A linear time algorithm for the inversion median problem in circular bacterial genomes

Enno Ohlebusch, Mohamed I. Abouelhoda, and Kathrin Hockel

Faculty of Engineering and Computer Sciences, University of Ulm, 89069 Ulm, Germany
Email: Enno.Ohlebusch@uni-ulm.de

Abstract

In the median problem, we are given a distance or dissimilarity measure d, three genomes G_1, G_2, and G_3, and we want to find a genome G (a median) such that the sum $\sum_{i=1}^{3} d(G, G_i)$ is minimized. The median problem is a special case of the multiple genome rearrangement problem, where one wants to find a phylogenetic tree describing the most "plausible" rearrangement scenario for multiple species. The median problem is NP-hard for both the breakpoint and the reversal distance. To the best of our knowledge, there is no approach yet that takes biological constraints on genome rearrangements into account. In this paper, we make use of the fact that in circular bacterial genomes the predominant mechanism of rearrangement are inversions that are centered around the origin or the terminus of replication and single gene inversions. These constraints simplify the median problem significantly. More precisely, we show that the median problem for the reversal distance can be solved in linear time for circular bacterial genomes.

Key words: median problem, reversal distance, inversions, circular genomes, genome rearrangements, comparative genomics

1 Introduction

During evolution, the genomic DNA sequences of organisms are subject to genome rearrangements such as *transpositions* (where a section of the genome is excised and inserted at a new position in the genome, without changing orientation) and *inversions* (where a section of the genome is excised, reversed in orientation, and re-inserted). In unichromosomal genomes, the most common rearrangements are inversions, which are usually called *reversals* in bioinformatics. In the following, we will focus on unichromosomal genomes and use the terms "inversion" and "reversal" synonymously. The study of genome rearrangements started more than 65 years ago [8], but interest on the subject

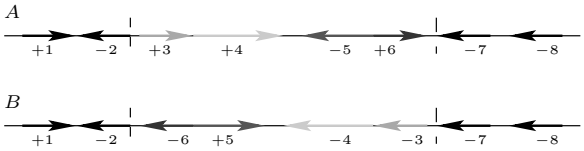

Fig. 1. Genome $(+1, -2, +3, +4, -5, +6, -7, -8)$ before and after the inversion $\rho(3, 6)$.

has flourished in the last decade because of the progress in large-scale sequencing. In the context of genome rearrangement, a genome G is typically viewed as a *signed permutation*, where each integer corresponds to a unique gene and the sign corresponds to its orientation. A + (−) sign means that the gene lies on the leading (lagging) DNA strand.

As usual in the context of genome rearrangement problems, we assume that orthologous genes between two genomes G_1 and G_2 have already been determined. That is, we model the genomes as permutations on the same set of orthologous genes $\{1, \ldots, n\}$ and do not consider the other genes (nor non-coding regions). So let $G_1 = (\pi_1, \ldots, \pi_n)$ and $G_2 = (\gamma_1, \ldots, \gamma_n)$ be permutations of genes $\{1, \ldots, n\}$. Two adjacent genes π_i and π_{i+1} in G_1 determine a *breakpoint* in G_1 w.r.t. G_2 if and only if neither π_i precedes π_{i+1} in G_2 nor $-\pi_{i+1}$ precedes $-\pi_i$ in G_2. The *breakpoint distance* $bd(G_1, G_2)$ between G_1 and G_2 is defined as the number of breakpoints in G_1 w.r.t. G_2 [17, 25]. This is clearly equal to the number of breakpoints in G_2 w.r.t. G_1. In other words, the breakpoint distance between G_1 and G_2 is the smallest number of places where one genome must be broken so that the pieces can be rearranged to form the other genome.

Given a genome $G = (\pi_1, \ldots, \pi_{i-1}, \pi_i \ldots, \pi_j, \pi_{j+1}, \ldots, \pi_n)$, a *reversal* $\rho(i, j)$ applied to G reverses the segment $\pi_i, \ldots, \pi_j$ and produces the permutation $G\rho(i, j) = (\pi_1, \ldots, \pi_{i-1}, -\pi_j, -\pi_{j-1}, \ldots, -\pi_{i+1}, -\pi_i, \pi_{j+1}, \ldots, \pi_n)$ (see Figure 1 for an illustration). Given two genomes G_1 and G_2, the *reversal distance* $rd(G_1, G_2)$ between them is defined as the minimum number of reversals required to convert one genome into the other. (The phrase *sorting by reversals* refers to the equivalent problem of finding the minimum number of reversals required to convert a permutation π into the identity permutation.) The study of the reversal distance was pioneered by Sankoff [20] and has received increasing attention in recent years. There are dozens of papers on the subject; see e.g. [1, 2, 10, 13] and the references therein.

