Reversal distance of signed circular chromosomes

João Meidanis[*]
Institute of Computing
University of Campinas-BRAZIL

Maria Emilia M. T. Walter[†]
Department of Computer Science
University of Brasilia-BRAZIL
Institute of Computing
University of Campinas-BRAZIL

Zanoni Dias[‡]
Institute of Computing
University of Campinas-BRAZIL

Campinas, march, 1997

Abstract

We study the problem of comparing two circular chromosomes, evolved from a common ancestor by reversals, given the order of the corresponding genes and their orientations. Determining the minimum number of reversals between the chromosomes is equivalent to look for the minimum number of reversals that transforms a circular sequence of signed integer numbers, defined in an appropriate manner, into another, where a reversal acts on a subsequence, reversing its order and flipping the signs. We carefully formalize the concepts of circular chromosome and circular reversal, and show that this problem is essentially equivalent to the analogous problem on linear chromosomes. As a consequence we derive polynomial time algorithms based on this observation. We also compute the reversal diameter for signed chromosomes, both linear and circular.

1 Introduction

The huge amount of data resulting from genome sequencing in Molecular Biology is giving rise to an increasing interest in the development of algorithms for comparing

[*]e-mail:meidanis@dcc.unicamp.br

[†]e-mail:emilia@dcc.unicamp.br

[‡]e-mail:zanoni@dcc.unicamp.br

genomes of related species. Particularly these data allowed studies on mutational events acting on large portions of the chromosomes, that can be used to compare genomes for which the traditional methods of comparing DNA sequences are not conclusive. There are several mutational events affecting large fragments of genomes of organisms, and among them, the *reversal* seems to be one of the commonest. A reversal replaces a sequence of an arbitrary region of the chromosome with the reverse complementary sequence. This reverses the gene order within the region, and changes the orientation of each gene. In this paper we study the comparison of two genomes, formed each by a single circular chromosome, on the basis of the order and orientation of their common genes, and in terms of the mutational event of reversal.

A circular chromosome can be seen as a circular arrangement of blocks of genes, where each block has an *orientation*. Figure 1 shows examples of circular chromosomes of two species of plants, where each number represents a block composed by one or more genes, and the arrows indicate the orientations of the blocks of one species relative to the other.

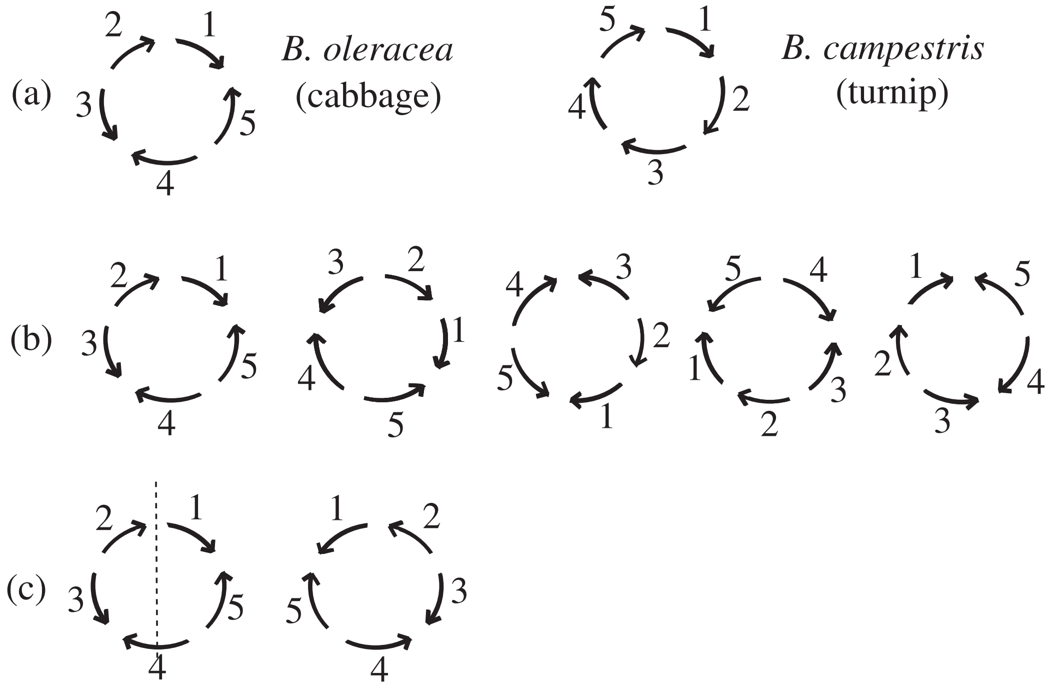

Figure 1: Examples of circular chromosomes of two species of plants. (a) The arrows indicate the orientations of a species relative to the other. (b) These examples show different representations of the same chromosome. (c) These examples show the same chromosome, considering the two possible forms to view the gene blocks of a circular chromosome. These two forms are considered equivalent, and these two chromosomes are obtained from one another by reflection relative to the axis shown in the figure.

In a circular chromosome, a reversal is defined by fixing two cut points in this chromosome, and reversing the order of the genes in one of the two regions delimited by these points (see Figure 2).

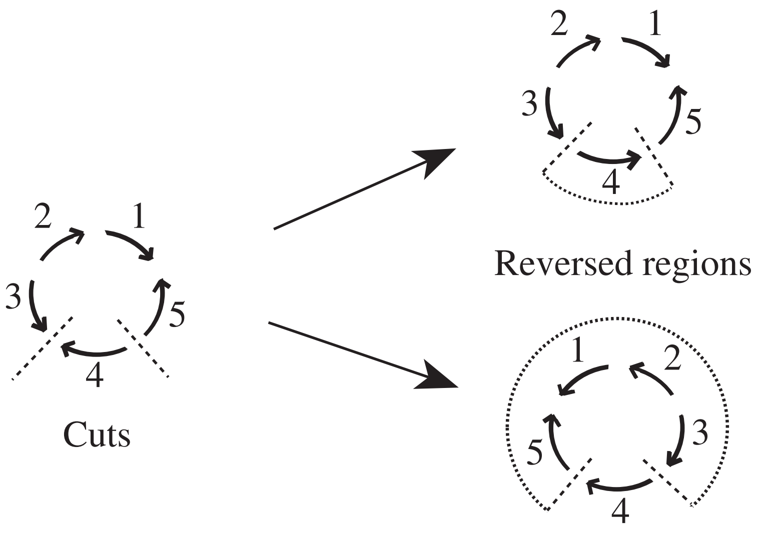

Figure 2: This example shows the two possibilities for reversal in a circular chromosome, given two cuts.

In general terms, the *problem of reversal distance of signed circular chromosomes* is formulated as follows. Given two circular chromosomes A and B, we want the shortest series of reversals that transforms A into B. This minimum number of reversals is called *reversal distance* between A and B. Figure 3 shows an example of a circular chromossome transformed into another with the minimum number of reversals.

Another version of this problem arises when the orientations of the genes on the chromosomes are not known. In that case, we have the *unsigned* version of the problem, where the reversals only reverse gene order. There are other versions of the same problem considering linear chromosomes, and other mutational events besides reversal. The literature on problems originated by different types of mutational events is growing very quickly in recent years. In the following, we briefly review other works studying reversal, observing that chromosomes are commonly represented by permutations in this context.

