle with only bad edges in its peri-cycle

FIG. 5. Examples of cycles in $\mathbb{C}_{\text{ann}}$, $\mathbb{C}_{\text{nona}}$, and $\mathbb{C}_{\text{evil}}$. In all the figures, the target genome B is a fragmented identity permutation (i.e., every gray edge is of the form $(i, i+1)$); pericentric edges are denoted by dotted lines.

A cycle is called *annoying* if: *(i)* it is contained in a leaf, *(ii)* its degree is 1, and *(iii)* a proper cent-mapping on its two pericentric edges satisfies $\Delta l = 1$ (i.e., one leaf is split into two leaves) (Fig. 5a). Thus a proper cent-mapping on an annoying cycle satisfies $\Delta(n - c + l) = 1$. On the other hand, any bad cent-mapping on a cycle contained in the span of a leaf (annoying or not) results in the elimination of that leaf. Thus, a cent-mapping on any two cycles in (two different) leaves satisfies $\Delta(n - c + l) = 1 + 1 - 2 = 0$. Let $\mathbb{C}_{\text{ann}}$ denote the set of annoying cycles and let $ann = |\mathbb{C}_{\text{ann}}|$. Let $\mathbb{C}_{\text{nona}}$ be the set of non-annoying cycles of degree 1 that are contained in the span of a leaf (Fig. 5b). Let $nona = |\mathbb{C}_{\text{nona}}|$.

5.2. Evil cycles

In this section we focus on cycles that are not in leaves. Let C be a cycle of degree at least 1 that is not in a leaf and let C^P be its peri-cycle. Let (b, g) be an edge in C^P. Denote by $V(b, g)$ the set of gray edges in the corresponding path between b and g in C. The edge (b, g) is *bad* if after a proper cent-mapping on b and g the edges in $V(b, g)$ belong to a leaf, otherwise it is *good*. For example, in Figure 3, the edge (b, g) where $V(b, g) = \{(1, 2), (2, 3)\}$ is bad.

Lemma 8. *The "badness" of edge (b, g) in a peri-cycle is unchanged by cent-mappings not involving b and g.*

Proof. Clearly the order in which we perform cent-mappings does not affect the final cycle graph. Let M be the component containing $V(b, g)$ in the cycle graph resulting from a proper cent-mapping on (b, g). If M does not contain any pericentric edge in its span, then clearly it is not affected by later cent-mappings. Suppose M contains a pericentric edge in its span. Thus, M must be external since it contains in its span centromeres of two different chromosomes in A. If M is not split by other cent-mappings, then clearly $V(b, g)$ remains in an external component. Suppose M is split into two components by a cent-mapping on pericentric edges b' and g'. In this case, each of the two new components contains in its span one of the two new black edges replacing b'. Hence, the component that contains $V(b, g)$ is guaranteed to remain external, since it contains in its span two different centromeres in A (corresponding to b and b'). ■

Lemma 9. *Let C be a cycle satisfying: (i) $deg(C) > 0$, and (ii) C contains a new gray edge, g_{new}, that was created by a cent-mapping. Let (b, g) be an edge in the peri-cycle of C such that $V(b, g)$ contains g_{new}. Then (b, g) is good.*

Proof. The edge g_{new} is adjacent to a vertex of a previously mapped centromere, $cen_1 \in CEN$. On the other hand, after a cent-mapping on (b, g), the path $V(b, g)$ will be adjacent to a vertex of a new mapped centromere, $cen_2 \in CEN$. These two centromeres belong to different chromosomes of A. Thus $V(b, g)$ must contain an external edge after any cent-mapping of b and g and hence (b, g) is good. ■

A path in a peri-cycle is *bad* if all the edges in it are bad. For a path P, let $len(P)$ denote the number of vertices in P. A cycle C is called *evil* if its peri-cycle contains a bad path P such that $len(P) > deg(C)$. For example, the single cycle in Figure 3 is evil since it contains a bad edge, which is a bad path of length 2, and its degree is 1. An example of an evil cycle with only bad edges in its peri-cycle is presented in Figure 5. Let $\mathbb{C}_{\text{evil}}$ denote the set of all evil cycles that are not in leaves. Define $evil = |\mathbb{C}_{\text{evil}}|$.

Lemma 10. *Let C be a cycle that does not belong to a leaf. There is a set of safe proper cent-mappings of all the pericentric edges in C iff C is not evil.*

Proof. Let C^P be the peri-cycle of C and let $k = deg(C)$. Suppose C is evil. Then PC contains a bad path P with $k+1$ vertices. There are $2k$ vertices in C^P, thus any proper cent-mapping of all the pericentric edges in C must match two vertices from P. It follows that there must be a proper cent-mapping on the two ends of an edge in P. Hence, by definition this cent-mapping is unsafe.

Suppose C is not evil. If $k = 1$ then the two edges in C^P are good and the proper cent-mapping of the two pericentric edges in C is safe. Suppose $k > 1$. Let $C^P = P_1, P_2$ where P_1 is a longest bad path in C^P. Let u be the first vertex in P_1 and let v be the last vertex in P_2. Then (u, v) is a good edge in C^P. Let C_1 and C_2 be the two cycles created by the proper cent-mapping on u and v, where C_1 contains $V(u, v)$. Obviously this proper cent-mapping is safe, $deg(C_1) = 0$ and $deg(C_2) = k-1$. It suffices to prove that C_2 is not evil. Let C_2^P be the peri-cycle of C_2. Then $C_2^P = P_1' P_2'$ where $len(P_1') = len(P_1)-1$, $len(P_2') = len(P_2)-1$, and P_1' and P_2' are connected by good edges (Lemma 9). Let p be the length of the longest bad path in C_2^P. Then *(i)* $p \leq len(P_1) \leq k$ (since P_1 is a longest bad path in C), *(ii)* $p \leq \max(len(P_1'), len(P_2')) = len(P_2')$, and *(iii)* $len(P_1) + len(P_2) = 2k$. It follows that $p \leq k - 1 = deg(C_2)$. Thus by definition C_2 is not evil. ■

Corollary 2. *Every proper cent-mapping satisfies* $\Delta(l + evil) \geq 0$.

We partition $\mathbb{C}_{\text{evil}}$ into three classes:

- $\mathbb{C}^1_{\text{evil}}$: Cycles of even degree and only bad edges in their peri-cycle.
- $\mathbb{C}^2_{\text{evil}}$: Cycles of odd degree and only bad edges in their peri-cycle.
- $\mathbb{C}^3_{\text{evil}}$: Cycles with at least one good edge in their peri-cycle.