The median problem is NP-hard for both the breakpoint and the reversal distance [5, 19]. That is the reason why researchers developed heuristics to solve the median and the multiple genome rearrangement problem. Very good heuristics exist for the breakpoint-based multiple genome rearrangement problems [3, 21]. These rely on the ability to solve the breakpoint median problem by reducing it to the Traveling Salesman Problem. Solutions to the reversal

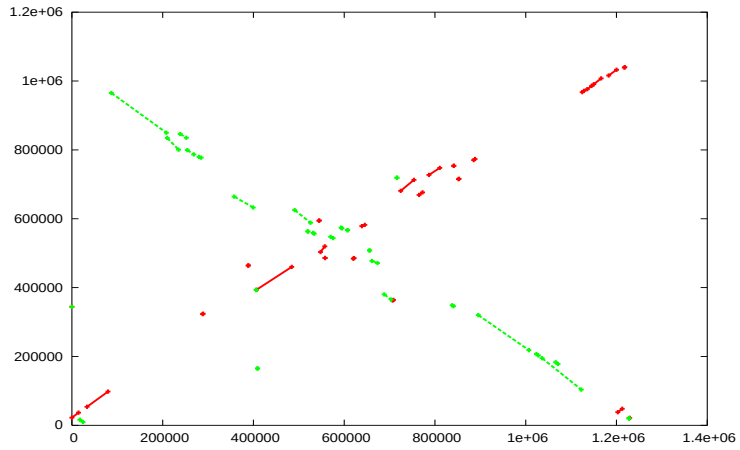

Fig. 2. An X-shaped pattern resulting from a genome comparison of the bacteria *Chlamydia pneumonia* (x-axis) and *Chlamydia trachomatis* (y-axis).

median problem can be found in [4, 6, 16, 22]. There is a dispute about the "right" distance in multiple genome rearrangement problems. While the authors of [3, 21] argue that the breakpoint distance is the better choice, in [16] it is conjectured that the usage of the reversal distance yields better phylogenetic reconstructions. Furthermore, [4] discusses some advantages of the reversal distance approach over the breakpoint distance approach.

To the best of our knowledge, there is no approach yet that takes biological constraints on genome rearrangements into account. In this paper, we make use of the fact that in circular bacterial genomes the predominant mechanism of rearrangement are inversions that are centered around the origin or the terminus of replication [9, 12, 23, 24] and single gene inversions [7, 14]. These constraints simplify the median problem significantly. More precisely, we show that the median problem for the reversal distance can be solved in linear time for circular bacterial genomes.

2 Inversions around the origin/terminus of replication and single gene inversions

In this paper, we study the median problem (unless stated otherwise, the term median problem refers to the reversal median problem) for circular bacterial genomes. In whole genome comparisons, an X-shaped pattern (see Figure 2) in plots of orthologous genes has been observed [9, 12, 23, 24], indicating that almost all long range inversions within closely related circular bacterial genomes are centered around the origin or the terminus of replication. Among the short range inversions, single gene inversions [7, 14] seem to be predominant. On the one hand, Tiller and Collins [24] have argued that a substantial proportion of rearrangements result from recombination sites that are determined by the po-

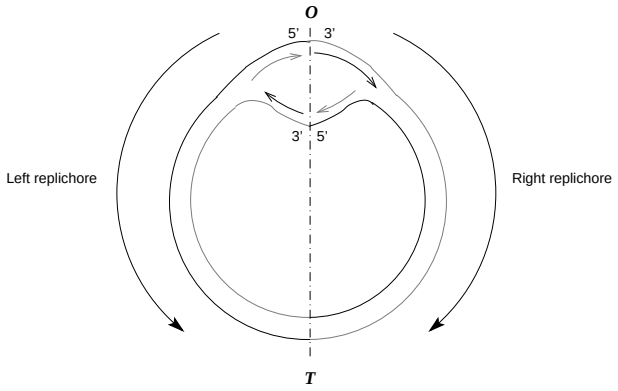

Fig. 3. A genome that replicates bi-directionally from a single origin.

sition of the replication forks. In genomes that replicate bi-directionally from a single origin, the two replication forks (see Figure 3) will be approximately equidistant from the origin, so that genes are inverted and translocated to the "opposite side" of the genome: a mirror-image position across the replication axis (defined by the origin O and terminus T of replication). On the other hand, Mackiewicz et al. [15] argued that selection may be mainly responsible. In their opinion, "selection pressure leads to the optimal position of genes with respect to the distance from the origin of replication." Furthermore, they write that another "selection force that could lead to biased rearrangements might be the trend towards keeping both replichores the same size." Moreover, according to Hughes [12], "a high frequency of recombination in the terminus region is related to the mechanism of chromosome separation after replication."

Whatever the reasons might be, the observations strongly indicate that inversions around the origin/terminus of replication and single gene inversions are the predominant rearrangements in prokaryotic genomes. In the following, we will take this into account.

As usual in the comparison of genomes on the gene level, we assume that the genomes have the same set $\{1, \ldots, n\}$ of unique genes and that inversions do not cut genes. As a consequence, genes may neither overlap on the same DNA strand nor on different DNA strands. In our model, in which inversions around the origin/terminus of replication and single gene inversions are the predominant mechanism of rearrangement, it is further assumed that in each genome, these n genes occur in the same order w.r.t. the distance to the origin of replication.

