The Median Problem for the Reversal Distance in Circular Bacterial Genomes

Enno Ohlebush, Mohamed Ibrahim Abouelhoda,
Kathrin Hockel, and Jan Stallkamp

Faculty of Computer Science, University of Ulm, 89069 Ulm, Germany
eo@informatik.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 [5, 14]. 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 [8, 10, 18]. This constraint simplifies 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.

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 [7], but interest on the subject 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.

Consider two genomes $G_1 = (\pi_1, \ldots, \pi_n)$ and $G_2 = (\gamma_1, \ldots, \gamma_n)$ on the same set 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

A. Apostolico, M. Crochemore, and K. Park (Eds.): CPM 2005, LNCS 3537, pp. 116–127, 2005.

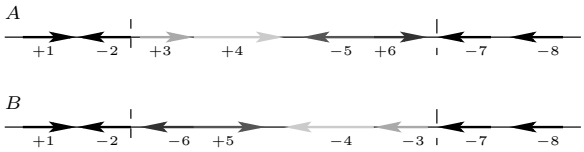

Fig. 1. Genome $(+1, -2, +3, +4, -5, +6, -7, -8)$ before and after the inversion $\rho(3, 6)$.

$-\pi_{i+1}$ 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 [13, 19]. 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 $G\rho(i, j) = (\pi_1, \ldots, \pi_{i-1}, -\pi_j, -\pi_{j-1}, \ldots, -\pi_{i+1}, -\pi_i, \pi_j, \pi_{j+1}, \ldots, \pi_n)$ (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 [15] and has received increasing attention in recent years. There are dozens of papers on the subject; see e.g. [1, 2, 9, 11] and the references therein.

As already mentioned, the median problem is NP-hard for both the breakpoint and the reversal distance [5, 14]. That is the reason why researchers developed heuristics to solve the median and the multiple genome rearrangement problem. For the breakpoint-based multiple genome rearrangement problems very good heuristics exist [3, 16]. These rely on the ability to solve the breakpoint median problem by reducing it to the Traveling Salesman Problem. Solutions to the reversal median problem can be found in [4, 6, 12, 17]. There is a dispute about the "right" distance in multiple genome rearrangement problems. While [3, 16] argue that the breakpoint distance is the better choice, [12] conjecture 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.

2 Inversions Around the Origin of Replication

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. As mentioned earlier, it has been observed [8, 10, 18] that inversions within circular bacterial genomes are centered around the origin or the terminus of replication. That is, the genes keep their distance to the origin O and the

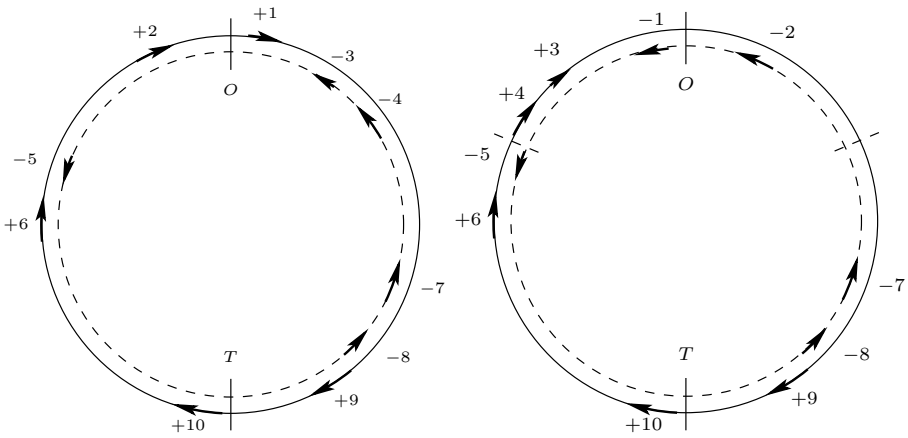

Fig. 2. Left: A cartoon representation of a circular bacterial genome. Of course, bacteria have hundreds, or even thousands, of genes. Moreover, a bacterial genome does not have long stretches of DNA without genes. Right: The same genome after the inversion $\overline{\rho}(4)$.

terminus T of replication under a reversal, but they are translocated to the opposite DNA strand and thus change their orientation.