Let $evil_1 = |\mathbb{C}^1_{\text{evil}}|$, $evil_2 = |\mathbb{C}^2_{\text{evil}}|$ and $evil_3 = |\mathbb{C}^3_{\text{evil}}|$. If $C \in \mathbb{C}_{\text{evil}}$ is of degree 1 then $C \in \mathbb{C}^3_{\text{evil}}$ (since otherwise it would be in a leaf). Every new evil cycle (i.e., an evil cycle created by a cent-mapping) contains a good edge (Lemma 9) and hence belongs to $\mathbb{C}^3_{\text{evil}}$. Let $C \in \mathbb{C}^3_{\text{evil}}$ and let (b, g) be an edge opposite to a good edge in the peri-cycle of C. A proper cent-mapping on b and g satisfies $\Delta l = 1$, $\Delta evil = -1$ and hence $\Delta(n - c + l + evil) = 0$. Such a cent-mapping can be viewed as a *replacement* of an evil cycle with a leaf. On the other hand, every proper cent-mapping on a cycle in $\mathbb{C}^1_{\text{evil}} \cup \mathbb{C}^2_{\text{evil}}$ satisfies $\Delta(n - c + l + evil) = \Delta(l + evil) = 1$. Thus by applying proper cent-mappings, a cycle in $\mathbb{C}^2_{\text{evil}} \cup \mathbb{C}^1_{\text{evil}}$ can be replaced by two leaves, where each leaf belongs to a different chromosome.

Lemma 11. *Let $C \in \mathbb{C}_{evil}$. There exists an improper cent-mapping on C for which $\Delta evil = -1$ iff $C \notin \mathbb{C}^1_{evil}$.*

Proof. Let $C \in \mathbb{C}_{\text{evil}}$ and let C^P be its peri-cycle. Suppose that $C \notin \mathbb{C}^1_{\text{evil}}$.

Case 1: deg(C) is odd. Let u and v be two opposite vertices in the peri-cycle of C. Thus u and v have opposite colors. Let C_1 be the cycle obtained from C after an improper cent-mapping between u and v. Then the peri-cycle of C_1 contains two opposite good edges (Lemma 9) and thus C_1 is not evil.

Case 2: deg(C) is even. Then $C \in \mathbb{C}^3_{\text{evil}}$. Let (b, g) be an edge opposite to a good edge in the peri-cycle of C. Let C_1 be the cycle obtained from C after performing an improper cent-mapping between b and g. Then the peri-cycle of C_1 has two opposite good edges and thus C_1 is not evil.

Suppose $C \in \mathbb{C}^1_{\text{evil}}$. Then $deg(C) = k$ is even and every edge in its peri-cycle is bad. Let C_1 be the result of an improper cent-mapping on C. Then $deg(C_1) = k - 1$ and the peri-cycle of C_1 must contain a bad path with at least k vertices. Thus C_1 is evil. ■

In other words: for every cycle in $\mathbb{C}^2_{\text{evil}} \cup \mathbb{C}^3_{\text{evil}}$ there exists an improper cent-mapping satisfying $\Delta(n - c + l + evil) = 0$; Every improper cent-mapping on a cycle in $\mathbb{C}^1_{\text{evil}}$ satisfies $\Delta(n - c + l + evil) = 1$. It follows that a cent-mapping on $C \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$ satisfies $\Delta(n - c + l + evil) = 0$ only if it is bad. Therefore, Corollary 2 and Lemma 11 imply:

Corollary 3. *For every cent-mapping* $\Delta(n - c + l + evil) \geq 0$.

5.3. A polynomial algorithm using at most opt + 2 translocations

In this section we present upper and lower bounds for the legal translocation distance. These bounds provide an intuition for the rather complicated formula for the legal translocation distance presented in the next section. The proof of the upper bound implies an approximation algorithm that sorts A into B using at most $d(A, B) + 2$ legal translocations.

Lemma 12. *Let $C_1, C_2 \in \mathbb{C}_{evil} \cup \mathbb{C}_{ann}$, where $deg(C_1) \leq deg(C_2)$. If $deg(C_1) = deg(C_2)$ then every bad cent-mapping on C_1 and C_2 satisfies $\Delta(l + evil) = -2$. If $deg(C_1) < deg(C_2)$ there exists a bad cent-mapping on C_1 and C_2 satisfying $\Delta(l + evil) = -2$ iff $C_2 \in \mathbb{C}^3_{evil}$.*

Proof. If $deg(C_1) = deg(C_2)$ then any bad cent-mapping on C_1 and C_2 results in a cycle whose peri-cycle contains two opposite good edges and hence non-evil. Suppose $k_1 = deg(C_1) < deg(C_2) = k_2$ and let C_1^P and C_2^P denote the peri-cycles of C_1 and C_2 respectively.

Case 1: $C_2 \in \mathbb{C}^3_{\text{evil}}$. Let (b, g) be the opposite edge of a good edge in C_2^P. Let C_3 be a result of a (bad) cent-mapping of the b and a vertex of an opposite color in C_2^P. Let P' be a longest bad path in the peri-cycle of C_3. Then $len(P') \leq \max\{k_2, 2k_1 - 1\} \leq k_2 + k_1 - 1 = deg(C_3)$.

Case 2: $C_2 \notin \mathbb{C}^3_{\text{evil}}$. In this case all the edges in C_2^P are bad. Let C_3 be the result of a bad cent-mapping on C_1 and C_2. Then the peri-cycle of C_3 contains a bad path with $2k_2 - 1$ vertices, while $deg(C_3) = k_1 + k_2 - 1 < 2k_2 - 1$. Thus C_3 is evil. ■

The *bad cent-mappings* graph, *BCM*, is defined as follows. It is a bipartite graph whose two parts are *DEG* and *CYC*, where:

$$DEG = \{i : |\{C : C \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}, \text{deg}(C) = i\}| \text{ is odd}\} \qquad CYC = \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}_{\text{nona}}$$

For example, if the degrees of the cycles in $\mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$ are $\{1, 2, 2, 2, 4, 4, 6, 8\}$ then $DEG = \{1, 2, 6, 8\}$. Vertices $i \in DEG$ and $C \in CYC$ are connected by an edge if $deg(C) \geq i$ (Fig. 6). Thus an edge (i, C) represents a bad cent-mapping operating on C and $C' \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $\text{deg}(C') = i$, for which $\Delta(n - c + l + evil) = 0$ and $\Delta|DEG| = -1$.

