Capítulo 6

An Alternative Algebraic Formalism For Genome Rearrangements *

João Meidanis

Zanoni Dias

Abstract

Here we relate the recent theory of genome rearrangements to the theory of permutation groups in a new way and hope to set the ground for further advances in the area. This work was motivated by the fact that many arguments in genome rearrangements are of the form "look at the figure", and lack more formal algebraic derivation. We intend to give the area a strong algebraic formalism, much as analytic geometry provided an alternative for geometric arguments based on pictures.

* ***Este trabalho corresponde a versão estendida do artigo apresentado no* Gene Order Dynamics, Comparative Maps and Multigene Families *(DCAF'2000) realizado na cidade de Le Chantecler, no Canadá, em setembro de 2000, e que também integra o livro* Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families *publicado em novembro de 2000.***

6.1 Introduction

In this paper we are concerned with the genome rearrangement problem viewed as a combinatorial problem. In the general formulation of this problem we are given two genomes (or parts of genomes), viewed as ordered lists of genes (or others markers), and a set of allowed mutation events (reversals, transpositions, etc). To solve the problem we must find the minimum number of events that lead from one genome to another. In general the solution is symmetric, that is, the same series of events, taken backward, will transform the second genome into the first. We will also restrict ourselves to the case of conservative events, that is, events that do not change the available gene pool. Thus events such as duplications or deletions will not be considered in this study.

Recent developments in this field include the polynomial solution to the signed reversal case [64], the NP-hardness of unsigned reversal distance [21], and partial results for transposition distance [9, 92], to name just a few. Many doctoral dissertations were devoted to this theme (see, for instance, the theses of [117], [30], and [119]). Transposition distance seems to be a harder problem, that has eluded researchers for many years now. Its computational complexity is still unknown. We feel that new, more powerful formal tools are needed to successfully attack this problem.

The mathematical formalization of genome rearrangements usually begins by representing genomes as *permutations*. Thus, a genome π consisting of genes π_1, π_2, π_3, ..., π_n in this order is written as:

$$\pi = (\pi_1\ \pi_2\ \pi_3\ \dots\ \pi_n) \tag{6.1}$$

meaning that π is the function (permutations are functions):

$$[1 \Rightarrow \pi_1, 2 \Rightarrow \pi_2, 3 \Rightarrow \pi_3, \dots, n \Rightarrow \pi_n]$$

that is, π maps 1 into π_1, 2 into π_2, and so on.

In this paper we will propose a different view of a genome as a permutation, namely, that Equation [6.1] denotes the function:

$$[\pi_1 \Rightarrow \pi_2, \pi_2 \Rightarrow \pi_3, \dots, \pi_{n-1} \Rightarrow \pi_n, \pi_n \Rightarrow \pi_1] \tag{6.2}$$

that is, π maps π_1 into π_2, π_2 into π_3, and so on. Note that the last gene π_n is mapped into the first gene π_1. This is necessary, because permutations are functions that map each element into some other, and they cannot repeat images. However, this implies a circular character to our genome. But circular genomes do exist, and, as we will see in subsequent sections, the study of rearrangements of linear genomes is really not much different from circular ones.

Our goal in this note is to convince the reader that interpretation (6.2) is much more sensible, for a number of reasons. First, it allows us to directly apply many long known results from permutation group theory. Important tools such as breakpoints, the breakpoint graph, cycles, good cycles, bad cycles, gray edges, black edges, which served as basic building blocks for most of the advances in the field can be algebraically defined instead of graphically defined as they have been until now. Therefore, arguments that relied on pictures can now be expressed completely in algebraic terms. We consider this a powerful step towards a massive attack on such problems, much like analytic geometry is a powerful way of looking into geometric problems.

In Section 6.2 we briefly review the basics on permutation groups. Section 6.3 contains the first steps in redefining genome rearrangements under the new formalism that we propose. In the Section 6.4 we use theory just developed to show some results that have been proved based on pictures. In Section 6.5 we apply the theory to reversal distance problem. Finally, we conclude in Section 6.6.

6.2 Permutation Groups

Permutations groups have been studied at least since the eighteenth century, when Galois wrote his much acclaimed theory for solving algebraic equations. Here we briefly recall a few classical results that are useful in genome rearrangements. For more information see references [70, 87].

Given a base set E, a *permutation* on E is a one-to-one function from E onto itself. Permutations are composed of one or more *cycles*. A cycle involving elements a, b, c, for instance, is written:

$$(a\ b\ c)$$

meaning that a is mapped into b, which is mapped into c, which in turn is mapped back into a. Cycles can be of any length. Cycles of length 1 are not explicitly written. Thus, if we write:

$$\alpha = (a\ b\ c)$$

we implicitly mean that all others elements are left in place by α, that is, $\alpha(x) = x$ for $x \neq a, b, c$. Note: $(a\ b\ c) = (b\ c\ a) = (c\ a\ b)$.

The *support* of a permutation α, $Supp(\alpha)$, is the set of elements not fixed by α:

$$Supp(\alpha) = \{x \in E_n | \alpha(x) \neq x\}.$$

The *size* of a cycle is the number of elements in its support.

The *product* (or *composition*) of two permutations α, β is denoted by $\alpha\beta$. In general $\alpha\beta \neq \beta\alpha$, but when α and β are disjoint cycles they commute: $\alpha\beta = \beta\alpha$. Every permutation can be written in an unique way as a product of disjoint cycles (apart from the order of the factors). We refer to this as the *cycle decomposition* of a permutation.

The identity permutation, that maps every element into itself, will be denoted by 1. Every permutation α has an inverse α^{-1} such that $\alpha\alpha^{-1} = \alpha^{-1}\alpha = 1$. For cycles, the inverse is obtained reverting the order of the elements: $(a\ b\ c)$ is the inverse of $(c\ b\ a)$. For a general permutation, invert every cycle in its cycle decomposition.