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 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 O, we enumerate them in increasing distance to the origin. 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 of replication is reached and it divides the circular genome into two halves. The clockwise traversal yields the right half and the counterclockwise traversal yields the left half. 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 labeled 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 $(+1 \mid 0)$.

$(+10, 0, 0, 0, +6, -5, 0, 0, +2, 0 \mid +1, 0, -3, -4, 0, 0, -7, -8, +9, 0)$ is a more complex example, which is shown in Figure 2.

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,

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

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

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

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 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) half of the circular bacterial genome. Furthermore, a + (resp. −) sign in the orientation vector at position p means that the gene lies on the leading (resp. legging) strand if it is in the right half (i.e., if there is a 1 at position p in the bit vector). Otherwise, if the gene is in the left half (i.e., there is a 0 at position p in the bit vector), a + (resp. −) sign at position p means that the gene lies on the legging (resp. leading) strand. With this definition, the orientation vector is invariant (i.e., it does not change) under inversions around O and T. In the following, the orientation vector will hence not be mentioned explicitly. Therefore, the preceding inversions are modeled by

$$(1,0,1,1,0,0,1,1,1,0)\,\overline{\rho}(4) = (0,1,0,0,0,0,1,1,1,0)$$
$$(1,0,1,1,0,0,1,1,1,0)\,\underline{\rho}(2) = (1,0,1,1,0,0,1,1,0,1)$$

Lemma 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 amount 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.

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

Lemma 2. *Let $\rho_1, \rho_2, \ldots, \rho_k$ be reversals. Then $G\rho_1 \cdot \rho_2 \cdots \rho_k = G'$ if and only if $G'\rho_1 \cdot \rho_2 \cdots \rho_k = G$.*

Proof. For $k = 1$, this follows from $G\rho_1 = G' \Leftrightarrow G\rho_1 \cdot \rho_1 = G'\rho_1 \Leftrightarrow G = G'\rho_1$. Now the claim follows by induction on k in conjunction with Lemma 1.

The inverse $inv(G)$ of a genome G is defined by $inv(G) := G\rho(n)$, where $\rho(n) := \overline{\rho}(n) = \underline{\rho}(n)$. Given reversal ρ, a reversal σ satisfying $\rho(n) \cdot \rho = \sigma$ is called the *complementary reversal* of ρ.

Lemma 3. *Every reversal ρ has a (unique) complementary reversal σ.*

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

```
1:1110:110:1
1:0001:110:0
```

Fig. 3. Breakpoints between two genomes, here depicted by colons.

3 The Reversal Distance

Let $(b_1, b_2, b_3, \ldots, b_n)$ be the bit vector representation of a circular bacterial genome G. 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 $1 \leq i \leq j \leq n$: $G[i] = b_i$ and $G[i..j] = (b_i, \ldots, b_j)$.

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

Definition 4. *Let $G = (b_1, b_2, b_3, \ldots, b_n)$ and $G' = (b'_1, b'_2, b'_3, \ldots, b'_n)$ be two circular genomes.*

1. *An interval $[i..j]$ of indices (where $1 \leq i \leq j \leq n$) is called a* strip *if $b_k = b'_k$ for all $i \leq k \leq j$, $b_{i-1} \neq b'_{i-1}$ if $i \neq 1$, and $b_{j+1} \neq b'_{j+1}$ if $j \neq n$.*
2. *If $[i..j]$ is a strip, then $(i-1, i)$ (if $i \neq 1$) and $(j, j+1)$ (if $j \neq n$) are* breakpoints *between G and G'.*

Figure 3 shows two genomes G and G' with three breakpoints. Note that if $G = inv(G')$, then there is no strip, hence no breakpoint between them. Thus, if there is no breakpoint between G and G', then either $G = G'$ or $G = inv(G')$.

