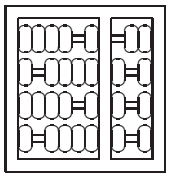

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Sorting by Block-Interchanges and Signed Reversals

Cleber V. G. Mira *João Meidanis*

Technical Report - IC-06-001 - Relatório Técnico

January - 2006 - Janeiro

Sorting by Block-Interchanges and Signed Reversals

Cleber Mira* João Meidanis†

Abstract

A *block-interchange* is a rearrangement event that exchanges two, not necessarily consecutive, contiguous regions in a genome, maintaining the original orientation. *Signed reversals* are events that invert and change the orientation of a region in a genome. Both events are important for the comparative analysis of genomes. For this reason, we propose a new measure that consists in finding a minimum sequence of block-interchanges and signed reversals that transforms a genome into another. For each event, we assign a weight related to its *norm* and we argue the adequacy of this parameter to indicate the power of each event.

We present a formula for the rearrangement measure and a polynomial time sorting algorithm for finding a sequence of block-interchanges and signed reversals that transforms a unichromosomal genome into another.

1 Introduction

Analyzing genome rearrangements by signed reversals is a well-known problem which was investigated in several works [8, 10, 2, 18]. The first polynomial-time algorithm for sorting by signed reversals was presented by Hannenhalli and Pevzner [8]. Several improvements were suggested to lower the algorithm's running-time [10, 2] until Tannier and Sagot [18] present a sub-quadratic algorithm based on data structures designed by Kaplan and Verbin [11]. Finding signed reversal distance can be done in linear time by using an algorithm from Bader et al [1].

Sorting by block-interchanges was proposed and solved by Christie [4]. Lin et al [12] presented a new solution to the problem by using the algebraic formalism developed by Meidanis and Dias [14].

We propose a new measure for the comparison of genomes based on both signed reversals and block-interchanges. A *block-interchange* is a rearrangement event that exchanges two, not necessarily consecutive, contiguous regions in a genome, maintaining the original orientation. *Signed reversals* are events that invert and change the orientation of a region in a genome. These rearrangement events, remarkably signed reversals [8], have been shown to be important in comparative analysis of genomes. Genome rearrangement measures involving several rearrangement events have been proposed earlier [5, 7], however it is not

*Institute of Computing, University of Campinas, 13081-970 Campinas, SP. Research supported by FAPESP, grant #03/00731-3

†Scylla Bioinformatics

clear which set of rearrangement events is the most appropriate biologically [19]. Another common matter that arises when one deals with a set of rearrangement events is how to assign a weight to each event in order to reflect its relative frequency in a parsimonious, evolutionary scenario. We propose a new parameter for the weight of a rearrangement event based on the *norm* of its representation as permutation in the algebraic formalism. The norm is a formal and systematic parameter applicable to any rearrangement event that can be represented as a permutation. Moreover, this parameter is in accordance with previous weight assignments [3]. On the other hand, a drawback in using the norm as a parameter is that it does not make distinction between some rearrangement events such as block-interchanges and transreversals.

The paper is organized as follows. In Section 2 we present a brief summary of the main concepts of the algebraic formalism and define genomes, signed reversals, and block-interchanges as permutations. In Section 3 the rearrangement measure based on signed reversals and block-interchanges is formally defined. In Section 4 we show a polynomial time algorithm for a minimum rearrangement event sequence that transforms a genome into another and we present a formula for the rearrangement measure, which can be quickly computed. The algorithm is based on results from the algebraic formalism proposed by Meidanis and Dias [14] and the analysis of the sorting by block-interchanges problem [12, 4]. We summarize the results in Section 5.