To compute the product of α and β, $\alpha\beta$, we must keep in mind that β will be applied first, and then α, as in $\alpha\beta(x) = \alpha(\beta(x))$. Therefore, to compute a product of non-disjoint cycles we need to proceed as follows. Take the example:

$$(a\,b\,c)(a\,b\,d)(c\,d\,b).$$

To compute this, we start with any element, say a, and compute its image. The element a is fixed by the rightmost cycle, then is mapped into b by the second cycle, and b is mapped into c by the leftmost cycle. So, the final destination of a is c. We then write:

$$(a\,b\,c)(a\,b\,d)(c\,d\,b) = (a\;c\;\ldots$$

and then proceed finding out the image of c: c goes to d, d goes to a, a goes to b, respectively, by the rightmost, middle, and leftmost cycle, so c is finally mapped into b. And so on. We reach the result:

$$(a\,b\,c)(a\,b\,d)(c\,d\,b) = (a\,c\,b)(d) = (a\,c\,b)$$

since singleton cycles do not need to be explicitly indicated.

One important operation is the *conjugation*. The conjugation of β by α is the permutation $\alpha\beta\alpha^{-1}$. This results in a permutation with the same cycle structure of β but the elements are changed by α. For instance, if $\beta = (b_1\ b_2\ \ldots\ b_k)$ then:

$$\alpha\beta\alpha^{-1} = (\alpha(b_1)\ \alpha(b_2)\ \ldots\ \alpha(b_k))$$

If β is a product of disjoint cycles, each one will be affected by α in the same way to form $\alpha\beta\alpha^{-1}$. Conjugations are so important that we will have a special notation for them: $\alpha \cdot \beta$ means the same as $\alpha\beta\alpha^{-1}$.

6.2.1 Short Cycles

A 2-cycle is a cycle of size 2. A 3-cycle is a cycle of order 3. It is important to know how products by 2- or 3-cycles affect an arbitrary permutation.

Let $\alpha = (a\ b)$ be a 2-cycle. Its effect on an arbitrary permutation β can be described as follows. If a and b are in the same cycle in β, this cycle is broken in two in $\alpha\beta$. If a and b are in two distinct cycles in β, these two cycles become one in $\alpha\beta$. Here and in the rest of the paper we say “cycle in β” meaning “cycle in the unique cycle decomposition of β”.

The same results are valid for $\beta\alpha$. Notice that $\beta\alpha$ and $\alpha\beta$ are conjugates: $\alpha(\beta\alpha)\alpha^{-1} = \alpha\beta$, and therefore have the same cycle structure.

Now take an arbitrary 3-cycle $\alpha = (a\ b\ c)$ and an arbitrary permutation β. Three cases appear:

1. If a, b, and c are in three different cycles in β, these three cycles become a single cycle in $\alpha\beta$.

2. If two of a, b, c, are in the same cycle, and the third element is in a different cycle in β, then these two cycles recombine into another two cycles in $\alpha\beta$. Thus, the total number of cycles is maintained.

3. If a, b, and c are all in the same cycle in β, the result depends on the orientation they have in this cycle of β. Selecting a as the starting point, this cycle can have the form $(a\ \ldots\ b\ \ldots\ c\ \ldots)$ or $(a\ \ldots c\ \ldots b\ \ldots)$. In the first case, the cycle becomes $(a\ \ldots c\ \ldots\ b\ \ldots)$ in $\alpha\beta$. In the second case, the cycle breaks into $(a\ \ldots)(b\ \ldots)(c\ \ldots)$ in $\alpha\beta$.

The same results (except for the exact format of the resulting cycles in case 3) are valid for $\beta\alpha$.

6.2.2 Norm and Divisibility

In past sections we used phrases of the form "*a*, *b*, *c* in this order in π". It turns out that there is an algebraic way of expressing the same meaning. The notion of norm comes to our rescue.

It is well known that every permutation can be written as a product of 2-cycles. For instance, the cycle $(a\,b\,c\,d)$ can be written as:

$$(a\,b\,c\,d) = (a\,b)(b\,c)(c\,d)$$

This can be easily generalized to all cycles. Since every permutation is a product of cycles, it is clear that every permutation is a product of 2-cycles.

There are in general many ways of writing a permutation α as a product of 2-cycles. Among all these products, there must be some that use as few 2-cycles as possible. Let $|\alpha|$ denote the minimum number k such that α can written as a product of k 2-cycles. We call $|\alpha|$ the **norm** of α. For instance, $|1| = 0$, $|(a\ b)| = 1$ and $|(a\ b\ c\ d)| \leq 3$ (we will see later that this is an equality).

Theorem 6.2.1 *For every permutations α and β, he have:*

1. $|\alpha| = 0 \Leftrightarrow \alpha = 1$
2. $|\alpha^{-1}| = |\alpha|$
3. $|\beta \cdot \alpha| = |\alpha|$
4. $|\alpha\beta| \leq |\alpha| + |\beta|$

Proof: Item 1 is obvious, since using no cycles all we can produce is the identity. For item 2: if $\alpha = \alpha_1 \ldots \alpha_k$, with each α_i being a 2-cycle, then $\alpha^{-1} = \alpha_k \ldots \alpha_1$. So, for each decomposition of $|\alpha|$ into 2-cycles we have a decomposition of α^{-1} with the same number of cycles. This proves that $|\alpha^{-1}| \leq |\alpha|$. By symmetry we conclude that $|\alpha| \leq |\alpha^{-1}|$, and this proves item 2.

