Available online at www.sciencedirect.com
SCIENCE DIRECT
Theoretical Computer Science 352 (2006) 322–328
Theoretical Computer Science
www.elsevier.com/locate/tcs

Note

On the complexity of unsigned translocation distance

Daming Zhu[a], Lusheng Wang[b,*]

[a] *School of Computer Science and Technology, Shandong University, Jinan 250100, PR China*
[b] *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

Received 4 November 2004; received in revised form 28 September 2005; accepted 28 September 2005

Communicated by A. Apostolica

Abstract

Translocation is one of the basic operations for genome rearrangement. Translocation distance is the minimum number of translocations required to transform one genome into the other. In this paper, we show that computing the translocation distance for unsigned genomes is NP-hard. Moreover, we show that approximating the translocation distance for unsigned genomes within ratio 1.00017 is NP-hard.

Keywords: Unsigned translocation; NP-hardness; Inapproximability

1. Introduction

Genome rearrangement is a common mode of molecular evolution in plants, mammals, viral, and bacteria [1,6–9,11,14]. Although the rearrangement process is very complicated, there are three basic operations for genome rearrangement, *reversal*, *translocation* and *transposition*. The reversal distance for signed genomes can be computed in polynomial time [2,9,10], whereas computing the reversal distance for unsigned genomes was proved to be NP-hard by Capara [4] and Max-SNP-hard by Berman and Karpinski [3].

The translocation distance computation for unsigned genomes was studied for the first time by Kececioglu and Sankoff [12], who gave a 2-approximation algorithm. This problem for signed genomes was solved in polynomial time by Hannenhalli [5]. The time complexity of Hannenhalli's algorithm is $O(n^3)$. An linear time algorithm for computing the signed distance value without the sequence of translocation operations was given in [13]. An $O(n^2 \log n)$ algorithm for computing an optimal sequence of translocation operations was given in [16]. Recently, an $O(n^2)$ algorithm for computing an optimal sequence of signed translocation operations was given in [15].

In this paper, we show that the translocation distance problem for unsigned genomes is NP-hard. This settles an open problem in this area. We further show that approximating the unsigned translocation distance within ratio 1.00017 is NP-hard.

* Corresponding author.
E-mail addresses: dmzhu@sdu.edu.cn (D. Zhu), cswangl@cityu.edu.hk (L. Wang).

0304-3975/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2005.09.078

2. Signed and unsigned translocation operations

A genome is a set of chromosomes and a chromosome $X = x_1, x_2, \ldots, x_p$ is a sequence of genes. For *signed* genomes, each gene, x_i, is represented as a signed integer. For *unsigned* genomes, each gene, x_i, is represented as a positive integer. x_i and x_{i+1} are *neighbors* in X and also *neighbors* in the genome containing X.

A translocation acts on two chromosomes. It swaps segments between two chromosomes and results in two new chromosomes. Let $X = x_1, \ldots, x_{b-1}, x_b, \ldots, x_p$ and $Y = y_1, \ldots, y_{c-1}, y_c, \ldots, y_q$ be two signed chromosomes in a signed genome. A *prefix–prefix* translocation $\rho(X, Y, b, c)$ produces two new chromosomes $X' = x_1, \ldots, x_{b-1}, y_c, \ldots, y_q$ and $Y' = y_1, \ldots, y_{c-1}, x_b, \ldots, x_p$. A *prefix–suffix* translocation $\rho(X, Y, b, c)$ produces two new chromosomes $X' = x_1, \ldots, x_{b-1}, -y_{c-1}, \ldots, -y_1$ and $Y' = -x_p, \ldots, -x_b, y_c, \ldots, y_q$.

A signed chromosome X is *identical* to chromosome Y if either $X = Y$ or $X = Y^R = -y_q, -y_{q-1}, \ldots, -y_1$. Genome A is *identical* to genome B if and only if the sets of chromosomes for A and B are the same. The *translocation distance* between two signed genomes A and B, denoted as $d(A, B)$, is the minimum number of translocations required to transform A into B.