Lemma 5. *Let G, G' and $\overline{\rho}(i)$ with $1 \leq i \leq n-1$ be given.*

1. *For all $(j, j+1)$ with either $1 \leq j < i$ or $i < j < n$ we have: $(j, j+1)$ is a breakpoint between G and G' if and only if $(j, j+1)$ is a breakpoint between $G\overline{\rho}(i)$ and G'.*
2. *$(i, i+1)$ is a breakpoint between G and G' if and only if $(i, i+1)$ is not a 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:

$$
\begin{aligned}
& (j, j+1) \text{ is a breakpoint between } G \text{ and } G' \\
\Leftrightarrow\ & \text{either } (b_j = b'_j \text{ and } b_{j+1} \neq b'_{j+1}) \text{ or } (b_j \neq b'_j \text{ and } b_{j+1} = b'_{j+1}) \\
\Leftrightarrow\ & \text{either } (inv(b_j) \neq b'_j \text{ and } inv(b_{j+1}) = b'_{j+1}) \\
& \text{or } (inv(b_j) = b'_j \text{ and } inv(b_{j+1}) \neq b'_{j+1}) \\
\Leftrightarrow\ & (j, j+1) \text{ is a breakpoint between } G\overline{\rho}(i) \text{ and } G'
\end{aligned}
$$

(2) This case follows by a similar reasoning as in (1).

Of course, a similar statement holds when $\overline{\rho}(i)$ is replaced with $\underline{\rho}(i)$. This is also true for the following corollary, which follows from the preceding lemma.

Corollary 6. *Let G, G' and $\overline{\rho}(i)$ with $1 \leq i \leq n-1$ be given.*

1. *If $(i, i+1)$ is a breakpoint between G and G', then the number of breakpoints between $G\overline{\rho}(i)$ and G' is one less than the number of breakpoints between G and G'.*
2. *If $(i, i+1)$ is not a breakpoint between G and G', then the number of breakpoints between $G\overline{\rho}(i)$ and G' is one more than the number of breakpoints between G and G'.*

First, we consider the case in which only inversions around the origin of replication are allowed. The following simple procedure $rd\_O(G, G')$ returns the reversal distance between two genomes G and G', using inversions around O only. (The procedure $rd\_T(G, G')$ that returns the reversal distance between G and G' using inversions around T only is defined similarly.)

procedure $rd\_O(G, G')$
  determine the breakpoints $(i_1, i_1+1), \ldots, (i_k, i_k+1)$ between G and G'
  if $G\overline{\rho}(i_1) \cdots \overline{\rho}(i_k) = G'$ **then** return k **else** return $k+1$

The correctness of procedure $rd\_O(G, G')$ is a direct consequence of Corollary 6. Each reversal $\overline{\rho}(i_1), \ldots, \overline{\rho}(i_k)$ removes one breakpoint, so that there is no breakpoint between $G\overline{\rho}(i_1) \cdots \overline{\rho}(i_k)$ and G'. Hence, we have $G\overline{\rho}(i_1) \cdots \overline{\rho}(i_k) = G'$ or $G\overline{\rho}(i_1) \cdots \overline{\rho}(i_k) = inv(G')$. In the latter case, k must be incremented by 1 because $\rho(n)$ has to be applied to make the genomes equal. It is easy to see that in both cases the algorithm returns the minimum number of inversions needed to transform G into G'.

Since the breakpoints $(i_1, i_1+1), \ldots, (i_k, i_k+1)$ between G and G' can be determined in $O(n)$ time and also the test as to whether two genomes are equal requires $O(n)$ time, the worst case running time of the procedure is $O(n)$.

Next, we consider the general case in which both inversions around the origin and the terminus of replication are allowed.

procedure $rd(G, G')$
  if G and G' do not have a breakpoint **then**
    if $G = G'$ **then** return 0 **else** return 1
  else
    choose a strip $[i..j]$
    $k_l := rd\_O(G[1..i-1], G'[1..i-1])$
    $k_r := rd\_T(G[j+1..n], G'[j+1..n])$
    return $(k_l + k_r)$

Procedure $rd(G, G')$ returns the minimum number of inversions needed to transform G into G' because each inversion removes one breakpoint. The transformed genome must be equal to G' (i.e., it cannot be $inv(G')$) because the chosen strip is not changed by the inversions. Furthermore, procedure $rd(G, G')$ runs in linear time because the procedures rd_O and rd_T do so.