Also, if $\alpha = \alpha_1 \ldots \alpha_k$, we have that:

$$\beta \cdot \alpha = (\beta \ \cdot \alpha_1)(\beta \ \cdot \alpha_2) \ldots (\beta \ \cdot \alpha_k)$$

with each $\beta \cdot \alpha_i$ being a 2-cycle. Again, this allows us to conclude that $|\beta \cdot \alpha| \leq |\alpha|$. Since this is true for every α and β, we also have $|\beta^{-1} \cdot (\beta \cdot \alpha)| \leq |\beta \cdot \alpha|$ or, equivalently, $|\alpha| \leq |\beta \cdot \alpha|$. This proves item 3.

Item 4:

$$|\alpha| = k \Leftrightarrow \alpha = \alpha_1 \ldots \alpha_k$$

$$|\beta| = l \Leftrightarrow \beta = \beta_1 \ldots \beta_l$$

So,

$$\alpha\beta = \alpha_1 \ldots \alpha_k \beta_1 \ldots \beta_l \Rightarrow |\alpha\beta| \leq k + l = |\alpha| + |\beta|$$

This proves item 4.

■

Corollary 6.2.1 $|\alpha\beta| = |\beta\alpha|$

Let us start with the observations in Section 2.1. One of them says that if $\alpha = (a\ b)$, and a e b are in the same cycle in β, this cycle is broken in two in $\beta\alpha$. Calling $c(\alpha)$ the number of cycles in the cycle decomposition of a permutation α (including singleton cycles), this can be represented as a and b in the same cycle of $\beta \Rightarrow c(\beta\alpha) = c(\beta) + 1$. The other observation is: a and b in different cycles of $\beta \Rightarrow c(\beta\alpha) = c(\beta) - 1$.

These two facts are important because they say that the number of cycles changes by at most one when we multiply by a 2-cycle. Since the identity permutation in a set E has exactly $|E|$ cycles, we have the following result:

Theorem 6.2.2 *For any permutation α, if $\alpha = \alpha_1 \ldots \alpha_k$, with each α_i being a 2-cycle, then $k \geq |E| - c(\alpha)$.*

Proof: Induction in *k*. For $k = 0$, we have $\alpha = 1$, $c(\alpha) = |E|$ and $|E| - c(\alpha) = 0$. For $k > 0$, let $\alpha' = \alpha_1 \ldots \alpha_{k-1}$. We know by induction hypothesis that $k - 1 \geq |E| - c(\alpha')$. But $c(\alpha) = c(\alpha'\alpha_k) \geq c(\alpha') - 1$. Then $k = (k-1)+1 \geq |E| - c(\alpha')+1 = |E| - (c(\alpha')-1) \geq |E| - c(\alpha)$.
■

Corollary 6.2.2 *For any permutation α, we have $|\alpha| \geq |E| - c(\alpha)$.*

In fact, this is an equality, as the next result shows.

Theorem 6.2.3 *For any permutation α, we have $|\alpha| \leq |E| - c(\alpha)$.*

Proof: Let $\alpha = \alpha_1 \dots \alpha_k$ be the cycle decomposition of α including the 1-cycles as well. If cycle α_i has size l_i, we can write it as a product of 2-cycles with $l_i - 1$ terms (see beginning of Section 6.2.2). We therefore end up with a product of $\sum_{i=1}^{k}(l_i - 1)$ 2-cycles for α. But:

$$\sum_{i=1}^{k}(l_i - 1) = \sum_{i=1}^{k} l_i - k = |E| - c(\alpha)$$

Therefore, we have $|\alpha| \leq |E| - c(\alpha)$. ∎

Corollary 6.2.3 *For any permutation α, we have $|\alpha| = |E| - c(\alpha)$.*

We say that a permutation α *divides* permutation β when $|\beta\alpha^{-1}| = |\beta| - |\alpha|$. We use the notation $\alpha|\beta$ to indicate this fact.

In the sequel, we present some basic propositions of the divisibility relation and later we use them to express algebraically the order of elements in a cycle.

Theorem 6.2.4 *The divisibility relation is an order relation, that is:*

1. $\alpha|\alpha$
2. $\alpha|\beta$ *and* $\beta|\alpha \Rightarrow \alpha = \beta$
3. $\alpha|\beta$ *and* $\beta|\delta \Rightarrow \alpha|\delta$

Proof:

1. $|\alpha\alpha^{-1}| = |1| = 0$ and $|\alpha| - |\alpha| = 0$.
2. We have that:

$$\alpha|\beta \Rightarrow |\beta\alpha^{-1}| = |\beta| - |\alpha|$$

$$\beta|\alpha \Rightarrow |\alpha\beta^{-1}| = |\alpha| - |\beta|$$

   However, $\alpha\beta^{-1} = (\beta\alpha^{-1})^{-1}$, and therefore $|\beta\alpha^{-1}| = |\alpha\beta^{-1}|$. But then we have $|\beta\alpha^{-1}| = 0 \Rightarrow \beta\alpha^{-1} = 1 \Rightarrow \alpha = \beta$.

3. In this case, we have:

$$\alpha|\beta \Rightarrow |\beta\alpha^{-1}| = |\beta| - |\alpha|,$$
$$\beta|\delta \Rightarrow |\delta\beta^{-1}| = |\delta| - |\beta|.$$

So,

$$|\beta\alpha^{-1}| + |\delta\beta^{-1}| = |\delta| - |\alpha|.$$

But

$$|\delta\alpha^{-1}| = |\delta\beta^{-1}\beta\alpha^{-1}| \leq |\delta\beta^{-1}| + |\beta\alpha^{-1}| \leq |\delta| - |\alpha|,$$

and

$$|\delta| = |\delta\alpha^{-1}\alpha| \leq |\delta\alpha^{-1}| + |\alpha| \Rightarrow |\delta\alpha^{-1}| \geq |\delta| - |\alpha|$$

We conclude that $|\delta\alpha^{-1}| = |\delta| - |\alpha|$, that is, $\alpha|\delta$.