Translocation for unsigned genomes follows from the signed case without caring about the direction of genes. Let $X = x_1, \ldots, x_{b-1}, x_b, \ldots, x_p$ and $Y = y_1, \ldots, y_{c-1}, y_c, \ldots, y_q$ be two unsigned chromosomes in an unsigned genome. A *prefix–prefix* translocation $\rho(X, Y, b, c)$ produces two new chromosomes $X' = x_1, \ldots, x_{b-1}, y_c, \ldots, y_q$ and $Y' = y_1, \ldots, y_{c-1}, x_b, \ldots, x_p$. A *prefix–suffix* translocation $\rho(X, Y, b, c)$ produces two new chromosomes $X' = x_1, \ldots, x_{b-1}, y_{c-1}, \ldots, y_1$ and $Y' = x_p, \ldots, x_b, y_c, \ldots, y_q$. The translocation ρ transforming genome A into A_1 is written as $A \cdot \rho = A_1$.

An unsigned chromosome X is *identical* to chromosome Y if either $X = Y$ or $Y^R = y_q, y_{q-1}, \ldots, y_1$. Genome A is *identical* to genome B if and only if the sets of chromosomes for A and B are the same. The *translocation distance* between two unsigned genomes A and B, denoted as $d(A, B)$, is the minimum number of translocations required to transform A into B.

2.1. Computation of the signed translocation distance

For a signed genome A, we will construct a graph G_A. For each chromosome $X = x_1, x_2, \ldots, x_p$ in genome A, we have $2p$ vertices in G_A, two vertices x_i^h, x_i^t for each gene x_i in X. The $2p$ vertices are arranged in a linear order from left to right as $l(x_1)r(x_1)l(x_2)r(x_2)\ldots l(x_p)r(x_p)$, where if x_i is a positive integer, then $l(x_i) = x_i^t$ and $r(x_i) = x_i^h$; and if x_i is a negative integer, then $l(x_i) = x_i^h$ and $r(x_i) = x_i^t$. For each $i \in \{1, 2, \ldots, p-1\}$, there is a black edge $(r(x_i), l(x_{i+1}))$, where $r(x_i)$ and $l(x_{i+1})$ are called *neighbors* in G_A.

Given two signed genomes A and B, we can construct the *cycle graph* G_{AB} from G_A by adding a *gray* edge to every pair of vertices u and v, where u and v are neighbors in G_B. Each vertex is in a unique cycle in G_{AB}. Moreover, each black edge in the cycle is followed by a gray edge and vice versa. In this paper, the so-called cycle always implies this kind and is called *alternate-color* cycle. A cycle is *long* if it contains at least two black edges. Otherwise, the cycle is *short*. If $A = B$, then all cycles in G_{AB} are short.

Let $X = x_1, x_2, \ldots, x_p$ be a chromosome in A. A *sub-permutation* (*SP*) is an interval $x_i, x_{i+1}, \ldots, x_{i+l}$ in X such that there is another interval of the same length $y_k, y_{k+1}, \ldots, y_{k+l}$ in a chromosome Y of B satisfying $\{|x_i|, |x_{i+1}|, \ldots, |x_{i+l}|\} = \{|y_k|, |y_{k+1}|, \ldots, |y_{k+l}|\}$, $y_k = x_i, y_{k+l} = x_{i+l}$, and, $x_{i+1}, \ldots, x_{i+l-1} \neq y_{k+1}, \ldots, y_{k+l-1}$. A minimal sub-permutation, *minSP* briefly, is a *SP* not containing any other *SP*. For example, let $A = \{X_1 = +1, -3, +2, +4, +5, -8, +6;\ X_2 = +7, +9\}$, $B = \{Y_1 = +1, +2, +3, +4, +5, +6;\ Y_2 = +7, +8, +9\}$. The interval $+1, -3, +2, +4$ in X_1 of A is a *minSP*. The interval $+1, -3, +2, +4, +5$ is not a *minSP* but a *SP*. The cycle graph is presented in Fig. 1.

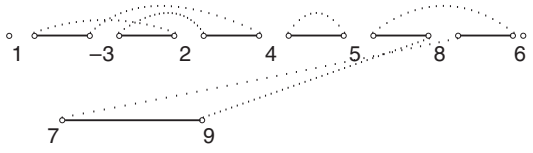

Fig. 1. The cycle graph G_{AB} and the sub-permutation 1,−3,2,4.