2 Algebraic Formalism for Block-Interchanges and Signed Reversals

A *permutation* is a bijective mapping from a set into itself. Given a permutation π over a set E, an element $x \in E$ is *fixed* by π when $\pi(x) = x$. In the sequel we will drop parentheses and represent $\pi(x)$ simply by πx. This shall not cause confusion since we use Greek letters for permutations and Roman letters for elements of E. The set of non-fixed elements in π is the *support* of π; i.e. $Supp(\pi) = \{x \in E \mid \pi x \neq x\}$. The *identity permutation* ι is the permutation such that $\iota x = x$ for all $x \in E$. The *orbit* of $x \in E$ under the permutation π, denoted by $orb(\pi, x)$, is the set $\{y \mid y = \pi^k x$ for an integer $k\}$. Denote by $Orb(\pi, E)$ the set of orbits of a permutation π over E. An orbit is called *nontrivial* when it has more than one element. Let $o(\pi, E)$ be the number of orbits in permutation π. A *cycle* is a permutation α over E such that it has at most one nontrivial orbit. A cycle α is an *r-cycle* if its nontrivial orbit contains r elements or an 1-*cycle* when $\alpha = \iota$. Two cycles are *disjoint* when their corresponding nontrivial orbits are disjoint, or when at least one of them is the identity. Any permutation can be represented uniquely as a product of disjoint cycles [6, 13]. A *k-cycle decomposition* of a permutation π is a representation of π as a product of k-cycles, not necessarily disjoint. The *norm* of π, denoted by $\|\pi\|$, is the minimum number of 2-cycles whose product is π. A permutation α *divides* a permutation β, denoted by $\alpha|\beta$, when $\|\beta\alpha^{-1}\| = \|\beta\| - \|\alpha\|$.

2.1 Block-Interchanges and Signed Reversals

Genomes and rearrangement events can be modeled as permutations in the algebraic formalism [14]. A DNA chromosome has two strands with complementary orientation to each other. Let E_+ be the set of blocks of genes (or other markers) in one of the strands of the chromosome. In the same way, we define E_- as the set of blocks of genes in the strand complementary to the previous one. Let $E = E_+ \cup E_-$. We define the permutation Γ as the function that associates each block of genes to its complementary block in E. So $E_- = \{\Gamma x \mid x \in E_+\}$.

Given the function Γ over E, a cycle α is called a *strand* when $x \in Supp(\alpha)$ if and only if $\Gamma x \notin Supp(\alpha)$, for each $x \in E$. The *conjugation* of a permutation α by β, denoted by $\beta \cdot \alpha$, is $\beta\alpha\beta^{-1}$ [14]. A *chromosome* is a product of two strands α and $\Gamma \cdot \alpha^{-1}$. Two chromosomes are *disjoint* when their supports are disjoint. A *genome* is a product of disjoint chromosomes. A fundamental property of genomes is: if π is a genome, then $\Gamma\pi\Gamma = \pi^{-1}$. In this text, we restrict our analysis to unichromosomal genomes, that is, genomes composed by a single chromosome. In addition, we deal with circular genomes instead of linear genomes since circular genomes are more naturally modeled as permutations. There are several works describing how to transform a kind of genome into another [9, 15].

Given a genome π and Γ, both over E, we define the following rearrangement events:

Definition 2.1 *1. A* block-interchange *is a rearrangement event ρ composed by the product of four 2-cycles*

$$(u\ x)(\pi\Gamma x\ \pi\Gamma u)(v\ y)(\pi\Gamma y\ \pi\Gamma v);$$

such that

(a) $u \neq x$,$u \neq y$ *and* $v \neq y$,

(b) $v, x, y \in orb(\pi, u)$,

(c) $(u\ v)(x\ y)|\pi$;

in this case we say that ρ is applicable *to π.*

2. A signed reversal *is a rearrangement event ρ composed by the product of two 2-cycles*

$$(u\ \pi\Gamma v)(v\ \pi\Gamma u);$$

such that $(u\ v)|\pi$ and $u \neq v$; in this case we say that ρ is applicable *to π.*

Given a rearrangement event ρ, its *weight*, denoted by $w(\rho)$, is $\|\rho\|/2$. The adoption of the norm of a genome rearrangement event as its weight, which is particularly important when evaluating genome rearrangement problems involving several, distinct rearrangement events, is supported by the similarity in weights assigned to rearrangement events in other works [3]. Particularly, block-interchanges and transpositions seem to be less frequent than signed reversals [16, 17]. To account for that, they are assigned the double of the weight of a signed reversal by the norm rule.