Because the genes keep their distance to the origin O, we enumerate them in increasing distance to O. That is, starting with the origin of replication, we simultaneously traverse both DNA strands of the circular genome in clockwise and counterclockwise order. This process ends when the terminus T of replication is reached and it divides the circular genome into two halves, called

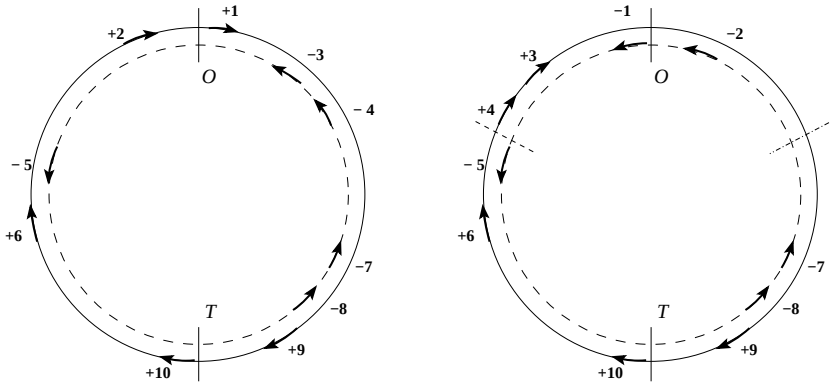

Fig. 4. Left: A cartoon representation of a circular bacterial genome. Of course, bacteria have hundreds, or even thousands, of genes. Moreover, a bacterial genome does not have long stretches of DNA without genes. Right: The same genome after the inversion $\overline{\rho}(4)$.

replichores. The clockwise traversal yields the right replichore and the counterclockwise traversal yields the left replichore. A gene encountered gets the next number (the first gene gets number 1). If this gene is lying on the leading strand, it is labeled with a + sign, otherwise it gets a − sign. If it was encountered in the clockwise (resp. counterclockwise) direction, its number is put to the right (resp. left) of the origin O and a 0 to the left (resp. right) of O, which for better readability will be denoted by the symbol |. For example, if the first gene is encountered in the counterclockwise direction and is lying on the leading strand, then this yields $(+1 \mid 0)$. A more complex example is $(+10, 0, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, +9, 0)$, which is shown in Figure 4.

In what follows, $\overline{\rho}(i)$ denotes an inversion centered around the origin of replication that acts on the ith nearest genes of O. Furthermore, we will use postfix notation to denote the application of a reversal to a genome. For example,

$$\begin{aligned}&(+10, 0, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, +9, 0)\ \overline{\rho}(4)\\ =\,&(+10, 0, 0, 0, +6, -5, +4, +3, 0, -1 \mid 0, -2, 0, 0, 0, 0, -7, -8, +9, 0)\end{aligned}$$

Similarly, $\underline{\rho}(i)$ denotes an inversion centered around the terminus of replication that acts on the ith nearest genes of T. As an example consider

$$\begin{aligned}&(+10, 0, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, +9, 0)\ \underline{\rho}(2)\\ =\,&(0, -9, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, 0, -10)\end{aligned}$$

Next, we will simplify the above representation without loosing any information. $(+10, 0, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, +9, 0)$, for example, will be represented by the bit vector $(1, 0, 1, 1, 0, 0, 1, 1, 1, 0)$ and the

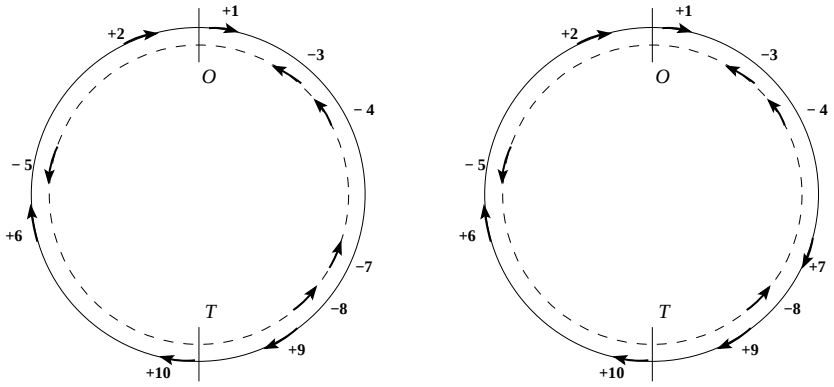

Fig. 5. Left: The cartoon representation of our circular bacterial genome. Right: The same genome after the single gene inversion $\sigma(7)$.

orientation vector $(+,+,-,-,-,+,-,-,+,+)$. In the bit vector, a 1 (resp. 0) at position p means that the gene with number p is located in the right (resp. left) replichore of the circular bacterial genome. Furthermore, a $+$ (resp. $-$) sign in the orientation vector at position p means that gene p lies on the leading (resp. lagging) strand. Therefore, the preceding inversions are modeled by

$$\begin{aligned}&(+1,+0,-1,-1,-0,+0,-1,-1,+1,+0)\ \overline{\rho}(4)\\ =&(-0,-1,+0,+0,-0,+0,-1,-1,+1,+0)\end{aligned}$$

$$\begin{aligned}&(+1,+0,-1,-1,-0,+0,-1,-1,+1,+0)\ \underline{\rho}(2)\\ =&(+1,+0,-1,-1,-0,+0,-1,-1,-0,-1)\end{aligned}$$

In the following, we will also consider single gene inversions. A single gene inversion $\sigma(i)$ flips the ith sign in the orientation vector because the ith gene is translocated to the opposite DNA strand and thus changes its orientation. However, a single gene inversion $\sigma(i)$ does not change the ith bit in the bit vector because the gene remains in its replichore. The following example is also depicted in Figure 5.