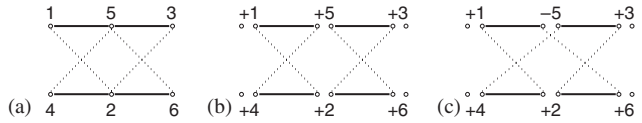

Fig. 2. The breakpoint graph and its two cycle decompositions. (a) The breakpoint graph B_{AB}, (b) the cycle graph $G_{\vec{A}_1\vec{B}}$ and (c) the cycle graph $G_{\vec{A}_2\vec{B}}$.

Let c_{AB} be the number of cycles in G_{AB}. The following theorem gives a very useful bound for translocation distance between two signed genomes.

Theorem 1. *The translocation distance between two signed genomes A and B satisfies $d(A, B) \geqslant n - m - c_{AB}$. If there is no sub-permutation for A and B, the translocation distance is $d(A, B) = n - m - c_{AB}$, where n is the number of genes and m is the number of chromosomes in the given genomes.*

Proof. A translocation increases at most one cycle for G_{AB}. If $A = B$, then G_{AB} has $n - m$ cycles. Thus it requires at least $n - m - c_{AB}$ translocations to transform A into B. That is $d(A, B) \geqslant n - m - c_{AB}$. If there is no *SP* for A and B, there is no *minSP*. Theorem 12 in [5] shows that if there is no *minSP* for A and $B \neq A$, there exists a translocation to increase a cycle for G_{AB} without producing any new *minSP*s. This leads to $d(A, B) = n - m - c_{AB}$. □

2.2. Breakpoint graph for unsigned genomes

Given two unsigned genomes A and B, the *breakpoint* graph B_{AB} is constructed as follows: (1) the vertex set is the set of genes in A in the linear order as in the chromosomes; (2) set a *black* edge between any two vertices that are neighbors in A and set a *gray* edge between any two vertices that are neighbors in B. A *nodal* vertex is the vertex for an end gene in a chromosome. Every *nodal* vertex in B_{AB} is incident to one black edge and one gray edge. Any non-nodal vertex is incident to two black and two gray edges.

Unlike cycle graphs, breakpoint graphs do not admit unique cycle decomposition. For each non-nodal vertex, there are two ways to pair the two black and two gray edges incident to the vertex. Once the choice for the paring of the two black and two gray edges is fixed, we have a decomposition of B_{AB} into alternate-color cycles. Any alternate-color cycle decomposition gives a direction of every gene in genomes A and B.

For example, let $A = \{1, 5, 3; 4, 2, 6\}$ and $B = \{1, 2, 3; 4, 5, 6\}$, both containing two chromosomes separated by a semicolon. The breakpoint graph B_{AB} is presented in Fig. 2(a). Assign every gene in B to positive direction to get $\vec{B} = \{+1, +2, +3; +4, +5, +6\}$. If $\vec{A}_1 = \{+1, +5, +3; +4, +2, +6\}$, the corresponding cycle graph $G_{\vec{A}_1\vec{B}}$ is as Fig. 2(b) and $d(\vec{A}_1, \vec{B}) = 2$. If $\vec{A}_2 = \{+1, -5, +3; +4, +2, +6\}$, the corresponding graph $G_{\vec{A}_2\vec{B}}$ is as Fig. 2(c). In this case, $d(\vec{A}_2, \vec{B}) = 3$.

Let ρ be a translocation that transforms genome A into A_1. There exists a translocation ρ_1 transforming A_1 into A. Translocation ρ_1 is called the *counter translocation* of ρ and ρ_1 is denoted as $\overline{\rho}$. Let $spin(A)$ be the set of all signed genomes obtained from A by assigning a direction to each gene in A. The following theorem gives the relationship between the unsigned and the signed translocation distances.

