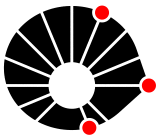

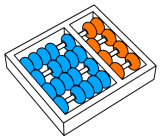

A Matrix-Based Theory for Genome Rearrangements

J. Meidanis *P. Biller* *J. P. P. Zanetti*

Technical Report - IC-18-10 - Relatório Técnico
August - 2018 - Agosto

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

A Matrix-Based Theory for Genome Rearrangements

João Meidanis[1] Priscilla do Nascimento Biller João Paulo Pereira Zanetti

[1]Support from FAPESP (Brazil) and NSERC (Canada).

Abstract

We started the work of reshaping these notes in 2016, when Meidanis was on sabbatical at the University of Ottawa, in David Sankoff's lab. We use the term "reshaping" because the main results shown here, on minimax genomes under the rank distance, were first presented in Biller's text for her PhD qualifying exam, written in 2014 [1].

The contents are divided in eight chapters, as follows. In Chapter 1 we introduce the first definitions, including genome matrices, distance, and orbits. We also derive an important formula for the distance based on orbits. In Chapter 2 we define operations on genomes, with focus on those with small rank. The chapter closes with the definition of basic operations, namely, cuts, joins, and double swaps. In Chapter 3 we study sorting scenarios going from a genome to another by basic operations. We show that from every genome we can reach any other with such scenarios. This also provides an alternative way of computing the distance. Intermediate genomes are the topic of Chapter 4. They can be characterized both as optimal scenario members, and as genomes for which the triangle inequality becomes an equality. They are also the medians of two genomes. A related notion is that of a minimax genome, explored in Chapter 5. We establish a lower bound for the minimax score, and show exactly the cases where it is possible to achieve such a score. In any case, it is always possible to find a genome within 1 unit of the lower bound. Chapter 6 deals with an interesting parity property of the rank distance. Chapter 7 tried to bridge the gap between mathematical definition and the real biological concepts. Finally, Chapter 8 contains exercises on the entire contents of the report, some of them with solutions or hints.

Contents

Foreword

We started the work of reshaping these notes in 2016, when Meidanis was on sabbatical at the University of Ottawa, in David Sankoff's lab. We use the term "reshaping" because the main results shown here, on minimax genomes under the rank distance, were first presented in Biller's text for her PhD qualifying exam, written in 2014 [1].

In this earlier text, genomes were represented as permutations of gene extremities. The text was, therefore, based on permutation group theory, especially on the notion of norm, which is essential for developing the genome rearrangement modeling. However, permutation norms are not familiar to most people, and we felt this impaired a wide dissemination of the ideas.

After an informal conversation with our colleague Luiz San Martin from the Math Department of the University of Campinas, we decided to look at genomes as matrices, and develop the theory from this point of view. Zanetti successfully used this approach to implement an approximate solution to the median problem [16]. It turns out that the algebraic distance is half the rank distance of the corresponding matrices. On top of that, matrices have two algebraic operations, addition and multiplication, whereas permutations have just multiplication. As we developed the theory, we discovered that many of the key concepts had an alternative, additive definition in the matrix realm, which may open new avenues for thought and may lead to new insights and results. For instance, in Section 4.2 we show that if B is an intermediate genome between A and C, it is possible to write $B = SA + (I - S)C$ for some matrix S. This is strikingly similar to a property of vector spaces, which states that a point in the line segment between two vectors u and v can be written as $\lambda u + (1 - \lambda)v$ for some scalar λ. It makes one wonder which other geometric ideas can be brought to the genomic realm. Being able to carry geometric intuition to comparative genomics may help solve important problems.

These notes are therefore our attempt at rewriting the minimax results in terms of the newly discovered matrix-based genome rearrangement theory. It builds everything from scratch, not depending on previous experience in genome rearrangements. Only a basic knowledge of linear algebra is required, say, at the level that most exact sciences students receive in their first university years. A slightly deeper understanding on eigenvalues and eigenvectors is needed for the section on parity, but this section is largely self-contained and does not significantly affect the rest of the presentation.

The contents are divided in six chapters, as follows. In Chapter 1 we introduce the first definitions, including genome matrices, distance, and orbits. We also derive an important formula for the distance based on orbits. In Chapter 2 we define operations on genomes, with focus on those with small rank. The chapter closes with the definition of basic operations, namely, cuts, joins, and double swaps. In Chapter 3 we study sorting scenarios going from a genome to another by basic operations. We show that from every genome we can reach any other with such scenarios. This also provides an alternative way of computing the distance. Intermediate genomes are the topic of Chapter 4. They can be characterized both as optimal scenario members, and as genomes for which the triangle inequality becomes an equality. They are also the medians of two genomes. A

related notion is that of a minimax genome, explored in Chapter 5. We establish a lower bound for the minimax score, and show exactly the cases where it is possible to achieve such a score. In any case, it is always possible to find a genome within 1 unit of the lower bound. Finally, Chapter 6 deals with an interesting parity property of the rank distance.

Our goal is that these notes serve as a guide for new students and researchers that want to know more about the subject. We also plan on evolving the material incrementally, say, at the rate of one new version per year, adding new results, exercises, better explanations, etc.

Context

To place this work within context, let's briefly review the steps that took us here. Genome rearrangements have been related to permutations at least since 1993, when Kececioglu and Sankoff studied the inversion distance [8]. A breakthrough occurred in 1995, when Hannenhalli and Pevzner solved the signed reversal distance problem polynomially, with a difficult and insightful algorithm [6].

The link between permutations and genome rearrangements had been established, but it was not completely satisfactory, for the following reason. Basic rearrangement operations, such as reversals, yielded permutations with supports of varying sizes in these early modelings. That is, a reversal could change 2, 3, 4, or even all the elements. This is at odds with the intuition that basic operations should change just a few elements. In addition, the processes of adding or removing genes, or considering a run of consecutive genes as a single "super gene", all require renumbering of other genes and become cumbersome.

An ailment to this discomfort comes when you view genomes not as a function associating a position to a gene, but rather as a function associating each gene to the next gene in the same chromosome. In principle, both views are equivalent, but, when you use the "next gene" method, a reversal is always a permutation with small support. Adding and removing genes is also simpler, as it involves local changes only. The same holds true for packing runs of consecutive genes into one "super gene", or the inverse operation.

This is the approach taken by Meidanis and Dias in their 2000 paper, but it only applied to circular genomes. Nevertheless, the genomes could have any number of chromosomes [10].

Yancopoulos *et al.* took this idea one step further and generalized Meidanis and Dias' idea to include linear chromosomes, creating the celebrated double cut and join (DCJ) distance. This distance gained much popularity because it is easy to compute, is capable of modeling many of the basic genome rearrangement operations, and applies to multichromosomal genomes with an arbitrary mixture of linear and circular chromosomes [15].

DCJ was so popular that people hardly bothered to think of alternatives. However, there is more than one way of incorporating linear chromosomes into the Meidanis and Dias scheme. Feijao and Meidanis presented one such way in 2012, leading to what we call the algebraic distance [3]. Meidanis and Yancopoulos compare the DCJ and algebraic distances, showing that their difference can be traced to the way paths are closed to form cycles in the breakpoint graph [11].

These notes are based on the rank distance, which is equal to twice the algebraic distance of Feijao and Meidanis.

Ottawa, March 2017

Joao Meidanis
Priscila Biller
Joao Paulo Pereira Zanetti

Additions in 2018

Since the last version, we started a chapter with more biological definitions of genes, chromosomes, and genomes. We added numerous exercises, some of them with solutions. Our goal is to provide solutions, or at least hints, for all problems. We also extended the parity result to orthogonal matrices. Small additions such as fixing typos, local presentation improvements, and the like were implemented as well.

Campinas, August 2018

Joao Meidanis
Priscila Biller
Joao Paulo Pereira Zanetti

Acknowledgments

The authors would like to acknowledge the financial support from Brazilian research funding agency FAPESP (grants 2012/14104-0 PNB, 2012/13865-7 JPPZ, 2016/01511-7 JM, and 2017/02748-3 JPPZ), and from Canadian research funding agency NSERC (Discovery Grant to David Sankoff).

Notation

A, B, C, D, E, G:	genome matrices
R, S, X, Y, Z:	generic matrices
$r()$:	rank
$d(,)$:	rank distance
x, y, z, w:	gene extremities
$\mathcal{E}$:	set of gene extremities
$\mathcal{S}$, $\mathcal{T}$, $\mathcal{V}$:	orbits, subspaces, vector sets in general
n:	number of genes
i, j, k, m:	generic indices
p, q:	number of orbits, number of even-sized orbits
$J(,), -J(,)$:	join, cut
$W(,,,)$:	double swap
P, Q:	generic genome operations
$\mathcal{L}$:	genome lists, scenarios
$w()$:	scenario weight
λ:	eigenvalue
v:	eigenvector, generic vector
$\sum(\mathcal{S})$:	notation for $\sum_{x \in \mathcal{S}} x$

Chapter 1

Introduction

In this chapter we introduce gene extremities, the representation of genomes as matrices, the rank distance, and orbits. We also study various subspaces related to the distance of two genomes, and derive an important formula for the distance based on orbits.

1.1 Genomes as Matrices

We can represent genomes as matrices. For a genome G involving n genes and therefore $2n$ gene extremities, we can choose an ordering for the extremities (any ordering is fine), and then define the corresponding **genome matrix** as follows:

$$G_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and extremities } i \text{ and } j \text{ are adjacent in } G, \text{ or} \\ & \text{if } i = j \text{ and extremity } i \text{ is a telomere in } G \\ 0 & \text{if } i \neq j \text{ and extremities } i \text{ and } j \text{ are } \textbf{not} \text{ adjacent in } G, \text{ or} \\ & \text{if } i = j \text{ and extremity } i \text{ is } \textbf{not} \text{ a telomere in } G \end{cases}$$

Let us see some examples. For genomes with just one gene, we have just two extremities. There are only two genomes: one with an adjacency linking these two extremities:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and the other with just telomeres:

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Here are some examples of genomes over two genes:

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Genome matrices are square matrices and have the following properties:

- They are **binary** matrices, i.e., have 0's and 1's only.
- They are **symmetric** matrices, that is, they satisfy $A^t = A$.

- They are **orthogonal** matrices, that is, they satisfy $A^t = A^{-1}$.
- They are **involutions**, that is, they satisfy $A^2 = I$.

It is easy to verify that any two of the last three properties implies the third one.

For binary matrices, being an orthogonal matrix is equivalent to having just one 1 in each row and in each column. Such binary matrices are called **permutation matrices**. We can then say that genome matrices are permutation matrices that are involutions.

1.2 Extremities as Column Vectors

To be compatible with our view of genomes as matrices, extremities will be seen as column unit vectors. For instance, considering genome C of Section 1.1, we see that x and y form an adjacency, where

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

Notice that $Cx = y$ and $Cy = x$, where Cx indicates the product of two matrices, C with size $2n \times 2n$ and x with size $2n \times 1$, giving a result of size $2n \times 1$ again, namely, y.

Extremities that are telomeres in a genome A, on the other hand, are characterized as column unit vectors z such that $Az = z$. For instance, considering the same matrix C as before, the vectors

$$z = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

are telomeres.

1.3 Distance

A distance can be defined for genomes over the same set of genes as follows. If A and B are genome matrices over the same genes, we write:

$$d(A, B) = r(B - A),$$

where r is the rank of a matrix.

There are several equivalent ways of defining matrix rank. In this text, we will use the following definition: $r(X)$ is the dimension of the subspace $\operatorname{im} X$, the image of X, which is the subspace of all vectors that can be written as Xu for some vector u.

This distance satisfies the properties required of a metric. To show this, notice first that $d(A, A) = r(0) = 0$, and the 0 matrix is the only one with rank 0, so if $d(A, B) = 0$ we have $B - A = 0$ and hence $A = B$. Also, $d(A, B) = d(B, A)$ since the rank does not change when the entire matrix is multiplied by a nonzero scalar (-1 in this case). The triangle inequality comes

from the fact that $r(X+Y) \leq r(X) + r(Y)$, because

$$\begin{aligned} d(A,C) &= r(C-A) \\ &= r((B-A)+(C-B)) \\ &\leq r(B-A)+r(C-B) \\ &= d(A,B)+d(B,C). \end{aligned}$$

Consider matrices A and B from Section 1.1. Their distance is:

$$d(A,B) = r(B-A) = r\left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right) = 1.$$

Now consider matrices C and D from the same section. Their distance is:

$$d(C,D) = r(D-C) = r\left(\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}\right) = 2.$$

If we also consider:

$$E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

we have:

$$d(C,E) = r(E-C) = r\left(\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\right) = 2.$$

Notice that sometimes you get zero rows in a matrix. These obviously do not add anything to the rank. But the number of nonzero rows is not the rank. In general, this number is greater than or equal to the rank.

1.4 Orbits

An important concept related to the distance between genome matrices A and B is that of the so-called AB-orbits (the name comes from the theory of permutation groups). Given two genomes A and B over the same genes, we define the following relation between gene extremities:

$$x \sim_{AB} y \text{ when } y - x \in \text{im}(B-A).$$

In this case, we say that x and y are AB**-equivalent**. When there is no risk of confusion, we will write simply $x \sim y$, omitting the subscript AB when it is clear from the context.

Observe that this is an equivalence relation. Notice that $x \sim_{AB} x$ because $x - x = 0$ which belongs to $\text{im}(B-A)$. Also, $x \sim_{AB} y$ implies $y \sim_{AB} x$, because when $y - x$ belongs to $\text{im}(B-A)$ its opposite vector $x - y$ also belongs to this linear subspace of $\mathbb{R}^{2n}$.

Finally, notice that $x \sim_{AB} y$ and $y \sim_{AB} z$ imply $y - x \in \text{im}(B-A)$ and $z - y \in \text{im}(B-A)$, which in turn imply that

$$z - x = (y - x) + (z - y) \in \text{im}(B-A)$$

again because $\mathrm{im}(B - A)$ is a linear subspace.

The classes of this equivalence relation are called AB**-orbits**. Here are some preliminary, relevant result on AB-orbits.

Lemma 1. *For any two genomes A and B over the same genes, and for any gene extremities x and y we have:*

1. $x \sim_{AB} BAx$
2. $x \sim_{AB} ABx$
3. $x \sim_{AB} y$ *implies* $x \sim_{AB} ABy$

Proof. For the first statement, notice that:

$$(B - A)(Ax) = BAx - AAx = BAx - x,$$

showing that $BAx - x \in \mathrm{im}(B - A)$ and therefore that $x \sim_{AB} BAx$.

For the second statement, notice that:

$$(B - A)(-Bx) = -BBx + ABx = ABx - x,$$

showing that $ABx - x \in \mathrm{im}(B - A)$ and therefore that $x \sim_{AB} ABx$.

For the third statement, just use the transitive property of AB-equivalence with $x \sim_{AB} y$ and $y \sim_{AB} ABy$. □

To proceed, we need a definition on matrices applied to orbits and other vector sets, and an auxiliary result.

Definition 1 (Matrix applied to set)**.** *If A is a matrix and $\mathcal{V}$ is a vector set, the notation $A\mathcal{V}$ denotes the following vector set:*

$$A\mathcal{V} = \{Av \mid v \in \mathcal{V}\}.$$

Lemma 2 (Orbit image)**.** *For any two genome matrices over the same genes A and B, and any extremities x and y, we have that $x \sim_{AB} y$ if and only if $Ax \sim_{AB} Ay$.*

Proof. If $x \sim_{AB} y$, then by definition $y - x = (B - A)u$ for some vector u. Then $Ay - Ax = A(y - x) = A(B - A)u = ABu - u$.

On the other hand, $(B-A)(-Bu) = -BBu+ABu = ABu-u$ as well, showing that $ABu-u \in \mathrm{im}(B - A)$. It follows that $Ay - Ax \in \mathrm{im}(B - A)$, supporting the conclusion that $Ay \sim_{AB} Ax$.