A cycle α is said to be *a cycle of* a permutation θ when α is one of the cycles in the unique disjoint cycle decomposition of θ. Given genomes π, σ, and function Γ, all over

E, a *pair* is a couple of cycles α and $(\pi\Gamma)\cdot\alpha^{-1}$ of $\sigma\pi^{-1}$. Let $c(\pi,\sigma)$ be the number of pairs of $\sigma\pi^{-1}$. The number of pairs of $\sigma\pi^{-1}$ is $c(\pi,\sigma) = (|E| - \|\sigma\pi^{-1}\|)/2$. We denote $c(\rho\pi,\sigma) - c(\pi,\sigma)$ by $\Delta c(\rho,\pi,\sigma)$ where ρ is a rearrangement event applicable to π. If ρ is a signed reversal applicable to π, then we have $\Delta c(\rho,\pi,\sigma) \in \{-1,0,1\}$ [8].

Given the genomes π and σ over E, they are called *equiorbital genomes* when their orbits are equal, that is, when $Orb(\pi,E) = Orb(\sigma,E)$. Given equiorbital genomes π and σ, and the function Γ, all over E, if ρ is a block-interchange applicable to π, then we have $\Delta c(\rho,\pi,\sigma) \in \{-2,0,2\}$ (Christie [4] presents a proof for the case when $\sigma = (1\ 2\ \ldots\ n)(-n\ \ldots\ -2\ -1)$ over the set $E = \{-n,\ldots,-1,1,\ldots,n\}$, function Γ is $(1\ -1)(2\ -2)\ldots(n\ -n)$, and π is a genome over E such that π and σ are equiorbital genomes, but the same proof can be easily extended to the general case.)

3 Genome Rearrangement Measure

The *algebraic rearrangement by block-interchanges and signed reversals problem* consists in finding a sequence with the minimum weight of block-interchanges and signed reversals to transform one circular genome with signals into another. In other words, we want to find a sequence of rearrangement events $\rho_1\ \rho_2\ \ldots\ \rho_k$, such that:

$$\sigma = \rho_k\ \rho_{k-1}\ \ldots\ \rho_1\ \pi$$

where each rearrangement event ρ_i is a block-interchange or signed reversal, and ρ_{i+1} is applicable to $\rho_i\ \rho_{i-1}\ \ldots\ \rho_1\ \pi$, and $\sum_{i=1}^{k} w(\rho_i)$ is minimum. We call this minimum $W(\pi,\sigma)$.

The *sorting by block-interchanges and signed reversals problem* differs from the previous problem just by assuming $\sigma = (1\ 2\ \ldots\ n)(-n\ \ldots\ -2\ -1)$.

We propose the parameter $W(\pi,\sigma)$ as a new measure in the comparison of genomes.

A *good event* for the pair (π,σ) is a block-interchange or a signed reversal ρ such that $\Delta c(\rho,\pi,\sigma) = w(\rho)$.

Lemma 3.1 *Given genomes π, σ, and the function Γ, all over E, a rearrangement event ρ applicable to π is a good event for the pair (π,σ) if and only if $\rho|\sigma\pi^{-1}$.*

Proof: If $\rho|\sigma\pi^{-1}$ then $\|\sigma\pi^{-1}\rho^{-1}\| = \|\sigma\pi^{-1}\| - \|\rho\|$. Manipulating the later formula:

$$\frac{\|\rho\|}{2} = \frac{\|\sigma\pi^{-1}\| - \|\sigma\pi^{-1}\rho^{-1}\|}{2} = \frac{|E| - \|\sigma\pi^{-1}\rho^{-1}\| - |E| + \|\sigma\pi^{-1}\|}{2} = c(\rho\pi,\sigma) - c(\pi,\sigma)$$

and since $c(\rho\pi,\sigma) - c(\pi,\sigma) = \Delta c(\rho,\pi,\sigma)$ then $\Delta c(\rho,\pi,\sigma) = w(\rho)$. Therefore ρ is a good event for the pair (π,σ).