A *matching* in a graph is a collection of edges no two of which share a common vertex. The size of a matching M, denoted $|M|$, is the number of edges in it. Finding a maximum matching in *BCM* is an easy task that can be completed in linear time by a greedy algorithm that iteratively matches vertices from *CYC* in increasing order of their degrees. Define *fbad* $= |DEG| - |M|$, where M is a maximum matching. For a matching M let F_M be the forest of internal components after performing a bad cent-mapping on every $C \in \mathbb{C}_{\text{ann}} \cup M$. In other words, F_M is obtained from F by the deletion of every component containing a cycle from either $\mathbb{C}_{\text{ann}}$ or $\mathbb{C}_{\text{nona}} \cap M$ in its span. In the following we prove that the cent-mappings produced by Algorithm 2 lead to a sorting scenario of at most $d(A, B) + 2$ legal translocations.

Observation 4. *Every cent-mapping satisfies $\Delta\lceil fbad/3\rceil \in \{-1, 0, 1\}$.*

Proof. Every cent-mapping involves at most three cycles (old and new). Hence $\Delta fbad \in [-3, 3]$. ■

Lemma 13. *Every cent-mapping satisfies $\Delta(n - c + l + evil + \lceil fbad/3\rceil) \geq 0$.*

Proof. Let $\Delta \equiv \Delta(n - c + l + evil + \lceil fbad/3\rceil)$. By Observation 4, if $\Delta(n - c + l + evil) > 0$ then $\Delta \geq 0$. Suppose $\Delta(n - c + l + evil) = 0$. We shall prove that $\Delta fbad \geq 0$:

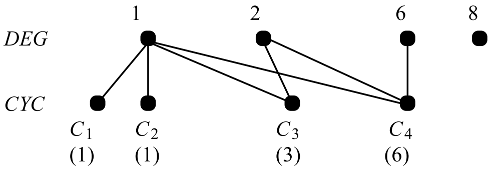


FIG. 6. An example for a bad cent-mappings (*BCM*) graph. $DEG = \{1, 2, 6, 8\}$, $CYC = \{C_1, C_2, C_3, C_4\}$. The degree of each cycle in *CYC* appears in brackets below the cycle.

Algorithm 2. Get_Mapping (a 2-additive approximation)

```
$M \leftarrow$ a maximum matching in $BCM$
Perform a bad cent-mapping on every $C_1, C_2 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $deg(C_1) = deg(C_2)$.
/* Now $|\mathbb{C}^1_{evil} \cup \mathbb{C}_{ann}| = |DEG|$ */
for all $(i, C) \in M$ do
    Perform a bad cent-mapping on $C$ and $C' \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $\text{deg}(C') = i$, such that $\Delta(l + evil) = -2$ (Lemma 12).
end for
while $|DEG| \geq 3$ do
    $C_1, C_2, C_3 \leftarrow 3$ cycles in $\mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $\text{deg}(C_1)$ is minimal.
    Perform a bad cent-mapping on $C_2$ and $C_3$ and let $C_4$ be the new evil cycle.
    Perform a bad cent-mapping on $C_1$ and $C_4$ such that $\Delta(l + evil) = -2$ (Lemma 12).
end while
if $|DEG| = 2$ then
    Perform a bad cent-mapping on $C, C' \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$. /* $DEG = 2 \rightarrow DEG = 1$ */
end if
if $|DEG| = 1$ then
    Perform an improper cent-mapping on $C \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$.
end if
/* Now $|\mathbb{C}^1_{evil}| = ann = 0$ */
Perform an improper cent-mapping on every $C \in \mathbb{C}_{\text{evil}}$ such that $\Delta evil = -1$ (Lemma 11).
/* Now $evil = 0$ */
Perform safe proper cent-mappings on every cycle of degree at least 1 (Lemma 10).
Perform a proper cent-mapping on every $C \in \mathbb{C}_{\text{nona}}$.
```

Case 1: $\Delta(n - c) = 0$ (i.e., proper cent-mapping). Then $\Delta(l + evil) = 0$ and thus either $\Delta l = 1$ and $\Delta evil = -1$, or $\Delta l = \Delta evil = 0$. Hence DEG is unchanged and $\Delta|CYC| \leq 0$. Therefore, $\Delta fbad \geq 0$.

Case 2: $\Delta(n - c) = 1$ (i.e., improper cent-mapping). Then $\Delta l = 0$ and $\Delta evil = -1$. Therefore DEG is unchanged, $\Delta|CYC| \leq 0$, and hence $\Delta fbad >= 0$.

Case 3: $\Delta(n - c) = 2$ (i.e., bad cent-mapping). Then $\Delta(l + evil) = -2$. Let C_1 and C_2 be the cycles on which the cent-mapping was performed. If C_1 and C_2 belong to the same class (e.g., $\mathbb{C}^1_{\text{evil}}$, $\mathbb{C}^3_{\text{evil}}$) then clearly DEG is unchanged and $\Delta|CYC| \leq 0$, hence $\Delta fbad \geq 0$. If C_1 and C_2 belong to different classes, then w.l.o.g. $C_1 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$ and $C_2 \in \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}_{\text{nona}}$. Hence, $\Delta fbad \geq 0$. ■

Lemma 14. *Every cent-mapping performed by Algorithm 2 satisfies* $\Delta(n - c + l + evil + \lceil fbad/3 \rceil) = 0$.

Theorem 3. *Let* $d = d(A, B)$ *and let* $f = n - N - c + l + evil + \lceil fbad/3 \rceil$. *Then* $d \in [f, f + 2]$. *In particular, Algorithm 2 produces* $\dot{A} \in \dot{\mathbb{A}}$ *and* $\dot{B} \in \dot{\mathbb{B}}$ *for which* $d(\dot{A}, \dot{B}) \leq d + 2$.