For the other direction, just notice that A is involutive. □

Lemma 3. *If $\mathcal{S}$ is an AB-orbit, then $A\mathcal{S}$ is an AB-orbit as well.*

Proof. To be an AB-orbit, $A\mathcal{S}$ must satisfy:

1. for any two extremities $x, y \in A\mathcal{S}$, we have $x \sim_{AB} y$, and
2. if $x \in A\mathcal{S}$ and $x \sim_{AB} y$ then $y \in A\mathcal{S}$.

Let us begin by proving the first statement. If $x, y \in A\mathcal{S}$, then $Ax, Ay \in \mathcal{S}$ and are, consequently, equivalent. We can then apply Lemma 2 to conclude that $x \sim_{AB} y$.

The second statement also follows from Lemma 2. If $x \sim_{AB} y$, then $Ax \sim_{AB} Ay$. Since $x \in A\mathcal{S}$, we have $Ax \in \mathcal{S}$, and therefore $Ay \in \mathcal{S}$ too, because $\mathcal{S}$ is an orbit, that is, an AB-equivalence class. We conclude that $y \in A\mathcal{S}$. □

Lemma 4. *If $\mathcal{S}$ is an AB-orbit, then $AB\mathcal{S} = \mathcal{S}$.*

Proof. From the fact that $x \sim_{AB} ABx$ (Lemma 1) we see that AB maps every extremity x of $\mathcal{S}$ into $\mathcal{S}$ itself. In other words, $AB\mathcal{S} \subseteq \mathcal{S}$. However, AB is a bijection, which means that $|AB\mathcal{S}| = |\mathcal{S}|$. We conclude that $AB\mathcal{S} = \mathcal{S}$. □

Notice that $AB\mathcal{S} = \mathcal{S}$ in turn implies that $B\mathcal{S} = AAB\mathcal{S} = A\mathcal{S}$. This observation will be useful in our analysis of rank 2 operations (see Section 2.2).

1.5 Symmetric Matrices

To continue our study, let us review some useful properties of symmetric matrices that we will use in the future. Recall the definition of the subspace $\mathcal{V}^\perp$ **orthogonal** to a given subspace $\mathcal{V}$:

Definition 2. *The subspace $\mathcal{V}^\perp$ is defined as:*

$$\mathcal{V}^\perp = \{u \mid v^t u = 0 \text{ for every } v \in \mathcal{V}\}.$$

It is known that $\dim \mathcal{V} + \dim \mathcal{V}^\perp = 2n$, for square matrices of size $2n \times 2n$.

Lemma 5. *If A is a symmetric matrix, then $(\operatorname{im} A)^\perp = \ker A$.*

Proof. Notice that if $u \in \operatorname{im} A$ and $v \in \ker A$, then $u = Aw$ for some vector w and

$$v^t u = v^t Aw = v^t A^t w = (Av)^t w = 0^t w = 0.$$

Hence, $\ker A \subseteq (\operatorname{im} A)^\perp$. Since these two subspaces have the same dimension, equality follows. □

Naturally, this last lemma applies to $B - A$ when A and B are genome matrices.

1.6 The Kernel

In this section we will examine the relationship of orbits to the kernel of $B - A$. Let's start by looking at the coefficients of equivalent extremities on vectors in this kernel.

Lemma 6 (Same coefficients)**.** *For any two genomes A and B over the same genes, if $v \in \ker(B-A)$ and $x \sim_{AB} y$, then $x^t v = y^t v$.*

Proof. Since $\operatorname{im}(B - A)$ and $\ker(B - A)$ are orthogonal, the hypotheses imply $(y - x)^t v = 0$. Therefore $x^t v = y^t v$. □

This lemma shows that, when we write a vector $v \in \ker(B - A)$ as a linear combination of gene extremities, as in:

$$v = \sum_{x \in \mathcal{E}} c_x x,$$

where $\mathcal{E}$ is the set of all gene extremities, then the coefficients c_x and c_y are equal when x and y are equivalent. This is because the coefficient c_x is nothing else than $x^t v$, and c_y is $y^t v$. With this result we can construct a basis for the subspace $\ker(B - A)$.

Definition 3. *For an AB-orbit $\mathcal{S}$, define:*

$$\sum \mathcal{S} = \sum_{x \in \mathcal{S}} x.$$

Lemma 7 (Kernel basis)**.** *For two genomes A and B over the same n genes, we have that the set*

$$\mathcal{V}_{AB} = \left\{ \sum \mathcal{S} \mid \mathcal{S} \text{ is an } AB\text{-orbit} \right\}$$

is a basis for $\ker(B - A)$.

Proof. Given an AB-orbit $\mathcal{S}$, consider the vector $\sum \mathcal{S}$. From Lemma 4 we learn that AB acts as a permutation in $\mathcal{S}$, and therefore $AB \sum \mathcal{S} = \sum \mathcal{S}$.

However, since A is a genome matrix, $ABv = v$ is equivalent to $Bv = Av$, that is, $v \in \ker(B-A)$. We conclude that $\sum \mathcal{S}$ belongs to $\ker(B - A)$ for every orbit $\mathcal{S}$.

On the other hand, if $v \in \ker(B - A)$, Lemma 6 implies that when we write

$$v = \sum_{x \in \mathcal{E}} c_x x$$

as a linear combination of extremities, there is a common coefficient $c_{\mathcal{S}}$ for all elements of orbit $\mathcal{S}$, allowing us to write instead:

$$v = \sum_{\mathcal{S}} c_{\mathcal{S}} \sum \mathcal{S},$$

that is, v is a linear combination of the p vectors $\sum \mathcal{S}$, where p is the number of AB-orbits. This shows that the $\sum \mathcal{S}$ over all AB-orbits $\mathcal{S}$ generate $\ker(B - A)$. Since the AB-orbits are disjoint, these vectors are linearly independent. Therefore, they form a basis of $\ker(B - A)$. □

1.7 Distance Formula

Now comes a result relating the AB-orbits to the distance $d(A, B)$.

Lemma 8 (Distance and orbits)**.** *For two genomes A and B over the same n genes, we have:*

$$d(A, B) = 2n - p,$$

where p is the number of AB-orbits.

Proof. We know that:

$$d(A, B) = r(B - A) = \dim \operatorname{im}(B - A) = 2n - \dim \ker(B - A), \tag{1.1}$$

since $2n$ is the number of extremities for n genes and the kernel and image of a matrix are complementary with respect to the domain space, which has dimension $2n$ in our case.

From Lemma 7 we learn that $p = \dim \ker(B - A)$. Plugging this on Equation 1.1 we get our result. □

Let us now establish a few more results about orbits that will be necessary in our forthcoming study on $\ker(B - A)$, and in the classification of the low-ranked operations studies in Chapter 2.

Lemma 9. *An extremity x belongs to a singleton AB-orbit if and only if x is either a common telomere of A and B or one end of a common adjacency of A and B.*

Proof. Suppose that $\mathcal{S} = \{x\}$ is a singleton orbit. From Lemma 4, we get $ABx = x$. Notice that $Bx = AABx = Ax$ in this case. If $Ax = x$, then $Bx = x$ as well and x is a common telomere of A and B. If $Ax = y \neq x$, then $Bx = Ax = y$ as well, and x is part of the common adjacency xy.

Conversely, if x is a common telomere or x is part of a common adjacency, then $Ax = Bx$, implying $x \in \ker(B - A)$. Let y be any extremity with $x \sim_{AB} y$. Applying Lemma 6 with x in place of v we reach:

$$x^t x = x^t y \Longrightarrow 1 = x^t y.$$

We conclude that in fact $y = x$, leading to the conclusion that $\{x\}$ is a singleton AB-orbit. □

1.8 The Image

We have seen earlier, with Lemma 6, a characterization of vectors in $\ker(B - A)$ in terms of their coefficients with respect to the extremities. Now we can use Lemma 5 to obtain a characterization for vectors in $\mathrm{im}(B - A)$.

Lemma 10. *Given two genomes A and B over the same genes, we have that $v \in \mathrm{im}(B - A)$ if and only if for any AB-orbit $\mathcal{S}$ the following equation holds:*

$$\sum_{x \in \mathcal{S}} x^t v = 0. \tag{1.2}$$

Proof. Suppose that $v \in \mathrm{im}(B - A)$. Given an AB-orbit $\mathcal{S}$, from Lemma 7 we know that $\sum \mathcal{S}$ belongs to $\ker(B - A)$. Since $(\mathrm{im}\, A)^\perp = \ker A$ (Lemma 5), it follows that

$$\sum_{x \in \mathcal{S}} x^t v = \left(\sum_{x \in \mathcal{S}} x \right)^t v = \left(\sum \mathcal{S} \right)^t v = 0.$$

Conversely, let's assume that Equation 1.2 holds for a certain vector v and all AB-orbits $\mathcal{S}$. We conclude that v is orthogonal to all the vectors in $\ker(B - A)$, that is:

$$v \in (\ker(B - A))^\perp.$$

However, $(\ker(B - A))^\perp = \mathrm{im}(B - A)$ by Lemma 5, and hence $v \in \mathrm{im}(B - A)$. □

1.9 Multiplicative characterization of orbits

We saw in Section 1.4 that the definition of AB-orbits relies on the difference matrix $B - A$. It turns out that these orbits could also be defined via successive multiplications by the matrix AB, as we will see in the present section.

Let us start by defining the following set:

$$\mathcal{T}_{AB}(x) = \{(AB)^m x \mid m \in \mathbb{Z}\}.$$

At this point, it is good to clarify what we mean by $(AB)^m$ when m is a negative number. Just observe that AB is invertible, and that its inverse is BA. Then, all you have to do is to think of $(AB)^m$ as being $(BA)^{-m}$ when m is a negative integer, and all will be fine. It does not hurt to mention also that $(AB)^0 = I$.

First notice that this set is actually finite.

Lemma 11 (Finite powers)**.** *For any two genome matrices A and B over the same genes, and any extremity x, we can write:*

$$\mathcal{T}_{AB}(x) = \{x, ABx, (AB)^2x, \ldots, (AB)^{k-1}x\},$$

where $k = |\mathcal{T}_{AB}(x)|$. Moreover, $(AB)^k x = x$.

Proof. The elements of $\mathcal{T}_{AB}(x)$ are all extremities. But, since there is only a finite number of extremities, we must have two *distinct* integers m_1 and m_2 such that $(AB)^{m_1}x = (AB)^{m_2}x$, which implies $(AB)^{m_1-m_2}x = x$ and also $(AB)^{m_2-m_1}x = x$. So, there is some *positive* integer m such that $(AB)^m x = x$.

Let k be the smallest such positive integer. Then $(AB)^k x = x$ and the extremities x, ABx, $(AB)^2x$, …, $(AB)^{k-1}x$ are all distinct, by definition of k, and hence $k \leq |\mathcal{T}_{AB}(x)|$.

On the other hand, any $y \in \mathcal{T}_{AB}(x)$ can be written as $(AB)^m x$, for some integer m. Performing an integer division we obtain $m = qk + r$, with q an integer and $0 \leq r < k$. We see then that $y = (AB)^m x = (AB)^{r+qk}x = (AB)^r(AB)^{qk}x = (AB)^r x$, since $(AB)^{qk}x = x$. This shows that actually $\mathcal{T}_{AB}(x) = \{x, ABx, (AB)^2x, \ldots, (AB)^{k-1}x\}$, and we conclude that $|\mathcal{T}_{AB}(x)| = k$. □

Another auxiliary lemma is needed here, showing that the sum of all elements of $\mathcal{T}_{AB}(x)$ belongs to the kernel of $B - A$.

Lemma 12. *For any two genome matrices A and B over the same genes, and any extremity x, we have:*

$$\sum_{m=0}^{k-1} (AB)^m x \in \ker(B - A),$$

where $k = |\mathcal{T}_{AB}(x)|$.

Proof. Notice that

$$A \sum_{m=0}^{k-1} (AB)^m x = Ax + \sum_{m=1}^{k-1} B(AB)^{m-1}x = Ax + \sum_{m=0}^{k-2} B(AB)^m x.$$

On the other hand,

$$B \sum_{m=0}^{k-1} (AB)^m x = \sum_{m=0}^{k-1} B(AB)^m x = \sum_{m=0}^{k-2} B(AB)^m x + B(AB)^{k-1}x.$$

Notice further that $Ax = B(AB)^{k-1}x$, because $(AB)^k x = x$. It follows that $Av = Bv$ for $v = \sum_{m=0}^{k-1}(AB)^m x$, and therefore $v \in \ker(B - A)$. □

We are now ready for our main characterization.

Theorem 1 (Multiplicative characterization of orbits)**.** *For any two genome matrices A and B over the same genes, two extremities x and y are AB-equivalent if and only if there is $m \in \mathbb{Z}$ such that $y = (AB)^m x$.*

Proof. If $y = (AB)^m x$, repeated application of Lemma 1 implies that $x \sim_{AB} y$.

For the other direction, assume that x and y are AB-equivalent. Taking $v = \sum_{m=0}^{k-1}(AB)^m x$, with $k = |\mathcal{T}_{AB}(x)|$, we have:

$$x^t v = x^t \sum_{m=0}^{k-1} (AB)^m x = 1,$$

because x is equal to exactly one of the extremities in the sum. By Lemma 6, we must have $y^t v = 1$ as well, given that $v \in \ker(B - A)$ by Lemma 12. Then,

$$1 = y^t v = y^t \sum_{m=0}^{k-1} (AB)^m x.$$

It follows that y must be equal to some extremity in this sum, that is, there actually is an $m \in \mathbb{Z}$ such that $y = (AB)^m x$. We can even pick m so that $0 \leq m \leq k - 1$. □

Chapter 2

Operations

In this chapter we define operations on genomes, with focus on those with small rank. The chapter closes with the definition of basic operations, namely, cuts, joins, and double swaps.

As we saw, the distance between two genomes is the rank of $B - A$. It makes sense to try to understand more about matrices of the form $B - A$, where A and B are genomes. For one thing, we can think of such a matrix as some kind of transformation that can be applied to genome A to yield a genome again. The following definitions captures this concept.

Definition 4. *A matrix P is **applicable** to a genome matrix A or can be **applied** to A when $A+P$ is also a genome matrix.*

Definition 5. *A matrix that is applicable to at least one genome is called an **operation**. The **weight** of an operation P is just its rank, $r(P)$.*

Some operations can be applied to several genomes. We will see examples in the sequel. The following result is simple, but it will be useful when we examine the effect of operations on the distance, later on in Chapter 4.

Lemma 13. *If P is an operation applicable to a genome A, we have:*

$$PA = -BP,$$

where $B = A + P$.

Proof. We have:

$$PA = (B - A)A = BA - A^2 = BA - I.$$

On the other hand,

$$-BP = -B(B - A) = -B^2 + BA = -I + BA.$$

Since both expressions agree, we have our result. □

We will be particularly interested in the simplest operations, that is, those with small rank. It turns out that, by using operations of rank 1 and 2 only, we can reach any genome from any other genome. This will be thoroughly demonstrated in Chapter 3 about scenarios, but here we will build the foundations for this work. It should be remarked that rank 1 operations are not enough to reach all other genomes. We do need the rank 2 operations in this result.

2.1 Operations of Weight 1

In this section we will study rank 1 operations. This is equivalent to looking at pairs of genomes A and B such that $d(A, B) = 1$, because in this case the difference $P = B - A$ will be a rank 1 operation. To uncover the format of such operations, we can reason as follows. If $d(A, B) = 1$, then Lemma 8 implies that there are $2n - 1$ AB-orbits. This is only possible if all AB-orbits are singletons, except for one 2-element orbit $\mathcal{S} = \{x, y\}$. Lemma 3 guarantees that $A\mathcal{S} = \mathcal{S}$, because $A\mathcal{S}$ must be an orbit as well and there are no other orbits with two elements. In particular, $Ax \in \mathcal{S}$.