With respect to linear chromosomes, Aigner and West [1] had studied the problem of sorting a permutation, considering the operation of reinsertion of the first element in the sequence of the permutation. The sorting diameter (the maximum distance between two permutations) in this case is $n - 1$, where n is the number of elements of the permutation. Kececioglu and Sankoff [6] had studied the problem of the reversal distance of unsigned linear permutations, and developed the first approximation algorithm for the problem. Their algorithm runs in $O(n^2)$ time and is guaranteed to use no more than two times the reversal distance. They also developed efficient bounds, used on a branch-and-bound algorithm, that solved to optimality or almost optimality permutations ranging from 30 to 50 elements. Bafna and Pevzner [2] afterwards introduced a new structure, the *breakpoint graph* of an initial permutation relative to a target permutation, that allowed to set up a more precise lower bound to the reversal distance, considering another parameter, based

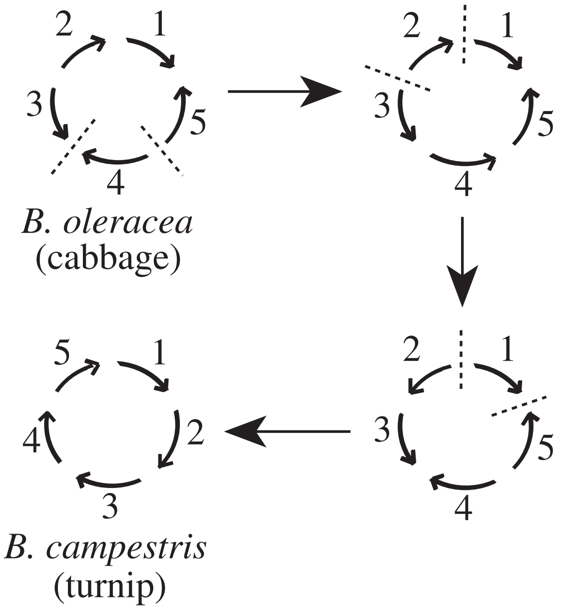

Figure 3: This example shows a series of reversals that transforms *B. oleracea* (cabbage) into *B. campestris* (turnip).

on a maximum alternating cycle decomposition, denoted by $c(\pi)$. Based on that graph, they devised an approximation algorithm with a performance guarantee of 7/4, and introduced an approximation algorithm for signed permutations with a guarantee of 3/2.

Hannenhalli and Pevzner [4] introduced two new parameters: the *number of hurdles* ($h(\pi)$) and an indicator of whether the breakpoint graph *is a fortress*, ($f(\pi)$). Together with the maximum number of cycles of the alternating cycles decomposition ($c(\pi)$) of the breakpoint graph, these parameters allowed the authors to demonstrate a *duality theorem*. Based on this theorem, they presented the first polinomial algorithm for the problem of the reversal distance of the signed linear permutations, with time complexity $O(n^4)$. Berman and Hannenhalli [3] introduced new data structures on that algorithm and lowered the complexity to $O(n^2\alpha(n))$. Finally, Kaplan, Shamir and Tarjan [5], based on the Hannenhalli and Pevzner theory, and using part of the Berman and Hannenhalli algorithm, showed a new algorithm with $O(n^2)$ complexity. We will call this last one *KST algorithm*.

With respect to circular chromosomes, Watterson and other authors [10] showed an algorithm, very simple, to find out the reversal distance of circular permutations, establishing a lower bound ($number\ of\ breakpoints/2$), and an upper bound ($n-2$) for the reversal distance. They presented also a stochastic algorithm for the problem. Kececioglu and Sankoff [7] presented an exact branch-and-bound algorithm for the problem of reversal distance of signed circular permutations. This algorithm, using simple methods to find the lower and upper bounds, found extremely precise values for the reversal distance in several experiments. The authors reported that they did not know reasons to justify the proximity of these limits. Now we know that the Hannenhalli and Pevzner theory justifies these results, because $h(\pi)$ and $f(\pi)$ are small for random permutations.

In this paper we present a formalism for circular chromosomes and for reversals acting on them. As a consequence we show polinomial algorithms for the problem of reversal distance of signed circular chromosomes. These algorithms are based on the theory for the linear problem given by Hannenhalli and Pevzner [4]. Besides, we calculate the reversal diameter for linear and circular chromosomes.

In Section 2 we first formalize a circular chromosome by an equivalence class, and next we show that there is an isomorphism between reversals acting on circular chromosomes and reversals acting on linear chromosomes. This result allow us to compute the reversal distance of signed circular chromosomes by computing the reversal distance of signed linear chromosomes with one less gene. In Section 3 we show some results concerning the reversal distances of signed circular and linear permutations of the same size. In Section 4 we calculate the reversal diameter of signed linear and circular permutations. Finally, the last section brings conclusions of this work and indicates some future directions.

2 A formalization for the problem

2.1 Linear Chromosomes

We begin by presenting a brief overview of some important results about signed linear chromosomes, due mainly to Bafna and Pevzner [2] and Hannenhalli and Pevzner [4]. A signed linear chromosome is represented by a signed permutation. A *signed permutation* is an ordinary permutation, except that each element has positive $(+)$ or negative $(-)$ sign, indicating the relative orientation of the block. In this case, a *reversal* ϱ of the interval $[i, j]$ is denoted by $\varrho(i, j)$ and we have

$$\varrho(i, j) \cdot \pi = (\pi_1 \dots \pi_{i-1} \overline{\pi}_j \overline{\pi}_{j-1} \dots \overline{\pi}_{i+1} \overline{\pi}_i \pi_{j+1} \dots \pi_n)$$

where $\overline{\pi}_k$ indicates the inversion of the sign of π_k.

The problem of the reversal distance of signed linear chromosomes is commonly formalized as follows. Given two permutations π and σ modeling two signed linear chromosomes, the *reversal distance problem* of π and σ is to find a series of reversals $\varrho_1, \varrho_2, \dots, \varrho_t$ such that $\varrho_t \cdot \varrho_{t-1} \cdot \dots \cdot \varrho_2 \cdot \varrho_1 \cdot \pi = \sigma$ and t is minimum. We call t the **reversal distance of** π **and** σ, denoted by $d(\pi, \sigma)$.

The algorithms of Bafna and Pevzner [2] and of Hannenhalli and Pevzner [4] are based on a structure called *breakpoint graph.* This graph is constructed from π and σ as follows. Each one of the signed integers is represented by an arrow, from left to right when the sign is $+$, and from right to left when the sign is $-$. The initial and final points of these arrows are the vertices of this graph. Besides, we add two reference points, one on the left of the sequence (labelled by L) and the other on its right (labelled by R). After that, we put *reality* edges joining extreme points of adjacent arrows in π, and *desire* edges joining extreme points of adjacent arrows in σ. Important properties of this graph are:

1. The resulting graph is formed by a collection of even cycles. When $\pi = \sigma$, the number of these cycles gets its maximum value, $n + 1$. For two different permutations, there are less than $n + 1$ cycles.

2. Each reality edge from a cycle whose size is larger than 2 represents a *breakpoint* in the permutation, that is, a point where a reversal will have to act in order

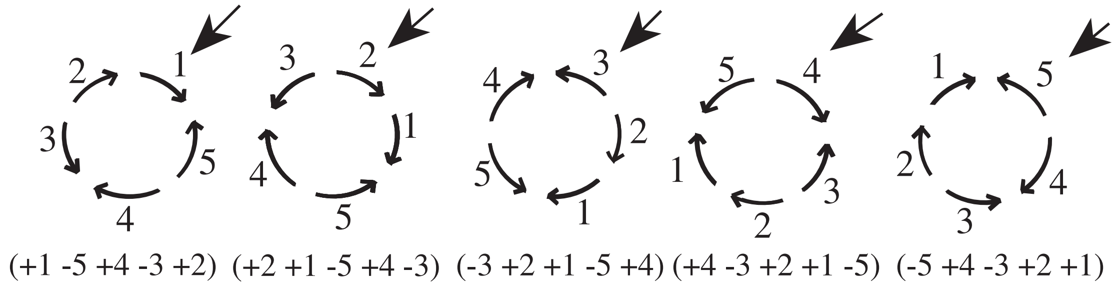

Figure 4: In a circular chromosome we can choose each one of the genes block as the first one. Then, all of these sequences are considered equivalent, and they represent the circular chromosome of *B. oleracea* shown in the Figure 1 (a).