Theorem 2. *Let A and B be two unsigned genomes. $\vec{B}$ is the signed genome obtained from B by setting the direction of every gene as positive. Then, $d(A, B) = \min_{\vec{A} \in spin(A)} d(\vec{A}, \vec{B})$.*

Proof. Let $\vec{A} \in spin(A)$, and the sequence of translocations $\rho_1, \rho_2, \ldots, \rho_t$ transform $\vec{A}$ into $\vec{B}$, $\vec{A} \cdot \rho_1 \cdot \rho_2 \cdot \ldots \cdot \rho_t = \vec{B}$. The same sequence of translocations must transform A into B, i.e., $A \cdot \rho_1 \cdot \rho_2 \cdot \ldots \cdot \rho_t = B$. Thus $d(A, B) \leqslant d(\vec{A}, \vec{B})$. Assume there is a sequence of translocations $\eta_1, \eta_2, \ldots, \eta_r$ to transform A into B. Consider the translocation sequence $\overline{\eta}_r, \overline{\eta}_{r-1}, \ldots, \overline{\eta}_1$ transforming $\vec{B}$ into $\vec{A}$: $\vec{B} \cdot \overline{\eta}_r \cdot \ldots \cdot \overline{\eta}_1 = \vec{A}$. $\vec{A}$ is a signed genome of A and $\vec{A} \cdot \eta_1 \cdot \eta_2 \cdot \ldots \cdot \eta_r = \vec{B}$, which means $d(A, B) \geqslant \min_{\vec{A} \in spin(A)} d(\vec{A}, \vec{B})$. Thus $d(A, B) = \min_{\vec{A} \in spin(A)} d(\vec{A}, \vec{B})$. □

Formally, the unsigned translocation distance problem(UT) is given as follows:

Instance. Two unsigned genomes A and B.

Question. Find a sequence of translocations $\rho_1, \rho_2, \ldots, \rho_k$ such that $A \cdot \rho_1 \cdot \rho_2 \cdot \ldots \cdot \rho_k = B$ and the number of translocations k is minimized.

3. NP-hardness proof

Caprara first suggested the problem of maximum alternate-color cycle decomposition for unsigned chromosomes(Max-Acd). The problem is given as follows:

Instance. Unsigned chromosomes X and Y, B_{XY} as the breakpoint graph with respect to X and Y.

Question. Find an alternate-color cycle decomposition of B_{XY} such that the number of cycles is maximized.

This problem was proved to be NP-hard in [4]. Given two unsigned genomes A and B. Consider the cycle decomposition of B_{AB}. If vertex x is split into x^t and x^h by a cycle decomposition of B_{AB}, each of x^t and x^h must be uniquely in one alternate-color cycle. Vertex x is *used* by cycle C if x^t or x^h is in C. Every cycle uses a vertex at most twice in a cycle decomposition of B_{AB}. A gray edge is refer to as *inside* X if its two ends are both in X. A gray edge *spans* X and Y if one of its end is in X and the other is in Y.

Theorem 3. *The unsigned translocation distance problem is NP-Hard.*

Proof. In fact, it is enough to prove that computing the unsigned translocation distance value is NP-hard. The reduction is from the Max-Acd problem. Let X and Y be the two unsigned chromosomes. Without loss of generality, let $X = g_1, g_2, \ldots, g_{n-1}, g_n$ and $Y = 1, 2, \ldots, n$, where $\{g_1, g_2, \ldots, g_n\} = \{1, 2, \ldots, n\}$, $g_1 = 1$, $g_n = n$. We construct two genomes $A = \{X_1, X_2\}$ and $B = \{Y_1, Y_2\}$ from X and Y.

There are $4n - 3 + (n-2)d$ genes in both genomes A and B, where genes in $\{1, 2, \ldots, n\}$ have been used in X and Y. The positive integer d is used to control the shape of the long cycles in the decomposition of B_{AB}. Its value will be discussed in Lemma 4. Chromosome X_1 of genome A is constructed by inserting $n - 1$ new genes into the midst of adjacent pairs of genes in chromosome X.

$$X_1 = 1, t_{1,1}, g_2, t_{1,2}, \ldots, g_{n-1}, t_{1,n-1}, n, \tag{1}$$

where, $t_{1,k} = 3n - 2 + k$, $1 \leqslant k \leqslant n - 1$.