Now if $Ax = x$, we must have $Ay = y$ to fulfill the condition $A\mathcal{S} = \mathcal{S}$. In this case, Bx must be equal to y since $B\mathcal{S} = \mathcal{S}$ as well and $ABx = y$. This characterizes an operation called a **join** in A, which will be studied more closely in the sequel.

The other possible case is when $Ax = y$. This implies $Ay = x$, $Bx = x$, and $By = y$. This operation is called a **cut** in A and will likewise be studied in more detail in the following sections. Cuts and joins are therefore the only rank 1 operations that exist.

2.1.1 Joins

Inspired by the above considerations, we define the first type of rank 1 operation.

Definition 6. *A **join** is a matrix of the form $J(x, y) = (x - y)(y^t - x^t)$ where x and y are distinct extremities.*

Notice that $J(x, y)^t = J(x, y)$ and $J(y, x) = J(x, y)$ for all joins. Also, the image of a join has dimension 1 because it is equal to the line generated by non-null vector $x - y$.

Lemma 14. *If P is a join, then $P^2 = -2P$.*

Proof. Let $P = J(x, y) = (x - y)(y^t - x^t)$. Then:

$$P^2 = (x - y)(y^t - x^t)(x - y)(y^t - x^t).$$

We can develop the center part as:

$$(y^t - x^t)(x - y) = y^t x - y^t y - x^t x + x^t y = 0 - 1 - 1 + 0 = -2.$$

Since this center part is a 1×1 matrix, we can bring it to the left as a scalar, yielding:

$$P^2 = -2(x - y)(y^t - x^t) = -2P.$$

□

The following result specifies the conditions a genome has to satisfy so that a join is applicable to it.

Lemma 15. *A join $P = J(x, y)$ is applicable to a genome matrix A if and only if $Ax = x$ and $Ay = y$. In this case the resulting genome $B = A + P$ satisfies $Bx = y$ and $By = x$.*

Proof. Since $B = A + P$, we have:

$$Bx = (A + P)x = Ax + Px = Ax + (x - y)(y^t - x^t)x = Ax + (x - y)(0 - 1) = Ax - x + y.$$

If $Ax \neq x$ then Bx is not an extremity, but rather a linear combination with non-null coefficients of at least two distinct extremities (x and y). Therefore, if $Ax \neq x$, P is not applicable to A.

By the same token, looking at the image of y under B, we get:

$$By = (A + P)y = Ay + Py = Ay + (x - y)(y^t - x^t)y = Ay + (x - y)(1 - 0) = Ay + x - y,$$

which shows that Ay has to be y if we want B to be a genome.

On the other hand, if $Ax = x$ and $Ay = y$, we have $Bx = y$ and $By = x$. In all extremities z other that x and y we have $Bz = Az$, so this is an extremity. We conclude that, if $Ax = x$ and $Ay = y$, B takes extremities to extremities, that is, it is a 0-1 matrix.

The resulting matrix B is also symmetric, as the sum of two symmetric matrices A and P. To be a genome, all we need to verify is that its square is I. Let's compute:

$$B^2 = (A + P)(A + P) = A^2 + AP + PA + P^2 = I + AP + PA - 2P,$$

in view of the fact that A is a genome and of Lemma 14.

Now

$$AP = A(x - y)(y^t - x^t) = (Ax - Ay)(y^t - x^t) = (x - y)(y^t - x^t) = P.$$

and

$$PA = P^t A^t = (AP)^t = P^t = P.$$

It follows that

$$B^2 = I + P + P - 2P = I.$$

□

2.1.2 Cuts

We analyze here the other type of rank 1 operation.

Definition 7. *A **cut** is a matrix of the form $-J(x, y) = (x-y)(x^t-y^t)$ where x and y are distinct extremities.*

Notice that opposite of a cut is a join and vice-versa. Cuts have also other interesting properties, as we can see in the next lemmas.

Lemma 16. *If P is a cut, then $P^2 = 2P$.*

Proof. The opposite of a cut is a join, so $Q = -P$ is a join. Then, using Lemma 14,

$$P^2 = (-Q)^2 = Q^2 = -2Q = 2P.$$

□

The following result specifies the conditions a genome has to satisfy so that a cut is applicable to it.

Lemma 17. *A cut $P = -J(x, y)$ is applicable to a genome matrix A if and only if $Ax = y$. In this case, the resulting genome $B = A + P$ satisfies $Bx = x$ and $By = y$.*

Proof. Since $B = A + P$, we have:

$$Bx = (A + P)x = Ax + Px = Ax + (x - y)(x^t - y^t)x = Ax + (x - y)(1 - 0) = Ax + x - y.$$

If $Ax \neq y$ then Bx is not an extremity, but rather a linear combination with non-null coefficients of at least two distinct extremities (x and y). Therefore, if $Ax \neq y$, P is not applicable to A.

On the other hand, if $Ax = y$, we have $Bx = x$. Likewise,

$$By = (A + P)y = Ay + Py = Ay + (x - y)(x^t - y^t)y = Ay + (x - y)(0 - 1) = x - x + y = y,$$

because $Ax = y$ implies $Ay = x$. In all extremities z other that x and y we have $Bz = Az$, so this is an extremity. We conclude that, if $Ax = y$, B takes extremities to extremities, that is, it is a 0-1 matrix.

The resulting matrix B is also symmetric, as the sum of two symmetric matrices A and P. To be a genome, all we need to verify is that its square is I. Let's compute:

$$B^2 = (A + P)(A + P) = A^2 + AP + PA + P^2 = I + AP + PA + 2P,$$

in view of the fact that A is a genome and of Lemma 16.

Now

$$AP = A(x - y)(x^t - y^t) = (Ax - Ay)(x^t - y^t) = (y - x)(x^t - y^t) = -P.$$

and

$$PA = P^t A^t = (AP)^t = (-P)^t = -P.$$

It follows that

$$B^2 = I - P - P + 2P = I.$$

□

The final conclusion is that genomes at distance 1 from A are just the ones where two telomeres in A are joined to form a new adjacency, or an adjacency in A is destroyed to create two telomeres.

2.1.3 Examples

Recall the matrices from Section 1.1:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \text{ and } C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

and the extremity vectors from Section 1.2:

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, z = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \text{ and } w = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

The difference $C - A$ is equal to the sum of a cut and a join, as follows. Consider the cut

$$-J(x, y) = -J\left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and the join

$$J(z, w) = J\left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

It is straightforward to verify that $C - A = -J(x, y) + J(z, w)$. Notice that we can either apply the cut first, and then the join, or the join first, and then the cut, since both operations are applicable to A.

The difference between A and B is also the sum of a cut and a join:

$$B - A = -J(x, y) + J(y, z).$$

However, in this case, the cut must be applied first, because the join is not applicable to A.

2.2 Operations of Weight 2

Of course, we can reach genomes at distance 2 from A by first getting a genome B at distance 1 from A and then looking for genomes at distance 1 from B. But we are interested in genomes at distance 2 from A that cannot be reached by this method.

Let's then think of a generic genome B at distance 2 from A. As we've done earlier, let's analyze the number of AB-orbits. For $d(A, B) = 2$ we must have $2n - 2$ orbits. The only possibilities are:

- all singleton orbits, except for two 2-element orbits;
- all singleton orbits, except for one 3-element orbit.

In the next sections we will contemplate each case.

2.2.1 Double Swaps

If we just have two non-singleton orbits $\mathcal{S}$ and $\mathcal{T}$, with $|\mathcal{S}| = |\mathcal{T}| = 2$, two things may happen: either $A\mathcal{S} = \mathcal{S}$ and $A\mathcal{T} = \mathcal{T}$ or $A\mathcal{S} = \mathcal{T}$ and $A\mathcal{T} = \mathcal{S}$. The first case does not add much because it leads to a rank 2 operation that is just a combination of two consecutive rank 1 operations acting on different extremities. Let's therefore concentrate our attention on the second case. If $\mathcal{S} = \{x, y\}$ and $\mathcal{T} = \{z, w\}$, without loss of generality we have $Ax = w$, which forces $Ay = z$, $Bx = z$, and $By = w$, since $B\mathcal{S} = \mathcal{T}$ and B cannot agree with A in x otherwise x would end up in an AB-orbit by itself. We conclude that A and B differ only on the following extremities:

$$\begin{aligned} Ax = w \quad &, \quad Bx = z, \\ Ay = z \quad &, \quad By = w, \\ Az = y \quad &, \quad Bz = x, \\ Aw = x \quad &, \quad Bw = y. \end{aligned}$$

In this case extremities x and y exchange partners going from A to B. This is a kind of distance-2 operation that cannot be achieved by two distance-1 operations. We would need 4 distance-1 operations: removing two adjacencies from A, and then adding two different adjacencies to the result to reach B.

We capture this operation type in our next definition.

Definition 8. *A **double swap** is a matrix of the form* $W(x, y, z, w) = (x - y)(z^t - w^t) + (z - w)(x^t - y^t)$ *where x, y, z, and w are distinct extremities.*

Notice that $W(x, y, z, w)^t = W(x, y, z, w)$ and $W(x, y, z, w) = W(y, x, w, z) = -W(y, x, z, w)$ for all double swaps. This means that the opposite of a double swap is another double swap. Special properties of double swaps are listed below.

Lemma 18. *The square of a double swap is a sum of cuts:*

$$W(x,y,z,w)^2 = -2J(x,y) - 2J(z,w).$$

Proof. Let $u = x - y$ and $v = z - w$. Then:

$$W(x,y,z,w)^2 = (uv^t + vu^t)(uv^t + vu^t) = uv^tvu^t + vu^tuv^t,$$

since the other terms vanish because u and v are orthogonal vectors, and hence $u^tv = v^tu = 0$. Also,

$$u^tu = v^tv = 2,$$

and these scalars can be moved to the beginning of the products, yielding:

$$W(x,y,z,w)^2 = (uv^t + vu^t)(uv^t + vu^t) = 2uu^t + 2vv^t,$$

which is just the sought formula, because $uu^t = -J(x,y)$ and $vv^t = -J(z,w)$. □

The following result specifies the conditions a genome has to satisfy so that a double swap is applicable to it.

Lemma 19. *A double swap $P = W(x,y,z,w)$ is applicable to a genome matrix A if and only if $Ax = w$ and $Ay = z$. In this case, the resulting genome $B = A + P$ satisfies $Bx = z$ and $By = w$.*

Proof. Since $B = A + P$, we have:

$$Bx = (A + P)x = Ax + Px = Ax + z - w.$$

If $Ax \neq w$ then Bx is not an extremity, but rather a linear combination with non-null coefficients of at least two distinct extremities (z and w). Therefore, if $Ax \neq w$, P is not applicable to A.

Likewise,

$$By = (A + P)y = Ay + Py = Ay + w - z,$$

so Ay must be z for this to produce an extremity.

On the other hand, if $Ax = w$ and $Ay = z$, we have $Bx = z$, $By = w$, $Bz = x$, and $Bw = y$. In all other extremities u we have $Bu = Au$, so B takes extremities to extremities, that is, it is a 0-1 matrix.

The resulting matrix B is also symmetric, as the sum of the two symmetric matrices A and P. To be a genome, all we need to verify is that its square is I. Let's compute:

$$\begin{aligned} B^2 &= (A + P)(A + P) \\ &= A^2 + AP + PA + P^2 \\ &= I + AP + PA - 2J(x,y) - 2J(z,w), \end{aligned}$$

in view of the fact that A is a genome and of Lemma 18.

Now

$$\begin{aligned} AP &= A(x - y)(z^t - w^t) + A(z - w)(x^t - y^t) \\ &= (w - z)(z^t - w^t) + (y - x)(x^t - y^t) \\ &= J(z,w) + J(x,y), \end{aligned}$$

and

$$PA = P^tA^t = (AP)^t = J(z,w) + J(x,y),$$

since AP is symmetric.

It follows that

$$B^2 = I + J(z,w) + J(x,y) + J(z,w) + J(x,y) - 2J(x,y) - 2J(z,w) = I.$$

□

2.2.2 Single Swaps

Let's now investigate the possibility of a single non-singleton, three-element AB-orbit $\mathcal{S}$. In this case, $A\mathcal{S} = \mathcal{S}$. Since adjacencies use two extremities, we must have at least one A-telomere in $\mathcal{S}$. We may have one or three A-telomeres in $\mathcal{S}$, but since B also has a telomere in $\mathcal{S}$, we cannot have three A-telomeres there, otherwise one of them would be a common telomere with B, contradicting the fact that $\mathcal{S}$ is an AB-orbit. Therefore, each of A, B has exactly one telomere in $\mathcal{S}$.

Say $\mathcal{S} = \{x, y, z\}$, where y is the B-telomere and z is the A-telomere. Then A and B differ only on the following extremities:

$$\begin{aligned} Ax = y \quad &, \quad Bx = z, \\ Ay = x \quad &, \quad By = y, \\ Az = z \quad &, \quad Bz = x. \end{aligned}$$

This is interesting but it is covered by doing two distance-1 operations: first remove adjacency xy and then add adjacency xz. We call this a **single swap**.

We conclude that the only distance-2 operations are the double swap and the ones obtained by applying distance-1 transformations twice.

2.2.3 Examples

Consider again the matrices used in Section 2.1.3. The difference $B - A$ mentioned in that section is in fact a single swap.

For an example of double swap, consider the following matrices:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Their difference is a double swap:

$$B - A = \begin{bmatrix} 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix}.$$

In terms of the vectors defined in Section 2.1.3, we can write:

$$B - A = W(x, w, z, y).$$

2.3 Basic Operations

Let's close the chapter with an important definition.

Definition 9. *A **basic** operation is an operation of one of the following kinds: cut, join, or double swap.*

We exclude the single swap because it can be replaced by a cut followed by a join.

Chapter 3

Sorting By Basic Operations

In this chapter, we will investigate the process of modifying a given genome by successive applications of basic operations. Can we reach any other genome by doing so? The answer is affirmative, and this provides an alternative way of computing the distance.

3.1 Sorting Scenarios and Weight

What is a sorting scenario? It is a list of matrices $\mathcal{L} = [B_0, B_1, \ldots, B_k]$ such that each matrix in the series differs from its neighbors by a basic operation (cut, join, or double swap). These operations have weights, which are equal to their ranks: cuts and joins have weight 1, whereas double swaps have weight 2. Each scenario also has a weight, also called its **total weight**, which is the sum of the weights of all consecutive operations in it. The weight of scenario $\mathcal{L}$ will be denoted by $w(\mathcal{L})$.

We say that a sorting scenario goes **from** A **to** C when $B_0 = A$ and $B_k = C$. Among all sorting scenarios from A to C, some will have minimum total weight. These will be called **optimal** sorting scenarios. The weight of an optimal sorting scenario from A to C will be denoted by $w(A, C)$. In principle, it is not even clear that there is a sorting scenario for any pair on genome matrices A and C. However, when there is one such scenario going from A to C, we can be sure that its weight is not smaller than the distance.

Theorem 2. *If there is a sorting scenario from A to C, its total weight is at least $d(A, C)$.*

Proof. Let $[B_0, B_1, \ldots, B_k]$ be a sorting scenario from A to C. We can write:

$$B_i = B_{i-1} + P_i,$$

for $i = 1, 2, \ldots, k$, where P_i is the operation transforming B_{i-1} into B_i. Therefore, each P_i is either a cut, a join, or a double swap. And the weight of P_i is its rank, $r(P_i)$. It follows that:

$$C - A = P_1 + P_2 + \ldots + P_k,$$

and hence

$$d(A, C) = r(C - A) = r(P_1 + P_2 + \ldots + P_k) \leq r(P_1) + r(P_2) + \ldots + r(P_k).$$

But $\sum_{i=1}^{k} r(P_i)$ is the weight of the scenario, so the desired conclusion follows. □

In fact, sorting scenarios not only exist for every pair of genome matrices A and C over the same genes but also the weight of optimal scenarios is always equal to the distance. To be able to prove that, we need a few preparatory results.