■

In addition, divisibility has some nice extra properties, and can be used to give an algebraic characterization of the fact that cycle α appears in the cycle decomposition of permutation β, a phrase that we usually abbreviate as "α is a cycle of β".

Theorem 6.2.5 *Let α be a cycle and β an arbitrary permutation. Then α is a cycle of β if and only if the following two conditions hold:*

1. $\alpha|\beta$
2. *there is no cycle α' with $\alpha|\alpha'|\beta$.*

Theorem 6.2.6 *If $\alpha|\beta$ then $\alpha^{-1}|\beta^{-1}$.*

We can use the divisibility relation to express concepts we use frequently, like the order of elements in a cycle or whether two elements belong to the same cycle or not in a permutation.

In view of the last corollary, we can say that $\alpha|\beta$ if and only if $c(\beta\alpha^{-1}) = c(\beta) + |\alpha|$, that is, if and only if the number of cycles in $\beta\alpha^{-1}$ increases by $|\alpha|$ compared to the number of cycles in β. In others words, α^{-1} causes maximum cycle breaking in β.

If α is the 2-cycle $(a\ b)$, α divides β if and only if a and b are in the same cycle of β. And if α is the 3-cycle $(a\ b\ c)$, α divides β if and only if a, b and c are in the same cycle of β, and appear in this order in it.

We summarize these observations in the following theorems.

Theorem 6.2.7 *Let $a, b, c \in E$ and α be a permutation. Then:*

1. *a, b are in the same cycle of $\alpha \iff (ab)|\alpha$*
2. *a, b, c in the same cycle of α and appear in this order in this cycle $\iff (abc)|\alpha$*

Theorem 6.2.8 *Let $\alpha = (ab), \beta$ be permutations. Then:*

1. *$\alpha|\beta \Longrightarrow$ a cycle is broken in $\alpha\beta$ (when compared to β)*
2. *$\alpha \not|\beta \Longrightarrow$ two cycles are joined in $\alpha\beta$ (when compared to β)*

Theorem 6.2.9 *Let $\alpha = (abc), \beta$ be permutations. Then:*

1. *$\alpha|\beta \iff$ a cycle is broken in three parts in $\beta\alpha^{-1}$ (when compared to β)*
2. *$\alpha \not| \beta$ but all 2-cycles (ab), (bc), (ac) divide $\beta \Longrightarrow$ a cycle has two regions transposed in $\beta\alpha^{-1}$ (when compared to β)*
3. *$\alpha \not| \beta$ and exactly two of (ab), (bc), (ac) divide $\beta \Longrightarrow$ two cycles are recombined in $\beta\alpha^{-1}$ (when compared to β)*
4. *$(ab), (bc), (ac)$ do not divide $\beta \Longrightarrow$ three cycles are joined in $\beta\alpha^{-1}$ (when compared to β)*

6.3 Genome Rearrangements

To formalize genome rearrangement problems we will use as base set for the permutations the set $E_n = \{-1, +1, -2, +2, \ldots, -n, +n\}$, where n is the number of genes. Thus, we will be modeling both strands of the underlying DNA molecule. Each element $+i$ or $-i$ represents a marker on the i^{th} gene, with its opposite meaning a marker in the same location in the opposite strand. We will first model circular genomes, which conform more naturally to the formalism, and will later comment on the necessary adaptations for linear genomes.

To begin with, let Γ be the permutation that maps each elements into its counterpart on the other strand. The permutation Γ can be written as:

$$\Gamma = (-1\ +1)(-2\ +2)\ldots(-n\ +n)$$

that is, a product of n disjoint 2-cycles. Notice that $\Gamma(a) \neq a$ for all $a \in E_n$, and $\Gamma^2(a) = \Gamma(\Gamma(a)) = a$ for all $a \in E_n$. In other words, $\Gamma^2 = 1$ or, equivalently, $\Gamma^{-1} = \Gamma$.

A cycle is *admissible* when it does not contain $-i$ and $+i$ for the same i. Thus, Γ is far from being an admissible cycle. An admissible cycle of size n is called a *genome strand*, because it models a strand of a genome formed by these n genes in some order. If we have an admissible cycle α, we can compute its reverse complement, as in the examples of the Table 6.1.

α	reverse complement
$(+3\ -1\ +7\ +5)$	$(-5\ -7\ +1\ -3)$
$(+2\ +4\ +6)$	$(-6\ -4\ -2)$

Table 6.1: Examples of admissible cycles α and their reverse complement.

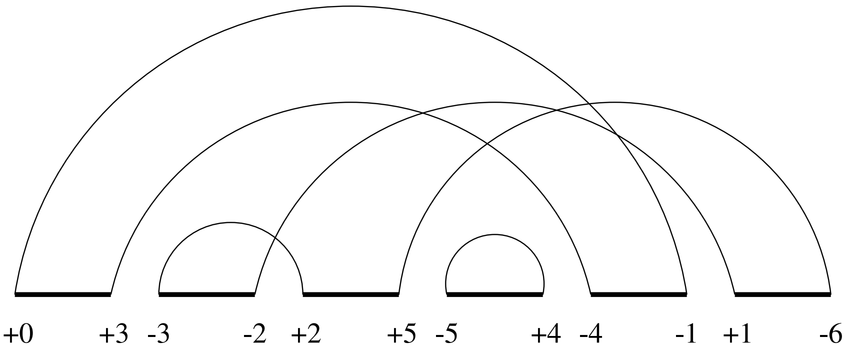

Figure 6.1: Breakpoint Graph for genomes $\pi = (-3\ +2\ -5\ -4\ +1)(-1\ +4\ +5\ -2\ +3)$ and $\sigma = (+1\ +2\ +3\ +4\ +5)(-5\ -4\ -3\ -2\ -1)$.