Proof. Let $\dot{A} \in \dot{\mathbb{A}}$. For every $\dot{B} \in \dot{\mathbb{B}}$, $evil(\dot{A}, \dot{B}) = fbad(\dot{A}, \dot{B}) = 0$ and thus by Theorem 1, $d_{\text{old}}(\dot{A}, \dot{B}) = f(\dot{A}, \dot{B}) + \delta(\dot{A}, \dot{B})$. By Lemma 13, $f(A, B) \leq \min_{\dot{B} \in \dot{\mathbb{B}}}\{f(\dot{A}, \dot{B})\}$. By Theorem 2, $d(A, B) = \min\{f(\dot{A}, \dot{B}) + \delta(\dot{A}, \dot{B}) : \dot{B} \in \dot{\mathbb{B}}\}$. Hence $f(A, B) \leq d(A, B)$. Let $\dot{B}$ be the genome defined by the cent-mappings produced by Algorithm 2. By Lemma 14, $f(A, B) = f(\dot{A}, \dot{B})$. Therefore, $d(A, B) \leq d_{\text{old}}(\dot{A}, \dot{B}) = f(A, B) + \delta(\dot{A}, \dot{B}) \leq f(A, B) + 2$. ■

6. A POLYNOMIAL ALGORITHM FOR THE LEGAL TRANSLOCATION DISTANCE

In this section we present an exact formula for the legal translocation distance, which leads to a polynomial algorithm for the problem. The proof, and subsequently the algorithm, is focused on finding an optimal mapping between the centromeres of genomes A and B (Step 2 in Algorithm 1). This requires an involved case analysis, which is deferred to an appendix. Let M be a maximum matching in the BCM graph. Denote by l_M be the number of leaves in F_M. Define $fgood(M) = |\mathbb{C}^3_{\text{evil}} \setminus M|$. Define $mbad = fbad \bmod 3$. Define $\delta' \in \{0, 1, 2\}$ as follows. $\delta' = 2$ iff all the following conditions are satisfied:

- $\mathbb{C}^2_{\text{evil}} = \mathbb{C}^3_{\text{evil}} = DEG = \emptyset$
- $|F_\emptyset| = 1$
- l and ann are even. If $ann > 0$ then $nona = 0$

If $\delta' \neq 2$ then $\delta' = 1$ iff for every maximum matching M all the following conditions are satisfied:

- $fgood(M) \in \{0, 1\}$
- l_M is even $\Rightarrow F_M = 1$
- (l_M is odd and $fgood(M) = 1$) $\Rightarrow C \in \mathbb{C}^3_{\text{evil}} \setminus M$ cannot be replaced by a leaf such that $|F_M| > 1$.
- $mbad = 1 \Rightarrow DEG = \{1\}$, $|F| = 1$, and ($l_\emptyset$ is odd $\Rightarrow evil_2 = 0$)
- $mbad = 2 \Rightarrow l_M$ is even and $fgood(M) = 0$

If $\delta' \neq 1, 2$ then $\delta' = 0$. Note that if $\delta' = 1$ and $mbad \in \{1, 2\}$ then $|F_M| = 1$.

Theorem 4. *The legal translocation distance between A and B is $d(A, B) = n - N - c(A, B) + l(A, B) + evil(A, B) + \lceil fbad(A, B)/3 \rceil + \delta'(A, B)$.*

The proof of Theorem 4, which appears in the appendix, is by a case analysis of the change in each of the parameters, $n - c$, l, $evil$, $fbad$ and δ', for each cent-mapping, and hence is quite involved. It leads to a polynomial time algorithm for finding an optimal mapping between the centromeres of A and B. This algorithm, which can be viewed as an extension of Algorithm 2, has the same time complexity as Algorithm 2.

Theorem 5. *LSRT can be solved in $O(\eta n + n^{3/2}\sqrt{\log(n)})$ time.*

Proof. Finding an optimal mapping between the centromeres of A and B can be done in $O(\eta n)$ in the following manner. The set of peri-cycles can be computed in $O(n)$. For every edge in a peri-cycle we compute its "badness" in $O(n)$ by simply performing the corresponding proper cent-mapping. Computing the badness of all the edges thus takes $O(\eta n)$. Computing $\mathbb{C}^1_{\text{evil}}$, $\mathbb{C}^2_{\text{evil}}$, $\mathbb{C}^3_{\text{evil}}$, $\mathbb{C}_{\text{ann}}$, $\mathbb{C}_{\text{nona}}$, and DEG requires a simple traversal of all the edges in every peri-cycle. Hence, it can be done in $O(\eta)$. Overall the algorithm performs $O(\eta)$ operations where each can be implemented in $O(n)$ time. ■

7. CONCLUSION

Computational studies in genome rearrangements have overlooked centromeres to date. In this study, we presented a new model for genomes that accounts for centromeres. Using this model, we defined the problem of legal sorting by reciprocal translocations (LSRT) and proved that it can be solved in polynomial time. Unfortunately, the legal translocation distance formula appears to be quite complex and it is an interesting open problem whether it or its proof can be simplified.

A solvable LSRT instance requires the two input genomes to be co-tailed and with the same set of elements (see Section 3.1). This requirement is a rather strong and unrealistic. Allowing for reversals, non-reciprocal translocations, fissions and fusions will cancel these restrictions. Under a centromere-aware model, fissions and fusions are legal if they are centric (Perry et al., 2004; Searle, 1998). In future work, we intend to study an extension of LSRT that allows for reversals, (centric) fusions and fissions. We expect an exact algorithm for this extended problem to bring us nearer to realistic rearrangement scenarios than can be done today.

8. APPENDIX

Proof of Theorem 4

The proof follows directly from Lemmas 15 and 16 below: Lemma 15 provides a lower bound for the legal distance while Lemma 16 proves this bound is tight.

Lemma 15. *Let* $\Delta = \Delta(n - c + l + evil + \lceil fbad/3 \rceil + \delta')$. *For every cent-mapping* $\Delta \geq 0$.

Proof. In the following "before" and "after" are used to define the state before and after the current cent-mapping respectively. However, unless specified otherwise, every condition refers to the state before the cent-mapping. For example, "l_M is odd" means "l_M is odd before." Let $\mathbb{C}_{\text{good}}$ be the set of cycles that are not in $\mathbb{C}_{\text{evil}} \cup \mathbb{C}_{\text{ann}} \cup \mathbb{C}_{\text{nona}}$. Following Lemma 13, if $\Delta\delta' \geq 0$ then $\Delta \geq 0$. Thus it suffices to prove $\Delta \geq 0$ only for $\delta' \in \{1, 2\}$.

Case 1: $\delta' = 2$. Then $\Delta fbad \geq 0$, since $DEG = \emptyset$.

Case 1.1: $\Delta(n - c) = 0$. Let C be the cycle on which the cent-mapping was performed. Since $\delta' = 2$ then $C \notin \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}^2_{\text{evil}}$.