3.2 The Effect of Operations on Orbits

To proceed we need to understand how orbits change when we apply an operation. In fact, we don't need to understand everything about this process. It suffices, for our purposes, to know that some orbits are left untouched, that is, some AC-orbits will still be BC-orbits when we transition from A to B by means of an operation P. Specifically, when an orbit $\mathcal{S}$ is nullified by an operation P, that is, $P\mathcal{S} = \{0\}$, then $\mathcal{S}$ is untouched by the application of P. Here the notation $P\mathcal{S}$ means matrix P applied to set $\mathcal{S}$ as in Definition 1. Let's incrementally build results that will culminate in this conclusion.

Lemma 20 (Nullified orbit)**.** *If A and C are genome matrices over the same genes, P is an operation applicable to A, and $\mathcal{S}$ is an AC-orbit such that $P\mathcal{S} = \{0\}$, then $PA\mathcal{S} = \{0\}$ and $PC\mathcal{S} = \{0\}$ as well.*

Proof. First notice that, in view of Lemma 13, calling $B = A + P$, we have:

$$PA\mathcal{S} = -BP\mathcal{S} = -B\{0\} = \{0\},$$

so $PA\mathcal{S} = \{0\}$. Now, since $\mathcal{S}$ is an AC-orbit, we have $\mathcal{S} = CA\mathcal{S}$ and then:

$$PC\mathcal{S} = PCCA\mathcal{S} = PA\mathcal{S}.$$

because C is a genome and therefore $C^2 = I$. But we have seen that $PA\mathcal{S} = \{0\}$, so it follows that $PC\mathcal{S} = \{0\}$ as well. □

With the next result we can examine what happens with powers of genome products applied to an orbit nullified by an operation.

Lemma 21. *If A and C are genome matrices over the same genes, P is an operation applicable to A, $\mathcal{S}$ is an AC-orbit such that $P\mathcal{S} = \{0\}$, then, for every extremity $u \in \mathcal{S}$, we have:*

$$(BC)^m u = (AC)^m u,$$

for any integer $m \geq 0$, where $B = A + P$.

Proof. By induction on m. If $m = 0$, then the result boils down to $u = u$, which is clearly true.

If $m \geq 1$, notice that:

$$(BC)^m u = BC(BC)^{m-1} u.$$

By the induction hypothesis, we can replace $(BC)^{m-1}u$ by $(AC)^{m-1}u$, so that:

$$(BC)^m u = BC(AC)^{m-1} u = AC(AC)^{m-1} u + PC(AC)^{m-1} u.$$

However, notice that $(AC)^{m-1}u$ belongs to $\mathcal{S}$, since it is in the same AC-orbit as u. Therefore, by using Lemma 20, we conclude that $PC(AC)^{m-1}u$ is zero. Then:

$$(BC)^m u = AC(AC)^{m-1} u = (AC)^m u,$$

which is the sought conclusion. □

The following result shows orbits that are left intact by P.

Lemma 22. *If A and C are genomes over the same genes, P is an operation applicable to A, and $\mathcal{S}$ is an AC-orbit such that $P\mathcal{S} = \{0\}$, then $\mathcal{S}$ is also a BC-orbit, where $B = A + P$.*

Proof. Let x be an extremity in $\mathcal{S}$, and let $\mathcal{T}$ be the BC-orbit of x. For any extremity y in $\mathcal{S}$, by Theorem 1, we have $y = (AC)^m x$ for some integer m. By Lemma 21, this implies $(BC)^m x = y$ as well, showing that $y \in \mathcal{T}$. It follows that $\mathcal{S} \subseteq \mathcal{T}$.

Conversely, if $y \in \mathcal{T}$, then $y = (BC)^m x$. By the same lemma, $y = (AC)^m x$, showing that $y \in \mathcal{S}$. It follows that $\mathcal{T} \subseteq \mathcal{S}$.

We conclude that in fact $\mathcal{T} = \mathcal{S}$, and therefore $\mathcal{S}$ is a BC-orbit as well. □

3.3 Finding Sorting Operations

We are now ready to state a result about situations where cuts, joins, or double swaps get closer to a final genome. The analysis will be divided into two cases, depending on whether or not A and C have the same telomeres. A new definition is in order.

Definition 10. *Two genomes A and C over the same genes are said to be **co-tailed** when A and C have exactly the same telomeres.*

3.3.1 Genomes Not Co-Tailed

Let's start with the cases where A and C are not co-tailed. Then there is a telomere in one of the genomes that is not a telomere of the other. Let's assume that C has a telomere x that is not a telomere of A. In this case, we have seen that the cut $P = -J(x, y)$ where $y = Ax$ is applicable to A, that is, $A + P$ is a valid genome matrix.

Lemma 23. *If genome C has a telomere x that is not a telomere of A, then the cut $-J(x, Ax)$ is applicable to A, and leads to a genome B such that:*

$$d(B, C) \leq d(A, C) - 1.$$

Proof. We need to compare AC-orbits with BC-orbits. First notice that the AC-orbit of x contains $y = Ax$:

$$ACx = Ax = y,$$

since $x \sim_{AC} ACx$ (Lemma 1) and x is a telomere of C. Any other AC-orbit is also a BC-orbit, according to Lemma 22, since $P = B - A = -J(x, y)$ maps to zero any extremity other than x and y.

On the other hand, $\{x\}$ is a BC-orbit, because x is a telomere of both B and C. Since x's AC-orbit contains at least one other element, we conclude that there are strictly more BC-orbits than there are AC-orbits, and therefore

$$d(B, C) \leq d(A, C) - 1,$$

in view of Lemma 8. □

3.3.2 Co-Tailed Genomes

When A and C are co-tailed, no cut or join applied to A will get closer to C, but certain double swaps involving A-adjacencies can move towards C. We know that a double swap $W(x, y, z, w)$ is applicable to A if and only if both xw and yz are adjacencies of A, that is, $Ax = w$ and $Ay = z$.

To find such double swaps, we need to study AC-orbits in more detail. In the sequel we have a neat result showing that adjacent extremities in the same orbit imply the presence of a telomere in this orbit. For co-tailed genomes, we will show that this cannot be, so adjacencies always link extremities from distinct orbits.

Lemma 24. *If A and C are genomes over the same genes, $Ax = y$, and x, y are in the same AC-orbit, then this orbit contains either an A-telomere or a C telomere.*

Proof. If x and y are in the same AC-orbit, then:

$$y = (AC)^m x.$$

If m is even, we can write $m = 2k$ and then:

$$Ax = y = (AC)^{2k}x = (AC)^k(AC)^k x,$$

or, recalling that $CA = (AC)^{-1}$:

$$(CA)^k Ax = (AC)^k x.$$

But $(CA)^k Ax = (CA)^{k-1}Cx$, so:

$$(CA)^{k-1}Cx = A(CA)^{k-1}Cx,$$

showing that $(CA)^{k-1}Cx$ is a telomere of A.

If m is odd, then $m = 2k + 1$ so that

$$Ax = y = (AC)^{2k+1}x = (AC)^{k+1}(AC)^k x.$$

Then, multiplying by $(CA)^{k+1}$ to the left,

$$(CA)^{k+1}Ax = (AC)^k x,$$

or,

$$C(AC)^k x = (AC)^k x,$$

showing that $(AC)^k x$ is a telomere of C. □

Lemma 25 (Adjacencies cross orbits)**.** *If A and C are co-tailed genomes, and $y = Ax$ is distinct from x, then x and y are in different AC-orbits.*

Proof. Assume for a moment that x and y are in the same AC-orbit. Then, by Lemma 24, there is an extremity z in this orbit that is either an A-telomere or a C-telomere. Since A and C are co-tailed, they have the same telomeres, and therefore z is a telomere of both. It follows that $ACz = z$.

On the other hand, x is in this orbit too, so $x = (AC)^m z$ for some integer m, implying that $x = z$. Likewise, y is in this orbit, and therefore $y = z$. This contradicts the hypothesis that $x \neq y$, and we are forced to conclude that x and y cannot be in the same orbit after all. □

We are now ready for the main result of the section.

Lemma 26. *If A and C are co-tailed, distinct genomes, then there is a double swap P applicable to A such that*

$$d(B, C) \leq d(A, C) - 2,$$

where $B = A + P$.

Proof. Since $A \neq C$, there is an extremity x such that $Ax \neq Cx$. Then, x is not a telomere of either A or C, otherwise it would be a common telomere implying $Ax = x = Cx$. Therefore $y = Ax$ and $z = Cx$ are different from x and from each other, and are not telomeres of either A or C. Let $w = Az$ and consider the double swap $P = W(w, x, y, z)$. It is applicable to A since $Aw = z$ and $Ax = y$.

Notice that x and w are in the same AC-orbit, as are y and z, since $w = ACx$ and $y = ACz$. However, these two AC-orbits are distinct, according to Lemma 25. All other AC-orbits are also BC-orbits, where $B = A + P$, because of Lemma 22. Moreover, $\{x\}$ and $\{z\}$ are also BC-orbits, since $BCx = x$ and $BCz = z$. In addition, $By = w$, implying that y and w are in different BC-orbits, in view of Lemma 25. This all means that from the elements of the two AC-orbits containing x, y, z, and w there are at least four BC-orbits, namely, $\{x\}$, $\{z\}$, one with y and one with w, all distinct. It follows that

$$d(B, C) \leq d(A, C) - 2.$$

□

3.4 Existence of Sorting Scenarios

We are now ready for an important result.

Theorem 3. *For any pair of genome matrices A and C over the same genes there is a sorting scenario $\mathcal{L}$ going from A to C satisfying:*

$$w(\mathcal{L}) \leq d(A, C).$$

Proof. We will prove this by induction on $d(A, C)$. For $d(A, C) = 0$ we have $A = C$ and then $[A]$ is a sorting scenario of weight zero, which is equal to the distance.

Let's now tackle the case $d(A, C) \geq 1$. Since $A \neq C$, one of three cases may occur: either C has a telomere which A doesn't, or A has a telomere not shared by C, or A and C are co-tailed, but distinct. Let's handle each case in its turn.

If there is a telomere x of C that is not a telomere of A, then, by Lemma 23, the cut $P = -J(x, Ax)$ yields a genome $B = A + P$ such that $d(B, C) \leq d(A, C) - 1$. We can then apply the induction hypothesis to B and C, and assume that there is a scenario $\mathcal{L}$ from B to C such that:

$$w(\mathcal{L}) \leq d(B, C).$$

Now consider the scenario $[P]{+}\mathcal{L}$ obtained by inserting operation P before all other operations in $\mathcal{L}$. (This notation is inspired by a similar notation in the programming language Python.) This new scenario goes from A to C and satisfies:

$$w([P]{+}\mathcal{L}) = 1 + w(\mathcal{L}) \leq 1 + d(B, C) \leq d(A, C),$$

so we have a suitable scenario from A to C.

The second case is when there is a telomere of A that is not a telomere of C. By the same Lemma 23, we know there is a cut P applicable to C such that $B = C + P$ satisfies:

$$d(B, A) \leq d(A, C) - 1.$$

The induction hypothesis applied to A and B yields a scenario $\mathcal{L}$ from A to B such that:

$$w(\mathcal{L}) \leq d(A, B).$$

Now consider the scenario $\mathcal{L}+[-P]$ obtained by inserting the join $-P$ after all other operations in $\mathcal{L}$. We can do that because $B + (-P) = C$, so $-P$ is applicable to B. This is a scenario from A to C that satisfies:

$$w(\mathcal{L}+[-P]) = w(\mathcal{L}) + 1 \leq d(A, B) + 1 \leq d(A, C),$$

and is therefore what we need in this case.

Finally, if A and C are distinct but co-tailed, Lemma 26 asserts that there is a double swap P applicable to A such that

$$d(B, C) \leq d(A, C) - 2,$$

where $B = A + P$. We can again apply the induction hypothesis to B and C yielding a scenario $\mathcal{L}$ with $w(\mathcal{L}) \leq d(B, C)$. Consider scenario $[P]+\mathcal{L}$ that goes from A to C and satisfies:

$$w([P]+\mathcal{L}) = 2 + w(\mathcal{L}) \leq 2 + d(B, C) \leq d(A, C).$$

This is in fact a scenario that proves the lemma for A and C in this case. Therefore, in all cases we have a suitable scenario, and the proof is complete. □

Corollary 1. *For any two genomes A and C over the same genes, there is at least one scenario that goes from A to C and furthermore:*

$$w(A, C) = d(A, C).$$

Proof. The existence of such a scenario is a consequence of Theorem 3. In addition, the scenario $\mathcal{L}$ given by this theorem satisfies:

$$w(\mathcal{L}) \leq d(A, C).$$

However, Theorem 2 states that the weight of any such scenario is at least $d(A, C)$:

$$d(A, C) \leq w(\mathcal{L}).$$

We conclude that $\mathcal{L}$ is then an optimal scenario and that

$$w(A, C) = w(\mathcal{L}) = d(A, C).$$

□

Chapter 4

Intermediate Genomes

In this chapter we study intermediate genomes. We prove here that they can be characterized both as optimal scenario members, and as genomes for which the triangle inequality becomes an equality. They are also the medians of two genomes.

4.1 Triangle Equality

We say that a genome B is **intermediate** between genomes A and C, all over the same genes, when equality occurs in the triangle inequality as follows:

$$d(A, C) = d(A, B) + d(B, C).$$

Intermediate genomes minimize $d(A, B) + d(B, C)$, and are therefore referred to as median genomes as well.

If we recall the way the triangle inequality was established, there are important conclusions to be drawn when it is an equality, as the next lemma makes clear.

Lemma 27. *If $d(A, C) = d(A, B) + d(B, C)$ for three genome matrices over the same genes A, B, and C, then $\mathrm{im}(B - A) \subseteq \mathrm{im}(C - A)$ and $\mathrm{im}(C - B) \subseteq \mathrm{im}(C - A)$.*

Proof. Recall that, in general,

$$\begin{aligned} d(A, C) &= \dim \mathrm{im}(C - A) \\ &\leq \dim(\mathrm{im}(B - A) + \mathrm{im}(C - B)) \\ &\leq \dim \mathrm{im}(B - A) + \dim \mathrm{im}(C - B) \\ &= d(A, B) + d(B, C). \end{aligned}$$

In order to have an equality above, all intermediate terms in these equations must be equal. In particular, this implies:

$$\mathrm{im}(B - A) + \mathrm{im}(C - B) = \mathrm{im}(C - A),$$

which implies both containments in the lemma's statement. □

But when the image of a matrix is contained in the image of another matrix, the matrices themselves are related by a product:

Lemma 28. *If X, Y are $k \times k$ matrices such that $\mathrm{im}\, X \subseteq \mathrm{im}\, Y$, then there is a $k \times k$ matrix R such that $X = YR$.*

Proof. Let $e_1, e_2, \ldots, e_k$ be a basis of R^k. Since $\operatorname{im} X \subseteq \operatorname{im} Y$, for every vector e_i there is a vector f_i such that $Xe_i = Yf_i$, for $1 \leq i \leq k$. Let R be the matrix that takes each e_i to f_i. Then:

$$YRe_i = Yf_i = Xe_i,$$

which shows that $YR = X$ in a basis. This is enough to guarantee that $X = YR$ everywhere, that is, they are the same matrix. □

4.2 Linear Combination

The results in the previous section lead to the following result.

Lemma 29 (Linear Combination)**.** *If B is an intermediate genome between A and C, then there is a matrix S such that:*

$$B = SA + (I - S)C.$$

Proof. By previous observations we know that $\operatorname{im}(C - B) \subseteq \operatorname{im}(C - A)$. From the previous lemma we conclude that there is a matrix R such that $C - B = (C - A)R$. Now:

$$B = B^t = (C + (A - C)R)^t = C^t + R^t(A - C)^t = R^tA + (I - R^t)C,$$

which is our result if we make $S = R^t$. □

4.3 Common Elements

The linear combination result in Lemma 29 has important consequences in terms of common elements in A and C: common images, common telomeres, and common adjacencies. All of these have to be shared among all intermediate genomes.