$$\begin{aligned}&(+1,+0,-1,-1,-0,+0,-1,-1,+1,+0)\ \sigma(7)\\ =&(+1,+0,-1,-1,-0,+0,+1,-1,+1,+0)\end{aligned}$$

Lemma 2.1 *The composition of inversions is commutative and associative.*

Proof Let ρ_1, ρ_2, and ρ_3 be inversions. We have $\rho_1 \cdot \rho_2 = \rho_2 \cdot \rho_1$ (commutativity) and $(\rho_1 \cdot \rho_2) \cdot \rho_3 = \rho_1 \cdot (\rho_2 \cdot \rho_3)$ (associativity) because every gene is inverted

the same number of times on either side of the respective equation. □

An important consequence of the preceding lemma is that reordering any sequence of inversions does not change the result. Thus, $G\rho_1\rho_2\cdots\rho_k$ (recall that an application of a reversal to a genome is denoted by postfix notation) is the genome obtained from G by applications of the reversals $\rho_1, \rho_2, \ldots, \rho_k$ in an arbitrary order.

Note that every reversal ρ has an inverse, viz. ρ itself because $\rho \cdot \rho = id$.

Let $\rho(n) := \overline{\rho}(n) = \underline{\rho}(n)$ be the inversion that inverts the whole genome. The reflection $flip(G)$ of genome G is defined by $flip(G) := \rho(n)$. A genome G is biologically equivalent to its reflection [11]. Given a reversal ρ around the origin/terminus of replication, a reversal τ around the origin/terminus of replication satisfying $\rho(n) \cdot \rho = \tau$ is called the *complementary reversal* of ρ.

Lemma 2.2 *Every reversal ρ around the origin/terminus of replication has a (unique) complementary reversal τ around the terminus/origin of replication.*

Proof If $\rho = \overline{\rho}(i)$, then $\tau = \underline{\rho}(n-i)$ because $\rho(n) \cdot \overline{\rho}(i) = \underline{\rho}(n-i)$. Otherwise, if $\rho = \underline{\rho}(i)$, then $\tau = \overline{\rho}(n-i)$ because $\rho(n) \cdot \underline{\rho}(i) = \overline{\rho}(n-i)$. □

The preceding lemma in conjunction with the fact that a genome and its reflection are equivalent implies that one can restrict solely to inversions around the origin of replication (or, by a symmetric argument, to inversions around the terminus of replication).

3 The reversal distance

Given two genomes G and G', we fix one of the genomes, say G', and try to transform G into G' or $flip(G')$ by as few inversions as possible.

Let $(\pm_1 b_1, \pm_2 b_2, \pm_3 b_3, \ldots, \pm_n b_n)$ be the oriented bit vector representation of a circular bacterial genome G. Here $\pm_i$ denotes the orientation of the i-th gene, i.e., $\pm_i = +$ ($\pm_i = -$) if gene i lies on the leading (lagging) DNA strand. In the rest of the paper, we will just speak of genome G, that is, we omit the phrase "circular bacterial." Furthermore, we will use the following notations for $1 \leq i \leq j \leq n$: $G[i] = \pm_i b_i$, $G_b[i] = b_i$, $G_o[i] = \pm_i$, $G[i..j] = (\pm_i b_i, \ldots, \pm_j b_j)$, $G_b[i..j] = (b_i, \ldots, b_j)$, and $G_o[i..j] = (\pm_i, \ldots, \pm_j)$. That is, G_b denotes the genes without their orientation, G_o denotes the orientations of the genes, and G denotes the genes with their orientation.

+1:+1+1+1+0:+1+1+0:+1

+1:-0-0+0-1:+1+1+0:-0

Fig. 6. Breakpoints between two genomes. The gene order breakpoints are depicted by colons, whereas the gene orientation breakpoints are underlined.

The next definition is a modification of the usual definition of a breakpoint. The distinction between gene order breakpoints and gene orientation breakpoints is crucial in our context.

Definition 3.1 Let two genomes $G = (\pm_1 b_1, \pm_2 b_2, \pm_3 b_3, \ldots, \pm_n b_n)$ and $G' = (\pm'_1 b'_1, \pm'_2 b'_2, \pm'_3 b'_3, \ldots, \pm'_n b'_n)$ be given. Two consecutive indices i and $i+1$ determine a *gene order breakpoint* if and only if neither $G_b[i..i+1] = G'_b[i..i+1]$ nor $G_b[i..i+1] = (flip(G'[i..i+1]))_b$. An index j is called *gene orientation breakpoint* if

- either $G_b[i] = G'_b[i]$ and $G_o[i] \neq G'_o[i]$
- or $G_b[i] \neq G'_b[i]$ and $G_o[i] = G'_o[i]$.

Figure 6 shows two genomes G and G' with three gene order breakpoints and one gene orientation breakpoint. Note that G and G' are equivalent (i.e., $G = G'$ or $G = flip(G')$) if and only if there is no (gene order nor gene orientation) breakpoint between G and G'. On the one hand, an inversion around the origin/terminus of replication can remove a gene order breakpoint, but a single gene inversion cannot. On the other hand, a single gene inversion can remove a gene orientation breakpoint, but an inversion around the origin/terminus of replication cannot. This is made precise in the following lemmata.