- $C \in \mathbb{C}_{\text{nona}}$. Then no other parameter is affected and $\Delta\delta' = 0$.
- $C \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$. Then $\Delta(l + evil) = 1$, $\Delta\lceil fbad/3 \rceil = 1$, and hence $\Delta \geq 0$.
- $C \in \mathbb{C}_{\text{good}}$. If $\Delta(l + evil) = 0$ then no other parameter is affected and $\Delta = 0$. If $\Delta(l + evil) = 2$ then clearly $\Delta \geq 0$. Suppose $\Delta(l + evil) = 1$. Note that DEG is unchanged (i.e., $DEG = \emptyset$ after). Hence $mbad = 0$ after. If $\Delta l = 1$ then after: $l_\emptyset$ is odd and $CYC = \emptyset$. If $\Delta evil = 1$ then after $l_\emptyset$ is even and $F|_\emptyset| = 1$ (since F is unchanged). Thus, in either case $\Delta = 0$.

Case 1.2: $\Delta(n - c) = 1$ (i.e., an improper move). Let C be the cycle on which the cent-mapping was performed.

- $C \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$. Then $\Delta(l + evil) = 0$, $\Delta\lceil fbad/3 \rceil = 1$, and hence $\Delta \geq 0$.
- $C \in \mathbb{C}_{\text{nona}}$. Then no other parameter is affected and hence $\Delta = 1$.
- $C \in \mathbb{C}_{\text{good}}$. Then $\Delta l = 0$, $\Delta evil \in \{0, 1\}$ and in either case $\Delta = 1$.

Case 1.3: $\Delta(n-c) = 2$. Let C_1 and C_2 be the two peri-cycles on which the cent-mapping was performed. If $\deg(C_1) = \deg(C_2)$ then C_1 and C_2 belong to the same class (either $\mathbb{C}^1_{\text{evil}}$ or $\mathbb{C}_{\text{ann}}$) and clearly $\Delta\delta' = 0$. Suppose $\deg(C_1) < \deg(C_2)$.

- $C_1, C_2 \in \mathbb{C}_{\text{good}}$. Then $\Delta = 2$.
- $C_1 \in \mathbb{C}_{\text{good}}$, $C_2 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$. Then $\Delta(l + evil) \in \{0, -1\}$. If $\Delta(l + evil) = 0$ then $C_2 \in \mathbb{C}^1_{\text{evil}}$ and hence $\Delta fbad = 0$. If $\Delta(l + evil) = -1$ then $\Delta fbad = 1$. Hence, in either case, $\Delta \geq 0$.
- $C_1 \in \mathbb{C}_{\text{good}}$, $C_2 \in \mathbb{C}_{\text{ann}} \cup \mathbb{C}_{\text{nona}}$. Then $\Delta l = -1$, $\Delta evil = 0$ (the new cycle is in $\mathbb{C}_{\text{good}}$). If $C_2 \in \mathbb{C}_{\text{ann}}$ then $\Delta fbad = 1$ and hence $\Delta \geq 0$. Suppose $C \in \mathbb{C}_{\text{nona}}$. Then $\Delta fbad = 0$, and after: $mbad = 0$, $l_\emptyset$ is odd, and $fgood(\emptyset) = evil_3 = 0$. Hence $\delta' = 1$ after and thus $\Delta = 0$.
- $C_1 \in \mathbb{C}^1_{\text{evil}}$, $C_2 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$ (different degrees). Then $\Delta(l + evil) = -1$, $\Delta\lceil fbad/3 \rceil = 1$ and hence $\Delta \geq 0$.
- $C_1 \in \mathbb{C}^1_{\text{evil}}$, $C_2 \in \mathbb{C}_{\text{nona}}$. Then $\Delta l = -1$, and the new resulting cycle, C_3 satisfies $C_3 \in \mathbb{C}^3_{\text{evil}}$ and $\deg(C_3) = \deg(C_1)$. Hence $\Delta evil = 0$, $\Delta fbad = 0$, and $\Delta\delta' = -1$. Hence $\Delta = 0$.

Case 2: $\delta' = 1$. If $\Delta(n - c + l + evil + \lceil fbad/3 \rceil) \geq 1$ then clearly $\Delta \geq 0$. We shall prove that if $\Delta(n - c + l + evil + \lceil fbad/3 \rceil) = 0$ then $\Delta\delta' \geq 0$ and thus $\Delta \geq 0$.

Case 2.1: $\Delta(n - c) = 0$. Then $\Delta(l + evil) \geq 0$ (Corollary 2), $\Delta(l + evil + \lceil fbad/3 \rceil) \geq 0$ (Lemma 13). If $\Delta(l + evil + \lceil fbad/3 \rceil) > 0$ then clearly $\Delta \geq 0$. Suppose $\Delta(l + evil + \lceil fbad/3 \rceil) = 0$.

- Suppose $\Delta(l + evil) = 0$. Then $\Delta\lceil fbad/3 \rceil = 0$, $C_1 \in \mathbb{C}_{\text{good}} \cup \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}_{\text{nona}}$. If $C \in \mathbb{C}_{\text{good}}$ then no parameter is affected and hence $\Delta = 0$. Suppose $C \in \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}_{\text{nona}}$. Then $\Delta fbad \in \{0, 1\}$ and $mbad \in \{0, 2\}$.

  —Suppose $fbad = 0$, $\Delta l = 1$. Then $C \in \mathbb{C}^3_{\text{evil}}$ and $\Delta evil = -1$. Thus for every maximum matching M after, there exists a maximum matching M' before satisfying $fgood(M') =$

$fgood(M) + 1$ and $l_{M'} = l_M - 1$. Since $\delta' = 1$ before it follows that $mbad = 0$ and $\Delta\delta' \geq 0$.

—Suppose $fbad = 0$, $\Delta l = 0$. If $C \in \mathbb{C}_{\text{nona}}$ then every maximum matching after is a maximum matching before, with the same properties. Suppose $C \in \mathbb{C}^3_{\text{evil}}$. Then C is replaced with an evil cycle C' of a smaller degree. Hence for every maximum matching M' after there exists a maximum matching M before, where C' is replaced by C, and which has the same properties as M. Hence in both cases $\Delta\delta' \geq 0$.

—Suppose $fbad = 1$. Then $mbad = 2$ before and $mbad = 0$ after.