Corollary 2 (Common image)**.** *If B is an intermediate genome between A and C and $Ax = Cx$ for a given vector x, then $Bx = Ax = Cx$.*

Proof. From the fact that $B = SA + (I - S)C$ and $Ax = Cx$ we get:

$$Bx = SAx + (I - S)Cx = SCx + Cx - SCx = Cx = Ax.$$

□

Corollary 3 (Common telomere)**.** *If x is a common telomere of genomes A and C over the same genes, then every intermediate genome B between A and C will have x as a telomere.*

Proof. It suffices to recall that a telomere is an extremity x such that $Ax = x$ and apply Corollary 2. □

Corollary 4 (Common adjacency)**.** *If xy is a common adjacency of genomes A and C over the same genes, then every intermediate genome B between A and C will have xy as an adjacency.*

Proof. It suffices to recall that an adjacency xy entails $Ax = y$ and apply Corollary 2. □

4.4 Intermediate Genomes as Optimal Scenario Members

Intermediate genomes are exactly those that participate in optimal scenarios. The following result proves this claim.

Theorem 4. *Given genomes A and C over the same genes, a genome B over the same genes is an intermediate genome between A and C if and only if it belongs to some optimal scenario going from A to C.*

Proof. Let's first prove that, if B is part of an optimal scenario going from A to C, then $d(A, C) = d(A, B) + d(B, C)$. Let $\mathcal{L} = [B_0, B_1, \ldots, B_k]$ be an optimal scenario with $B_0 = A$ and $B_k = C$, and suppose $B = B_i$ for some i between 0 and k. Then $\mathcal{L}_1 = [B_0, \ldots, B_i]$ is a scenario going from A to B. It is an optimal scenario, but we don't need this fact. Just by being a scenario it satisfies:

$$d(A, B) \leq w(\mathcal{L}_1),$$

as stated by Theorem 2. Likewise, there is a scenario $\mathcal{L}_2 = [B_i, \ldots, B_k]$ from B to C, and hence

$$d(B, C) \leq w(\mathcal{L}_2),$$

by the same theorem. Adding these last two equations, we get:

$$d(A, B) + d(B, C) \leq w(\mathcal{L}_1) + w(\mathcal{L}_2) = w(\mathcal{L}) = d(A, C),$$

because $\mathcal{L}$ is optimal (Theorem 1). Using the triangle inequality, we conclude that B is in fact an intermediate genome.

Conversely, if B is an intermediate genome, then

$$d(A, C) = d(A, B) + d(B, C).$$

Since $d(A, B) = w(A, B)$ (Theorem 1), there is an optimal scenario $\mathcal{L}_1$ from A to B such that:

$$d(A, B) = w(\mathcal{L}_1).$$

Similarly, there is an optimal scenario $\mathcal{L}_2$ from B to C such that:

$$d(B, C) = w(\mathcal{L}_2).$$

Combining these two scenarios, we get a third scenario $\mathcal{L}$, obtained from $\mathcal{L}_1 + \mathcal{L}_2$ by removing one of the copies of the repeated B, such that:

$$w(\mathcal{L}) = w(\mathcal{L}_1) + w(\mathcal{L}_2) = d(A, B) + d(B, C) = d(A, C).$$

This last equation shows that $\mathcal{L}$ is actually an optimal scenario going from A to C. Since B is in $\mathcal{L}$, we have our result. □

Chapter 5

Minimax Genomes

The notion of minimax genome is explored in this chapter. We establish a lower bound for the minimax score, and show exactly the cases where it is possible to achieve such a score. We also show that, in any case, it is always possible to find a genome within 1 unit of the lower bound.

5.1 Definitions

Given three genomes A, B, and C, a **minimax** genome for the three is a genome M that minimizes:

$$\mathrm{sc}(M; A, B, C) = \max(d(A, M), d(B, M), d(C, M)),$$

where $\mathrm{sc}(M; A, B, C)$ is called the **score** of M relative to A, B, and C.

We can also look for minimax genomes of two input genomes A and B, minimizing:

$$\mathrm{sc}(M; A, B) = \max(d(A, M), d(B, M)).$$

The triangle inequality gives almost immediately a lower bound on the minimax score:

Lemma 30. *For any three genomes A, B, and M we have:*

$$\mathrm{sc}(M; A, B) \geq \frac{d(A, B)}{2}.$$

Proof. Notice that:

$$d(A, B) \leq d(A, M) + d(B, M) \leq 2 \max(d(A, M), d(B, M)) = 2\,\mathrm{sc}(M; A, B).$$

From this, the statement easily follows. □

In fact, since the score is always an integer, we can strengthen this result and claim that:

$$\mathrm{sc}(M; A, B) \geq \left\lceil \frac{d(A, B)}{2} \right\rceil.$$

Our goal in this chapter is to find minimax genomes for A and B.

5.2 Examples

Consider the following pairs of matrices, for any given dimension. In this section, we will look at matrices that are not necessarily genome matrices, because they may not have an even dimension. Notice however that the definitions of distance, intermediate matrices, and minimax matrices work for them. Here are the matrices:

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, B_3 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$A_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, B_5 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and so on. They have an interesting pattern around the diagonal. The distance between A_k and B_k is $k-1$.

Let us first write general formulas for A_k and B_k. Assume that the k extremities involved are called $e_1, e_2, \ldots, e_k$. We have:

$$\begin{aligned} A_k e_1 &= e_1 \\ A_k e_2 &= e_3 \\ A_k e_3 &= e_2 \\ &\ldots \\ A_k e_{2j} &= e_{2j+1} \\ A_k e_{2j+1} &= e_{2j} \\ &\ldots \\ A_k e_k &= \begin{cases} e_k & \text{if } k \text{ is even} \\ e_{k-1} & \text{if } k \text{ is odd} \end{cases} \end{aligned}$$

and

$$\begin{aligned} B_k e_1 &= e_2 \\ B_k e_2 &= e_1 \\ &\ldots \\ B_k e_{2j} &= e_{2j-1} \\ B_k e_{2j+1} &= e_{2j+2} \\ &\ldots \\ B_k e_k &= \begin{cases} e_{k-1} & \text{if } k \text{ is even} \\ e_k & \text{if } k \text{ is odd} \end{cases} \end{aligned}$$

As a result, $B - A$ can be written as:

$$\begin{aligned}
(B_k - A_k)e_1 &= e_2 - e_1 \\
(B_k - A_k)e_2 &= e_1 - e_3 \\
&\dots \\
(B_k - A_k)e_{2j} &= e_{2j-1} - e_{2j+1} \\
(B_k - A_k)e_{2j+1} &= e_{2j+2} - e_{2j} \\
&\dots \\
(B_k - A_k)e_k &= \begin{cases} e_{k-1} - e_k & \text{if } k \text{ is even} \\ e_k - e_{k-1} & \text{if } k \text{ is odd} \end{cases}
\end{aligned}$$

Notice that there is a single AB-orbit in this example, because of the following fact. From the definition of orbits, we see that all odd-numbered extremities e_{2j+1} belong to the same AB-orbit, because of the images of the even-numbered extremities e_{2j} under $B - A$. Also, all even-numbered extremities e_{2j} are AB-equivalent because of the images of odd-numbered extremities e_{2j+1} under $B - A$. On top of that, e_1 and e_2 are in the same orbit, as witnessed by the image of e_1 under $B - A$. It follows that all extremities are in the same, large AB-orbit, and hence the distance is $d(A, B) = k - 1$.

Now let's compute the distance between I_k, the identity matrix on k dimensions, with A_k and B_k. For A_k and I_k we have:

$$\begin{aligned}
(I_k - A_k)e_1 &= e_1 - e_1 = 0 \\
(I_k - A_k)e_2 &= e_2 - e_3 \\
&\dots \\
(I_k - A_k)e_{2j} &= e_{2j} - e_{2j+1} \\
(I_k - A_k)e_{2j+1} &= e_{2j+1} - e_{2j} \\
&\dots \\
(I_k - A_k)e_k &= \begin{cases} e_k - e_k = 0 & \text{if } k \text{ is even} \\ e_k - e_{k-1} & \text{if } k \text{ is odd} \end{cases}
\end{aligned}$$

We notice that e_1 is on a singleton orbit by itself, and then $\{e_{2j}, e_{2j+1}\}$ are additional orbits for $2 \leq 2j \leq k - 1$. Finally, if k is even, e_k will form another singleton orbit, or, if k is odd, e_k was already counted in an orbit with e_{k-1}. We have therefore 1 orbit for e_1, at least $\lfloor (k-1)/2 \rfloor$ additional orbits, and perhaps one last orbit for e_k if k is even, which can be included in the additional orbits if we round $(k-1)/2$ up to the nearest integer instead of rounding it down. The final orbit count is then $1 + \lceil (k-1)/2 \rceil$. The distance is then $d(A_k, I_k) = k - 1 - \lceil (k-1)/2 \rceil$ or $\lfloor (k-1)/2 \rfloor$.

For B_k and I_k we have:

$$\begin{aligned}
(I_k - B_k)e_1 &= e_1 - e_2 \\
(I_k - B_k)e_2 &= e_2 - e_1 \\
&\dots \\
(I_k - B_k)e_{2j} &= e_{2j} - e_{2j-1} \\
(I_k - B_k)e_{2j+1} &= e_{2j+1} - e_{2j+2} \\
&\dots \\
(I_k - B_k)e_k &= \begin{cases} e_k - e_{k-1} & \text{if } k \text{ is even} \\ e_k - e_k = 0 & \text{if } k \text{ is odd} \end{cases}
\end{aligned}$$

k	$d(A_k, B_k)$	$d(A_k, I_k)$	$d(B_k, I_k)$
2	1	0	1
3	2	1	1
4	3	1	2
5	4	2	2

Table 5.1: Distances between minimax genome I_k and matrices A_k, B_k for $k = 2$ to 5.

We notice that $\{e_{2j-1}, e_{2j}\}$ are orbits for $2 \leq 2j \leq k$. Also, if k is odd, e_k will form a singleton orbit, or, if k is even, e_k was already counted in an orbit with e_{k-1}. We have therefore at least $\lfloor k/2 \rfloor$ orbits, and perhaps one last orbit for e_k if k is odd, which can be counted with the other orbits if we round $k/2$ up to the nearest integer instead of rounding it down. The final orbit count is then $\lceil k/2 \rceil$. The distance is then $d(B_k, I_k) = k - \lceil k/2 \rceil$ or $\lfloor k/2 \rfloor$.

Table 5.1 summarizes the results up to $k = 5$.

We realize that for these matrices there is a minimax genome which is also an intermediate genome between A_k and B_k, for every k, namely, the genome I_k.

5.3 A Conjecture

As we said earlier, in the beginning of Section 1.1, any ordering of the extremities can be used to create a genome matrix. This means that, by changing the ordering of extremities, we get other matrices representing essentially the same genome. This raises an interesting question: maybe all genomes will, by choosing a suitable ordering, produce matrices just like the ones in Section 5.2. If this is true our problem is solved, since for this case we already have minimax matrices satisfying the lower bound.

Let's recall that changing the extremity ordering is equivalent to taking a genome matrix A and computing PAP^{-1}, where P is a permutation matrix, that is, P is a square matrix of zeros and ones that satisfies:

$$PP^t = I,$$

where I is the identity matrix and P^t indicates the transpose of P, that is, the matrix obtained from P by exchanging rows and columns. Notice that if $PP^t = I$, then P^tP is also I, because of a general property of matrices stating that their "left inverse" must be equal to their "right inverse". Notice also that, since $P^{-1} = P^t$ for a permutation matrix, we can write PAP^t instead of PAP^{-1} and get the same result. This is also true for odd-dimensional matrices, which cannot represent genomes.

Lemma 31. *If A is a genome and P is a permutation matrix, then PAP^t is also a genome.*

Proof. We need to verify that:

- PAP^t is a 0-1 matrix
- PAP^t is symmetric
- PAP^t squared is equal to I

To ensure that PAP^t is a 0-1 matrix, recall that both A and P (and hence P^t as well) are 0-1 matrices with exactly one 1 in each row, and exactly one 1 in each column. Looking at the product PA, we see that each entry in this product is the scalar product of a row from P with a column from A. If the position of the 1 in these row and column match, the result will be 1; otherwise, it will be zero. Therefore, each entry in PA is 0 or 1. Moreover, since each row of A has just one 1, a row in the product will have exactly one 1. The same reasoning works for columns.

We conclude that PA is a 0-1 matrix with exactly one 1 in each row and in each column. A repetition of this argument shows us that the same will be true for PAP^t.

For the second part, notice that $(PAP^t)^t = (P^t)^t A^t P^t = PAP^t$, since A is symmetric. Hence, PAP^t is symmetric.

Fir the last part, notice that $(PAP^t)^2 = PAP^tPAP^t = PAAP^t = PP^t = I$, since $A^2 = I$ and $PP^t = I$. Hence, PAP^t is a bona fide genome matrix. □

However, our conjecture is not true. Just notice that if a matrix A has a 1 in the diagonal, this means that for some extremity x we have:

$$Ax = x,$$

that is, x is a telomere. Then PAP^t will have a 1 in the diagonal too, in the column corresponding to Px, because:

$$PAP^tPx = PAIx = PAx = Px,$$

since $P^tP = I$. Then extremity Px is a telomere of PAP^t, and PAP^t will have a 1 in the diagonal too. But since there are genome matrices without 1's in the diagonal (see for instance some of the B_k's), we conclude that the examples of Section 5.2 do not represent all genomes by extremity reordering.

5.4 Minimax Lower Bound: Not Always Attainable

We are interested in working on the following conjecture in this section:

Conjecture 1. *For any two genomes A and B over the same genes, there is at least one genome M over the same genes that satisfies:*

$$d(A, M) = \lceil d(A, B)/2 \rceil$$

and

$$d(B, M) = \lfloor d(A, B)/2 \rfloor.$$

This genome would of course be a minimax genome, since it would attain the lower bound established in Section 5.1. However, this is false, as can be seen from the following example representing genomes that differ by a double swap:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

To compute their distance, let's subtract B from A:

$$A - B = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{bmatrix}.$$

This matrix has rank 2. Both A and B are circular genomes, since they do not have telomeres. Now for a circular genome such as A, the only genomes at distance 1 from it are the ones obtained by cutting an adjacency, since no extra adjacencies can be added to A. Genome A has only two adjacencies, so there are just two genomes at distance 1 from it, namely:

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

However, it can be readily verified that none of these two genomes is at distance 1 from B. In fact, they are both at distance 3 from B. We conclude that the minimax conjecture is **not** true.

5.5 Finding Minimax Genomes

In this section we will find minimax genomes for every pair of input genomes A and C. The discussion will be divided in two cases, according to the genomes A and C being or not co-tailed.

5.5.1 Co-Tailed Genomes

When A and C are co-tailed, we not always get a minimax genome satisfying the lower bound, but we can get within 1 unit of it. Let's begin by studying properties of intermediate genomes between two co-tailed ones.

Lemma 32. *If A and C are co-tailed genomes and B is an intermediate genome between A and C, then B is co-tailed with A and C.*

Proof. It suffices to show that B is co-tailed with A. Suppose for a moment that B is not co-tailed with A. Then either A has a telomere that B doesn't, or B has a telomere that A doesn't. The first case is ruled out by Corollary 3, because a telomere of A would also be a telomere of C, since they are co-tailed, and would have to be shared by all intermediate genomes.

So let's assume that B has a telomere x not shared by A. Since B is intermediate between A and C, according to Lemma 27, we can write:

$$\mathrm{im}(B - A) \subseteq \mathrm{im}(C - A).$$

This observation leads to

$$(B - A)x = x - y \in \mathrm{im}(B - A) \subseteq \mathrm{im}(C - A),$$

where $y = Ax$ is an extremity distinct from x.