Lemma 3.2 *Let* $G = (\pm_1 b_1, \pm_2 b_2, \ldots, \pm_n b_n)$, $G' = (\pm'_1 b'_1, \pm'_2 b'_2, \ldots, \pm'_n b'_n)$, *and* $\overline{\rho}(i)$ *with* $1 \leq i \leq n-1$ *be given.*

(1) For all j *with either* $1 \leq j < i$ *or* $i < j < n$ *we have:* $(j, j+1)$ *is a gene order breakpoint between* G *and* G' *if and only if* $(j, j+1)$ *is a gene order breakpoint between* $G\overline{\rho}(i)$ *and* G'.

(2) If $(i, i+1)$ *is a gene order breakpoint between* G *and* G', *then* $(i, i+1)$ *is not a gene order breakpoint between* $G\overline{\rho}(i)$ *and* G' *and vice versa.*

(3) For all k, $1 \leq k \leq n$, *index* k *is a gene orientation breakpoint between* G *and* G' *if and only if* k *is a gene orientation breakpoint between* $G\overline{\rho}(i)$ *and* G'.

Proof (1) If $i < j < n$, then there is nothing to show because $\overline{\rho}(i)$ has no effect on the genes j and $j+1$. Suppose $1 \leq j < i$. The following equivalences hold:

$(j, j+1)$ is a gene order breakpoint between G and G'

$\Leftrightarrow$ either $(b_j = b'_j$ and $b_{j+1} \neq b'_{j+1})$ or $(b_j \neq b'_j$ and $b_{j+1} = b'_{j+1})$
$\Leftrightarrow$ either $(flip(b_j) \neq b'_j$ and $flip(b_{j+1}) = b'_{j+1})$
or $(flip(b_j) = b'_j$ and $flip(b_{j+1}) \neq b'_{j+1})$
$\Leftrightarrow (j, j+1)$ is a gene order breakpoint between $G\overline{\rho}(i)$ and G'

(2) This case follows by a similar reasoning as in (1).
(3) This is because $\overline{\rho}(k)$ either changes both $G_b[k]$ and $G_o[k]$ or it has no effect on both of them. □

Lemma 3.3 *Let* $G = (\pm_1 b_1, \pm_2 b_2, \ldots, \pm_n b_n)$, $G' = (\pm'_1 b'_1, \pm'_2 b'_2, \ldots, \pm'_n b'_n)$, *and* $\sigma(i)$ *with* $1 \leq i \leq n$ *be given.*

(1) For all j *with either* $1 \leq j < i$ *or* $i < j < n$ *we have:* j *is a gene orientation breakpoint between* G *and* G' *if and only if* j *is a gene orientation breakpoint between* $G\sigma(i)$ *and* G'.
(2) If i *is a gene orientation breakpoint between* G *and* G', *then* i *is not a gene orientation breakpoint between* $G\sigma(i)$ *and* G' *and vice versa.*
(3) For all k *with* $1 \leq k < n$ *we have:* $(k, k+1)$ *is a gene order breakpoint between* G *and* G' *if and only if* $(k, k+1)$ *is a gene order breakpoint between* $G\sigma(i)$ *and* G'.

Proof Straightforward. □

In particular, any reversal can remove at most one breakpoint. The following simple procedure $rd(G, G')$ returns the reversal distance d between two genomes G and G', as well as d inversions that transform G into G'.

procedure $rd(G, G')$
determine the gene order breakpoints $(i_1, i_1+1), \ldots, (i_k, i_k+1)$ of G and G'
determine the gene orientation breakpoints $j_1, \ldots, j_\ell$ of G and G'
return the reversal distance $k + \ell$ and $\overline{\rho}(i_1), \ldots, \overline{\rho}(i_k), \sigma(j_1), \ldots, \sigma(j_\ell)$

The correctness of procedure $rd(G, G')$ is a direct consequence of the preceding lemmata. This can be seen as follows. By Lemma 3.2, each reversal $\overline{\rho}(i_1), \ldots, \overline{\rho}(i_k)$ removes one gene order breakpoint, so that there is no gene order breakpoint between $\tilde{G} := G\overline{\rho}(i_1) \cdots \overline{\rho}(i_k)$ and G'. Furthermore, according to Lemma 3.3, each single gene inversion $\sigma(j_1) \ldots \sigma(j_\ell)$ removes one gene orientation breakpoint, so that there is no breakpoint at all between $\tilde{G}\sigma(j_1) \ldots \sigma(j_\ell)$ and G'. Because no reversal can remove more than one breakpoint, the reversal distance between G and G' is $k + \ell$.

Clearly, the worst case running time of the procedure $rd(G, G')$ is $O(n)$.