Conversely, if ρ is a good event for the pair (π,σ) then $\Delta c(\rho,\pi,\sigma) = w(\rho)$, that is, we have $c(\rho\pi,\sigma) - c(\pi,\sigma) = \|\rho\|/2$. By definition of $c(\pi,\sigma)$ we have

$$\frac{\|\sigma\pi^{-1}\| - \|\sigma\pi^{-1}\rho^{-1}\|}{2} = \frac{\|\rho\|}{2}.$$

Therefore $\|\sigma\pi^{-1}\rho^{-1}\| = \|\sigma\pi^{-1}\| - \|\rho\|$ and hence $\rho|\sigma\pi^{-1}$. □

Lemma 3.2 *Given genomes π, σ, and the function Γ, all over E, for any sequence of rearrangement events $\rho_1, \ldots, \rho_k$, such that $\rho_k \ldots \rho_1 \pi = \sigma$ and ρ_i is applicable to the genome $\rho_{i-1} \ldots \rho_1 \pi$, we have:*

1. $\sum_{j=1}^{k} w(\rho_j) \geq \frac{\|\sigma\pi^{-1}\|}{2}$;
2. $\sum_{j=1}^{k} w(\rho_j) = \frac{\|\sigma\pi^{-1}\|}{2}$ *if and only if each rearrangement event ρ_i is a good event for the pair $(\rho_{i-1} \ldots \rho_1 \pi, \sigma)$ for $1 \leq i \leq k$.*

Proof:

1. Let $\rho_1, \ldots, \rho_k$ be a sequence of rearrangement events such that $\rho_k \ldots \rho_1 \pi = \sigma$ and ρ_i is applicable to $\rho_{i-1} \ldots \rho_1 \pi$ for $1 \leq i \leq k$. Therefore $\rho_k \ldots \rho_1 = \sigma\pi^{-1}$, and we get the following upper bound for $\|\sigma\pi^{-1}\|$.

$$\begin{aligned} \|\sigma\pi^{-1}\| &= \|\rho_k \ldots \rho_1\| \\ &\leq \|\rho_k\| + \ldots + \|\rho_1\| \\ &= 2\sum_{j=1}^{k} \frac{\|\rho_j\|}{2} \\ &= 2\sum_{j=1}^{k} w(\rho_j) \end{aligned}$$

   Therefore we have $\sum_{j=1}^{k} w(\rho_j) \geq \frac{\|\sigma\pi^{-1}\|}{2}$

2. Firstly, we prove the "if" part, that is, we assume that each rearrangement event ρ_i is a good event for the pair $(\rho_{i-1} \ldots \rho_1 \pi, \sigma)$ for $1 \leq i \leq k$. By definition of weight, we have:

$$\sum_{j=1}^{k} w(\rho_j) = \frac{\|\rho_1\| + \ldots + \|\rho_k\|}{2}. \tag{1}$$

   Since rearrangement event ρ_i is a good event for the pair $(\rho_{i-1} \ldots \rho_1 \pi, \sigma)$ for $1 \leq i \leq k$ then

$$\frac{\|\rho_i\|}{2} = c(\rho_i \ldots \rho_1 \pi, \sigma) - c(\rho_{i-1} \ldots \rho_1 \pi, \sigma).$$

   By definition of number of pairs $c(,)$ we have

$$\frac{\|\rho_i\|}{2} = \frac{\|\sigma\pi^{-1}\rho_1^{-1} \ldots \rho_{i-1}^{-1}\|}{2} + \frac{\|\sigma\pi^{-1}\rho_1^{-1} \ldots \rho_i^{-1}\|}{2}. \tag{2}$$

   Using Equation 1 and Equation 2 and some manipulation:

$$\sum_{j=1}^{k} w(\rho_j) = \frac{\|\sigma\pi^{-1}\| - \|\sigma\pi^{-1}\rho_1^{-1} \ldots \rho_k^{-1}\|}{2}$$

$$\begin{aligned} &= \frac{\|\sigma\pi^{-1}\| - \|\sigma\sigma^{-1}\|}{2} \\ &= \frac{\|\sigma\pi^{-1}\|}{2}. \end{aligned}$$

Therefore $\sum_{j=1}^{k} w(\rho_j) = \frac{\|\sigma\pi^{-1}\|}{2}$.