> to transform π into σ. When two vertices belong to a cycle of size 2, that is, are joined by two parallel edges, exactly one reality and one desire edge, we say that *there is not* a break in that position.

From this graph we can compute three parameters that allow us to compute the reversal distance of π and σ: the number of cycles $c(\pi, \sigma)$, the number of hurdles $h(\pi, \sigma)$ and a parameter $f(\pi, \sigma)$ indicating whether the graph is a fortress, where this last value can be equal to 1 or 0 only. The reversal distance is then given by:

$$d(\pi, \sigma) = n + 1 - c(\pi, \sigma) + h(\pi, \sigma) + f(\pi, \sigma).$$

We refer the readers to the important works [4, 3, 5] or to the introdutory text of Setubal and Meidanis [9] for a more detailed explanation on these parameters. We will not need details on hurdles and fortresses until Section 4, where we will review these concepts.

2.2 Circular Chromosomes

Now we give a formalization of a circular chromosome by an equivalence class.

Intuitively, a circular chromosome is a circular arrangement of signed blocks (see Figure 1). A *block* of genes of the chromosome will be modelled by a signed integer. The sign "+" indicates an arrow in clockwise direction in Figure 1, and the sign "−" indicates an arrow in counterclockwise direction. Given an *initial block*, we can represent a circular chromosome by a sequence as follows. Through convention, we always read the blocks in clockwise direction. Walk around the chromosome in clockwise direction, beginning at the initial block, and write down the signed integers corresponding to the blocks found. Then, $\pi = (\pi_1 \pi_2 \ldots \pi_n)$ will denote the circular chromosome, with n blocks of genes. As an example, the chromosome of *B. oleracea* of Figure 1a can be represented by the sequence $(+1\ {-5}\ {+4}\ {-3}\ {+2})$.

We can choose each one of the blocks as the first one, and therefore we can have many different sequences representing the same chromosome (see Figure 4). All of these sequences are considered equivalent. Besides, two sequences where one of them is obtained by the other by reflection are considered equivalent, and in particular $\pi = (\pi_1 \pi_2 \ldots \pi_n)$ and $s \cdot \pi = (\overline{\pi}_n \overline{\pi}_{n-1} \ldots \overline{\pi}_2 \overline{\pi}_1)$ are considered equivalent sequences (see Figure 5).

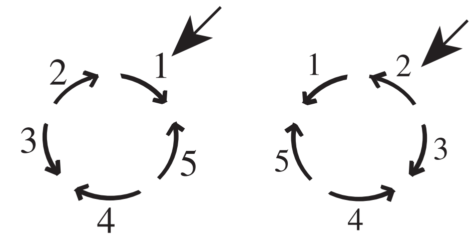

Figure 5: In a circular chromosome two sequences where one of them is obtained from the other by reflection are considered equivalent. The circular chromosome represented is *B. oleracea* shown in Figure 1 (a).

This way, a sequence modeling a circular chromosome is a representative of an equivalence class in the set of all sequences. Below we define the *rotation* and the *reflection* operations, that will formalize the two characteristics described above. From these operations we will define an equivalence relation between two sequences, and an equivalence class that will represent a circular chromosome.

We will call S_n the set of all possible sequences of distinct signed integers, where each sequence has size n. These integers must belong to the interval $[1..n]$. Observe that $|S_n| = 2^n n!$. Let us take $\pi = (\pi_1 \pi_2 \dots \pi_n)$, a sequence of S_n. We will define two types of operations acting in π as follows:

- *Rotations.* We will denote by r the basic rotation that moves the permutation elements one position to the left:

  $$r \cdot \pi = (\pi_2 \pi_3 \dots \pi_n \pi_1).$$

  We will define r^i for every $i \in Z$ in the usual way: r^i is the composition of r i times for $i > 0$ and r^{-i} is the inverse of r^i. Besides, r^0 is the identity. We have the following important relations:

  $r^n = r^0$, or more generally, $r^i = r^j$ if $i \equiv j \pmod{n}$ for all $i, j \in Z$.

  $r^i r^j = r^{i+j}$ for all $i, j \in Z$.

  The operations r^i are called *rotations.*

- *Reflections.* We will denote by s the basic reflection that inverses the order of the permutation and also the signs. So,

  $$s \cdot \pi = (\overline{\pi}_n \overline{\pi}_{n-1} \dots \overline{\pi}_2 \overline{\pi}_1).$$

  We will define s^i for all $i \in Z$ as follows: s^i is the composition of s i times for $i > 0$ and s^{-i} is the inverse of s^i. Note that $s^{-i} = s^i$. Besides, s^0 is the identity. We have the following important relations:

  $s^2 = s^0$, or more generally, $s^i = s^j$ if $i \equiv j \pmod{2}$ for all $i, j \in Z$.

$s^i s^j = s^{i+j}$ for all $i, j \in Z$.

We can apply r and s to a sequence, using the above definitions. Then, $rs\pi = r(s\pi) = r(\overline{\pi}_n \overline{\pi}_{n-1} \ldots \overline{\pi}_2 \overline{\pi}_1) = (\overline{\pi}_{n-1} \overline{\pi}_{n-2} \ldots \overline{\pi}_2 \overline{\pi}_1 \overline{\pi}_n)$.

We have the following relation:

$$rs = sr^{-1}. \tag{1}$$

Generically, the operations sr^i are called *reflections*. Each reflection is equal to its own inverse.

Now we will define an equivalence relation between two sequences π and γ.

Definition 2.1 *Given two sequences π and γ, we define*

$$\pi \sim \gamma$$

if and only if there are $i, j \in Z$ such that $\gamma = r^i s^j \cdot \pi$.

The above relation is an equivalence relation. The proof of this result is simple. Equation (1) can be used in this proof.

From this equivalence relation, we can define an equivalence class of the sequence π, denoted by $[\pi]$, which represents a signed circular chromosome, as follows

$$[\pi] = \{\gamma \in S_n | \pi \sim \gamma\}$$

This formalization is interesting biologically, because it does not fix the first element of the sequence, and then each one of the genes block can be the first, it is sufficient to apply rotation. Besides, two sequences where one of them is obtained from the other by reflection can be produced applying the s operator.

2.3 Circular Reversals

We model now how a reversal will act in a class A representing a circular chromosome. First we note that there are two possibilities for a reversal acting on a circular chromosome, given the two points where the cuts have occurred (see Figure 6).

Suppose the two cuts occur between $i \ominus 1, i$ and $j, j \oplus 1$, with $1 \leq i \leq j \leq n$. Here $\ominus$ and $\oplus$ are the usual operations of subtraction and addition, respectively, except that we take the result modulo n and choose n rather than zero as the representative of the class of multiples of n. We will assume that these cuts are distinct, therefore $i \neq (j \oplus 1)$.

Also if we choose i and j such that $i > j$, we can change i and j without problems because both are just pointers to the cuts.

Then we have the following lemma.