4 The Median Problem for the Reversal Distance

Recall that in the median problem we want to find a genome G (a median) such that $\sum_{i=1}^{3} rd(G, G_i)$ is minimized. In the following, let $d_m(G_1, G_2, G_3) = \min\{\sum_{i=1}^{3} rd(G, G_i) \mid G \text{ is a genome}\}$. Furthermore, for $b^1, b^2, b^3 \in \{0, 1\}$ let

$$majority(b^1, b^2, b^3) = \begin{cases} 1 \text{ if } \sum_{j=1}^{3} b^j \geq 2 \\ 0 \text{ otherwise} \end{cases}$$

Again, we first consider the case in which only inversions around the origin of replication are allowed. In this case, the following procedure *median_O* returns a median, as shown in Theorem 7. (The procedure *median_T* that returns a median using inversions around T only is defined analogously.)

procedure $median\_O(G_1, G_2, G_3)$ / $\star$ where $G_j = (b_1^j, b_2^j, b_3^j, \ldots, b_n^j)$ $\star$ /
  $d := 0$
  for $i := n$ **downto** 1 **do**
    $b := majority(b_i^1, b_i^2, b_i^3)$
    if there is a j, $1 \leq j \leq 3$, such that $b_i^j \neq b$ **then**
      $G_j := G_j\overline{\rho}(i)$
      $d := d + 1$
  return (G_1, d)

If we would really apply the reversals to the genomes (in line 5 of the procedure), then $median\_O(G_1, G_2, G_3)$ would take quadratic time. However, a linear time implementation is possible by simply counting the number of times a gene i was inverted in genome G_j. If it was flipped an even number of times giving G'_j, then $G'_j[1..i] = G_j[1..i]$. Otherwise, if it was flipped an odd number of times, then $G'_j[1..i] = inv(G_j[1..i])$.

Theorem 7. *If procedure* $median\_O(G_1, G_2, G_3)$ *returns the pair* (G, d)*, then* G *is a median of the three genomes* $G_j = (b_1^j, b_2^j, b_3^j, \ldots, b_n^j)$*,* $1 \leq j \leq 3$*, using inversions around* O *only, and* d *is the number of required reversals.*

Proof. We proceed by induction on the length n of the genomes. The case $n = 1$ is trivial. According to the inductive hypothesis, procedure *median_O* returns a median of three genomes of size $n - 1$. For $1 \leq j \leq 3$, let $G'_j = (b_1^j, b_2^j, b_3^j, \ldots, b_{n-1}^j)$. If $b_n^1 = b_n^2 = b_n^3$, then an application of the inductive hypothesis to G'_1, G'_2, and G'_3 proves the theorem. Otherwise, there is a bit, say b_n^3, such that $b_n^1 = b_n^2 \neq b_n^3$. Hence procedure *median_O* first applies $\rho(n)$ to G_3, i.e., it inverts G_3, and then computes a median $G' = (\hat{b}_1^j, \hat{b}_2^j, \hat{b}_3^j, \ldots, \hat{b}_{n-1}^j)$ of G'_1, G'_2, and $inv(G'_3)$. Let $d' = rd(G', G'_1) + rd(G', G'_2) + rd(G', inv(G'_3))$ and $G = (\hat{b}_1^j, \hat{b}_2^j, \hat{b}_3^j, \ldots, \hat{b}_{n-1}^j, b_n^1)$. Clearly, $\sum_{j=1}^{3} rd(G, G_j) = d' + 1$.

In order to prove that G is a median of G_1, G_2, and G_3, it suffices to show that the bit representation of nth gene of a median cannot be b_n^3. For an indirect proof, suppose the contrary. Then, in an optimal sequence of inversions that

transforms G_1 (G_2) into a median, there must be one that inverts the whole genome. According to Lemma 1, we may assume that this inversion is the first in the sequence. Procedure *median_O* applied to $inv(G_1'), inv(G_2')$, and G_3' gives a median $\tilde{G}' = (\tilde{b}_1^j, \tilde{b}_2^j, \tilde{b}_3^j, \ldots, \tilde{b}_{n-1}^j)$ of these. It is not difficult to see that

$$rd(\tilde{G}', inv(G_1')) + rd(\tilde{G}', inv(G_1')) + rd(\tilde{G}', G_3') = d'$$

because the two problems under consideration are equivalent (inverting all genes in one problem yields the other problem). $\tilde{G} = (\tilde{b}_1^j, \tilde{b}_2^j, \tilde{b}_3^j, \ldots, \tilde{b}_{n-1}^j, b_n^3)$ cannot be a median of G_1, G_2, and G_3 because $\sum_{j=1}^3 rd(\tilde{G}, G_j) = d' + 2 > d' + 1 = \sum_{j=1}^3 rd(G, G_j)$. This contradiction shows that the bit representation of the nth gene of a median cannot be b_n^3.