4 The median problem for the reversal distance

In the median problem we want to find a genome G (a median) for G_1, G_2, and G_3 such that $\sum_{i=1}^{3} rd(G, G_i)$ is minimized. In the following, let

$$d_m(G_1, G_2, G_3) = \min\{\sum_{i=1}^{3} rd(G, G_i) \mid G \text{ is a genome}\}$$

With G_1, G_2, and G_3 we associate a labeled, weighted graph (V, E) defined as follows. The set of vertices $V = \{1, \ldots, n\}$ coincides with the set of genes. For every i, $1 \leq i \leq n-1$, there is an edge $(i, i+1) \in E$ with weight $w_c(i, i+1)$, where $w_c(i, i+1)$ is the number of times the genes i and $i+1$ are on the same replichore in G_1, G_2, and G_3. Obviously, genes i and $i+1$ occur $w_d(i, i+1) = 3 - w_c(i, i+1)$ times on different replichores. If $w_c(i, i+1) = 1$, then we further label the edge $(i, i+1)$ with the genome G_j for which $G_j[i]$ and $G_j[i+1]$ are on the same replichore. Analogously, if $w_d(i, i+1) = 1$ (i.e., $w_c(i, i+1) = 2$), then we label the edge $(i, i+1)$ with the genome G_j for which $G_j[i]$ and $G_j[i+1]$ are on different replichores. Moreover, a vertex $i \in V$ can also get a label. In what follows, let $G_j[i] = \pm_i^j b_i^j$ for $1 \leq j \leq 3$.

- If $b_i^1 = b_i^2 = b_i^3$, then we set $sign := majority(\pm_i^1, \pm_i^2, \pm_i^3)$. If there is a j such that $\pm_i^j \neq sign$, where $1 \leq j \leq 3$, then we label vertex i with G_j. If there is no such j (i.e., $\pm_i^1 = \pm_i^2 = \pm_i^3$), then vertex i remains unlabeled.
- Otherwise, if there is a bit, say b_i^3, which differs from the other two bits b_i^1 and b_i^2 (this implies $b_i^1 = b_i^2$), then we first flip $G_3[i]$, so that $b_i^1 = b_i^2 = flip(b_i^3)$, and then determine the label of vertex i as in the previous case.

An example graph can be found in Fig. 7.

Lemma 4.1 *Three genomes G_1, G_2, and G_3 are pairwise equivalent if and only if their associated graph has no label.*

Proof The following equivalences hold true.

- G_1, G_2, and G_3 are pairwise equivalent.
- For all $k, \ell \in \{1, 2, 3\}$, all $1 \leq i < n$, and all $1 \leq j \leq n$:
  $(i, i+1)$ is not a gene order breakpoint between G_k and G_ℓ and index j is not a gene orientation breakpoint between G_k and G_ℓ.
- For all $1 \leq i < n$ either $w_c(i, i+1) = 3$ or $w_c(i, i+1) = 0$ and for all $1 \leq j \leq n$ vertex j has no label.
- The graph associated with G_1, G_2, and G_3 has no label.

□

	1	2	3	4	5	6
G_1	+1	+1	+1	+0	+1	−0
G_2	+1	−0	+0	−1	+1	−0
G_3	+1	+1	+1	+1	+1	+0

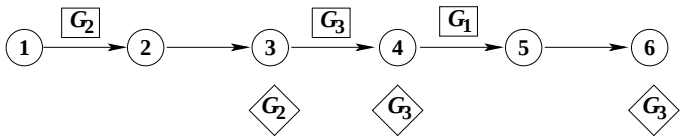

Fig. 7. Three genomes G_1, G_2, and G_3 and their associated graph (V, E). The labels within the boxes belong to the edges, while the labels within the diamonds belong to the vertices. Procedure $median(G_1, G_2, G_3)$ returns $(G_1\overline{\rho}(4), \overline{\rho}(4), \overline{\rho}(1)\sigma(3), \overline{\rho}(3)\sigma(4)\sigma(6))$.

Consequently, to transform G_1, G_2, and G_3 into a median, their associated graph must be transformed into an unlabeled graph. The following lemma characterizes the effect of an inversion on the labels.

Lemma 4.2 *Let (V, E) be the labeled, weighted graph associated with the three genomes G_1, G_2, and G_3 of length n. Suppose an inversion ρ is applied to one of the genomes, say G_3, and let (V', E') be the labeled, weighted graph associated with the three genomes G_1, G_2, and $G_3\rho$. Then the following statements hold:*

(1) If $\rho = \overline{\rho}(i)$, where $1 \leq i \leq n-1$, then (V, E) and (V', E') coincide, except for the label of the edge $(i, i+1)$ (and its weight $w_c(i, i+1)$):
- *If the edge $(i, i+1)$ is labeled with G_3 in (V, E), then it has no label in (V', E').*
- *If the edge $(i, i+1)$ has no label in (V, E), then it is labeled with G_3 in (V', E').*
- *If the edge $(i, i+1)$ is labeled with G_1 (resp. G_2) in (V, E), then it is labeled with G_2 (resp. G_1) in (V', E').*

(2) If $\rho = \sigma(i)$, where $1 \leq i \leq n$, then (V, E) and (V', E') coincide, except for the label of vertex i:
- *If vertex i is labeled with G_3 in (V, E), then it has no label in (V', E').*
- *If vertex i has no label in (V, E), then it is labeled with G_3 in (V', E').*
- *If vertex i is labeled with G_1 (resp. G_2) in (V, E), then it is labeled with G_2 (resp. G_1) in (V', E').*