On the other hand, if $\sum_{j=1}^{k} w(\rho_j) = \|\sigma\pi^{-1}\|/2$ then $\|\sigma\pi^{-1}\| = \sum_{j=1}^{k} \|\rho_j\|$.

By the triangular inequality property, we have

$$\|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| \leq \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| + \|\rho_i\|,$$

for $1 \leq i \leq k$, in other words we have

$$0 \leq \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| - \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| + \|\rho_i\|,$$

for $1 \leq i \leq k$. Since each term $\|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| - \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| + \|\rho_i\|$ is nonnegative and manipulating the expanded sum we get

$$\begin{aligned} 0 &\leq \sum_{i=1}^{k} \Big(\|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| - \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| + \|\rho_i\| \Big) \\ &= \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_k^{-1}\| - \|\sigma\pi^{-1}\| + \sum_{i=1}^{k} \|\rho_i\|. \end{aligned}$$

But since $\sigma\pi^{-1}\rho_1^{-1} \dots \rho_k^{-1} = \iota$ and $\sum_{i=1}^{k} \|\rho_i\| = \|\sigma\pi^{-1}\|$ then

$$\sum_{i=1}^{k} \Big(\|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| - \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| + \|\rho_i\| \Big) = 0.$$

Then $\|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| - \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| + \|\rho_i\| = 0$ for $1 \leq i \leq k$; i.e. we have

$$\|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_i^{-1}\| = \|\sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}\| - \|\rho_i\|$$

for $1 \leq i \leq k$. Therefore, we have $\rho_i | \sigma\pi^{-1}\rho_1^{-1} \dots \rho_{i-1}^{-1}$ for $1 \leq i \leq k$ and by Lemma 3.1 each ρ_i applicable to $\rho_{i-1} \dots \rho_1\pi$ is a good event for the pair $(\rho_{i-1} \dots \rho_1\pi, \sigma)$.

□

Lemma 3.3 *Given distinct, equiorbital genomes π, σ, and the function Γ, all over E, there is a block-interchange ρ applicable to the genome π such that ρ is a good event for the pair (π, σ) and $\rho\pi$ and σ are equiorbital genomes.*

Proof: Since genomes π and σ are distinct, consider the elements x, y, $\sigma\pi^{-1}x$ and $\sigma\pi^{-1}y$ where $\sigma\pi^{-1}x \neq x$, $\sigma\pi^{-1}x \neq \sigma\pi^{-1}y$, $x \neq y$, and $x \neq \sigma\pi^{-1}y$, such that:

1. $y \in orb(\pi, x)$;
2. $\pi\sigma\pi^{-1}x \neq x$;
3. $\sigma\pi^{-1}y \neq y$ and $\pi\sigma\pi^{-1}y \neq y$;
4. $(y\ \sigma\pi^{-1}y)(x\ \sigma\pi^{-1}x) \nmid \pi$.

In Section 3 of Lin et al [12] there is a demonstration that elements x and y exist for any pair of distinct, equiorbital genomes.

Since $\pi\sigma\pi^{-1}x \neq x$, $\pi\sigma\pi^{-1}y \neq y$, and $(x\ \sigma\pi^{-1}x)(y\ \sigma\pi^{-1}y) \nmid \pi$ then permutation $(x\ \sigma\pi^{-1}y)(y\ \sigma\pi^{-1}x)$ divides π by the rules of product by 2-cycles [14]. Therefore permutation

$$\rho = (x\ \sigma\pi^{-1}x)(\pi\Gamma\sigma\pi^{-1}x\ \pi\Gamma x)(y\ \sigma\pi^{-1}y)(\pi\Gamma\sigma\pi^{-1}y\ \pi\Gamma y)$$

is a block-interchange applicable to π since $(x\ \sigma\pi^{-1}y)(y\ \sigma\pi^{-1}x)|\pi$. In addition, the block-interchange ρ is a good event for the pair (π, σ) because $\rho|\sigma\pi^{-1}$ by the choice of w, z, a, b. Since ρ is a block-interchange then $\rho\pi$ and σ are equiorbital because block-interchanges do not change the elements in each strand and, consequently, in each orbit of π. Therefore ρ is a block-interchange applicable to the genome π such that ρ is a good event for the pair (π, σ) and $\rho\pi$ and σ are equiorbital genomes.