Next, we consider the median problem in which both inversions around the origin and the terminus of replication are allowed. We distinguish between two cases: (a) G_1, G_2, and G_3 have a common bit and (b) G_1, G_2, and G_3 do not have a common bit.

Definition 8. *We say that i is a* common bit *of the genomes G_1, G_2, and G_3 if $G_1[i] = G_2[i] = G_3[i]$.*

Lemma 9. *Suppose $G'\rho_1 \cdot \rho_2 \cdots \rho_k = G$ and $G'[i] = G[i]$, that is, i is a common bit of G and G'. Then there are inversions $\rho_1' \cdot \rho_2' \cdots \rho_k'$ such that $G'\rho_1' \cdot \rho_2' \cdots \rho_k' = G$ and each ρ_j' does not invert the ith gene.*

Proof. If there is an inversion that inverts the ith gene, then there must be an inversion that inverts it back. If both are inversions around O (a similar statement holds if both act around T), say $\overline{\rho}(p)$ and $\overline{\rho}(q)$, then they can be replaced by the inversions $\underline{\rho}(n-p)$ and $\underline{\rho}(n-q)$ around T. These do not invert the ith gene; see Figure 4.

If one is an inversion around O, say $\overline{\rho}(q)$, and the other is an inversion around T, say $\underline{\rho}(p)$, then they can be replaced with the inversions $\overline{\rho}(n-p)$ and $\underline{\rho}(n-q)$. These do not invert the ith gene; see Figure 4. Now the lemma follows by induction on the number of inversions in $\rho_1 \cdot \rho_2 \cdots \rho_k$ that invert the ith gene.

If G_1, G_2, and G_3 have a common bit, the following procedure computes a median; see Theorem 10.

procedure *median_cb*(G_1, G_2, G_3)
  determine a common bit i of G_1, G_2, and G_3
  $(G_l, d_l) :=$ *median_O*$(G_1[1..i-1], G_2[1..i-1], G_3[1..i-1])$
  $(G_r, d_r) :=$ *median_T*$(G_1[i+1..n], G_2[i+1..n], G_3[i+1..n])$
  return $(G_l G_1[i] G_r, d_l + d_r)$

Procedure *median_cb*(G_1, G_2, G_3) runs in linear time because the procedures *median_O* and *median_T* do so.

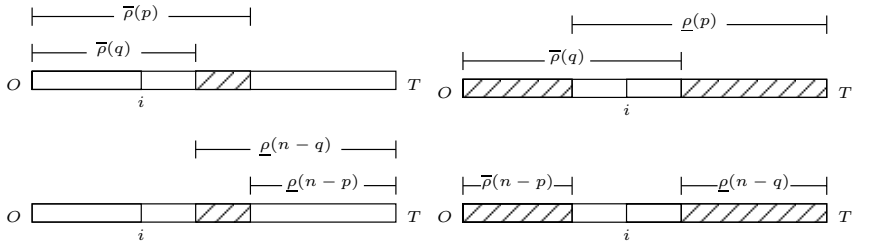

Fig. 4. Left: Two inversions around the origin of replication inverting the same gene i can be replaced by two inversions around the terminus, both not changing gene i. Right: Two inversions (one around the origin and the other around the terminus) that both invert the same gene i can be replaced by two different inversions (again one around the origin and the other around the terminus) that both do not invert gene i.

Theorem 10. *If G_1, G_2, and G_3 have a common bit, then median_cb(G_1, G_2, G_3) returns a pair (G, d) such that G is a median of the three genomes and $d = d_m(G_1, G_2, G_3)$.*

Proof. We claim that if there is a median G' of G_1, G_2, and G_3, such that $G'[i] \neq G_1[i] = G_2[i] = G_3[i]$, then there is another median G such that $G[i] = G_1[i] = G_2[i] = G_3[i]$.