Let $z = Cx$. We know that $y \sim_{AC} z$, since $y = ACz$. According to Lemma 6, vectors in $\mathrm{im}(C - A)$ have the same coefficients with respect to AC-equivalent extremities. Therefore:

$$y^t(x - y) = z^t(x - y)$$

or

$$-1 = z^t x - z^t y.$$

Notice that $z \neq x$, otherwise x would be a telomere of C and not a telomere of A, contradicting the fact that A and C are co-tailed. This implies $z^t x = 0$. Also, $z \neq y$, otherwise xy would be

a common adjacency of A and C, and then B, being an intermediate genome between A and C, would have to have this adjacency also, which is not the case since $Bx = x$. This implies $z^t y = 0$ as well. Then the last formula becomes

$$-1 = 0,$$

which is impossible. This contradiction shows that B cannot have telomeres not shared with A and C. □

Only double swaps occur in optimal sorting scenarios of co-tailed genomes. This leads to a parity restriction.

Lemma 33. *If A and C are co-tailed genomes, and $\mathcal{L} = [B_0, B_1, \ldots, B_k]$ is an optimal scenario going from A to C, then $d(A, C) = 2k$.*

Proof. According to Lemma 32, all B_i's are co-tailed with A, so none of the operations $B_{i+1} - B_i$ can be cuts or joins. Therefore, we have $r(B_{i+1} - B_i) = 2$ for $0 \leq i \leq k - 1$. But then

$$d(A, C) = w(\mathcal{L}) = \sum_{i=0}^{k-1} r(B_{i+1} - B_i) = \sum_{i=0}^{k-1} 2 = 2k.$$

□

Corollary 5. *If A and C are co-tailed genomes, then $d(A, C)$ is even.*

It is easy to find minimax genomes if the parity is right. However, when the parity does not help, we are forced to take the second best, which is 1 unit off the lower bound.

Lemma 34. *If A and C are co-tailed genomes and $d(A, C)/2$ is even, then there is a genome B satisfying the minimax lower bound.*

Proof. Let $[B_0, B_1, \ldots, B_k]$ be an optimal scenario going from A to C. We know that $d(A, C) = 2k$ from Lemma 33. Since $k = d(A, C)/2$ is even, we can write $k = 2m$ for some integer m. It is then straightforward to verify that B_m is the sought intermediate genome satisfying the minimax lower bound. □

Lemma 35. *If A and C are co-tailed genomes and $d(A, C)/2$ is odd, then there is no genome B satisfying the minimax lower bound.*

Proof. If such a genome B existed, then we would have:

$$d(A, B) = d(B, C) = d(A, C)/2.$$

This implies that B would be an intermediate genome between A and C. By Lemma 32, B would be co-tailed with A. But then, by Corollary 5, $d(A, B) = d(A, C)/2$ would have to be even, contradicting the hypothesis. □

Lemma 36. *For any two genomes A and C, there is an intermediate genome B such that*

$$\lceil d(A, C)/2 \rceil \leq d(A, B) \leq \lceil d(A, C)/2 \rceil + 1.$$

Proof. If $A = C$ the result is clear taking $B = A = C$. If $A \neq C$, let $[B_0, B_1, \ldots, B_k]$ be an optimal scenario going from A to C and take i as the smallest index such that $d(A, B_i) \geq \lceil d(A, C)/2 \rceil$. We claim that $B = B_i$ is the sought genome. Notice that B is an intermediate genome between A and C because it is a member of an optimal scenario going from A to C. Moreover, the first inequality in the lemma statement is satisfied because of the choice if i.

For the second equality, notice that, by the minimality of i, we have:

$$d(A, B_{i-1}) < \lceil d(A, C)/2 \rceil.$$

Genome B_{i-1} exists since $A \neq C$ implies $\lceil d(A, C)/2 \rceil \geq 1$, so i cannot be zero. Given that in any scenario the steps have weight 1 or 2, we know that $d(B_{i-1}, B_i) \leq 2$. It follows that

$$d(A, B_i) \leq d(A, B_{i-1}) + d(B_{i-1}, B_i) \leq d(A, B_{i-1}) + 2 < \lceil d(A, C)/2 \rceil + 2$$

or

$$d(A, B_i) \leq \lceil d(A, C)/2 \rceil + 1,$$

since both sides are integers. □

5.5.2 Genomes Not Co-Tailed

If A and C are not co-tailed, then there are intermediate genomes at any feasible distance between A and C. To ascertain that, we need a few preliminary lemmas on operation switch and other properties.

Lemma 37. *Let A be a genome, P a cut applicable to A, and Q a double swap applicable to $A + P$. Then Q is applicable to A.*

Proof. Let $Q = W(x, y, z, w)$. We know that Q is applicable to $A + P$, which means that $A + P$ has adjacencies xw and yz. Since P is a cut, which only removes adjacencies, xw and yz must have been present in A as well, leading to the conclusion that Q can be applied to A. □

An analogous result is valid for joins, saying that joins can be brought back through double swaps, but we won't need it now.

Lemma 38. *Let A and C be two genomes not co-tailed. Then, for every integer i such that $0 \leq i \leq d(A, C)$ there is an intermediate genome B between A and C with $d(A, B) = i$.*

Proof. By induction on $d(A, C)$. The base case is $d(A, C) = 1$, because A and C are not co-tailed and hence cannot be equal. The statement is clearly true for $d(A, C)$ because in this case we only have two possibilities for i, namely, $i = 0$ or $i = 1$, and we can take $B = A$ for $i = 0$ and $B = C$ for $i = 1$.

Now assume $d(A, C) \geq 2$ and consider an integer i such that $0 \leq i \leq d(A, C)$. Since A and C are not co-tailed, there is either a telomere in A not shared by C or a telomere in C not shared by A. Without loss of generality, we may assume that there is a telomere in C not shared by A, otherwise we can just exchange A and C and i with $d(A, C) - i$.

Given that there is a telomere in C that is not an A-telomere, Lemma 23 gives us a cut P applicable to A such that $A + P$ is an intermediate genome between A and C. If $A + P$ is not co-tailed with C, we can apply the induction hypothesis to $A + P$ and C and get intermediate genomes at an arbitrary distance j from $A + P$, provided that $0 \leq j \leq d(A + P, C) = d(A, C) - 1$, which will be at distance $j + 1$ from A. This covers all the distances we need except 0, for which we can take $B = A$.

Now if $A+P$ is co-tailed with C, then they are distinct, since $d(A, A+P) = 1$ and $d(A, C) \geq 2$. Lemma 26 guarantees that there is a double swap Q applicable to A yielding an intermediate genome $A+P+Q$ between $A+P$ and C. However, according to Lemma 37, a cut can go forward past a double swap, which means that Q is applicable to A. The resulting genome, $A+Q$, is intermediate between A and C because $A+Q+P$ is just another way of getting to $A+P+Q$, which we know is intermediate between A and C. We can then apply the induction hypothesis to $A+Q$ and C, which are not co-tailed since $A+Q$ is co-tailed with A, obtaining intermediate genomes at distances i from A for $2 \leq i \leq d(A, C)$. For $i = 0$ we have A, and for $i = 1$ we have $A+P$. This completes the induction step and the proof of our lemma. □

5.6 Main Result

Theorem 5. *Let A and C be arbitrary genome matrices over the same genes. Then:*

1. *If A and C are not co-tailed, then there is a genome matrix B such that:*

$$d(A, B) = \left\lceil \frac{d(A, C)}{2} \right\rceil$$

and

$$d(B, C) = \left\lfloor \frac{d(A, C)}{2} \right\rfloor.$$

2. *If A and C are co-tailed and $d(A, C)$ is a multiple of 4, then there is a genome matrix B such that:*

$$d(A, B) = \frac{d(A, C)}{2}$$

and

$$d(B, C) = \frac{d(A, C)}{2}.$$

3. *If A and C are co-tailed and $d(A, C)$ is not a multiple of 4, then there is no genome matrix B such that:*

$$d(A, B) = \frac{d(A, C)}{2}$$

and

$$d(B, C) = \frac{d(A, C)}{2}.$$

However, there is a genome matrix B such that:

$$d(A, B) = \frac{d(A, C)}{2} + 1$$

and

$$d(B, C) = \frac{d(A, C)}{2} - 1.$$

Proof. Part 1 is a consequence of Lemma 38, since $0 \leq \lceil d(A, C)/2 \rceil \leq d(A, C)$. Part 2 is a consequence of Lemma 34. Part 3 is a consequence of Lemmas 35 and 36. □

Chapter 6

Parity

In this chapter we show an interesting property of the rank distance for genome matrices. For any three genome matrices, the sum of their pairwise distances has to be an even number. In fact, we will also see that this property extends to arbitrary orthogonal matrices as well.

6.1 Parity Property

To prove the intriguing parity result, we will need a few auxiliary lemmas. The proof that we will offer relies on properties of AB-orbits. We already know that the number of AB-orbits is related to the distance $d(A, B)$. These orbits are also related to the determinant of AB in a manner that will be explained in the sequel.

6.2 The Determinant

In this section we recall some well-know properties of eigenvalues and eigenvectors of permutation matrices. Garca-Planas and Magret provide a recent, clear account of the subject [5]. See also the earlier work by Wieand [14].

Let us start with a lemma relating AB-orbits to eigenvalues of AB. Notice that in this lemma we will consider A and B as matrices with complex coefficients, although the coefficients can be thought of as being taken from any field. But we need the field of complex numbers here to be able to encompass all eigenvalues of AB.

Lemma 39. *If $\mathcal{S} = \{x, ABx, (AB)^2x, \ldots, (AB)^{k-1}x\}$ is an AB-orbit with k elements, then the subspace generated by $\mathcal{S}$ contains k eigenvectors of AB, each corresponding to an eigenvalue of the form λ_k^j, for $j = 0, 1, \ldots, k-1$, where $\lambda_k = e^{2\pi i/k}$ and i is the imaginary unit.*

Proof. Consider the vectors v_j, for $j = 0, 1, \ldots, k-1$, defined as:

$$v_j = \sum_{m=0}^{k-1} \lambda_k^{-jm} (AB)^m x.$$

These vectors certainly belong to $\langle \mathcal{S} \rangle$, the subspace generated by $\mathcal{S}$, since they are linear com-

binations with complex coefficients of extremities in $\mathcal{S}$. Let's compute ABv_j:

$$\begin{aligned} ABv_j &= AB\sum_{m=0}^{k-1}\lambda_k^{-jm}(AB)^m x \\ &= \sum_{m=0}^{k-1}\lambda_k^{-jm}(AB)^{m+1}x \\ &= \lambda_k^j\sum_{m=0}^{k-1}\lambda_k^{-j(m+1)}(AB)^{m+1}x \\ &= \lambda_k^j\sum_{m=1}^{k}\lambda_k^{-jm}(AB)^m x. \end{aligned}$$

But since the last term of this last sum is $\lambda_k^{-jk}(AB)^k x = (\lambda_k^k)^{-j}x = x$, we have:

$$ABv_j = \lambda_k^j v_j,$$

showing that v_j is actually an eigenvector of AB for the eigenvalue λ_k^j. □

The next lemma is just an auxiliary computation on complex roots of the unit.

Lemma 40. *For any integer $k \geq 1$, the product of all k-th roots of 1:*

$$p_k = \prod_{j=0}^{k-1}\lambda_k^j$$

is equal to -1 if k is even and equal to $+1$ if k is odd, where $\lambda_k = e^{2\pi i/k}$ and i is the imaginary unit.

Proof. The product p_k can be written as:

$$\prod_{j=0}^{k-1}\lambda_k^j = \lambda_k^{\sum_{j=0}^{k-1} j} = \lambda_k^{(k-1)k/2}.$$

Now if k is even, we have:

$$\lambda_k^{k/2} = e^{\pi i} = -1$$

and hence

$$\lambda_k^{(k-1)k/2} = (\lambda_k^{k/2})^{k-1} = (-1)^{k-1} = -1,$$

since $k-1$ is odd.

On the other hand, if k is odd, then $(k-1)/2$ is an integer and therefore:

$$\lambda_k^{(k-1)k/2} = (\lambda_k^k)^{(k-1)/2} = (+1)^{(k-1)/2} = +1.$$

□

We are now ready for one of the main results, relating orbits to the determinant.

Lemma 41. *For any two genomes A and B we have:*

$$\det(AB) = (-1)^q,$$

where q is the number of even-sized AB-orbits.

Proof. The determinant $\det(AB)$ is the product of all eigenvalues of AB considered with their respective multiplicities. For each AB-orbit $\mathcal{S}$, the product of the eigenvalues related to the eigenvectors generated by $\mathcal{S}$ is either -1 if $|\mathcal{S}|$ is even, or $+1$ if $|\mathcal{S}|$ is odd, according to Lemmas 39 and 40. It follows that the product of all eigenvalues of AB with their multiplicities will be $(-1)^q$, where q is the number of orbits that contribute with -1 factors to this product, that is, the even-sized ones. □

6.3 Determinant and Distance

We are now ready advance our developments, relating the determinant of AB to the distance $d(A, B)$.

Lemma 42. *For any two genomes A and B we have:*

$$\det(AB) = (-1)^{d(A,B)}.$$

Proof. Let p be the number of AB-orbits, and q be the number of even-sized AB-orbits. Notice first that since the total size of all AB-orbits is $2n$, an even number, we must have an even number of odd-sized AB-orbits, otherwise the total wouldn't be even. This shows that p and q have the same parity, because their difference (the number of odd-sized AB-orbits) is an even number.

It follows that we can write:

$$(-1)^q = (-1)^p.$$

Now:

$$\det(AB) = (-1)^q = (-1)^p = (-1)^{2n-2p+p} = (-1)^{2n-p} = (-1)^{d(A,B)}.$$

□

And now the main result of the chapter.

Theorem 6. *For any three genomes A, B, and C we have:*

$$d(A, B) + d(A, C) + d(C, B) \equiv 0 \pmod 2.$$

Proof. Since the three matrices are involutions, we have $ABBCCA = I$, and therefore:

$$1 = \det(ABBCCA) = \det(AB)\det(BC)\det(CA) = (-1)^{d(A,B)+d(B,C)+d(C,A)}.$$

Since each distance is an integer, it follows that

$$d(A, B) + d(B, C) + d(C, A) \equiv 0 \pmod 2.$$

□

6.4 Extension to orthogonal matrices

In this section, we will generalize Theorem 6 to orthogonal matrices. All genome matrices are orthogonal matrices, but there are many orthogonal matrices that are not genomic matrices.

We begin by giving a general overview of the proof for orthogonal matrices. It is known that every orthogonal matrix T is the product of certain special orthogonal matrices called **Householder transformations**, defined in the sequel. This is a special case of the celebrated Cartan-Dieudonné theorem [4].

Perhaps less widely known it the fact that we can always use exactly k factors in this product, where $k = d(T, I)$. This is Theorem 7 below. Since each Householder transformation has determinant -1, as we will see in a moment, it follows that $\det(T) = (-1)^{d(T,I)}$, or, for arbitrary orthogonal matrices A and B, $\det(A^{-1}B) = (-1)^{d(A,B)}$. Here it is important to notice that $d(A, B) = d(A^{-1}B, I)$. With this result, the idea of the proof given for Theorem 6 extends easily to orthogonal matrices: it suffices to write $I = A^{-1}BB^{-1}CC^{-1}A$ and then follow essentially the same steps as in the original proof.

To make this overview concrete, let us now proceed with the definition of Householder transformations.

Definition 11. *Let $v \neq 0$ be an $n \times 1$ vector. The **Householder transformation along** v is defined as the matrix*

$$H(v) = I - 2vv^T/v^Tv.$$

*Any matrix of the form $H(v)$ for a non-zero vector v is called a **Householder transformation**.*

Householder transformations have a number of interesting properties:

- $H(v) = H(\alpha v)$, for any non-zero scalar α
- $H(v)u = u$ if and only if $u \perp v$
- $H(v)^T = H(v)$
- $d(H(v), I) = 1$
- $H(v)^2 = I$

These properties are all well-known and easy to prove from the definitions, with straightforward calculations. We proceed now with an important result related to the factoring of an orthogonal matrix as a product of such hyperplane reflections.