X_2 contains two types of new genes, denoted as $t_{2,l}$ and s_i, respectively.

$$\begin{aligned} X_2 = {} & t_{2,1}, t_{2,2}, s_1, s_2, \ldots, s_d, \\ & t_{2,3}, t_{2,4}, s_{d+1}, \ldots\ldots, s_{2d}, \\ & \vdots \\ & t_{2,2(n-2)-1}, t_{2,2(n-2)}, s_{(n-3)d+1}, \ldots, s_{(n-2)d}, \\ & t_{2,2(n-1)-1}, t_{2,2(n-1)}, \end{aligned} \tag{2}$$

where $t_{2,l} = n + l$, $1 \leqslant l \leqslant 2(n-1)$, $s_i = 4n - 3 + i$, and $1 \leqslant i \leqslant (n-2)d$.

Now construct genome $B = \{Y_1, Y_2\}$. Let $t_{1,k}$, $t_{2,l}$, and s_i be the same integers as used in A. Chromosome Y_1 is identical to Y, $Y_1 = 1, 2, \ldots, n-1, n$. Y_2 is constructed from X_2 by inserting $t_{1,k}$ into the midst of $t_{2,2k-1}$ and $t_{2,2k}$ in X_2.

$$\begin{aligned} Y_2 = {} & t_{2,1}, t_{1,1}, t_{2,2}, s_1, s_2, \ldots, s_d, \\ & t_{2,3}, t_{1,2}, t_{2,4}, s_{d+1}, \ldots\ldots, s_{2d}, \\ & \vdots \end{aligned}$$

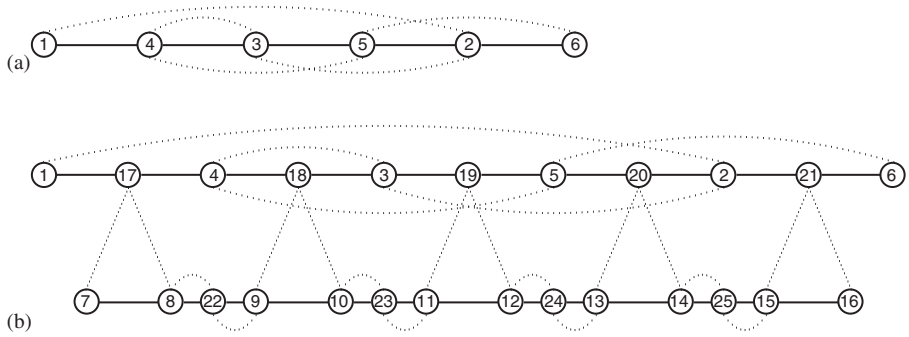

Fig. 3. The breakpoint graphs with respect to chromosomes X and Y and genomes A and B. (a) The breakpoint graph G_{XY} and (b) the breakpoint graph B_{AB}.

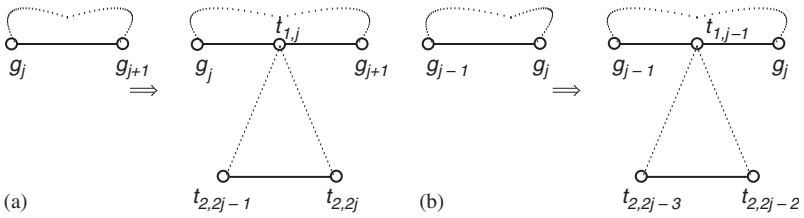

Fig. 4. Two cases for replacing (u_{2i-1},u_{2i}) by path $P_{2i-1,2i}$. (a) Replacement of (g_j, g_{j+1}) and (b) replacement of (g_j, g_{j-1}).

$$t_{2,2(n-2)-1}, t_{1,n-2}, t_{2,2(n-2)}, s_{(n-3)d+1}, \ldots, s_{(n-2)d},$$
$$t_{2,2(n-1)-1}, t_{1,n-1}, t_{2,2(n-1)}. \tag{3}$$

Example. Suppose $X = 1, 4, 3, 5, 2, 6$ and $Y = 1, 2, 3, 4, 5, 6$. Then the breakpoint graph G_{XY} is shown in Fig. 3(a). For simplicity, set $d = 1$. The two genomes are constructed as $A = \{X_1 = 1, 17, 4, 18, 3, 19, 5, 20, 2, 21, 6;\ X_2 = 7, 8, 22, 9, 10, 23, 11, 12, 24, 13, 14, 25, 15, 16\}$, and $B = \{Y_1 = 1, 2, 3, 4, 5, 6;\ Y_2 = 7, 17, 8, 22, 9, 18, 10, 23, 11, 19, 12, 24, 13, 20, 14, 25, 15, 21, 16\}$. The corresponding breakpoint graph B_{AB} are shown in Fig. 3(b).