Let G' be such a median of G_1, G_2, and G_3. Then, for $1 \leq j \leq 3$, there are inversions such that $G_j\rho_1^j \cdot \rho_2^j \cdots \rho_{\ell_j}^j = G'$ and $\ell_1 + \ell_2 + \ell_2$ is minimal. Because $G'[i] \neq G_1[i] = G_2[i] = G_3[i]$, in each $G_j\rho_1^j \cdot \rho_2^j \cdots \rho_{\ell_j}^j = G'$, $1 \leq j \leq 3$, the ith gene must have been inverted. By Lemma 1, we may assume that $\rho_{\ell_j}^j$ inverts the ith gene. Moreover, according to Lemma 9, we may assume that none of the other inversions inverts the ith gene.

(a) Suppose that $\rho_{\ell_1}^1$, $\rho_{\ell_2}^2$, and $\rho_{\ell_3}^3$ all act around O, say $\rho_{\ell_j}^j = \overline{\rho}(l_j)$ (for T, the reasoning is verbatim the same). Then inverting the complementary regions yields a median G with $G[i] = G_1[i] = G_2[i] = G_3[i]$; see Figure 5. To be precise, $G = G_j\rho_1^j \cdot \rho_2^j \cdots \rho_{\ell_j-1}^j \cdot \underline{\rho}(n - l_j)$ for every j with $1 \leq j \leq 3$.

(b) Suppose that $\rho_{\ell_1}^1$, $\rho_{\ell_2}^2$, and $\rho_{\ell_3}^3$ do not act around the same spot, say $\rho_{\ell_1}^1 = \overline{\rho}(l_1)$ and $\rho_{\ell_2}^2 = \overline{\rho}(l_2)$ act around O, but $\rho_{\ell_3}^3 = \underline{\rho}(r_3)$ acts around T. Again, inverting the complementary regions yields a median G with $G[i] = G_1[i] = G_2[i] = G_3[i]$; see Figure 5. More precisely, $G = G_j\rho_1^j \cdot \rho_2^j \cdots \rho_{\ell_j-1}^j \cdot \underline{\rho}(n - l_j)$ for $1 \leq j \leq 2$ and $G = G_3\rho_1^3 \cdot \rho_2^3 \cdots \rho_{\ell_3-1}^3 \cdot \overline{\rho}(n - r_3)$.

Thus, there is also a median G of G_1, G_2, and G_3 such that $G[i] = G_1[i] = G_2[i] = G_3[i]$. Then, according to Lemma 9, G_j can be converted to G by (the same number of) inversions that do not invert the ith gene. That is, some of the inversions act only on the genes left to index i, while the others act only on the genes right to index i. In other words, $G = G_l G_1[i] G_r$, where G_l (G_r) is a median of $G_1[1..i-1]$, $G_2[1..i-1]$, and $G_3[1..i-1]$ obtained by inversions around O ($G_1[i+1..n]$, $G_2[i+1..n]$, and $G_3[i+1..n]$ obtained by inversions

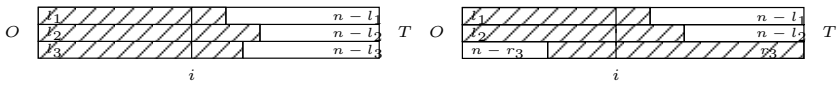

Fig. 5. Left: Case (a) of the proof of Theorem 10. Right: Case (b) of that proof.

around T). Therefore, the correctness of procedure *median_cb* is a consequence of the correctness of the procedures *median_O* and *median_T*.

Now we consider the last case, in which G_1, G_2, and G_3 do not have a common bit. It can be shown (see Theorem 13) that in this case the following procedure $median\_ncb(G_1, G_2, G_3)$ returns a median of the three genomes. Moreover, the procedure runs in linear time because the procedures *median_cb* and *rd* do so.

procedure $median\_ncb(G_1, G_2, G_3)$
  if two genomes coincide, say $G_i = G_j$ with $i \neq j$ **then** return $(G_i, 1)$
  else if one of the genomes is the inverse of another, say $G_i = inv(G_j)$ with $i \neq j$
    then return $(G_i, 1 + rd(G_i, G_k))$ where $k \in \{1, 2, 3\} \setminus \{i, j\}$
  else / ⋆ $G_i \neq G_j$ and $G_i \neq inv(G_j)$ for all $i \neq j$ ⋆ /
    $(G', d') := median\_cb(inv(G_1), G_2, G_3)$
    $d'_1 := rd(inv(G_1), G_2) + rd(inv(G_1), G_3)$
    if $d'_1 = d'$ **then** return (G_1, d')
    else return (G', d')

Due to space limitations, the proofs of the following lemmata are omitted.