Lemma 43. *If $T \neq I$ is an orthogonal matrix, there exists a Householder transformation H such that $d(HT, I) < d(T, I)$.*

Proof. Take any unitary vector u in $\ker(T - I)^\perp$, which is not equal to $\{0\}$ by hypothesis, and define $H = H(v)$, where $v = Tu - u$. Now, if $w \in \ker(T - I)$, then $Tw = w$ and we can prove that $v^Tw = 0$, which implies $Hw = w$ and $HTw = Hw$. We conclude that $\ker(T - I) \subseteq \ker(HT - I)$.

Notice that $u \notin \ker(T - I)$, otherwise v would be zero.

On the other hand, notice that $Hu = Tu$, since:

$$Hu = (I - 2vv^T/v^Tv)u = u - 2vv^Tu/v^Tv = u - (u - Tu) = Tu.$$

This is because $2vv^Tu = (2 - 2u^TTu)(u - Tu)$ and $v^Tv = (2 - 2u^TTu)$. But if $Hu = Tu$, then $HTu = u$, showing that $u \in \ker(HT - I)$.

We conclude that u belongs to $\ker(HT - I)$ but not to $\ker(T - I)$, leading to

$$\dim \ker(T - I) < \dim \ker(HT - I),$$

which in turn implies that $d(HT, I) = \dim \operatorname{im}(HT - I) < \dim \operatorname{im}(T - I) = d(T, I)$, exactly what we want to prove. □

The next result shows that we can decompose orthogonal matrices into $k = d(T, I)$ Householder transformations.

Theorem 7. *Any orthogonal matrix T can be written as a product of k Householder transformations, where $k = d(T, I)$.*

Proof. We show this by induction on k. If $k = 0$, then $T = I$ is indeed the product of zero Householder transformations. If $k > 1$, then there is at least one unitary vector in $\ker(T - I)^{\perp}$, and therefore Lemma 43 tells us that there is a Householder transformation H such that $d(HT, I) < d(T, I)$. Notice that $d(HT, I)$ cannot be smaller than $d(T, I) - 1$, because of the triangle inequality:

$$d(T, I) \leq d(T, HT) + d(HT, I) = 1 + d(HT, I),$$

since $d(T, HT) = d(I, H) = 1$ due to the fact that T is invertible. It follows that $d(HT, I) = k - 1$ and we can apply the induction hypothesis to HT, which is orthogonal because both H and T are, obtaining $k - 1$ Householder transformations H_1, H_2, ..., H_{k-1} such that

$$HT = H_1 H_2 \dots H_{k-1}.$$

Since $H^2 = I$, we conclude that $T = H H_1 H_2 \dots H_{k-1}$ is a product of k Householder transformations. □

Lemma 42 also has a version for orthogonal matrices, as follows.

Lemma 44. *For any two orthogonal matrices A and B we have:*

$$\det(A^{-1}B) = (-1)^{d(A,B)}.$$

Proof. Notice that $A^{-1}B$, being an orthogonal matrix, can be written as a product of k Householder transformations $H_1, H_2, \dots, H_k$:

$$A^{-1}B = H_1 H_2 \dots H_k,$$

where $k = d(A^{-1}B, I)$. But every Householder transformation $H = H(v)$ can be diagonalized with one eigenvector relative to eigenvalue -1 (in the direction of v), and $n - 1$ eigenvectors relative to eigenvalue $+1$ (in the hyperplane orthogonal to v). Therefore, $\det(H) = -1$. It follows that

$$\det(A^{-1}B) = \det(H_1 H_2 \dots H_k) = \prod_{i=1}^{k} \det(H_i) = (-1)^k.$$

But $k = d(A^{-1}B, I) = d(A, B)$, and we have our result. □

And now the generalization of the parity result to orthogonal matrices.

Theorem 8. *For any three orthogonal matrices A, B, and C we have:*

$$d(A, B) + d(A, C) + d(C, B) \equiv 0 \pmod 2.$$

Proof. We can write $I = A^{-1}BB^{-1}CC^{-1}A$, and therefore:

$$\begin{aligned} 1 &= \det(A^{-1}BB^{-1}CC^{-1}A) \\ &= \det(A^{-1}B)\det(B^{-1}C)\det(C^{-1}A) \\ &= (-1)^{d(A,B)+d(B,C)+d(C,A)}. \end{aligned}$$

Since each distance is an integer, it follows that

$$d(A,B) + d(B,C) + d(C,A) \equiv 0 \pmod 2.$$

□

Chapter 7

Genes and Chromosomes

This chapter introduces genes and chromosomes, which are important structures that genomes are made of.

7.1 Genes

Up to this point, we have talked much about genomes, but we have not said anything about genes or chromosomes. Everything we did does not depend on the fact that the extremities come from genes. Indeed, it is surprising to see how far we could advance without mentioning these elements, which are the main components of genomes.

Our treatment of genomes in this text uses a simplified view of genomes. A genome is in fact a very complex structure. It contains all the genetic information passed to an organism from its parents, or, at the cellular level, from a dividing cell to its daughter cells. Here, we will just look at **genomes** as sets of **chromosomes**, with each chromosome being a succession of oriented **genes**. Figure 7.1 shows a schematic view of a fictitious chromosome with three genes. Figure 7.2 shows one of its matrix representations.

The chromosome depicted in Figure 7.1 is linear, but everything that we will say applies equally well to circular chromosomes. The only difference between linear and circular chromosomes is that a circular chromosome does not have telomeres, i.e., free gene extremities. Every gene extremity in a circular chromosome is adjacent to some other gene extremity. But let us not advance too much with jargon. It is important to define the proper terms first.

A gene is an **oriented** entity that corresponds to a certain contiguous stretch in a chromosome. In Figure 7.1, genes are represented as arrows. The tip of the arrow is called the **head** of the gene and is denoted by adding an h subscript to the gene name. Thus, gene a has head a_h. The other extremity of a gene is its **tail**, denoted by adding a t subscript to the gene name. For instance, gene a has tail a_t. Every gene has a head and a tail.

In our modeling, we assume that genes do not overlap in a chromosome. In fact, gene overlap is relatively common in prokaryotes [7], but rare in mammals and other vertebrates [9, 12]. In cases where this overlap does exist, it is possible to replace the overlapping genes involved by some other oriented marker in the same position in the chromosome, sometimes including several consecutive or overlapping genes, or parts of genes, and choose the markers so that they do not overlap. Nonetheless, we will still keep using the term *gene* to refer to these non overlapping entities, even if some of them are other types of markers in the chromosome.

Assuming that genes do not overlap, each gene extremity is immediately followed or preceded by an extremity of a different gene, except in the ends of linear chromosomes. For instance, in

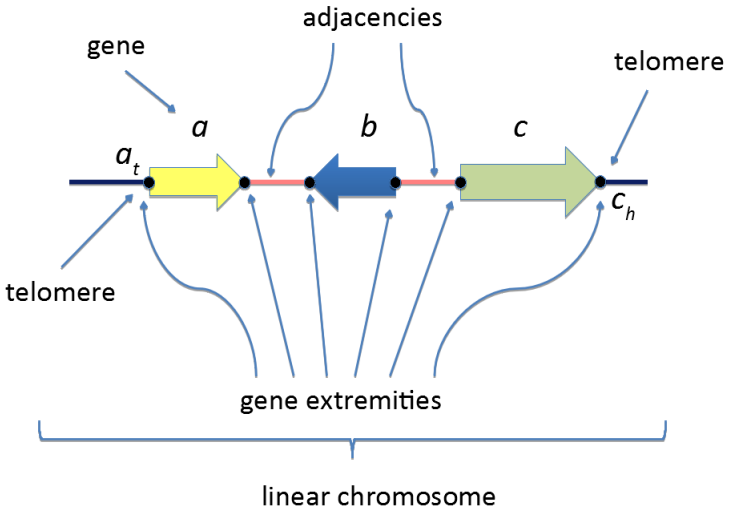

Figure 7.1: A schematic view of a fictitious, linear chromosome with three genes: a, b, and c. Each gene g contributes two extremities, its tail g_t and its head g_h to the extremity pool. Adjacencies are pairs of neighboring extremities in consecutive genes. In this case, $a_h b_h$ and $b_t c_t$ are adjacencies. Telomeres are extremities not involved in adjacencies. In the figure, a_t and c_h are telomeres.

$$
\begin{array}{c}
 \\
a_t \\
a_h \\
b_t \\
b_h \\
c_t \\
c_h
\end{array}
\begin{array}{c}
\begin{array}{cccccc} a_t & a_h & b_t & b_h & c_t & c_h \end{array} \\
\left[\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]
\end{array}
$$

Figure 7.2: Genome matrix corresponding to genome depicted in Figure 7.1.

Figure 7.1 we have:

- a_t is at the end of a linear chromosome
- a_h is followed by b_h in gene b
- b_h is preceded by a_h in gene a
- b_t is followed by c_t in gene c
- c_t is preceded by b_t in gene b
- c_h is at the end of a linear chromosome.

When an extremity is immediately preceded or followed by another extremity, and this other extremity belongs to a different gene, we say that these two extremities form an **adjacency**. In Figure 7.1, the adjacencies are: $a_h b_h$, and $b_t c_t$. Notice that $a_t a_h$ is *not* and adjacency, because, even though a_h immediately follows a_t, the fact that they belong to the same gene precludes the pair being called an adjacency. The order of extremities is not important in adjacencies. Thus, adjacency $a_t a_h$ could be written also as $a_h a_t$.

When an extremity is at the end of a linear chromosome, and therefore is not adjacent to any other extremity, we call it a **telomere**. In Figure 7.1, there are two telomeres: a_t and c_h.

7.2 Multi-genome breakpoint graph

Given a genome, we can draw a graph representing it. In the literature, we can find many, non-equivalent definitions for graphs. For instance, some definitions permit loops and parallel edges, others do not. Also, there is the question of directed and undirected graphs. For us, it is important to be able to represent parallel edges. In our case, when we compare multiple genomes, parallel edges mean that the same adjacency is present in several distinct genomes.

Douglas West defines graphs in a way that permits loops and parallel edges [13]. We will adopt his definition here, which we reproduce below. We will also adopt his general notation on graphs, for things such as degree, etc.

Definition 12 (West, 2001)**.** *A **graph** G is a triple consisting of a **vertex set** $V(G)$, and **edge set** $E(G)$, and a relation that associates with each edge two vertices (not necessarily distinct) called its **endpoints**.*

For a given genome A, we define a graph representation $BG(A)$ for it, called the **multi-genome breakpoint graph** of A, as follows. The vertex set $V(BG(A))$ is the set of all gene extremities of A. The edge set $E(BG(A))$ is the set of all adjacencies of A. And the relation between edges and vertices associates each adjacency with its two gene extremities. As an example, Figure 7.3 shows the graphical representation of the genome depicted in Figure 7.1.

The reason this structure is called multi-genome breakpoint graph is because one can represent several genomes with it, provided they are all over the same genes. Sometimes, edges from distinct genomes are drawn with different colors, or different drawing styles, but this is not technically necessary.

We can denote linear chromosomes as lists of genes, to indicate their relative order in the chromosome, as in $[-b, c]$. Likewise, we can denote circular genomes as lists within parentheses, to stress the fact that it is a circular chromosome, as in $(a, -d)$. A negative sign in a gene indicates that its orientation is reversed with respect to the unsigned (or positive) genes. Figure 7.4 shows

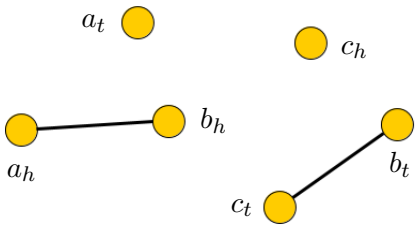

Figure 7.3: Graphical representation of genome in Figure 7.1.

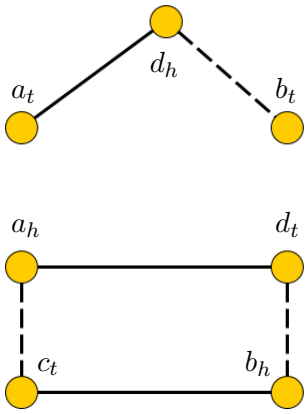

Figure 7.4: Example of multi-genome breakpoint graph with two genomes, namely $A = \{a_t d_h, a_h d_t, b_h c_t\}$ and $B = \{b_t d_h, b_h d_t, a_h c_t\}$. Adjacencies from A are solid, while adjacencies from B are dashed.

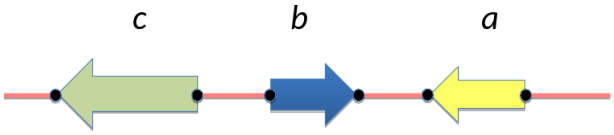

Figure 7.5: The same chromosome as in Figure 7.1, flipped over.

the multi-genome breakpoint graph of genomes A and B, where $A = \{(a, d), [b, c]\}$ and $B = \{[a, c], (b, d)\}$.

If a graph G is the multi-genome breakpoint graph of some genome A, we say that G is a **genome graph**. Notice that genome graphs are **matchings**. Conversely, every matching in a complete graph with vertex set E corresponds to a genome, where E consists of all extremities from a given gene set. Notice also that the connected components of $BG(A, B)$ for any two genomes A and B over the same genes are either paths or cycles.

7.3 Chromosomes

As said earlier, a genome is a collection of chromosomes. If we know the gene names, the adjacencies, and the telomeres of a genome, we are able to reconstruct the structure of each chromosome, and consequently of the entire genome. By "structure" here we mean the order of genes and their relative orientation. We say "relative orientation" because genes only have an orientation with respect to one another in the same chromosome, and not an absolute, say left-to-right, orientation. For instance, the chromosome in Figure 7.1 could be just as well be represented as in Figure 7.5. Notice that the adjacencies and telomeres do not change when we flip a chromosome over, although genes that went left-to-right now go right-to-left and vice-versa.

If we are interested in the number of chromosomes of a genome A, and in whether each chromosome is linear or circular, then we only need a matching defining which head corresponds to which tail to form a gene. This can be encoded in a genome matrix G containing the head-tail associations. This matching is maximum, that is, it covers all extremities. Interestingly enough, G can be seen as a genome, one with only one gene per circular chromosome.

Notice that, the multi-genome breakpoint graph $BG(A, G)$ will have just paths and cycles as connected components, because each vertex will have degree one or two. But the interesting thing is that each one of these components will be a chromosome of A. This fact can be also translated into our matricial interpretation of matchings, as follows.

Theorem 9. *A chromosome of a genome A corresponds to a minimal extremity set $\mathcal{X}$ such that $A\mathcal{X} = \mathcal{X}$ and $G\mathcal{X} = \mathcal{X}$, where G is the head-tail association. Conversely, every such extremity set corresponds to a chromosome of A.*

Also, every chromosome is either equal to a single AG-orbit, or it is the union of exactly two AG-orbits of the same size.

Proof. Let us start by observing that chromosomes are connected components in the multigenome breakpoint graph formed from A and G. As such, if an extremity x belongs to a chromosome, the Ax and Gx must belong to it, too. This observation implies that a chromosome $\mathcal{S}$ must satisfy $A\mathcal{S} \subseteq \mathcal{S}$ and $G\mathcal{S} \subseteq \mathcal{S}$. These containments are only possible if they are equalities, because both A

and G are one-to-one mappings. Moreover, we must require minimality to ensure that unions of several chromosomes are excluded from the definition. With this, we establish the characterization of chromosomes as minimal extremity sets $\mathcal{S}$ such that $A\mathcal{S} = \mathcal{S}$ and $G\mathcal{S} = \mathcal{S}$.