* Suppose $C \in \mathbb{C}^3_{\text{evil}}$. If $\Delta l = 1$ (and hence $\Delta evil = -1$) then every maximum matching M after satisfies l_M is odd and $fgood(M) = 0$. If $\Delta l = 0$ then every maximum matching M after satisfies either (l_M is even and $|F_M| = 1$) or (l_M is odd and $fgood(M) = 0$). Hence, in any case $\Delta\delta' \geq 0$.
* Suppose $C \in \mathbb{C}_{\text{nona}}$. Then every maximum matching M after satisfies l_M is odd and $fgood(M)$ is even. Hance $\delta' = 1$ after.

- Suppose $\Delta(l + evil) = 1$. Then $\Delta\lceil fbad/3 \rceil = -1$.

—Suppose $\Delta fbad = -1$. Then $mbad = 1$ before and thus $evil_3 = nona = 0$ and $C \in \mathbb{C}_{\text{good}} \cup \mathbb{C}^2_{\text{evil}}$. It follows that every maximum matching M after satisfies either (l_M is even and $|F_M| = 1$) or (l_M is odd and $fgood(M) = 0$). (The later happens only if $C \in \mathbb{C}^2_{\text{evil}}$ and $\Delta l = 1$.) Hence $\Delta\delta' \geq 0$.

—Suppose $\Delta fbad = -2$. Then $mbad = 2$ before and $C \in \mathbb{C}^2_{\text{evil}} \cup \mathbb{C}^1_{\text{evil}}$. Moreover, if $C \in \mathbb{C}^1_{\text{evil}}$ then $\deg(C) \in DEG$. Then for every maximum matching M after either (l_M is even and $|F_M| = 1$) or (l_M is odd and $fgood(M') = 0$). (The latter case may happen only if $C \in \mathbb{C}^1_{\text{evil}}$.) Hence $\Delta\delta' = 0$.

—Suppose $\Delta fbad = -3$. Then $C \in \mathbb{C}^1_{\text{evil}}$ and for every maximum matching M after there exists a maximum matching M' before with the same properties. Hence $\Delta\delta' = 0$.

Case 2.2: Suppose $\Delta(n-c) = 1$. Then $\Delta l = 0$ and $\Delta(evil + \lceil fbad/3 \rceil) \geq -1$. If $\Delta(evil + \lceil fbad/3 \rceil) \geq 0$ then clearly $\Delta \geq 0$. Suppose $\Delta(evil + \lceil fbad/3 \rceil) = -1$. Let C the cycle on which the cent-mapping was performed.

- Suppose $\Delta evil = -1$. Then $\Delta\lceil fbad/3 \rceil = 0$, $C \in \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}^2_{\text{evil}}$, F is unchanged. If $C \in \mathbb{C}^2_{\text{evil}}$ then clearly $\Delta\delta' \geq 0$. Suppose $C \in \mathbb{C}^3_{\text{evil}}$. Then $\Delta fbad \in \{0, 1\}$.

—Suppose $\Delta fbad = 0$. Then for every maximum matching M after there exists a maximum matching M' before such that $F_M = F_{M'}$ and $fgood(M) = fgood(M') - 1$. Hence $\Delta\delta' \geq 0$.

—Suppose $\Delta fbad = 1$. Then before $mbad = 2$. It follows that after: $mbad = 0$ and every maximum matching M satisfies $|F_M| = 1$ and l_M is even. Hence $\delta' = 1$ after.

- Suppose $\Delta evil = 0$. Then $\Delta\lceil fbad/3 \rceil = -1$, $C \in \mathbb{C}^2_{\text{evil}} \cup \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$.

—Suppose $C \in \mathbb{C}^2_{\text{evil}}$. Then before $mbad = 1$ and hence after: $mbad = 0$ and the single maximum matching satisfies l_M is even and $|F_M| = 1$. Hence $\delta' = 1$ after.

—Suppose $C \in \mathbb{C}^1_{\text{evil}}$. Then $\deg(C) \in DEG$, F is unchanged, and $mbad = 2$ before. Hence after: $mbad = 0$ and every maximum matching M satisfies l_M is even and $|F_M| = 1$. Hence $\delta' = 1$ after.

—Suppose $C \in \mathbb{C}_{\text{ann}}$. Then $mbad = 1$ before. Therefore after $DEG = \emptyset$ and $\Delta\delta' \geq 0$.

Case 2.3: $\Delta(n-c) = 2$. Let C_1 and C_2 be the cycles on which the cent-mapping was performed. In this case $\Delta|F| \leq 0$, $\Delta(l+evil) \geq -2$, $\Delta(l+evil+\lceil fbad/3 \rceil) \geq -2$. If $\Delta(l+evil+\lceil fbad/3 \rceil) \geq -1$ then clearly $\Delta \geq 0$. Suppose $\Delta(l + evil + \lceil fbad/3 \rceil) = -2$.

- Suppose $\Delta(l + evil) = -1$. Then $\Delta\lceil fbad/3 \rceil = -1$.

—Suppose $\Delta fbad = -1$. Then $mbad = 1$ before, $C_1 \in \mathbb{C}_{\text{ann}}$, $C_2 \in \mathbb{C}_{\text{good}} \cup \mathbb{C}^2_{\text{evil}}$. Hence after: $mbad = 0$, $DEG = \emptyset$, $|F_\emptyset| = 1$ ($F_\emptyset$ is unchanged). If $l_\emptyset$ is even then clearly $\Delta\delta' \geq 0$. Suppose $l_\emptyset$ is odd. Then $C \in \mathbb{C}_{\text{good}}$ and hence $fgood(\emptyset) = 0$ after. Therefore $\Delta\delta' \geq 0$.

—Suppose $\Delta fbad = -2$. Then $mbad = 2$ before and $mbad = 0$ after. Note that before F_M is fixed for every maximum matching M (i.e., $F_M = F'$). Let M be a maximum matching after. Then either $F_M = F'$ (i.e., as before), or l_M is odd and $fgood(M) = 0$.

(The latter may happen only if $nona > 0$ and $C_1 \in \mathbb{C}_{\text{ann}} \cup \mathbb{C}_{\text{nona}}$.) In both cases $\delta' = 1$ after.

—Suppose $\Delta fbad = -3$. Then $C_1, C_2 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, $\deg(C_1), \deg(C_2) \in DEG$, and for every maximum matching M after, there exists a maximum matching M' before, such that $F_M = F_{M'}$ and $fgood(M) = fgood(M')$, hence $\delta' = 1$ after.

- Suppose $\Delta(l + evil) = -2$. Then $\Delta\lceil fbad/3\rceil = 0$ and only the following cases are possible.