There is an algebraic way of obtaining the reverse complement. If α is an admissible cycle, α^{-1} is its reverse; $\Gamma \cdot \alpha = \Gamma\alpha\Gamma$ is its complement. The reverse complement is when we do both: $(\Gamma \cdot \alpha)^{-1}$ or $\Gamma \cdot (\alpha^{-1})$, which results in the same expression $\Gamma\alpha^{-1}\Gamma$.

Given a genome strand π_1, its reverse complement $\pi_2 = \Gamma\pi_1^{-1}\Gamma$ forms the complementary strand of the same genome. We represent this genome as the product of the two strands: $\pi = \pi_1\pi_2$. Since the strands form two disjoint cycles it does not matter in which order we take the product: $\pi_1\pi_2 = \pi_2\pi_1$. Also, it does not matter which strand we call π_1: had we started with π_2 we would have computed its reverse complement π_1 and the final genome would have been the same. This is just as DNA should be: no matter which strand you pick, when you let it pair with its reverse complement, you get the same DNA molecule.

Formally, we define a *genome* as a permutation that can be written as $\pi_1\Gamma\pi_1^{-1}\Gamma$, for some genome strand π_1. Note that $\Gamma\pi\Gamma = \pi^{-1}$ for every genome π. The general genome rearrangement problem then becomes: given two genomes π and σ and a class of operations, find the minimum number of events (operations) that transform π into σ. This minimum number is called the *distance* between π and σ.

We will talk about classes of operations later, but for any of the several problems obtained by choosing a different set of operations, the *breakpoint graph* plays an important role. Classically, the breakpoint graph is constructed as in Figure 6.1.

Details of the construction have been described previously several times and will not be repeated here. See for instance [64, 106]. Our objective is to obtain this graph, or an equivalent structure, by algebraic manipulations. The breakpoint graph is used when we want to transform

π into a constant genome $\sigma = (+1\ +2\ +3\ +4\ +5)(-5\ -4\ -3\ -2\ -1)$. It then depends on both π and σ. In fact, it has been defined for linear genomes, and to adapt to that we need consider "extended" versions of π and σ: take $\pi_1 = (+0\ -3\ +2\ -5\ -4\ +1)$ and identify -0 with -6.

The breakpoint graph is composed of black edges, which depend only on π, and of gray edges, which depend only on σ. Its turns out that $\Gamma\pi$ is a product of 2-cycles that correspond exactly to the black edges. And $\Gamma\sigma$ corresponds to the gray edges in the same way. In the preceding example, we have:

$$\begin{aligned}\Gamma\pi &= (-0\ +0)(-1\ +1)(-2\ +2)(-3\ +3)(-4\ +4)(-5\ +5)\\ &\quad (+0\ -3\ +2\ -5\ -4\ +1)(-1\ +4\ +5\ -2\ +3\ -0)\\ &= (+0\ +3)(-3\ -2)(+2\ +5)(-5\ +4)(-4\ -1)(+1\ -0),\end{aligned}$$

exactly the black edges. And $\Gamma\sigma$ will give the gray edges:

$$\Gamma\sigma = (+0\ -1)(+1\ -2)(+2\ -3)(+3\ -4)(+4\ -5)(+5\ -0).$$

In the classical theory of genome rearrangements the cycle structure of the breakpoint graph plays an important role. Although we could not obtain the cycles of the breakpoint graph themselves, we derived an algebraic expression for the *square* of the cycles. This expression is just the product $(\Gamma\pi)(\Gamma\sigma) = \Gamma\pi\Gamma\sigma$. In the example, we have:

$$\Gamma\pi\Gamma\sigma = (+0\ -4)(+3\ -1)(-3\ +5\ +1)(-2\ -0\ +2)(-5)(+4)$$

For each cycle of the breakpoint graph we have two cycles in $\Gamma\pi\Gamma\sigma$. If the cycle in the breakpoint graph is $(a_1\ a_2\ \ldots\ a_{2k})$, we have $(a_1\ a_3\ \ldots\ a_{2k-1})$ and $(a_{2k}\ a_{2k-2}\ \ldots\ a_2)$ in $\Gamma\pi\Gamma\sigma$. Therefore, this is not exactly the square of a breakpoint cycle, because one of them is reversed. Strictly speaking, we cannot model as permutations the cycles of the breakpoint graph, since they have no orientation. This in part explains why one cycle in the square is reversed. Had we taken $\Gamma\pi\Gamma\sigma$ the other cycle would have been reversed. Notice that $\Gamma\pi\Gamma\sigma = \pi^{-1}\sigma$, and $\Gamma\sigma\Gamma\pi = \sigma^{-1}\pi$.

In any case, these constructions allow us to rephrase technical properties of breakpoint graphs in algebraic terms. For instance, how many breakpoints π has with respect to σ? This is just the number of elements not fixed by $\Gamma\pi\Gamma\sigma$, divided by 2:

$$b(\pi, \sigma) = \frac{|Supp(\Gamma\pi\Gamma\sigma)|}{2}.$$

Likewise, the number of cycles in the breakpoint graph is half the number of cycles in $\pi^{-1}\sigma$. We can define also the length (size) of the cycles, good or bad cycles, cycles that can be broken by certain operations. We hope to be able to define interleaving cycles, hurdles, fortresses, all in algebraic terms.

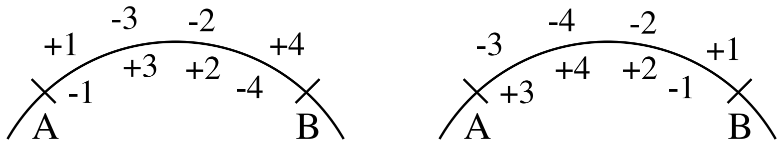

Figure 6.2: Linear genomes with fixed extremes.