Lemma 2.1 *Given a sequence π from an equivalence class A which models a circular chromosome, and two integers i and j with $1 \leq i \leq j \leq n$ and $i \neq (j \oplus 1)$ such that these cuts occur between $i \ominus 1, i$, and $j, j \oplus 1$, the sequences resulting from the two possible ways of reversing the circular chromosome between these cuts belong to the same equivalence class.*

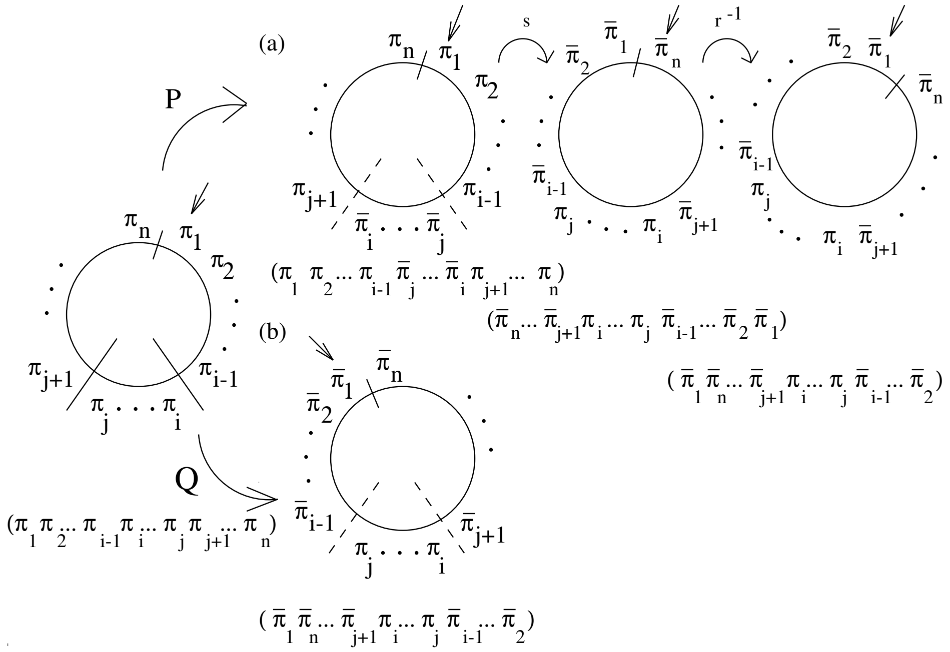

Figure 6: This figure shows that the two circular chromosomes resulting from the reversal are represented by two sequences that belong to the same equivalence class. Note that the arrow, before the reversion, indicates the first block of the sequence chosen from the equivalence class which represents the circular chromosome. The portion of the chromosome suffering the reversal can include or not the arrow. (a) In this case, the reversal does not include the arrow. The sequence resulting from the reversal is shown. (b) In that case, the reversal includes the arrow. We can apply reflection and rotation in the sequence resulting from the reversal in order to obtain the same sequence as in case (a). The sequences resulting from each operation are shown.

Proof:

We will denote by P and Q the two possible ways of reversing the circular chromosome (see Figure 6). Taking sequence $\pi = (\pi_1 \ldots \pi_{i-1}\pi_i \ldots \pi_j\pi_{j+1} \ldots \pi_n)$ from class A, and applying P on A we have

$$P \cdot [(\pi_1 \ldots \pi_{i-1}\pi_i \ldots \pi_j\pi_{j+1} \ldots \pi_n)] = [(\pi_1 \ldots \pi_{i-1}\overline{\pi}_j \ldots \overline{\pi}_i\pi_{j+1} \ldots \pi_n)]$$

Applying Q on A we have

$$Q \cdot [(\pi_1 \ldots \pi_{i-1}\pi_i \ldots \pi_j\pi_{j+1} \ldots \pi_n)] = [(\overline{\pi}_1\overline{\pi}_n \ldots \overline{\pi}_{j+1}\pi_i \ldots \pi_j\overline{\pi}_{i-1} \ldots \overline{\pi}_2)]$$

But applying r^{-1} and s on $(\pi_1 \ldots \pi_{i-1}\overline{\pi}_j \ldots \overline{\pi}_i\pi_{j+1} \ldots \pi_n)$ we have

$$r^{-1}s \cdot (\pi_1 \ldots \pi_{i-1}\overline{\pi}_j \ldots \overline{\pi}_i\pi_{j+1} \ldots \pi_n) = (\overline{\pi}_1\overline{\pi}_n \ldots \overline{\pi}_{j+1}\pi_i \ldots \pi_j\overline{\pi}_{i-1} \ldots \overline{\pi}_2)$$

So

$$[P \cdot A] = [Q \cdot A]$$

□

Now we can enunciate the problem of finding the minimal number of reversals acting on circular chromosomes with known relative orientations.

$[\pi] = \{(+1\ -5\ +4\ -3\ +2)\ (-5\ +4\ -3\ +2\ +1)\ (+4\ -3\ +2\ +1\ -5)$

$(-3\ +2\ +1\ -5\ +4)\ (+2\ +1\ -5\ +4\ -3)$

$(-2\ +3\ -4\ +5\ -1)\ (+3\ -4\ +5\ -1\ -2)\ (-4\ +5\ -1\ -2\ +3)$

$(+5\ -1\ -2\ +3\ -4)\ (-1\ -2\ +3\ -4\ +5)\}$

$\mathrm{can}\ ([\pi]) = (+1\ -5\ +4\ -3\ +2)$

Figure 7: Example of an equivalence class $[\pi]$ and its canonical representative $can([\pi])$.

Given two equivalence classes A and B, representing two circular chromosomes with known relative orientations, the **problem of reversal distance of signed circular chromosomes** is to find a series of reversals $P_1, P_2, \ldots, P_u$ such that $P_u \cdot P_{u-1} \cdot \ldots \cdot P_2 \cdot P_1 \cdot A = B$ and u is minimum. We call u the **reversal distance** of A and B, denoted by $d^c(A, B)$.

2.4 Relating Circular Chromosomes to Linear Chromosomes

In the formalization of circular chromosomes, we would like to use some results from the linear case. A linear reversal $\varrho(i, j)$ acts as described in Section 2.1. It would be tempting to define a corresponding circular reversal $\varrho^c(i, j)$ by

$$\varrho^c(i, j) \cdot [\pi] = [\varrho(i, j) \cdot \pi]$$

However this definition does not make sense, because different choices of sequences π inside an equivalence class A lead to non-equivalent right-hand members. So, it will not be permitted a random choice of the sequence in A in which the reversal will act.

We will define a *canonical representative* of A, denoted by $can(A)$, with the characteristics of having the 1 block fixed as the first element of the sequence, and with the + orientation (see Figure 7). Note that each equivalence class has a unique canonical representative. For the formalism, a reversal will be applied only in the canonical representative. Thus, given a linear reversal $\varrho(i, j)$ with $1 \leq i \leq j \leq n$, we define a circular reversal $\varrho^c(i, j)$ by the formula

$$\varrho^c(i, j) \cdot A = [\varrho(i, j) \cdot can(A)]$$

Notice that the case $(i, j) = (1, n)$ is excluded from consideration as mentioned in Section 2.3.

The next theorem tells us that every circular reversal is of the form $\varrho^c(i, j)$ for some i, j. Moreover, we can always choose the indices from 2 to n.