Lemmas 4 and 5 will be used to clarify the corresponding relationships between maximum cycle decomposition of G_{XY} and the unsigned translocation distance between genomes A and B.

Lemma 4. *Assume that $d \geqslant n-1$. There exists a decomposition of G_{XY} into J alternate-color cycles if and only if there exists a decomposition of B_{AB} into not less than $(n-2)(d+1)+J$ alternate-color cycles.*

Proof. ($\rightarrow$) There are always even number of vertices in any cycle. Thus every cycle can be represented by its vertex list: $u_1, u_2, \ldots, u_{2k-1}, u_{2k}$, where (u_{2i-1}, u_{2i}) is a black edge and (u_{2i}, u_{2i+1}) is a gray edge, $1 \leqslant i \leqslant k$, $u_{2k+1} = u_1$.

Assume there is a decomposition of G_{XY} into J alternate-color cycles. For each of the J cycles $C = u_1, u_2, \ldots, u_{2k-1}, u_{2k}$, if $u_{2i-1} = g_j$, then $u_{2i} = g_{j+1}$ or $u_{2i} = g_{j-1}$. A new cycle C' in B_{AB} can be obtained by replacing each black edge (u_{2i-1}, u_{2i}) with path $P_{2i-1,2i}$, where,

$$P_{2i-1,2i} = \begin{cases} g_j, t_{1,j}, t_{2,2j-1}, t_{2,2j}, t_{1,j}, g_{j+1} & \text{if } u_{2i-1} = g_j, u_{2i} = g_{j+1}, \\ g_j, t_{1,j-1}, t_{2,2j-3}, t_{2,2j-2}, t_{1,j-1}, g_{j-1} & \text{if } u_{2i-1} = g_j, u_{2i} = g_{j-1}. \end{cases} \tag{4}$$

Fig. 4 describes the method of the replacement. By the replacement, we get J long cycles from B_{AB}. The remaining edges in B_{AB} can form $(n-2)(d+1)$ short cycles. Thus B_{AB} can be decomposed into $(n-2)(d+1)+J$ alternate-color cycles.

($\leftarrow$) Let $\mathcal{C}$ be the set of $(n-2)(d+1)+J$ cycles decomposed from B_{AB}. Only the edges $(t_{2,2j}, s_{(j-1)d+1}), \ldots, (s_{jd-1}, s_{jd}), (s_{jd}, t_{2,2j+1})$ for $j \in \{1, \ldots, n-2\}$ can form short cycles. Thus there are at most $(n-2)(d+1)$ short

cycles in $\mathcal{C}$. An alternate-color cycle using a vertex in X_1 must contain at least two black edges in X_1 and at least two gray edges spanning X_1 and X_2. Thus there are at most $n-1$ cycles using vertices in X_1. Since $d \geqslant n-1$, the $(d+2)$ consecutive vertices in X_2, $t_{2,2j}, s_{(j-1)d+1}, \ldots, s_{jd}, t_{2,2j+1}$, cannot be involved in one cycle in $\mathcal{C}$ for $1 \leqslant j \leqslant n-2$. Otherwise, the number of short cycles will be reduced by $d+1 \geqslant n$ and the total number of cycles cannot be greater than

$$(n-3)(d+1)+n-1 \leqslant (n-2)(d+1)-1 < (n-2)(d+1)+J. \tag{5}$$

Thus, if a cycle in $\mathcal{C}$ contains one of the gray edges $(t_{1,j}, t_{2,2j-1})$ and $(t_{1,j}, t_{2,2j})$, it must contain both of them. Moreover, if a long cycle in $\mathcal{C}$ uses two consecutive vertices in the vertex sequence $t_{2,2j}, s_{(j-1)d+1}, \ldots, s_{jd}, t_{2,2j+1}$, it must contain both of the black and the gray edges between the two vertices. Thus we can re-decompose the long cycle to increase the number of short cycles. Therefore, we can assume that all the $(n-2)(d+1)$ short cycles are in $\mathcal{C}$.