Lemma 11. *If G and G' are two genomes such that neither $G = G'$ nor $inv(G) = G'$, then $rd(G, G') = rd(inv(G), G')$.*

Lemma 12. *Let G_1, G_2, and G_3 be genomes such that $G_i \neq G_j$ and $G_i \neq inv(G_j)$ for all $i \neq j$. Then the following statements are equivalent for $\{i, j, k\} = \{1, 2, 3\}$:*

1. *$d_m(inv(G_i), G_j, G_k) = rd(inv(G_i), G_j) + rd(inv(G_i), G_k)$*
2. *$inv(G_i)$ is a median of $inv(G_i)$, G_j, and G_k*
3. *G_i is a median of G_i, G_j, and G_k.*

Theorem 13. *If the three genomes G_1, G_2, G_3 do not have a common bit, then procedure $median\_ncb(G_1, G_2, G_3)$ returns a pair (G, d) such that G is a median of the three genomes and $d = d_m(G_1, G_2, G_3)$.*

Proof. As in the procedure, we proceed by case analysis.

if-statement: If two genomes coincide, say $G_i = G_j$ with $i \neq j$, then it follows $G_i = inv(G_k)$ where $k \in \{1, 2, 3\} \setminus \{i, j\}$. Clearly, inverting G_k yields the median G_i.

else if-statement: Suppose G_1, G_2, and G_3 are pairwise distinct but one of the genomes is the inverse of another, say $G_i = inv(G_j)$ with $i \neq j$. Let $k \in \{1, 2, 3\} \setminus \{i, j\}$. Because G_i, G_j, and G_k are pairwise distinct and $G_i = inv(G_j)$, Lemma 11

implies that $rd(G_i, G_k) = rd(G_j, G_k)$. Thus, for $d_i := rd(G_i, G_i) + rd(G_i, G_j) + rd(G_i, G_k)$, we have $d_i = 1 + rd(G_j, G_k)$.

We must show that G_i is a median of the three genomes. For an indirect proof, suppose that G_i is not a median. Let G be a median of G_i, G_j, and G_k, i.e., there are reversals such that

$$G_i\rho_1^i \cdot \rho_2^i \cdots \rho_{\ell_i}^i = G, \quad G_j\rho_1^j \cdot \rho_2^j \cdots \rho_{\ell_j}^j = G, \text{ and } G_k\rho_1^k \cdot \rho_2^k \cdots \rho_{\ell_k}^k = G$$

It follows from the last two equations in combination with Lemma 2 that $G_j\rho_1^j \cdot \rho_2^j \cdots \rho_{\ell_j}^j \cdot \rho_1^k \cdot \rho_2^k \cdots \rho_{\ell_k}^k = G_k$. Consequently, $rd(G_j, G_k) \leq \ell_j + \ell_k$. On the other hand, since G is a median and G_i is not, we have $\ell_i + \ell_j + \ell_k < d_i = 1 + rd(G_j, G_k)$ and hence $\ell_i + \ell_j + \ell_k < 1 + \ell_j + \ell_k$. We conclude that $\ell_i = 0$, that is, $G = G_i$. This contradiction proves that G_i is a median of the three genomes.