Theorem 2.1 *For any circular reversal P, there are integers i and j with $2 \le i \le j \le n$ such that*

$$P \cdot A = [\varrho(i,j) \cdot can(A)].$$

Proof:

A circular reversal P must be applied only in the canonical representative of the equivalence class A representing the circular chromosome. There are two possible forms for a reversal acting on any sequence of A, but both of them produce sequences that belong to the same equivalence class (Lemma 2.1). As we can choose any of these forms we will pick the form not including $\pi_1 = +1$. This way $P \cdot A$ will produce a sequence which is also a canonical representative. In other words, the canonical representative of the equivalence class which models the circular chromosome before the reversal is carried to a canonical representative of the equivalence class which represents the circular chromosome after the reversal. In this case, in terms of the linear representation, the reversal acts in the canonical representative like a linear reversal $\varrho(i,j)$. Then,

$$\varrho(i,j) \cdot (+1\ \pi_2 \ldots \pi_i \ldots \pi_j \ldots \pi_n) = (+1\ \pi_2 \ldots \overline{\pi}_j \ldots \overline{\pi}_i \ldots \pi_n)$$

with $2 \le i \le j \le n$. This comes from the definition of linear reversal. As the right sequence is canonical we have

$$\varrho(i,j) \cdot can(A) = can(P \cdot A)$$

from where

$$[\varrho(i,j) \cdot can(A)] = [can(P \cdot A)] = P \cdot A.$$

□

We will see now that there is an isomorphism between reversals acting on circular chromosomes and reversals acting on linear chromosomes. To prove this, we will initially define two bijections. Recall that S_n is the set of all signed linear permutations on n elements. Let R_n be the set of all linear reversals on n elements, and S_n^c, R_n^c the analogous sets for the circular case. Define

$$\varphi : S_n^c \longrightarrow S_{n-1}$$

so that

$$\varphi(A) = \text{ take } can(A)\text{, remove } +1\text{, subtract 1 from the others}$$

and

$$\theta : R_n^c \longrightarrow R_{n-1}$$

so that

$$\theta(P) = \varrho(i-1, j-1)$$

where $P = \varrho^c(i,j)$, $2 \le i \le j \le n$.

We enunciate the result.

Theorem 2.2 *Given the two bijections φ and θ defined above, we have*

$$\varphi(P \cdot A) = \theta(P) \cdot \varphi(A)$$

Proof:

First we have

$$\varphi(P \cdot A) = \text{ take } can(P \cdot A), \text{ remove } +1, \text{ subtract } 1 \text{ from the others}$$

Let (by Theorem 2.1) $P = \varrho^c(i,j)$, with $2 \le i \le j \le n$, and $A = [\pi]$, where $\pi_1 = +1$. Then

$$\varphi(P \cdot A) = \text{ take } \varrho(i,j) \cdot \pi, \text{ remove } +1, \text{ subtract } 1 \text{ from the others}$$

On the other side, since $can(A) = \pi$, we have

$$\varphi(A) = \text{ take } \pi, \text{ remove } +1, \text{ subtract } 1 \text{ from the others}$$

and

$$\theta(P) = \varrho(i-1, j-1)$$

Then we have the result, because $\theta(P)$ will act on the same elements as $\varrho(i,j)$.

□

Note that $\mid S_n^c \mid = \mid S_{n-1} \mid = 2^{n-1}(n-1)!$ and $\mid R_n^c \mid = \mid R_{n-1} \mid = ((n-1)n)/2$.
From Theorem 2.2 we have immediately

Corollary 2.1 *Given any two classes A and B modeling two circular chromosomes, and the bijection φ defined above,*

$$d^c(A,B) = d(\varphi(A), \varphi(B))$$

From Corollary 2.1 we can derive an algorithm to the problem of signed circular chromosomes. Basically it consists in running any algorithm solving the problem of signed linear chromosomes, taking as inputs two permutations, obtained from applying the bijection φ in the two classes representing the circular chromosomes.

In particular, if we take the KST algorithm, the complexity of the algorithm is $O(n^2)$ (to find out the input sequences costs $O(n)$ and the KST algorithm has complexity $O(n^2)$), where n is the number of genes blocks of the circular chromosomes.

We can also obtain the circular reversals used, just applying the inverse of θ on each step of the algorithm for the linear chromosome. It does not affect the complexity of the above algorithm because it takes $O(1)$.

3 Relating circular chromosomes to linear chromosomes of the same size

In the previous section we saw that there is a distance preserving correspondence between circular chromosomes and linear chromosomes of size one unit smaller. Here we will derive similar results for circular and linear chromosomes of the same size.

First of all, we would like to know what is the relation between $d(\pi, \sigma)$ and $d^c([\pi], [\sigma])$ for any π and σ. As we will see in Theorem 3.1, $d^c([\pi], [\sigma]) \le d(\pi, \sigma)$. Before this, we need three technical lemmas.

Lemma 3.1 *Given two linear permutations σ and π, such that $\sigma = q \cdot \pi$ where $q = r$ or s, then for every reversal ϱ there is a reversal ϱ' such that $\varrho \cdot \pi \sim \varrho' \cdot \sigma$.*

Proof:

Let $\pi = (\pi_1 \dots \pi_n)$.
We have two possibilities for q.

- Suppose $\sigma = r \cdot \pi = (\pi_2\pi_3 \dots \pi_n\pi_1)$ and $\varrho = \varrho(i,j)$ so that
$$\varrho(i,j) \cdot \pi = (\pi_1 \dots \pi_{i-1}\overline{\pi}_j \dots \overline{\pi}_i\pi_{j+1} \dots \pi_n)$$
We have three cases.
  1. $i = 1, j = n$:
  $$\varrho(1,n) \cdot \pi = (\overline{\pi}_n \dots \overline{\pi}_1)$$
  In this case $\varrho = s$. Take $\varrho' = s$ also. We have
  $$\varrho \cdot \pi = s \cdot \pi \sim \pi \sim \sigma \sim s \cdot \sigma = \varrho' \cdot \sigma$$
  2. $i = 1, j < n$: In this case
  $$\varrho(1,j) \cdot \pi = (\overline{\pi}_j \dots \overline{\pi}_1\pi_{j+1} \dots \pi_n).$$
  Then:
  $$\varrho(j,n-1) \cdot \sigma = (\pi_2\pi_3 \dots \pi_{j-1}\pi_j\overline{\pi}_n \dots \overline{\pi}_{j+1}\pi_1)$$
  $$s \cdot \varrho(j,n-1) \cdot \sigma = (\overline{\pi}_1\pi_{j+1} \dots \pi_n\overline{\pi}_j \dots \overline{\pi}_3\overline{\pi}_2)$$
  $$r^{n-j+1} \cdot s \cdot \varrho(j,n-1) \cdot \sigma = (\overline{\pi}_j \dots \overline{\pi}_3\overline{\pi}_2\overline{\pi}_1\pi_{j+1} \dots \pi_n)$$
  $$\varrho(1,j) \cdot \pi = r^{n-j+1} \cdot s \cdot \varrho(j,n-1) \cdot \sigma$$
  Therefore, taking $\varrho' = \varrho(j,n-1)$ we have
  $$\varrho \cdot \pi \sim \varrho' \cdot \sigma.$$
  3. $i > 1, j \leq n$: In this case
  $$\varrho(i,j) \cdot \pi = (\pi_1 \dots \pi_{i-1}\overline{\pi}_j \dots \overline{\pi}_i\pi_{j+1} \dots \pi_n).$$
  Then:
  $$\varrho(i-1,j-1) \cdot \sigma = (\pi_2\pi_3 \dots \overline{\pi}_j \dots \overline{\pi}_i \dots \pi_n\pi_1)$$
  $$r^{-1} \cdot \varrho(i-1,j-1) \cdot \sigma = (\pi_1\pi_2\pi_3 \dots \pi_{i-1}\overline{\pi}_j \dots \overline{\pi}_i \dots \pi_n)$$
  $$\varrho(i,j) \cdot \pi = r^{-1} \cdot \varrho(i-1,j-1) \cdot \sigma$$
  Therefore, taking $\varrho' = \varrho(i-1,j-1)$ we have
  $$\varrho \cdot \pi \sim \varrho' \cdot \sigma.$$
- Suppose $\sigma = s \cdot \pi = (\overline{\pi}_n \dots \overline{\pi}_1)$ and $\varrho = \varrho(i,j)$ so that
$$\varrho(i,j) \cdot \pi = (\pi_1 \dots \pi_{i-1}\overline{\pi}_j \dots \overline{\pi}_i\pi_{j+1} \dots \pi_n)$$
Then:
$$\varrho(n+1-j, n+1-i) \cdot \sigma = (\overline{\pi}_n \dots \overline{\pi}_{j+1}\pi_i \dots \pi_j\overline{\pi}_{i-1} \dots \overline{\pi}_1)$$
$$s \cdot \varrho(n+1-j, n+1-i) \cdot \sigma = (\pi_1 \dots \pi_{i-1}\overline{\pi}_j \dots \overline{\pi}_i\pi_{j+1} \dots \pi_n)$$
$$\varrho(i,j) \cdot \pi = s \cdot \varrho(n+1-j, n+1-i) \cdot \sigma$$
Therefore, taking $\varrho' = \varrho(n+1-j, n+1-i)$, we have
$$\varrho \cdot \pi \sim \varrho' \cdot \sigma.$$