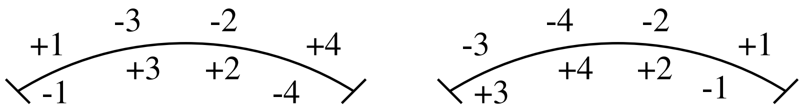

Figure 6.3: Linear genomes without fixed extremes.

6.3.1 Linear Genomes

The theory developed so far fits nicely with circular genomes. In this section we will briefly examine the case of linear genomes.

First, we must recognize that there are actually two kinds of linear genome: with free and with fixed extremes. Let us define each kind, starting with the one with fixed extremes.

When we compare two regions of two different genomes, and these regions are flanked by conserved parts, we need to use the fixed-extreme case (Figure 6.2). In this case, we add an extra dummy gene BA, which represents the fixed extremities of the regions, and proceed as in the circular case.

When we compare two entire linear genomes, we need to take into account that there is a free reversal that can be applied, so the distance in this case becomes:

$$d_{free} = \min(d_{fixed}(\pi, \sigma), d_{fixed}(\Gamma \cdot \pi^{-1}, \sigma))$$

More details on the relationship between linear and circular genome rearrangement problems can be found in the references [119, 92].

6.3.2 Operations

We will define in this section the events (operations) of reversal, transposition (or block move), and block interchange, some of them in the their signed and unsigned version. We will do a very detailed job for reversals, and then just state the results for the others, to save space.

Given a genome π, to perform a reversal on it we need to choose two distinct markers u and v, in the same strand of π, and then replace the path from u to v (including u but excluding v) by its reverse complement. Of course, a similar operation will be performed on the other strand, to make sure the final result is still a valid genome. Figure 6.4 shows what is meant.

We want to write the resulting genome σ as $\rho\pi$, where ρ is a permutation that will represent

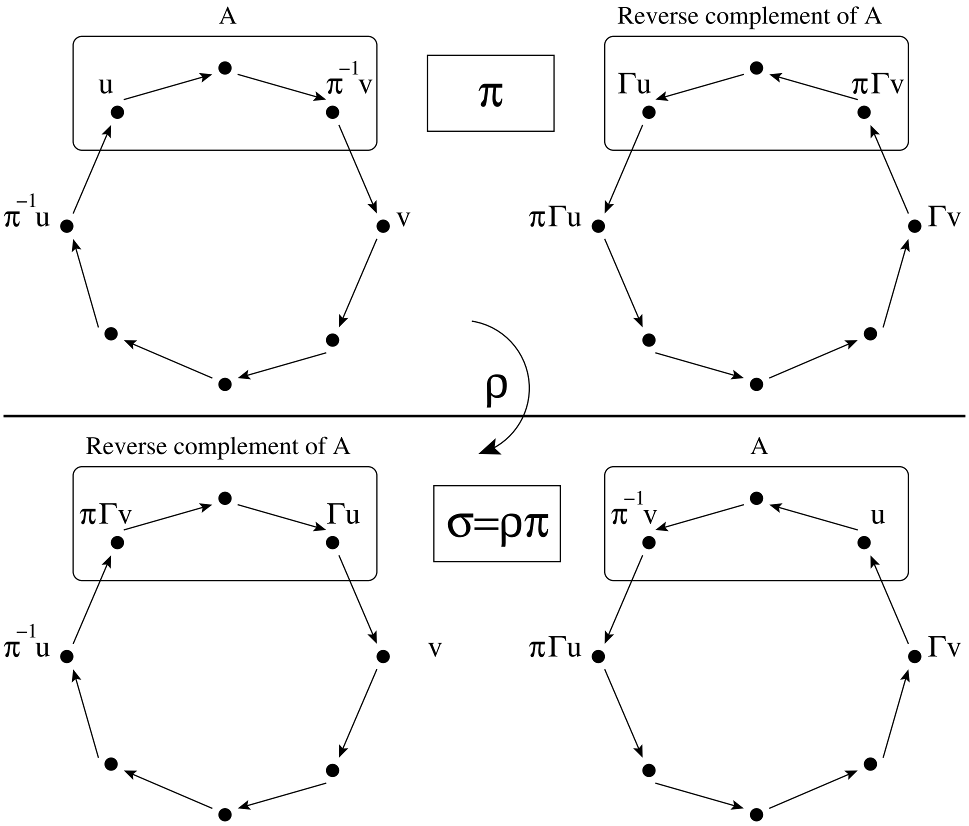

Figure 6.4: A reversal ρ applied to genome π.

the reversal. With some work, we see that σ differs from π only in the following mappings:

$$\sigma\pi^{-1}u = \pi\Gamma v,\ \sigma\Gamma u = v,\ \sigma\Gamma v = u,\ \sigma\pi^{-1}v = \pi\Gamma u.$$

Therefore, $\rho = \sigma\pi^{-1}$ maps:

$$\rho\pi\Gamma v = \sigma\pi^{-1}\pi\Gamma v = \sigma\Gamma v = u$$

$$\rho v = \sigma\pi^{-1}v = \pi\Gamma u$$

$$\rho u = \sigma\pi^{-1}u = \pi\Gamma v$$

$$\rho\pi\Gamma u = \sigma\pi^{-1}\pi\Gamma u = \sigma\Gamma u = v$$

with all other elements fixed by ρ. Or, written as a product of disjoint cycles:

$$\rho = (u\ \pi\Gamma v)(v\ \pi\Gamma u).$$

This is then the general formula of a reversal applicable to π, where u and v are two elements in same strand of π. We say that π and σ *differ by a reversal* when there is such a reversal ρ

with $\sigma = \rho\pi$. Notice that the definition of a reversal depends on π. There is no way to define a class of permutations that will be "the reversals", valid for all genomes. Each genome has a particular set of reversals that can be applied to it, and this sets varies from one genome to another.