else-statement: We have $G_i \neq G_j$ and $G_i \neq inv(G_j)$ for all $i \neq j$. This implies that $inv(G_1)$, G_2, G_3 have a common bit, as can be seen as follows. If both G_1, G_2, G_3 and $inv(G_1)$, G_2, G_3 would not have a common bit, then it would follow that $G_2[\ell] = G_3[\ell]$ for all $1 \leq \ell \leq n$. In other words, $G_2 = inv(G_3)$. This contradiction shows that $inv(G_1)$, G_2, G_3 must have a common bit. Therefore, we can apply procedure *median_cb* to compute a median G' of $inv(G_1)$, G_2, and G_3. Let $d'_1 = rd(inv(G_1), G_2) + rd(inv(G_1), G_3)$. If $d'_1 = d'$, then G_1 is a median of G_1, G_2, and G_3 by Lemma 12. We will show that $d'_1 \neq d'$ (or, equivalently, $d'_1 > d'$) implies that G' is also a median of G_1, G_2, and G_3. According to Lemma 12, neither is G_1 a median of G_1, G_2, and G_3 nor is $inv(G_1)$ a median of $inv(G_1)$, G_2, and G_3. Since G' is a median of $inv(G_1), G_2$, and G_3, there are reversals such that $inv(G_1)\rho_1'^1 \cdot \rho_2'^1 \cdots \rho_{\ell'_1}'^1 = G'$, $G_2\rho_1'^2 \cdot \rho_2'^2 \cdots \rho_{\ell'_2}'^2 = G'$, $G_3\rho_1'^3 \cdot \rho_2'^3 \cdots \rho_{\ell'_3}'^3 = G'$, and $d' = \ell'_1 + \ell'_2 + \ell'_3$. Moreover, $\ell'_1 > 0$ because $inv(G_1) \neq G'$. It follows from $inv(G_1)\rho_1'^1 \cdot \rho_2'^1 \cdots \rho_{\ell'_1}'^1 = G_1\rho(n)\rho_1'^1 \cdot \rho_2'^1 \cdots \rho_{\ell'_1}'^1$ in conjunction with Lemma 3 that there is an inversion σ^1 such that $G_1\sigma^1 \cdot \rho_2'^1 \cdots \rho_{\ell'_1}'^1 = G'$. Therefore, $d \leq d'$, where $d := d_m(G_1, G_2, G_3)$. We show that $d < d'$ is impossible. For an indirect proof, suppose that $d < d'$ holds. Let G be a median of G_1, G_2, and G_3, i.e., there are reversals such that $G_1\rho_1^1 \cdot \rho_2^1 \cdots \rho_{\ell_1}^1 = G$, $G_2\rho_1^2 \cdot \rho_2^2 \cdots \rho_{\ell_2}^2 = G$, $G_3\rho_1^3 \cdot \rho_2^3 \cdots \rho_{\ell_3}^3 = G$, and $d = \ell_1 + \ell_2 + \ell_3$. Since $G \neq G_1$, we have $\ell_1 > 0$. It follows from $inv(G_1)\rho(n) \cdot \rho_1^1 \cdot \rho_2^1 \cdots \rho_{\ell_1}^1 = G_1\rho_1^1 \cdot \rho_2^1 \cdots \rho_{\ell_1}^1 = G$ that there is an inversion σ such that $inv(G_1)\sigma \cdot \rho_2^1 \cdots \rho_{\ell_1}^1 = G$. Thus, $rd(G, inv(G_1)) + rd(G, G_2) + rd(G, G_3) \leq \ell_1 + \ell_2 + \ell_3 = d < d'$. This contradicts the definition of $d' = \min\{rd(G, inv(G_1)) + rd(G, G_2) + rd(G, G_3) \mid G \text{ is a genome}\}$. In conclusion, $d' = d$, i.e., G' is a median of G_1, G_2, and G_3.

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. T. Dobzhansky and A.H. Sturtevant. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23:28–64, 1938.
8. 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.
9. 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.
10. D. Hughes. Evaluating genome dynamics: The constraints on rearrangements within bacterial genomes. *Genome Biology*, 1(6):1–8, 2000.
11. H. Kaplan, R. Shamir, and R.E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.*, 29(3):880–892, 1999.
12. 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.
13. J.H. Nadeau and B.A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 81(3):814–818, 1984.
14. I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. Technical Report TR98-071, Electronic Colloquium on Computational Complexity, 1998.
15. 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.
16. D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5(3):555–570, 1998.
17. 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.
18. E.R.M. Tiller and R. Collins. Genome rearrangement by replication-directed translocation. *Nature Genetics*, 26:195–197, 2000.
19. G.A. Watterson, W.J. Ewens, T.E. Hall, and A. Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99:1–7, 1982.