—$C_1 \in \mathbb{C}^3_{\text{evil}}$, $C_2 \in \mathbb{C}^2_{\text{evil}}$. Then $\Delta fbad \in \{0, 1\}$. If $\Delta fbad = 0$ then for every maximum matching M after there exists a maximum matching M' before such that $F_M = F_{M'}$ and $fgood(M) = fgood(M') - 1$, hence $\Delta\delta' \geq 0$. Suppose $\Delta fbad = 1$. Then $\Delta mbad = 2$ before. Hence after: $mbad = 0$, and every maximum matching satisfies l_M is even and $|F_M| = 1$, hence $\Delta\delta' = 0$.

—$C_1 \in \mathbb{C}^3_{\text{evil}}$, $C_2 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$.

* $\deg(C_2) \in DEG$. Then $\Delta fbad \in \{0, 1\}$. If $\Delta fbad = 0$ then clearly $\Delta \geq 0$. Suppose $\Delta fbad = 1$. Then $mbad = 2$ before and after: $mbad = 0$, and either (l_M is even and $|F_M| = 1$, or (l_M is odd and $fgood(M) = 0$). Hence $\Delta\delta' = 0$.

* $\deg(C_2) \notin DEG$. Then $\Delta fbad \in \{0, 1\}$ again. In both cases $C_2 \in \mathbb{C}_{\text{ann}}$, and after $mbad = 0$ and every maximum matching M after satisfies $(1, C') \in M$, where $C' \in \mathbb{C}_{\text{nona}}$, l_M is odd and $fgood(M) = 0$ (since $\delta' = 1$ before). Hence $\Delta\delta' = 0$.

—$C_1 \in \mathbb{C}^3_{\text{evil}}$, $C_2 \in \mathbb{C}_{\text{nona}}$. Then $\Delta fbad \in \{0, 1\}$.

* $\Delta fbad = 0$. Then if $1 \in DEG$ then $nona \geq 2$. Hence for every maximum matching M after there exists a maximum matching M' before such that $l_M = l_{M'} - 1$ and $fgood(M) = fgood(M') - 1$. Thus before: $mbad = 0$ and every maximum matching M' for which $fgood(M') = 1$ satisfied l_M is even. Thus $\Delta\delta' \geq 0$.

* $\Delta fbad = 1$. Then before: $mbad = 2$ and thus $nona = 1$. It follows that $1 \notin DEG$ and hence after: $mbad = 0$, and every maximum matching M satisfies l_M is odd and $fgood(M) = 0$. Thus $\delta' = 1$ after.

—$C_1, C_2 \in \mathbb{C}^2_{\text{evil}}$, or $C_1, C_2 \in \mathbb{C}^1_{\text{evil}}$, or $C_1, C_2 \in \mathbb{C}_{\text{ann}}$. Then clearly $\Delta\delta' \geq 0$.

—$C_1 \in \mathbb{C}_{\text{ann}}$, $C_2 \in \mathbb{C}_{\text{nona}}$. If $1 \in DEG$ then clearly $\Delta\delta' \geq 0$. Suppose $1 \notin DEG$. Then $\Delta fbad \in \{0, 1\}$ and for every maximum matching before $F_M = F'$ and $fgood(M) = fgood'$ are fixed (i.e., independent of M).

* $nona > 1$ before. Then $|F'| > 1$ and hence $mbad = 0$, $l(F')$ is odd and $fgood' = 0$. Thus after, every maximum matching M satisfies: $l_M = l(F') - 2$ is odd and $fgood(M) = fgood' = 0$, and thus $\delta' = 1$.

* $nona = 1$ before. Then after: $nona = 0$ and for every maximum matching M, $F_M = F''$ (i.e., independent of M) and $l(F'') = l(F') - 1$. There there are two possible cases. In the first case $fgood' = 0$ before, and then $\Delta fbad = 1$, and hence $mbad = 2$ before. In the second case $fgood = 1$, and then $\Delta fbad = 0$, $mbad = 0$ and $l(F')$ is even (since F' contains a non-annoying leaf). It follows that in both cases after: $mbad = 0$, $fgood'' = 0$ and $l(F'')$ is odd. Hence $\delta' = 1$ after.

—$C_1, C_2 \in \mathbb{C}_{\text{nona}}$. If $1 \notin DEG$ or $nona > 2$ then clearly $\Delta\delta' \geq 0$. We shall prove that no other case is not possible. Suppose $1 \in DEG$ and $nona = 2$. It follows that before for every maximum matching M, $(1, C) \in M$ where $C \in \mathbb{C}_{\text{nona}}$, l_M is odd and $fgood(M) = 0$. Hence $mbad = 0$ before and $\Delta fbad = 1$, a contradiction to $\Delta\lceil fbad/3\rceil = 0$. ■

Lemma 16. *Let $\Delta = \Delta(n - c + l + evil + \lceil fbad/3\rceil + \delta')$. There exists a sequence of η cent-mappings where each satisfies $\Delta = 0$.*

Proof. Below we present Algorithm 3, which satisfies $\Delta = 0$ for every cent-mapping. Moreover, after the run of this algorithm the following conditions are satisfied: *(i)* $DEG = \emptyset$, *(ii)* $\delta' = 0 \Rightarrow l_\emptyset$ is even and $F_\emptyset \neq 1$, and *(iii)* $\delta' = 1 \Rightarrow l_\emptyset$ is odd. It follows that if we apply Algorithm 2 after Algorithm 3, then every cent-mapping performed by the latter algorithm satisfies $\Delta = 0$. (Note that in this case Steps 3–16 in Algorithm 2 are skipped, since $DEG = \emptyset$.) ■