□

Lemma 3.4 *Given genomes π, σ that are not equiorbital, and the function Γ, all over E, there is a signed reversal ρ applicable to π such that ρ is a good event for the pair (π, σ).*

Proof: Since genomes π and σ are not equiorbital then there exists an element $x \in E$ such that $\sigma\pi^{-1}x \notin orb(\pi, x)$. We are going to show that the signed reversal $\rho = (x\ \sigma\pi^{-1}x)(\pi\Gamma\sigma\pi^{-1}x\ \pi\Gamma x)$ is applicable to π and it is a good event for the pair (π, σ). Elements x and $\sigma\pi^{-1}x$ belong to distinct strands of the unichromosomal genome π and we have $\Gamma\pi^{-1}\sigma\pi^{-1}x = \pi\Gamma\sigma\pi^{-1}x$ by the fundamental property $\Gamma\pi\Gamma = \pi^{-1}$, so $\pi\Gamma\sigma\pi^{-1}x \in orb(\pi, x)$ and therefore $(x\ \pi\Gamma\sigma\pi^{-1}x)|\pi$. In addition, we have $x \neq \pi\Gamma\sigma\pi^{-1}x$ because otherwise $x = \pi\Gamma\sigma\pi^{-1}x$ implies $\pi\Gamma x = \sigma\Gamma\pi\Gamma x$ and then there is an element z such that $\sigma\Gamma z = z$, which contradicts the definition of a genome and the function Γ. Therefore, reversal $(x\ \sigma\pi^{-1}x)(\pi\Gamma\sigma\pi^{-1}x\ \pi\Gamma x)$ is applicable to π. Moreover, since $\sigma\pi^{-1}$ is a product of companion cycles $\alpha\pi\Gamma\alpha^{-1}\pi\Gamma$ then for any cycle $(a_1 \ldots a_m)$ of $\sigma\pi^{-1}$ there is a cycle $(\pi\Gamma a_m \ldots \pi\Gamma a_1)$ of $\sigma\pi^{-1}$, and therefore $\pi\Gamma x \notin orb(\sigma\pi^{-1}, x)$. Finally, by the choice of x and since $\sigma\pi^{-1}x \in orb(\sigma\pi^{-1}, x)$, $\pi\Gamma\sigma\pi^{-1}x \in orb(\sigma\pi^{-1}, \pi\Gamma x)$, and $\pi\Gamma x \notin orb(\sigma\pi^{-1}, x)$ then $\rho|\sigma\pi^{-1}$, and by Lemma 3.1 reversal ρ is a good event for the pair (π, σ). □

4 Algorithm

In this section we present an algorithm for finding a sequence of good events and the weight, which takes $O(n^2)$ running time.

```
Algorithm SRBISort
 1.        r = 0;
 2.        θ = π;
 3.        W = 0;
 4.          while (θ ≠ σ) do {
 5.                r++;
 6.                if (Orb(θ, E) = Orb(σ, E) )
 7.                    find a permutation ρ_r such that
 8.                    ρ_r is a block-interchange
 9.                    ρ_r applicable to θ
10.                    ρ_r is a good event for the pair (θ, σ).
11.                else
12.                    find a permutation ρ_r such that
13.                    ρ_r is a signed reversal
14.                    ρ_r applicable to θ
15.                    ρ_r is a good event for the pair (θ, σ).
16.                θ = ρ_r θ;
17.                W = W + w(ρ_r);
18.          }
19.       return ρ_1, ..., ρ_r and W;
```

Lemma 4.1 *Given π, σ, and Γ over E, SRBISort algorithm presents a sequence of good events for the pair (π, σ) with minimum weight $W(\pi, \sigma)$ that transforms genome π into σ in $O(n^2)$ running time, where $n = |E|/2$.*

Proof: We show that the algorithm SRBISort is correct by defining the following loop invariant over the parameter r: $\theta = \rho_r \dots \rho_1 \pi$ and rearrangement event ρ_i is a good event for the pair $(\rho_{i-1} \dots \rho_1 \pi, \sigma)$ applicable to $\rho_{i-1} \dots \rho_1 \pi$, for $1 \leq i \leq r$.