$\square$

Lemma 3.2 *Given two linear permutations π and σ, such that $\pi \sim \sigma$ then for every reversal ϱ there is a reversal ϱ' such that $\varrho \cdot \pi \sim \varrho' \cdot \sigma$.*

Proof:

Take $\sigma = q_v \cdot q_{v-1} \cdot \ldots \cdot q_1 \cdot \pi$, where $v \geq 0$, and $q_i = r$ or s, for $1 \leq i \leq v$.
This proof will be made by induction on v.

- $v = 0$: just make $\varrho' = \varrho$
- $v > 0$: Take
$$\sigma' = q_{v-1} \cdot \ldots \cdot q_1 \cdot \pi$$
Given ϱ, we want to obtain ϱ' such that
$$\varrho \cdot \pi \sim \varrho' \cdot \sigma$$
By the induction hiphotesis, we have
$$\varrho \cdot \pi \sim \varrho'' \cdot \sigma'$$
But, $\sigma = q_v \cdot \sigma'$ and then, using Lemma 3.1, there is ϱ' such that
$$\varrho'' \cdot \sigma' \sim \varrho' \cdot \sigma$$
Then,
$$\varrho \cdot \pi \sim \varrho' \cdot \sigma$$

$\square$

Lemma 3.3 *Given a permutation π and a reversal ϱ, then*
$$[\varrho \cdot \pi] = P \cdot [\pi]$$
where $P = I$, the identity transformation, or P is a circular reversal.

Proof:

Let σ be the canonical representative of $[\pi]$:
$$\pi \sim \sigma = can([\pi])$$
Lemma 3.2 says that given ϱ there is ϱ' such that
$$\varrho \cdot \pi \sim \varrho' \cdot \sigma,$$
hence
$$[\varrho \cdot \pi] = [\varrho' \cdot \sigma]$$
But $\varrho' = \varrho(i, j)$ with $1 \leq i \leq j \leq n$. Then we have two cases:

1. $i = j \oplus 1$. Then, $\varrho' = s$ and

$$[\varrho \cdot \pi] = [s \cdot \sigma] = [\sigma] = [\pi],$$

so $P = I$ works in this case.

2. $i \neq j \oplus 1$. Because σ is canonical,

$$[\varrho(i,j) \cdot [\sigma]] = \varrho^c(i,j) \cdot [\pi]$$

hence $P = \varrho^c(i,j)$ works in this case.

$\square$

Now we show that there are fewer reversals in the circular case than in the linear case when both chromosomes have the same size.

Theorem 3.1 *Given any two permutations π and σ,*

$$d(\pi, \sigma) \geq d^c([\pi], [\sigma])$$

Proof:
Take $t = d(\pi, \sigma)$. Then,

$$\varrho_t \cdot \varrho_{t-1} \cdot \ldots \cdot \varrho_1 \cdot \pi = \sigma$$

$$[\varrho_t \cdot \varrho_{t-1} \cdot \ldots \cdot \varrho_1 \cdot \pi] = [\sigma]$$

Using Lemma 3.3 we have

$$P'_t \cdot P'_{t-1} \cdot \ldots \cdot P'_1 \cdot [\pi] = [\sigma]$$

where P'_i is either a circular reversal or the identity. Then,

$$d^c([\pi], [\sigma]) \leq t = d(\pi, \sigma)$$

$\square$

We note that it is not true that $d(\pi, \sigma) = d^c([\pi], [\sigma])$, for any π and σ. It is enough to take $\pi = (-2 + 3 + 1)$ and $\sigma = (+1 + 2 + 3)$. We have $d^c([\pi], [\sigma]) = 1$ because $d^c([\pi], [\sigma]) = d(can([\pi]), can([\sigma])) = 1$, where $can([\pi]) = (+1 - 2 + 3)$. But $d(\pi, \sigma) = 3$. To make this computation, it is sufficient to construct the breakpoint graphs of π and σ, and use the formula presented by Hannenhalli and Pevzner [4].

Following we demonstrate another theorem that solves the problem of reversal distance for signed circular chromosomes.

Theorem 3.2 *Given two circular chromosomes represented by classes A and B we have*

$$d^c(A, B) = d(can(A), can(B))$$

Proof:

First we will show that

$$d^c(A, B) \leq d(can(A), can(B))$$

By Theorem 3.1 we know that $d(\pi, \sigma) \geq d^c([\pi], [\sigma])$. In particular, taking $\pi = can(A)$ and $\sigma = can(B)$, we immediately have the result.

Secondly, we will show that

$$d^c(A, B) \geq d(can(A), can(B))$$

To solve the problem of the reversal distance of signed circular chromosome, we use reversals in the interval $[2, n]$, that act always in the canonical representative sequence. Considering the linear chromosome $can(A)$, initially, $\pi_1 = +1$ is in its correct position, and this is not modified throughout the process. Thus, these reversals supply a series of reversals for the linear case too.

□

From Theorem 3.2 we can derive another algorithm for the problem of signed circular chromosomes that consists in running any algorithm solving the problem of signed linear chromosomes giving as input the canonical representatives of A and B.

Let us take the two input permutations α and β, where α is a permutation of the A class which represents one of the circular chromosomes, and β a permutation of the B class wich represents the other circular chromosome. The canonical representatives are obtained traversing the two permutations α and β finding the position k of the 1 block. If it has sign $+$ we just apply r^{k-1}, and if it has sign $-$ we apply r^{k-n} followed by s.

In particular, if we take the KST algorithm, the complexity of the algorithm is $O(n^2)$ (to find out the canonical representatives costs $O(n)$ and the KST algorithm has complexity $O(n^2)$), where n is the number of gene blocks of the circular chromosomes.

Finally we prove a theorem that allow us to say that the canonical representatives of the classes modeling the circular chromosomes provide a minimum distance, among all permutations belonging to those two classes.

Theorem 3.3 *Given any two classes A and B modeling circular chromosomes, we have*

$$d(can(A), can(B)) = \min_{\substack{\pi \in A \\ \sigma \in B}} \{d(\pi, \sigma)\}.$$

Proof:

To begin with, notice that we have $d^c(A, B) = d(can(A), can(B))$ (from Theorem 3.2). From Theorem 3.1, we have each value $d(\pi, \sigma)$ greater than or equal to $d^c([\pi], [\sigma])$.

□

A question arises here. Which sequences, from the two equivalence classes modeling the circular chromosomes, lead to a minimum reversal distance? Our results showed that the canonical representatives from the classes certainly do. But they are not the only ones. An example found in an article of Palmer and co-authors [8] did not have the characteristics of our canonical representatives, but led to a minimum distance. The sequences in that case were $(-8\ -7\ -6\ -5\ -4\ -3\ -2\ -1\ -11\ -10\ -9C\ -9B\ -9A)$ and $(-4\ +3\ -2\ +8\ +7\ -1\ -5\ -6\ -11\ +10\ +9A\ -9B\ +9C)$. If we call *optimal representatives* of two classes modeling circular chromosomes, two permutations, one for each class, that lead to a minimum reversal distance, we would like to know how to characterize this set of optimal representatives.