Proof (1) Let $\rho = \overline{\rho}(i)$, where $1 \leq i \leq n-1$. It is an immediate consequence of the definition of the label of a vertex that labels of the respective vertices in (V, E) and (V', E') are the same. Furthermore, because $\overline{\rho}(i)$ does not affect

the genes $i+1, \ldots, n$, the labels (and weights) of the edges $(p, p+1)$, where $i+1 \leq p \leq n-1$, are also the same in (V, E) and (V', E'). Thus consider the genes p and $p+1$, where $1 \leq p \leq i-1$. Because $p < i$, the genes p and $p+1$ are on the same replichore in G_3 if and only if they are on the same replichore in $G_3\overline{\rho}(i)$. Obviously, this implies that the labels (and weights) of the edges $(p, p+1)$ coincide in (V, E) and (V', E').
Now let us consider the edge $(i, i+1)$. If it is labeled with G_3, then genes i and $i+1$ are on the same replichore in G_3 but on different replichores in G_1 and G_2 or vice versa. Clearly, after the application of reversal $\overline{\rho}(i)$ to G_3 genes i and $i+1$ are either on the same replichore in all three genomes or they are on different replichores in all three genomes, i.e., the edge $(i, i+1)$ has no label in (V', E'). The other statements are proven in a similar fashion.
(2) Because $\sigma(i)$ does not affect the bit representation of G_3, the labels (and weights) of the edges $(i, i+1)$ coincide in (V, E) and (V', E'). Clearly, $\sigma(i)$ does only affect gene i, so that the labels of a vertex $p \neq i$ coincide in (V, E) and (V', E'). If vertex i is labeled with G_3 in (V, E), then we have $\pm_i^1 = \pm_i^2 \neq \pm_i^3$ (the case in which we first have to flip gene i in G_3 is treated similarly). Obviously, after the application of reversal $\sigma(i)$ to G_3 the orientation of gene i is the same in all three genomes, i.e., vertex i has no label in (V', E'). The other statements are proven similarly.
□

In particular, any reversal can remove at most one label. The following procedure $median(G_1, G_2, G_3)$ relies on this fact. It returns a median of the genomes G_1, G_2, and G_3, as well as the inversions that transform each of the genomes into the median.

procedure $median(G_1, G_2, G_3)$
  construct the graph (V, E)
  for $m := 1$ **to** 3 **do**
    determine the edges $(i_1^m, i_1^m + 1), \ldots, (i_{k_m}^m, i_{k_m}^m + 1)$ that are labeled with G_m
    determine the vertices $j_1^m, \ldots, j_{\ell_m}^m$ that are labeled with G_m
  return a median $G = G_1\overline{\rho}(i_1^1) \cdots \overline{\rho}(i_{k_1}^1)\sigma(j_1^1) \cdots \sigma(j_{\ell_1}^1)$ and the reversals
    $\overline{\rho}(i_1^1), \ldots, \overline{\rho}(i_{k_1}^1), \sigma(j_1^1), \ldots, \sigma(j_{\ell_1}^1)$,
    $\overline{\rho}(i_1^2), \ldots, \overline{\rho}(i_{k_2}^2), \sigma(j_1^2), \ldots, \sigma(j_{\ell_2}^2)$,
    $\overline{\rho}(i_1^3), \ldots, \overline{\rho}(i_{k_3}^3), \sigma(j_1^3), \ldots, \sigma(j_{\ell_3}^3)$

The graph associated with G_1, G_2, and G_3 has $\sum_{m=1}^{3}(i_{k_m}^m + j_{\ell_m}^m)$ labels. Each reversal in $G_m\overline{\rho}(i_1^m) \cdots \overline{\rho}(i_{k_m}^m)\sigma(j_1^m) \cdots \sigma(j_{\ell_m}^m)$ removes one label. Therefore, upon termination of procedure $median(G_1, G_2, G_3)$, the graph associated with the genomes $G_1\overline{\rho}(i_1^1) \cdots \overline{\rho}(i_{k_1}^1)\sigma(j_1^1) \cdots \sigma(j_{\ell_1}^1)$, $G_2\overline{\rho}(i_1^2) \cdots \overline{\rho}(i_{k_2}^2)\sigma(j_1^2) \cdots \sigma(j_{\ell_2}^2)$, and $G_3\overline{\rho}(i_1^3) \cdots \overline{\rho}(i_{k_3}^3)\sigma(j_1^3) \cdots \sigma(j_{\ell_3}^3)$ has no label. Because no reversal can remove more than one label, the genome G returned by procedure *median* is a median of the genomes G_1, G_2, and G_3.

If we would really apply the reversals $\overline{\rho}(i_1^1) \cdots \overline{\rho}(i_{k_1}^1)$ around the origin of replication to genome G_1, then $median(G_1, G_2, G_3)$ would take quadratic time. However, a linear time implementation is possible. According to Lemma 2.1, we may assume w.l.o.g. that $i_1^1 > i_2^1 > \ldots > i_{k_1}^1$. We observe that for each pair of reversals $\overline{\rho}(i_p^1)$ and $\overline{\rho}(i_{p+1}^1)$, where p is an odd number, the application of both $\overline{\rho}(i_p^1)$ and $\overline{\rho}(i_{p+1}^1)$ has the effect that just the genes i with $i_{p+1}^1 < i \leq i_p^1$ are flipped. In other words, the application of all reversals can be mimicked in linear time.