For $r = 0$, before the main loop in the line 4, we have $\theta = \pi$ and the invariant is trivially valid.

Suppose that the invariant is valid for $r = k$, that is, we have $\theta = \rho_k \dots \rho_1 \pi$ and ρ_i is a good event for the pair $(\rho_{i-1} \dots \rho_1 \pi, \sigma)$ applicable to $\rho_{i-1} \dots \rho_1 \pi$ for $1 \leq i \leq k$. In the next iteration of the loop in line 6 we have two cases: θ and σ are either equiorbital or not. In both cases there is a good event for the pair (θ, σ) by Lemma 3.3 and Lemma 3.4. Therefore the invariant remains valid before the next iteration of the loop in line 4.

If $\theta = \sigma$ then the condition in the line 4 is false and the algorithm executes the code in the line 19. At this point in the execution then we have $\theta = \rho_r \dots \rho_1 \pi$ such that each ρ_i is a good event for the pair $(\rho_{i-1} \dots \rho_1 \pi, \sigma)$ applicable to $\rho_{i-1} \dots \rho_1 \pi$, for $1 \leq i \leq r$ since the loop invariant is valid. Therefore $\sigma = \rho_r \dots \rho_1 \pi$ and $\rho_1, \ \dots, \ \rho_r$ is a sequence of good events, such that ρ_i is a good event for the pair $(\rho_{i-1} \dots \rho_1 \pi, \sigma)$ and it is applicable to $\rho_{i-1} \dots \rho_1 \pi$, where $1 \leq i \leq r$, transforming genome π into genome σ.

In the worst case, just one element in each strand of the genome will be placed in its proper position per iteration of the loop in line 4, i.e. the block *while* will be executed $O(n)$ times. A sequence of signed reversals is an example of a worst case instance in the time complexity. For each step in the *while* loop it is verified whether genomes θ and σ are equiorbital in line 6. This verification can be accomplished in $O(n)$ running time by maintaining two labels for each element in the set E representing the strand the element belongs to in θ and σ, respectively. Genomes θ and σ are equiorbital when every element with the same value for the θ label (and those elements only) have the same value for σ label. Labels must be updated for the pair of genomes $\rho_r\theta$ and σ where ρ_r is a signed reversal (block-interchanges do not affect orbits). The update process can be achieved in $O(n)$ running time since a signed reversal may change the orientation, and conquently the label value of $n - 1$ elements in each strand of θ. Finding a block-interchange that is a good event for the pair (θ, σ) can be performed in $O(n)$ time by representing genomes using a simple array data structure and a rotation of elements as suggested by Lin et al [12]. Signed reversals that are good events can be found in $O(n)$ by using the same data structure employed on verifying whether two genomes are equiorbital. Given an element $x \in E$ one must confirm if x and $\sigma\theta^{-1}x$ belong to distinct strands of θ; and, in this case, the signed reversal $\rho_r = (x \ \sigma\theta^{-1}x)(\theta\Gamma\sigma\theta^{-1}x \ \theta\Gamma x)$ is a good event. Since there are n pairs x and $\sigma\theta^{-1}x$, then such verification takes $O(n)$ running time. Therefore the total time complexity is quadratic in n. □

Theorem 4.2 *Given genomes π, σ, and the function Γ, all over E, we have*

$$W(\pi, \sigma) = \frac{\|\sigma\pi^{-1}\|}{2}.$$

Proof: Given genomes π and σ over E, Lemma 4.1 guarantees the existence of a sequence of rearrangement events $\rho_1, \dots, \rho_k$ such that $\rho_k \dots \rho_1 \pi = \sigma$ and ρ_i is applicable to $\rho_{i-1} \dots \rho_1 \pi$ and ρ_i is a good event for $(\rho_{i-1} \dots \rho_1 \pi, \sigma)$ for $1 \leq i \leq k$. In addition, by Lemma 3.2, we have $W(\pi, \sigma) \geq \|\sigma\pi^{-1}\|/2$ and $\sum_{i=1}^{k} w(\rho_i) = \|\sigma\pi^{-1}\|/2$. Therefore $W(\pi, \sigma) = \|\sigma\pi^{-1}\|/2$. □