For this reason, we cannot view the genome rearrangement problem directly as a "group generators" problem, where a class of generators of the symmetric group is given and we seek the minimum number of generators to write a given permutation. Nevertheless, it can be viewed indirectly as a group generators problem. We intend to pursue this approach in a future paper.

The reversal distance problem is: given two genomes π and σ, find the minimum k such that there are genomes $\gamma_0, \gamma_1, \ldots, \gamma_k$ with $\pi = \gamma_0$, $\sigma = \gamma_k$ and γ_i differs from γ_{i+1} by a reversal, for $i = 0, \ \ldots, \ (k-1)$.

In a previous version of this manuscript we stated that unsigned reversals could also be handled, but this is false, as far as we can see, because unsigned reversals do not exist in nature. They are artifacts created to deal with the lack of information on orientation. However, the following operations can be handled by this formalism.

A *transposition* is defined as:

$$\tau = (\pi u \ \pi v \ \pi w)(\Gamma w \ \Gamma v \ \Gamma u)$$

where u, v and w are three distinct elements in the same strand in π, appearing in this order (u, v, w) in the strand.

A *reversal+transposition* is defined as:

$$\tau = (\pi u \ \pi v \ \Gamma w)(\pi w \ \Gamma v \ \Gamma u)$$

where u, v and w are distinct elements in the same strand in π, appearing in this order (u, v, w) in this strand. A reversal+transposition models the event in which a block detaches itself from a genome and reappears elsewhere, in the same strand but the block is reversed.

A *block interchange* is defined as:

$$\beta = (\pi u \ \pi w)(\pi x \ \pi v)(\Gamma u \ \Gamma w)(\Gamma x \ \Gamma v)$$

where u, v, w and x are four distinct elements in the same strand of π, appearing in the order (u, v, w, x) in this strand.

Each one (or a group of) of these types of events can be used to define a genome rearrangement problem: given two genomes π and σ, find the minimum k such that there are genomes γ_0,γ_1,. . .,γ_k with $\pi = \gamma_0$, $\sigma = \gamma_k$ and γ_i differs from γ_{i+1} by the specific operation (or a group of), for $i = 0, \ \ldots, \ (k-1)$.

6.4 Using the Theory

We will use the theory developed to show two results whose proof was based on pictorial representations. The first result appears in Christie's proof that a block interchange cannot create three cycles [28]. The other result is that there is only one way for a transposition to break a cycle, proved by Walter & colleagues by reference to a picture [120].

Theorem 6.4.1 *Let π and σ be two genomes, and β a block interchange on π. Then the number of cycles in $\sigma(\beta\pi)^{-1}$ is not higher than 4 plus the number of cycles in $\sigma\pi^{-1}$.*

Proof: We have $\sigma(\beta\pi)^{-1} = \sigma\pi^{-1}\beta^{-1}$, which is $\sigma\pi^{-1}$ multiplied by β^{-1}. The cycle structure of β^{-1} is the same as β's: four 2-cycles. By the classical results about products by 2-cycles it is immediate that multiplying by four 2-cycles we cannot create more than 4 extra cycles. ∎

Theorem 6.4.2 *Let π and σ be two genomes, and $\tau = (u\ v\ w)(\pi\Gamma w\ \pi\Gamma v\ \pi\Gamma u)$ a transposition on π, where u, v, w appear in this order in the same cycle of π. Then $\sigma(\tau\pi)^{-1}$ has four more cycles than $\sigma\pi^{-1}$ if and only if u, v, w are in the same cycle in $\sigma\pi^{-1}$ and appear in the order (u, v, w) in this cycle.*

Proof: Transpositions do not mix genome strands, and therefore we know that the elements of a strand of π will form a strand in $\tau\pi$ (possibly in different order). Let π_1 be the strand that contains u, v and w, and σ_1, the corresponding strand in σ. We then have $\pi = \pi_1\Gamma\pi_1^{-1}\Gamma$, $\sigma = \sigma_1\Gamma\sigma_1^{-1}\Gamma$, and $\sigma\pi^{-1} = \sigma_1\pi_1^{-1}(\Gamma\sigma_1^{-1}\Gamma)(\Gamma\pi_1\Gamma)$. Then $\sigma(\tau\pi)^{-1} = \sigma\pi^{-1}\tau^{-1}$ will be the product of disjoint permutations $\sigma_1\pi_1^{-1}(w\ v\ u)$, and $(\Gamma\sigma_1^{-1}\Gamma)(\Gamma\pi_1\Gamma)(\pi\Gamma u\ \pi\Gamma v\ \pi\Gamma w)$.

In the first component $\sigma_1\pi_1^{-1}(w\ v\ u)$ we have a product of a 3-cycle by $\sigma_1\pi_1^{-1}$. We know from Theorem 6.2.9) that this produces two extra cycles if and only if $(uvw)|\sigma_1\pi_1^{-1}$, that is, if and only if u, v, w appear in the same cycle of $\sigma_1\pi_1^{-1}$ in the order (u, v, w), as stated. ∎

6.5 Reversal Distance

In this section we will apply our theory to the reversal distance problem.