Algorithm 3. Improve δ'

```
if $mbad = 2$ then
  Let $M$ be a maximum matching, let $C_1, C_2 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $\deg(C_1), \deg(C_2) \notin M$, and $\deg(C_1) \neq \deg(C_2)$. Perform a bad cent-mapping on $C_1$ and $C_2$
else if $mbad = 1$ then
  $i \leftarrow \max\{j : j \in DEG\}$
  if $i > 1$ then
    Let $M$ be a maximum matching where $i$ is not matched. Let $C \in \mathbb{C}^1_{\text{evil}}$ satisfying $\deg(C) = i$
    if $l_M$ is even then
      Perform 2 proper cent-mapping on $C$ such that $\Delta l = 2$ and $\Delta evil = -1$
    else
      Perform an improper cent-mapping on $C$ followed by a proper cent-mapping satisfying $\Delta l = 1$, $\Delta evil = -1$ and $\Delta |F_M| > 1$ after
    end if
  else
    if $l_\emptyset = 0$ then
      Let $C \in \mathbb{C}_{\text{ann}}$, let $C_1 \neq C$ be any other cycle satisfying $\deg(C_1) > 0$.
      if $C_1 \in \mathbb{C}_{\text{good}} \cup \mathbb{C}^2_{\text{evil}}$ then
        Perform a bad cent-mapping on $C$ and $C_1$
      else
        Then $C_1 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$. Let $C_2$ be a cycle of the same class as $C_1$, different from $C$ and $C_1$, satisfying $\deg(C_2) = \deg(C_1)$. Perform a bad cent-mapping on $C_1$ and $C_2$. Let $C_3$ be new cycle. Perform a bad cent-mapping on $C$ and $C_3$
      end if
    else if $|F| > 1$ then
      Depending on the parity of $l_\emptyset$: perform either a proper or an improper cent-mapping on a cycle from $\mathbb{C}_{\text{ann}}$ such that after: $l_\emptyset$ is even and $|F_\emptyset| > 1$
    else if $l_\emptyset$ is odd then
      if $evil_2 > 0$ then
        Let $C' \in \mathbb{C}^2_{\text{evil}}$. Perform a bad cent-mapping on $C$ and $C'$
      else
        Perform a proper cent-mapping on $C$
      end if
    else
      Perform an improper cent-mapping on $C$
    end if
  end if
end if
call Procedure 4
```

Procedure 4. Handle $mbad = 0$

```
if $1 \in DEG$ and $nona > 0$ then
  Let $M$ be a maximum matching in $BCM$ satisfying $(1, C_1) \in M$, where $C_1 \in \mathbb{C}_{\text{nona}}$
  if $|F_M| = 1$ and $nona \geq 2$ then
    Let $M$ be a maximum matching in $BCM$ satisfying $(1, C_2) \in M$ where $C_1 \neq C_2 \in \mathbb{C}_{\text{nona}}$
  end if
else
  Let $M$ be any maximum matching in $BCM$
end if
```

(*continued*)

Procedure 4. (*Continued*)

```
if $l_M$ is odd, and $fgood(M) = 1$, and after $C \in \mathbb{C}^3_{\text{evil}} \setminus M$ is replaced by a leaf $|F_M| = 1$ then
    if there exists $i \in DEG$ such that $i \leq \deg(C)$ then
        Update $M$ such that $(i, C) \in M$
    end if
end if
if $l_M$ is odd and there exists $C \in \mathbb{C}^3_{\text{evil}} \setminus M$ that can be replaced by a leaf such that $|F_M| > 1$ after then
    Perform this replacement
else if $l_M$ is even and $|F_M| = 1$ then
    if $fgood(M) \geq 2$ then
        Replace two unmatched cycles in $\mathbb{C}^3_{\text{evil}}$ by two leaves (each cycle is replaced by one leaf)
    else if $evil_2 > 0$ then
        Replace a cycle in $\mathbb{C}^2_{\text{evil}}$ by two leaves
    else if $fbad > 0$ then
        Let $i_1, i_2, i_3 \in DEG \setminus M$, where $i_1 < i_2 < i_3$. Let $C_1, C_2, C_3 \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $\deg(C_j) = j$ for
        $j = 1, 2, 3$. Perform a bad cent-mapping on $C_1$ and $C_2$. Replace $C_3$ by two leaves
    else if $|M| > 0$ then
        Choose $C \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, $C' \in \mathbb{C}^3_{\text{evil}} \cup \mathbb{C}_{\text{nona}}$ such that $(\deg(C), C') \in M$
        if $\deg(C) = 1$ then
            Perform an improper cent-mapping on $C$
            if $C' \in \mathbb{C}^3_{\text{evil}}$ then
                Replace $C$ by a leaf
            end if
        else
            Replace $C$ by two leaves
        end if
    else if $ann > 0$ and $nona > 0$ then
        Let $C_1, C_2 \in \mathbb{C}_{\text{ann}}$, $C_3 \in \mathbb{C}_{\text{nona}}$. Perform a proper cent-mapping on $C_1$. Perform a bad cent-mapping on $C_2$
        and $C_3$
    else if $\mathbb{C}^3_{\text{evil}} >$ then
        Replace $C \in \mathbb{C}^3_{\text{evil}}$ by a leaf
    end if
end if
for all $(i, C) \in M$ do
    Perform a bad cent-mapping on $C$ and a $C' \in \mathbb{C}^1_{\text{evil}} \cup \mathbb{C}_{\text{ann}}$, where $\deg(C') = i$, such that $\Delta(l + evil) = -2$
    (Lemma 12).
end for
```

ACKNOWLEDGMENTS

This study was supported in part by the Israeli Science Foundation (grant 309/02).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bader, D., Moret, B.M., and Yan, M. 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.* 8, 483–491.

Bergeron, A., Mixtacki, J., and Stoye, J. 2006. On sorting by translocations. *J. Comput. Biol.* 13, 567–578.

Hannenhalli, S. 1996. Polynomial algorithm for computing translocation distance between genomes. *Discrete Appl. Math.* 71, 137–151.

Kaplan, H., Shamir, R., and Tarjan, R.E. 2000. Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.* 29, 880–892.

Ozery-Flato, M., and Shamir, R. 2006a. An $O(n^{3/2}\sqrt{\log(n)})$ algorithm for sorting by reciprocal translocations. *Lect. Notes Comput. Sci.* 4009, 258–269.

Ozery-Flato, M., and Shamir, R. 2006b. Sorting by translocations via reversals theory. *Lect. Notes Comput. Sci.* 4205, 87–98.

Ozery-Flato, M., and Shamir, R. 2007. Rearrangements in genomes with centromeres—part I: translocations. *Lect. Notes Comput. Sci.* 4453, 339–353.

Perry, J., Slater, H., and Choo, K.A. 2004. Centric fission–simple and complex mechanisms. *Chromosome Res.* 12, 627–640.

Searle, J. 1998. Speciation, chromosomes, and genomes. *Genome Res.* 8, 1–3.

Sullivan, B., Blower, M., and Karpen, G. 2001. Determining centromere identity: cyclical stories and forking paths. *Nat. Rev. Genet.* 2, 584–596.

Address reprint requests to:
Michal Ozery-Flato
School of Computer Science
Tel-Aviv University
Tel-Aviv 69978, Israel

E-mail: ozery@post.tau.ac.il