From the above results, it can be shown that Corollary 2.1 and Theorem 3.2 are equivalent, in the following sense:

Theorem 3.4 *Given two classes A and B modeling two linear chromosomes and the bijection φ defined above, then*

$$d(\varphi(A), \varphi(B)) = d(can(A), can(B))$$

4 The reversal diameter of signed chromosomes

The **circular reversal diameter**, denoted by $D^c(n)$, of the equivalence classes on S_n, with respect to the circular reversal distance, is the maximum distance between two equivalence classes. Similarly, the **linear reversal diameter**, denoted by $D(n)$, of the n element permutations of the set S_n, with respect to the linear reversal distance, is the maximum distance between two permutations. We show now that the reversal diameter for signed linear and circular chromosomes are respectively $n+1$ and n (except in a few cases). This corrects a statement from Kececioglu and Sankoff [7] that said that $n-2 \leq D(n) \leq n-1$.

Now we need some definitions and facts about hurdles and fortresses, as mentioned earlier. A cycle C is **bad** when for any reversal ϱ acting on two reality edges of C we have

$$c(\pi, \sigma) = c(\varrho \cdot \pi, \sigma).$$

Otherwise, the cycle is **good**.

Two cycles are **interleaving** when there are two desire edges, one from each cycle, that cross. A cycle C is **contained** in another cycle D when C and D are not interleaving and C is contained in at least one desire edge of D.

The following facts will be important in this section:

- If a bad cycle C does not interleave with and does not contain any other cycle, then C forms a **hurdle** just by itself. We should point out that these are not the only types of hurdles that can exist in a breakpoint graph, but this will suffice for our purposes.

- In a **fortress** there is at least one cycle that does not belong to a hurdle. Again, we point out that this condition is not sufficient to define fortresses.

Theorem 4.1 *The reversal diameter of linear chromosomes is*

$$D(n) = \max_{\substack{\pi \in S_n \\ \sigma \in S_n}} \{d(\pi, \sigma)\} = \begin{cases} n & \text{if } n = 1 \text{ or } n = 3 \\ n+1 & \text{otherwise} \end{cases}$$

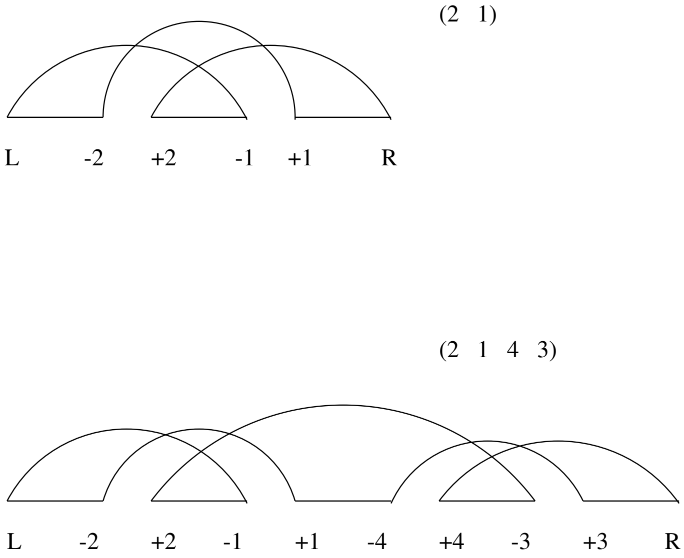

Figure 8: The breakpoint graph for $n = 2$ e 4 with respect to ι_n.

Proof:

We will show two sequences, π_n and σ_n, that give $d(\pi_n, \sigma_n) = n + 1$, for each n. The construction depends on n being even or odd. All the integers will have $+$ sign in our examples, so we omit them in the proof.

- For n even, $n \geq 2$, consider

$$\pi_n = (2\ 1\ 4\ 3\ 6\ 5\ \ldots\ n-4\ n-5\ n-2\ n-3\ n\ n-1)$$

  and

$$\sigma_n = \iota_n = (1\ 2\ 3\ 4\ \ldots\ n-1\ n)$$

  The breakpoint graph for π_n with respect to ι_n is formed by exactly one cycle, of size $n + 1$, involving all labels. This is a bad cycle and therefore a hurdle. Figure 8 shows examples of breakpoint graphs for $n = 2$ and 4 with respect to ι_n.

  In this case, using the Hannenhalli and Pevzner formula [4], and by construction of the breakpoint graph $G(\pi_n, \sigma_n)$ of the sequence π_n with respect to ι_n,

$$d(\pi_n, \iota_n) = (n + 1)\ -\ 1\ +\ 1\ +\ 0 = n + 1$$

- For n odd, $n = 2 \cdot k + 1$ with $k \geq 0$

  Initially we observe that for $n = 1$ we have just two permutations with distance 1 between them, so $D(1) = 1$.

  For $n = 3$ we have from a theorem of Kececioglu and Sankoff [7] that the greedy algorithm sorts any permutation π with at least one negative element in at most $n - 1$ steps. Then, only the permutations with all elements positive are candidates to have $D(n) = n + 1$. Using this fact and constructing the breakpoint graphs for all possible sequences with all their elements positive

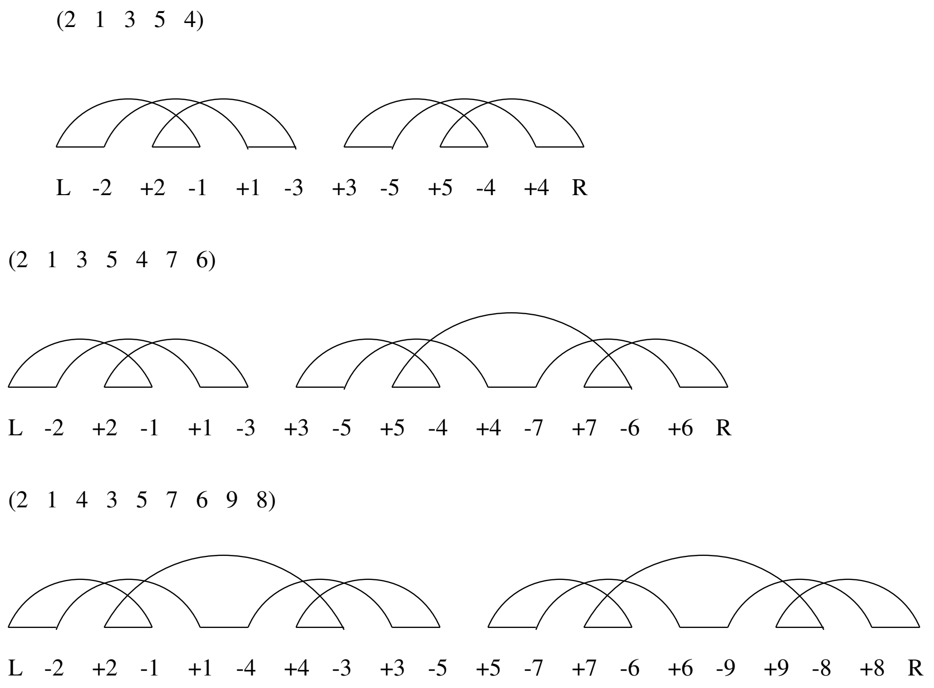

Figure 9: The breakpoint graphs for $n = 5, 7$ and 9 with respect to ι_n.

for $n = 3$, we conclude that $d(\pi_3, \iota_3) \leq 3$. On the other hand, $\pi_3 = (3\ 2\ 1)$ and $\iota_3 = (1\ 2\ 3)$ give $d(\pi_3, \iota_3) = 3$, so $D(3) = 3$.