Any long cycle decomposed from B_{AB} cannot only contain gray edges inside X_1. Thus each long cycle in $\mathcal{C}$ must take the shape $P_{1,2}, \ldots, P_{2k-1,2k}$, where $P_{2i-1,2i} = u_{2i-1}, t_{1,j}, t_{2,2j-1}, t_{2,2j}, t_{1,j}, u_{2i}$, and $\{u_{2i-1}, u_{2i}\} = \{g_j, g_{j+1}\}$. Replacing the path $P_{2i-1,2i}$ with black edge (u_{2i-1}, u_{2i}), we obtain an alternate-color cycle in G_{XY}. Thus G_{XY} can be decomposed into J alternate-color cycles. □

Lemma 5. *There exists a decomposition of B_{AB} into $(n-2)(d+1)+J$ alternate-color cycles if and only if $d(A,B) \leqslant 3n-3-J$.*

Proof. ($\rightarrow$) If B_{AB} can be decomposed into $(n-2)(d+1)+J$ cycles, there must exist a decomposition of B_{AB} into at least $(n-2)(d+1)+J$ cycles such that every long cycle has a gray edge spanning X_1 and X_2. This is because a cycle containing black edges inside X_1 must contain a gray edge spanning X_1 and X_2 and thus must be long. Moreover, a long cycle only containing edges inside X_2 can be re-decomposed into multiple short cycles. Thus every long cycle decomposed from B_{AB} must have a gray edge spanning X_1 and X_2. This implies that there are no sub permutations for $\vec{A}$ and $\vec{B}$, where $\vec{A}$ and $\vec{B}$ are the signed genomes obtained from the cycle decomposition. Thus, $d(A,B) \leqslant d(\vec{A},\vec{B}) = 4n-3+(n-2)d-2-((n-2)(d+1)+J) = 3n-3-J$ by Theorem 1.

($\leftarrow$) Let $d(A,B) \leqslant 3n-3-J$. It follows that there exist directed genomes $\vec{A}$ and $\vec{B}$ obtained from A and B such that $d(\vec{A},\vec{B}) \leqslant 3n-3-J$. Theorem 1 implies $d(\vec{A},\vec{B}) \geqslant 4n-3+(n-2)d-2-c_{\vec{A}\vec{B}}$. Thus $c_{\vec{A}\vec{B}} \geqslant (n-2)(d+1)+J$. This ensures that the maximum number of alternate-color cycles decomposed from B_{AB} is not less than $(n-2)(d+1)+J$. □

By Lemmas 4 and 5, we have a reduction such that there is a decomposition of G_{XY} into J alternate-color cycles if and only if there exist at most $3n-3-J$ translocations to transform A into B.

If the instance of Max-Acd has n genes, then the corresponding instance of unsigned translocation distance has $4n-3+(n-2)d$ genes. Setting $d = n-1$, the reduction is polynomial. This completes the proof of Theorem 3. □

4. Inapproximability proof

In this section, we study the hardness of approximating the unsigned translocation distance. Consider the breakpoint graph decomposition (BGD) problem [3]. The instance of BGD is the same as Max-Acd, but the objective of BGD is to minimize $cost(\mathcal{C}) = b - |\mathcal{C}|$, where b is the number of black edges of the breakpoint graph and $\mathcal{C}$ is the set of alternate-color cycles. In [3], Berman and Karpinski proved.

Lemma 6. *For any $\varepsilon > 0$, it is NP-hard to decide if an instance of BGD with $2240p$ breakpoints has the minimum cost of alternate-color cycle decomposition below $(1236+\varepsilon)p$ or above $(1237-\varepsilon)p$.*

Now we use the reduction of Theorem 3 to show that approximating the unsigned translocation distance within a factor 1.00017 is difficult.