6.5.1 Good and Bad Cycle Pairs

Let us start by examining the cycle structure of $\sigma\pi^{-1}$, where σ and π are two genomes. It turns out that every cycle α in $\sigma\pi^{-1}$ has a *companion* cycle $(\pi\Gamma)\cdot\alpha^{-1}$ also in $\sigma\pi^{-1}$. Their product will be called a *cycle pair* in $\sigma\pi^{-1}$

Every cycle α in $\pi^{-1}\sigma$ has a companion cycle $(\Gamma\pi)\cdot\alpha^{-1}$ also in $\pi^{-1}\sigma$. The cycle α is good if and only if $(\Gamma\pi)\cdot\alpha^{-1}$ is good.

First, we note that there are four ways of specifying the pivot points u, v of a reversal $\rho(u, v)$, because:

$$\rho(u, v) = \rho(v, u) = \rho(\pi\Gamma u, \pi\Gamma v) = \rho(\pi\Gamma v, \pi\Gamma u).$$

According to the relative positions of u, v in $\sigma\pi^{-1}$ we have three cases:

1. u, v in the same cycle of $\sigma\pi^{-1}$
2. u, v in companion cycles of $\sigma\pi^{-1}$
3. u, v in non-companion cycles of $\sigma\pi^{-1}$

Notice that any of these conditions, if true for u and v, is also true for v, u, for $\pi\Gamma u, \pi\Gamma v$, and for $\pi\Gamma v, \pi\Gamma u$, that is, they are independent of the particular pivot points chosen and are therefore characteristics of the reversal itself.

In case 1 the number of cycles of $\sigma\pi^{-1}$ does not change, because we have:

$$\sigma\pi^{-1} = \alpha(u \ldots v \ldots)(\ldots \pi\Gamma v \ldots \pi\Gamma u)$$

Then,

$$\begin{aligned}
\sigma\pi^{-1}\rho^{-1} &= \sigma\pi^{-1}\rho \\
&= \alpha(u \ldots v \ldots)(\ldots \pi\Gamma v \ldots \pi\Gamma u)(u\ \pi\Gamma v)(v\ \pi\Gamma u) \\
&= \alpha(u \ldots \pi\Gamma u \ldots)(v \ldots \pi\Gamma v \ldots)
\end{aligned}$$

In case 2 the number of cycles of $\sigma\pi^{-1}$ increases by 2:

$$\sigma\pi^{-1} = \alpha(u \ldots \pi\Gamma v \ldots)(\ldots v \ldots \pi\Gamma u)$$

Then,

$$\begin{aligned}
\sigma\pi^{-1}\rho^{-1} &= \sigma\pi^{-1}\rho \\
&= \alpha(u \ldots \pi\Gamma v \ldots)(\ldots v \ldots \pi\Gamma u)(u\ \pi\Gamma v)(v\ \pi\Gamma u) \\
&= \alpha(u \ldots)(v \ldots)(\pi\Gamma u \ldots)(\pi\Gamma v \ldots)
\end{aligned}$$

In case 3 the number of cycles of $\sigma\pi^{-1}$ decreases by 2:

$$\sigma\pi^{-1} = \alpha(u \ldots)(\ldots \pi\Gamma u)(v \ldots)(\ldots \pi\Gamma v)$$

Then,

$$\begin{aligned}
\sigma\pi^{-1}\rho^{-1} &= \sigma\pi^{-1}\rho \\
&= \alpha(u \ldots)(\ldots \pi\Gamma u)(v \ldots)(\ldots \pi\Gamma v)(u\ \pi\Gamma v)(v\ \pi\Gamma u) \\
&= \alpha(u \ldots \pi\Gamma v \ldots)(v \ldots \pi\Gamma u \ldots)
\end{aligned}$$

Based on these observations, we say that a reversal $\rho(u, v)$ that can be applied to genome π is a *cycle twister* in case 1, a *cycle breaker* in case 2, and a *cycle merger* in case 3. In the first two cases we say that $\rho(u, v)$ *acts on the cycle pair* that contains u and v. Because there are alternative ways of picking the pivot points for a given reversal, it is necessary to refer to the cycle pair as a whole and not just to the cycle individually.

Take an arbitrary cycle α of $\sigma\pi^{-1}$. Let k be the length of this cycle, and let k_1 and k_2 be the number of elements in this cycle that are in strands π_1 and π_2 of π. Define the *imbalance* of α as $|k_1 - k_2|$. This number has the same parity as k and lies between 0 and k. It is in fact a property of the cycle pair of α, because the companion cycle $(\pi\Gamma) \cdot \alpha$ will have the same imbalance.

We define a **bad** $\sigma\pi^{-1}$ cycle pair as a cycle pair that has imbalance equal to k, where k is the length of its cycles. A **good** $\sigma\pi^{-1}$ cycle pair is a cycle pair with imbalance less than k.

Theorem 6.5.1 *A cycle pair γ is good if and only if there is a cycle breaker that acts on γ.*

Proof: If γ is good then there are u, v in the same cycle of γ and in different strands of π. The reversal $\rho(u, \pi\Gamma v)$ is a cycle breaker acting on γ.

Conversely, let $\rho(u, v)$ be a cycle breaker of γ. Then u, v are in companion cycles. But then u, $\pi\Gamma v$ are in the same cycle of γ and in different strands of π. ■

6.5.2 Interleaving Cycles

Two cycle pairs γ_1 and γ_2 **interleave** when there are reversals ρ_1 acting on γ_1 and ρ_2 acting on γ_2 such that $\rho_1\rho_2\pi \neq \rho_2\rho_1\pi$. We can construct a graph where each cycle pair is a vertex, and interleaving cycle pairs are joined by an edge. A connected component of this graph is *good* when there is at least one good cycle pair in it. Otherwise it is a *bad* component.

Theorem 6.5.2 *If all components are good, then*

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

6.6 Conclusions

We propose a new way of looking of genomes as permutations, one that is more comfortable for those that have experience in permutation groups. Much remains do be done, but we feel this is the right way to attack difficult problems.