Given the genomes π and σ over E, Theorem 4.2 offers a simple formula to the measure $W(\pi, \sigma)$. The measure $W(\pi, \sigma)$ can be obtained in $O(n)$, which is the time complexity for compute $\sigma\pi^{-1}$ and finding its norm, although finding a sequence of good events that transforms the genome π into σ is quadratic.

5 Conclusion

Genome Rearrangement analysis involving signed reversals and block-interchanges may be an important technique for unichromosomal genome comparison. We showed that a measure based on *genome rearrangement by signed reversals and block-interchanges* can be properly achieved through the algebraic formalism. In addition, we presented a simple algorithm to find a minimum sequence of signed reversals and block-interchanges that transforms a genome into another.

Dealing with multichromosomal genomes and recombination events (e.g. translocations) is a promising direction for future work. This approach was already explored in the work of Yancopoulos et al [19], although we believe that a further simplification is possible by using the algebraic formalism.

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(5):483–491, 2001.

[2] A. Bergeron. A very elementary presentation of the hannenhalli-pevzner theory. *Discrete Applied Mathematics*, 146(2):134–135, 2005.

[3] M. Blanchette, T. Kunisawa, and D. Sankoff. Parametric genome rearrangement. *Journal of Computational Biology*, 172:11–17, 1996.

[4] D. A. Christie. Sorting permutations by block-interchanges. *Information Processing Letters*, 60(4):165–169, November 1996.

[5] Z. Dias and J. Meidanis. Genome rearrangements distance by fusion, fission, and transposition is easy. In *Proceedings of the String Processing and Information Retrieval (SPIRE'2001)*, pages 250–253, Laguna de San Rafael, Chile, November 2001. IEEE Computer Society.

[6] I. Grossman and W. Magnus. *Groups and Their Graphs*. The Mathematical Association of America, 1992.

[7] S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS'95)*, pages 581–592, Los Alamitos, USA, Oct. 1995. IEEE Computer Society Press.

[8] S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1):1–27, Jan. 1999.

[9] T. Hartman and R. Shamir. A simpler and faster 1.5-approximation algorithm for sorting by transpositions. In *Proceedings of CPM'03*, pages 156 – 169, 2003. extended version.

[10] H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29(3):880–892, Jan. 2000.

[11] H. Kaplan and E. Verbin. Efficient data structures and a new randomized approach for sorting signed permutations by reversals. In *Combinatorial Pattern Matching, 14th Annual Symposium, CPM 2003, Morelia, Michocán, Mexico, June 25-27, 2003, Proceedings*, volume 2676 of *Lecture Notes in Computer Science*. Springer, 2003.

[12] Y. C. Lin, C. L. Lu, , H. Chang, and C. Y. Tang. An efficient algorithm for sorting by block-interchanges and its application to the evolution of vibrio species. *Journal of Computational Biology*, 12:102–112, 2005.

[13] S. MacLane and G. Birkhoff. *Algebra*. The Macmillan Company, London, sixth printing edition, 1971.

[14] J. Meidanis and Z. Dias. An alternative algebraic formalism for genome rearrangements. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics: Empirical and Analyitical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*, pages 213–223. Kluwer Academic Publishers, Nov. 2000.

[15] J. Meidanis, M. E. M. T. Walter, and Z. Dias. Reversal distance of signed circular chromosomes. Technical Report IC-00-23, Institute of Computing - University of Campinas, Dec. 2000.

[16] J. D. Palmer and L. A. Herbon. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, 27:87–97, 1988.

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

[18] E. Tannier and M.-F. Sagot. Sorting by reversals in subquadratic time. Technical Report 5097, INRIA, Institut National de Recherche en Informatique et en Automatique, January 2004.

[19] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, June 2005.