Theorem 7. *For any $\varepsilon > 0$, it is NP-hard to decide if an instance of the unsigned translocation distance problem can be approximated within factor $5717/5716-\varepsilon$, i.e., $1.00017-\varepsilon$.*

Proof. The reduction of Theorem 3 implies that an instance of Max-Acd with n vertices has the alternate-color cycles of maximum number J, if and only if the translocation distance of the corresponding UT instance is $3n - 3 - J$. Consider the reduction from the Max-Acd instance with $n = 2240p + 1$ vertices or $2240p$ breakpoints. Recall that the Max-Acd instance must be an instance of BGD. The constructed instance of UT, genomes A and B, have $4(2240p + 1) - 3 + (2240p - 1)d$ vertices or genes. For $d \geqslant 2240p$, the objective function of UT is $d(A, B) = 3(2240p + 1) - 3 - |\mathcal{C}| = 4480p + cost(\mathcal{C})$, where $cost(\mathcal{C})$ is the cost of the alternate-color cycle decomposition of BGD instance. From Lemma 6, it is NP-hard to decide if an instance of UT has the translocation distance above $(5717 - \varepsilon)p$ or below $(5716 + \varepsilon)p$. □

Acknowledgments

Daming Zhu is supported by NSFC:60273032. Lusheng Wang is fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project no. CityU 1070/02E]. We thank the referees for their helpful suggestions.

References

[1] V. Bafna, P. Pevzner, Sorting by reversals: genome rearrangements in plant organelles and evolutionary history of x chromosome, Mole. Biol. Evol. 12 (1995) 239–246.
[2] V. Bafna, P. Pevzner, Genome rearrangement and sorting by reversals, SIAM J. Comput. 25 (2) (1996) 272–289.
[3] P. Berman, M. Karpinski, On some tighter inapproximability results, ECCC Report No. 65, University of Trier, 1998.
[4] A. Caprara, Sorting by reversal is difficult, Proc. First Annu. Internat. Conf. on Computational Molecular Biology 1997, pp. 75–83.
[5] S. Hannenhalli, Polynomial time algorithm for computing translocation distance between genomes, Discrete Appl. Math. 71 (1996) 137–151.
[6] S. Hannenhalli, C. Chappey, E.V. Koonin, P. Pevzner, Genome sequence comparison and scenarios for gene rearrangements: a test case, Genomics 30 (1995) 299–311.
[7] S. Hannenhalli, P. Pevzner, Towards a computational theory of genome rearrangement, Lecture Notes in Comput. Sci. 1000 (1995) 184–202.
[8] S. Hannenhalli, P. Pevzner, To cut or not to cut (applications of comparative physical maps in molecular evolution), Proc. Seventh Annu. ACM-SIAM Symp. on Discrete Algorithms, January 1996, pp. 304–313.
[9] S. Hannenhalli, P. Pevzner, Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals, J. ACM 46 (1) (1999) 1–27.
[10] H. Kaplan, R. Shamir, R.E. Tarjan, Faster and simpler algorithm for sorting signed permutations by reversals, Proc. Eighth ACM-SIAM Symp. on Discrete Algorithms, ACM Press, New York, 1997.
[11] J. Kececioglu, R. Ravi, Of mice and men: algorithms for evolutionary distances between genomes with translocation, Proc. Eighth Annu. ACM-SIAM Symp. on Discrete Algorithms, 1995, pp. 604–613.
[12] J. Kececioglu, D. Sankoff, Exact and approximation algorithms for the inversion distance between two permutations, Proc. Fourth Annu. Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science, Vol. 684, Springer, Berlin, 1993, pp. 87–105.
[13] G. Li, X. Qi, X. Wang, B. Zhu, A linear time algorithm for computing translocation distance between signed genomes, Proc. CPM'2004, Lecture Notes in Computer Science, Vol. 3109, Springer, Berlin, 2004.
[14] D. Sankoff, G. Leduc, N. Antoine, B. Paquine, B.F. Lang, R. Cedergen, Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome, Proc. Natl. Acad. Sci. USA 89 (1992) 6575–6579.
[15] L.S. Wang, D.M. Zhu, X.W. Liu, S.H. Ma, An $O(n^2)$ algorithm for signed translocation, J. Comput. Syst. Sci. 70 (2005) 284–299.
[16] D.M. Zhu, S.H. Ma, An improved polynomial time algorithm for translocation sorting problems, J. Comput. 25 (2) (2002) 189–196 (in Chinese).