5 Conclusions

In this paper, we have shown that—under the assumption that in circular bacterial genomes the predominant mechanism of rearrangement are inversions around the origin/terminus of replication and single gene inversions—the median problem for the reversal distance can be solved in linear time. Because the median problem for the reversal distance is in general NP-hard, our result nicely demonstrates that it is worthwhile to make use of biological constraints. We consider this "message" to be the main contribution of this paper. From an algorithmic point of view, our method is rather simple. We would like to mention that this method can directly be extended to more than three genomes.

Remark
A preliminary version of this paper appeared in [18]. The paper at hand extends the model presented in [18] by single gene inversions. Moreover, the presentation is considerably simplified because a genome is considered to be equivalent to its reflection.

Acknowledgment
The authors were supported by DFG-grants Oh 53/4-1 and Oh 53/5-1. We thank the anonymous reviewers for their comments that helped to improve the article.

References

[1] D.A. Bader, B.M.E. Moret, and M. Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology*, 8:483–491, 2001.

[2] A. Bergeron, J. Mixtacki, and J. Stoye. Reversal distance without hurdles and fortresses. In *Proc. 15th Annual Symposium on Combinatorial Pattern*

Matching, volume 3109 of *Lecture Notes in Computer Science*, pages 388–399. Springer-Verlag, 2004.

[3] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In *Proc. Genome Informatics Workshop*, pages 25–34. Univ. Academy Press, 1997.

[4] B. Bourque and P.A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research*, 12(1):26–36, 2002.

[5] A. Caprara. Formulations and hardness of multiple sorting by reversals. In *Proc. 3rd Annual International Conference on Research in Computational Molecular Biology*, pages 84–94. ACM Press, 1999.

[6] A. Caprara. On the practical solution of the reversal median problem. In *Proc. 1st International Workshop on Algorithms in Bioinformatics*, volume 2149 of *Lecture Notes in Computer Science*, pages 238–251. Springer-Verlag, 2001.

[7] D.A. Dalevi, N. Eriksen, K. Eriksson, and S.G.E. Andersson. Measuring genome divergence in bacteria: A case study using chlamydian data. *Journal of Molecular Evolution*, 55:24–36, 2002.

[8] T. Dobzhansky and A.H. Sturtevant. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23:28–64, 1938.

[9] J.A. Eisen, J.F. Heidelberg, O. White, and S.L. Salzberg. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology*, 1(6):1–9, 2000.

[10] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Journal of the ACM*, 48:1–27, 1999.

[11] T. Hartman and R. Sharan. A 1.5-approximation algorithm for sorting by transpositions and transreversals. In *Proc. of 4th International Workshop on Algorithms in Bioinformatics*, volume 3240 of *Lecture Notes in Computer Science*, pages 50–61. Springer-Verlag, 2004.

[12] D. Hughes. Evaluating genome dynamics: The constraints on rearrangements within bacterial genomes. *Genome Biology*, 1(6):1–8, 2000.

[13] H. Kaplan, R. Shamir, and R.E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29(3):880–892, 1999.

[14] J.F. Lefebvre, N. El-Mabrouk, E.R.M. Tillier, and D. Sankoff. Detection and validation of single gene inversions. *Bioinformatics*, 19:i190–i196, 2003.

[15] P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, and S. Cebrat. Flip-flop around the origin and terminus of replication in prokaryotic genomes. *Genome Biology*, 2(12):1–4, 2001.

[16] B.M.E. Moret, A.C. Siepel, J. Tang, and T. Liu. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *Proc. 2nd International Workshop on Algorithms in Bioinformatics*, volume 2542 of *Lecture Notes in Computer Science*, pages 521–536. Springer-Verlag, 2002.

[17] J.H. Nadeau and B.A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. National Academy of Science USA*, 81(3):814–818, 1984.

[18] E. Ohlebusch, M.I. Abouelhoda, K. Hockel, and J. Stallkamp. The median problem for the reversal distance in circular bacterial genomes. In *Proc. 16th Annual Symposium on Combinatorial Pattern Matching*, volume 3537 of *Lecture Notes in Computer Science*, pages 116–127, Berlin, 2005. Springer-Verlag.

[19] I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. Technical Report TR98-071, Electronic Colloquium on Computational Complexity, 1998.

[20] D. Sankoff. Edit distance for genome comparison based on non-local operations. In *Proc. 3rd Annual Symposium on Combinatorial Pattern Matching, 3rd Annual Symposium*, volume 644 of *Lecture Notes in Computer Science*, pages 121–135. Springer-Verlag, 1992.

[21] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5(3):555–570, 1998.

[22] A.C. Siepel and B.M.E. Moret. Finding an optimal inversion median: Experimental results. In *Proc. 1st International Workshop on Algorithms in Bioinformatics*, volume 2149 of *Lecture Notes in Computer Science*, pages 189–203. Springer-Verlag, 2001.

[23] M. Suyama and P. Bork. Evolution of prokaryotic gene order: Genome rearrangement in closely related species. *TRENDS in Genetics*, 17(1):10–13, 2001.

[24] E.R.M. Tiller and R. Collins. Genome rearrangement by replication-directed translocation. *Nature Genetics*, 26:195–197, 2000.

[25] G.A. Watterson, W.J. Ewens, T.E. Hall, and A. Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99:1–7, 1982.