Now we will present sequences π_n such that $d(\pi_n, \iota_n) = n + 1$ for the other cases of n odd, $n = 2 \cdot k + 1$ with $k \geq 2$, that is, $n \geq 5$. We consider now the *remainder* between $n + 1$ and 3. We have three cases:

- *remainder* $= 0$: Consider the sequence

$$(2\ 1\ 3\ 5\ 4\ 6\ 8\ 7\ 9\ \cdots\ n-6\ n-7\ n-5\ n-3\ n-4\ n-2\ n\ n-1)$$

  The breakpoint graph for π_n with respect to ι_n with $n \geq 5$ is formed by exactly $(n+1)/3$ cycles of size 3, with $n \geq 5$, constructed one beside the other. These are bad cycles and therefore hurdles. Figure 9 shows an example of a breakpoint graph for $n = 5$ with respect to ι_5.

  In this case, using Hannenhalli and Pevzner formula [4], and by construction of the breakpoint graph $G(\pi_n, \sigma_n)$ of π_n with respect to ι_n,

$$d(\pi_n, \iota_n) = (n+1)\ -\ (n+1)/3\ +\ (n+1)/3\ +\ 0 = n + 1$$

- *remainder* $= 1$: Consider the sequence

$$(2\ 1\ 3\ 5\ 4\ 6\ \cdots\ n-15\ n-13\ n-14\ n-12\ n-10\ n-11$$

$$n-9\ n-7\ n-8\ n-5\ n-6\ n-4\ n-2\ n-3\ n\ n-1)$$

  The breakpoint graph for π_n with respect to ι_n with $n \geq 9$ is formed by exactly $(n-9)/3$ cycles of size 3, and 2 cycles of size 5, with $n \geq 9$, constructed one beside the other. These are bad cycles, and therefore hurdles. We note that the restriction $n \geq 9$ does not eliminate any n such that $(n+1) \bmod 3 = 1$, because $n = 9$ is the first odd number satisfying this condition. Figure 9 shows an example of a breakpoint graph for $n = 9$ with respect to ι_9.

In this case, using the Hannenhalli and Pevzner formula [4], and by construction of the breakpoint graph $G(\pi_n, \sigma_n)$ of π_n with respect to ι_n,

$$d(\pi_n, \iota_n) = (n+1) \ - \ ((n-9)/3+2) \ + \ ((n-9)/3+2) \ + \ 0 = n+1$$

– *remainder* $= 2$: Consider the sequence

$$(2\ 1\ 3\ 5\ 4\ 6\ \cdots\ n-11\ n-12\ n-10\ n-8\ n-9\ n-7\ n-5\ n-6$$

$$n-4\ n-2\ n-3\ n\ n-1)$$

The breakpoint graph for π_n with respect to ι_n with $n \geq 5$ is formed by exactly $(n-4)/3$ cycles of size 3, and 1 cycle of size 5, with $n \geq 5$, constructed one beside the other. These are bad cycles, and so hurdles. We note that the restriction $n \geq 5$ does not eliminate any n such that $(n+1) \bmod 3 = 2$, because $n = 7$ is the first odd number satisfying this condition. Figure 9 shows an example of a breakpoint graph for $n = 7$ with respect to ι_7.

In this case, using the Hannenhalli and Pevner formula [4], and by construction of the breakpoint graph $G(\pi_n, \sigma_n)$ of π_n with respect to ι_n,

$$d(\pi_n, \iota_n) = (n+1) \ - \ ((n-4)/3+1) \ + \ ((n-4)/3+1) \ + \ 0 = n+1$$

Then, we proved that $D(n) \geq n+1$. We yet have to prove that $D(n) < n+2$, to obtain the wanted result.

We have, by the Hannenhalli and Pevner formula [4],

$$d(\pi_n, \iota_n) = (n+1) - c(\pi_n, \iota_n) + h(\pi_n, \iota_n) + f(\pi_n, \iota_n)$$

First, we have $h(\pi_n, \iota_n) \leq c(\pi_n, \iota_n)$, by definition of $h(\pi_n, \iota_n)$. So, if $h(\pi_n, \iota_n) = c(\pi_n, \iota_n)$, then we have $d(\pi_n, \iota_n) \leq (n+1)+1$, that is, $d(\pi_n, \iota_n) \leq n+2$. But if $f(\pi_n, \iota_n) = 1$, then necessarily $h(\pi_n, \iota_n) < c(\pi_n, \iota_n)$, and then $d(\pi_n, \iota_n) < n+2$.

This proves the linear case.

□

From the bijections defined earlier, we have the following result.

Lemma 4.1

$$D^c(n) = D(n-1)$$

From this lemma, we have the following theorem showing the circular reversal diameter of the equivalence classes on S_n.

Theorem 4.2 *The reversal diameter of circular chromosomes is*

$$D^c(n) = \max_{\substack{A \in S_n^c \\ B \in S_n^c}} \{d^c(A,B)\} = \begin{cases} n-1 & \textit{if } n=1,\ n=2 \textit{ or } n=4 \\ n & \textit{otherwise} \end{cases}$$

5 Conclusions

In this work, we attempted to start a systematic study of the theory of the reversal distance problem for signed circular chromosomes. To do this, we gave some contributions, described as follows. First we formalized circular chromosomes by equivalence classes. This is interesting because it includes the different forms to visualize a signed circular chromosome, obtained by rotations and reflections. We also defined circular reversals using the known definitions of linear reversals, which allowed to solve the reversal distance problem of signed circular chromosomes by using polinomial algorithms that solve the reversal distance problem of signed linear chromosome, giving as input suitable sequences from the equivalence classes. Besides, we presented some results concerning the linear and circular chromosomes of the same size. Finally, we determined the signed reversal diameter for linear ($D(n) = n + 1$) and circular chromosomes ($D^c(n) = n$), correcting a result of Kececioglu and Sankoff [7] on the linear reversal diameter $D(n)$.

To finish, a question arising from these studies is which permutations from the equivalence classes lead to a minimum reversal distance, that is, we would like to know how to characterize precisely the set of optimal representatives.

References

[1] M. Aigner and D. B. West. Sorting by insertion of leading element. *Journal of Combinatorial Theory*, 45:306–309, 1987.

[2] V. Bafna and P. Pevzner. Genome rearrangements and sorting by reversals. In *34th Annual IEEE Symposium on Foundations of Computer Science*, pages 148–157, 1993. Extended version with the same title appeared on SIAM Journal of Computing, 25(2):272-289,1996.

[3] P. Berman and S. Hannenhalli. Fast sorting by reversals. In *Proceedings of Combinatorial Pattern Matching - CPM'96*, 1996.

[4] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 178–189, 1995.

[5] H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. *SODA'97*, 1997.

[6] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. In *Combinatorial Pattern Matching, Proc. 4th Annual Symposium (CPM'93) de Lecture Notes in Computer Science*, volume 684, pages 87–105, Berlin, 1993. Springer-Verlag. Extended version with the title: "Exact and approximation algorithms for the inversion distance between two permutations" appeared on Algorithmica, 13: 180-210, 1995.

[7] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. *Lecture Notes in Computer Science*, 807:307–325, 1994.

[8] J.D. Palmer, B. Osorio, and W.F. Thompson. Evolutionary significance of inversions in legume chloroplast dnas. *Current Genetics*, 14:65–74, 1988.

[9] J. C. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing Co, 1997.

[10] G. A. Watterson, W. J. Ewens, and T. E. Hall. The chromosome inversion problem. *J. Theor. Biol.*, 99:1–7, 1982.