Regarding the relationship between chromosomes and AG-orbits, we may reason as follows. Since a chromosome is characterized as an extremity set stable by both A and G, every chromosome contains an AG-orbit. Recall that, for an AG-orbit $\mathcal{X}$, we have that $A\mathcal{X}$ is also an AG-orbit. If $A\mathcal{X} = \mathcal{X}$, then $G\mathcal{X} = A\mathcal{X} = \mathcal{X}$ as well, and $\mathcal{X}$ is a chromosome. If $A\mathcal{X} = \mathcal{Y} \neq \mathcal{X}$, then $A\mathcal{Y} = \mathcal{X}$ because $A^2 = I$. In this case, $\mathcal{X} \cup \mathcal{Y}$ is a chromosome, and $\mathcal{X}$ has the same size as $\mathcal{Y}$. □

In fact, there is more to say about chromosomes. The fixed ones are linear, and the double-orbit ones are circular.

Theorem 10. *Extremity set $\mathcal{X}$ is a linear chromosome of genome A if and only if $\mathcal{X}$ is an AG-orbit with $A\mathcal{X} = \mathcal{X}$, where G is the head-tail association.*

Proof. If $\mathcal{X}$ is an AG-orbit such that $A\mathcal{X} = \mathcal{X}$, consider any extremity $x \in \mathcal{X}$. Because $A\mathcal{X} = \mathcal{X}$, we have $y = Ax \in \mathcal{X}$. From Lemma 24, we learn that there is a telomere of A or G in $\mathcal{X}$. Since G has no telomeres, we conclude that there is an A-telomere in $\mathcal{X}$. Now $\mathcal{X}$ is an AG-orbit such that $A\mathcal{X} = G\mathcal{X} = \mathcal{X}$, so by Theorem 9 we see that $\mathcal{X}$ is a chromosome. The presence of an A-telomere in it guarantees that it is a linear chromosome.

Conversely, if $\mathcal{X}$ is a linear chromosome, then there is an A-telomere $x \in \mathcal{X}$. Because $Ax = x$, the AG-orbit $\mathcal{X}'$ of x satisfies $A\mathcal{X}' = \mathcal{X}'$, which implies $G\mathcal{X}' = \mathcal{X}'$. Now $\mathcal{X}' \subseteq \mathcal{X}$, because chromosomes are composed of one or two orbits. But $\mathcal{X} \subseteq \mathcal{X}'$ as well, because $\mathcal{X}'$ is an extremity set closed under the mappings A and G, so it contains all the chromosomes not disjoint with it. We conclude that $\mathcal{X}' = \mathcal{X}$, i.e., $\mathcal{X}$ is an AG-orbit with an A-telomere. □

7.3.1 Graph components

There are analogs of Theorems 9 and 10 for the connected components of a breakpoint graph $BG(A, B)$ of two chromosomes.

Theorem 11. *A connected component of $BG(A, B)$ corresponds to a minimal extremity set $\mathcal{X}$ such that $A\mathcal{X} = \mathcal{X}$ and $B\mathcal{X} = \mathcal{X}$. Conversely, every such extremity set corresponds to a connected component of $BG(A, B)$.*

Also, every connected component of $BG(A, B)$ is either equal to a single AB-orbit, or it is the union of exactly two AB-orbits of the same size.

The proof is analogous to the one we presented for Theorem 9. There is also an analogous of Theorem 10.

Theorem 12. *Extremity set $\mathcal{X}$ is a path in $BG(A, B)$ if and only if $\mathcal{X}$ is an AB-orbit with $A\mathcal{X} = \mathcal{X}$.*

Again, its proof closely follows the proof of Theorem 10.

7.3.2 Rank distance from graph elements

It is possible to compute the rank distance using graph parameters. To begin with, notice that the number of edges in $BG(A)$ is the distance between A and I:

$$d(A, I) = |E(BG(A))|.$$

To be able to write the rank distance in terms of graph elements, we need to count the number of paths and cycles in a graph. Let's denote by $P(G)$ the number of connected components in a graph G that are paths. Connected components consisting of a single isolated vertex are considered paths, too. Define also $C(G)$ as being the number of connected components in a graph G that are cycles. For loopless graphs, cycles have at least two vertices.

For the distances, we need to be more specific and count the number of paths and cycles with a certain length. Recall that the **length** of a path or cycle is its number of edges. Therefore, we extend our notation, writing $P_k(G)$ for the number of connected components that are paths with length exactly k in a graph G. For instance, $P_0(G)$ is the number of isolated vertices in G. If G is a genome graph, then $P_0(G)$ is just the number of telomeres in it.

Similarly, we write $C_k(G)$ for the number of connected components that are cycles with length exactly k in a graph G. These quantities are related by the following formulas:

$$P(G) = \sum_{k=0}^{\infty} P_k(G), \tag{7.1}$$

$$C(G) = \sum_{k=1}^{\infty} C_k(G), \tag{7.2}$$

since there are no cycles of length 0.

For multi-genome breakpoint graphs of two genomes, notice that every connected component is a path or a cycle. Moreover, the cycles have always an even number of edges. Also, every vertex in $BG(A, B)$ belongs to exactly one connected component. Therefore, we have:

$$|V(BG(A,B))| = \sum_{k=0}^{\infty}(k+1)P_k(BG(A,B)) + \sum_{k=2}^{\infty} kC_k(BG(A,B)), \tag{7.3}$$

since there are $k+1$ vertices in a path of length k and k vertices in a cycle of length k. Also, there are no cycles of length 1 (or of any other odd length, for that matter). In other words, $C_k(BG(A, B)) = 0$ for odd k.

Two formulas for the rank distance are given below. The first one in terms of the number of paths and cycles. The second one in terms of contribution of each type of component.

Theorem 13. *The rank distance between genomes A and B over the n genes is:*

$$d(A, B) = 2n - 2C - P,$$

where C is the number of cycles in $BG(A, B)$ and P is the number of paths in $BG(A, B)$.

Proof. Lemma 8 states that:

$$d(A, B) = 2n - p,$$

where p is the number of AB-orbits. However, the number of AB-orbits p can be written as $2C+P$, since a pair of circular orbits corresponds to each cycle, and linear orbits are in 1-1 correspondence with paths, by Theorems 11 and 12. □

Theorem 14. *The rank distance between genomes A and B over the same genes is:*

$$d(A,B) = \sum_{k=1}^{\infty} kP_k(BG(A,B)) + \sum_{k=4}^{\infty}(k-2)C_k(BG(A,B)).$$

Proof. From Theorem 13 we have:

$$d(A, B) = 2n - 2C - P,$$

where C is the number of cycles in $BG(A, B)$ and P is the number of paths in $BG(A, B)$. Then Equation 7.3 gives $2n = |V(BG(A, B))|$ as a sum (or two sums). Furthermore, P and C can also be written as sums using Equations 7.1 and 7.2.

Putting all this together, and noting that the cycle sum can start at 2 because there are no odd cycles, we arrive at the desired formula. □

7.3.3 DCJ distance from graph elements

DCJ is defined as the minimum number of cut, join, or double swap operations needed to transform A into B. This distance also has formulas relating it to the numbers of (certain) paths and cycles in a breakpoint graph, and stressing the contribution of each component, as follows.

Theorem 15 (Yancopoulos *et al.*, 2005 [15])**.** *The DCJ distance between genomes A and B over n genes is:*

$$d_{DCJ}(A, B) = n - C - \frac{1}{2}P_{even},$$

where C is the number of cycles in $BG(A, B)$ and P_{even} is the number of paths in $BG(A, B)$ with an even number of edges.

Theorem 16. *The DCJ distance between genomes A and B over the same genes is:*

$$d_{DCJ}(A, B) = \sum_{k=1}^{\infty} \lceil k/2 \rceil P_k(BG(A, B)) + \sum_{k=4}^{\infty} (k/2 - 1) C_k(BG(A, B)).$$

7.3.4 SCJ distance from graph elements

SCJ is defined as the size of the symmetric difference between A and B when viewed as adjacency sets [2].

Theorem 17. *The SCJ distance between genomes A and B over the same genes is:*

$$d_{SCJ}(A, B) = \sum_{k=1}^{\infty} k P_k(BG(A, B)) + \sum_{k=4}^{\infty} k C_k(BG(A, B)).$$

7.4 The Effect of Operations on the Breakpoint Graph

In Section 3.2 we studied the effect of basic operations on orbits. Here we will study their effect on the breakpoint graph.

A cut means the removal of an A-adjacency (edge) from the graph. A join means the creation of an A-adjacency (edge) between two extremities (nodes) that are A-telomeres. Finally, a double swap means exchanging two A-adjacencies (edges) by two other A-adjacencies (edges) using the same four nodes, but in a way that the A-adjacencies remain a matching.

To prove Theorem 15, we need two results:

- to show that, for any pair of genomes A, B over the same genes, there is a sorting scenario with that many steps;

- to show that, for any pair of genomes A, B over the same genes, and for any basic operation applicable to A, applying this operation on A will not modify the formula by more than 1 unit.

Chapter 8

Exercises

1. A **basis** is an $n \times k$ matrix with rank k, where $k \leq n$. If B is a basis, its column vectors form a basis of $\text{im}(B)$. Show that, when B is a basis, the orthogonal projection onto $\text{im}(B)$ is given by $B(B^tB)^{-1}B^t$. Conversely, given a projection matrix P with rank k, any set of k linearly independent columns from it forms a basis of $\text{im}(P)$.

2. Show that, if P is an orthogonal projection onto a subspace V, then $2P - I$ is the orthogonal reflection through V, matrix $I-P$ is the orthogonal projection onto $V^\perp$, and $I-2P$ is the orthogonal reflection through $V^\perp$.

8.1 Intermediate Genomes

3. Show that in the formula $B = SA + (I - S)C$ for intermediate matrices B between A and C we can choose S as being a projection, that it, such that $S^2 = S$.

4. True or false: B is intermediate between A and C if and only if $V_4 = V_5 = \{0\}$.

5. Show that if M is an intermediate matrix between A and B, then $A + B - M$ is also an intermediate matrix between A and B.

Solution:
If $\text{sc}(X; A, B) = d(A, X) + d(B, X)$, notice that $\text{sc}(A + B - M; A, B) = \text{sc}(M; A, B)$, because $d(A, A + B - M) = d(B, M)$ and $d(B, A + B - M) = d(A, M)$.

6. Show that if $(A + B)/2$ is an intermediate matrix between A and B, then $A = B$.

Solution:
Notice that $d(A, (A + B)/2) = r((B - A)/2) = r(B - A) = d(A, B)$. Likewise, $d(B, (A + B)/2) = d(A, B)$. If $(A + B)/2$ is intermediate, then $d(A, (A + B)/2) + d(B, (A + B)/2) = d(A, B)$ implies $2d(A, B) = d(A, B)$, that is, $d(A, B) = 0$.

7. Show that if $\mu A+(1-\mu)B$ is an intermediate matrix between A and B, with μ a real number different from 0 and 1, then $A = B$.

Solution:
Very similar to the solution of Problem 6. If $M = \mu A+(1-\mu)B$, then $d(A, M) = r((1-\mu)(A-B)) = r(A - B) = d(A, B)$ if $\mu \neq 1$. Likewise, $d(B, M) = r(\mu(A - B)) = r(A - B) = d(A, B)$ if $\mu \neq 0$. If M is intermediate, we have $d(A, M) + d(B, M) = d(A, B)$, which leads to $2d(A, B) = d(A, B)$ and then to $d(A, B) = 0$.

8. Let A and C be genomes over the same genes, and x a gene extremity that is not a telomere of A or C. True or false: no intermediate genome between A and C can satisfy $BAx = Cx$.

Solution 1:
(Partial) If A and C are circular genomes, this certainly cannot happen. If it does, then $(C-B)Cx =$

$x - BCx = x - Ax$, so $x - Ax \in \text{im}(C - B)$. By Lemma 27, $\text{im}(C - B) \subseteq \text{im}(C - A)$, since B is an intermediate genome between A and C. Therefore, $x - Ax \in \text{im}(C - A)$. It follows that x and Ax are in the same AC-orbit. However, by Lemma 24, this implies that this orbit contains a telomere, a contradiction with the circularity of A and C.

8.2 Components

9. Let A and B be two genomes. Show that a linear component $\mathcal{C}$ of the multi-genome breakpoint graph of A, B contributes $|\mathcal{C}| - 1$ units to the distance $d(A, B)$, while a circular component $\mathcal{C}$ contributes $|\mathcal{C}| - 2$ units to this distance.

10. True or false: let A and B be genomes, $P = -J(x, y)$ a cut applicable to A, and $\mathcal{C}$ the component of x in the multi-genome breakpoint graph $BG(A, B)$. If $\mathcal{C}$ is circular, then $d(A+P, B) = d(A, B) + 1$, but if $\mathcal{C}$ is linear, then $d(A + P, B) = d(A, B) - 1$.

11. True or false: if A and B are genomes, at least one minimax genome of A, B does not link components of $BG(A, B)$.

Solution 1:
Let M be a genomic median of A, B, and C. Let M' be the genome obtained from M by removing all edges (adjacencies) that cross components. We claim that $\text{sc}(M'; A, B, C) \leq \text{sc}(M; A, B, C)$.

To prove this, consider one of the input genomes and call it X, that is, X is either A, or B, or C. If we draw the breakpoint graph $BG(X, M)$, we have paths and cycles. We know that a k-path (path with k edges) contributes k units to the rank distance, and that a k-cycle (cycle with k edges) contributes $k - 2$ units to $d(X, M)$.

Let us see how each component of $BG(X, M)$ is affected by the removal of the crossing edges, that is, let us see what this component becomes in $BG(X, M')$. A component that is a k-path becomes a collection of paths of sizes $k_1, k_2, \ldots, k_s$, with $s \geq 1$ and $\sum_{i=1}^{s} k_i \leq k$. Therefore, the contribution of these s paths to $d(X, M')$ is not larger than the contribution of the original k-path to $d(X, M)$. In this component, then, M' is not worse off than M with respect to its rank distance to X.

Now consider a k-cycle in $BG(X, M)$. We claim that this component cannot have exactly one crossing edge. Indeed, if there is just one crossing edge, connecting, say, connected components S and T of the graph, its removal would leave at least two components, one inside S and one inside T. But removing one edge from a cycle leaves just one component, a contradiction. Therefore, this k-cycle either has zero crossing edges or two or more crossing edges.

If it has zero crossing edges, its contribution to $d(X, M)$ is the same as its contribution to $d(X, M')$. If is has two or more crossing edges, then the k-cycle becomes a collection of s paths of sizes $k_1, k_2, \ldots, k_s$, with $s \geq 2$ and $\sum_{i=1}^{s} k_i \leq k - 2$. It follows that the contribution of these s paths to $d(X, M')$ is not larger than the contribution of the original k-cycle to $d(X, M)$. In this component, again, M' is not worse off than M with respect to the rank distance.

Since every component of $BG(X, M)$ is either a path or a cycle, we conclude that $d(X, M') \leq d(X, M)$ for $X = A$, B, and C, showing that M' is also a median.

12. True or false: if A and B are genomes, no minimax genome of A, B links components of $BG(A, B)$.

8.3 Medians

13. True or false: if x and y are two AB-equivalent extremities, then $B + P$ is closer to A than B, where $P = (By - Bx)(x - y)^t$.

14. True or false: Let A, B, and C be three genomes over the same genes. If $Ax = Bx$ for a certain extremity x, then any median genome M of A, B, and C satisfies $Mx = Ax$.

Note: a median genome is a genome M that minimizes $\mathrm{sc}(M; A, B, C)$ over all *genomes*. This is not always the same as a matrix median M of A, B, and C that happens to be a genome, although the latter certainly implies the former.

Solution:

We think this is true. If $Ax = y \neq x$ and x, y are telomeres in M, then $M + J(x, y)$ has a better score than M.

If $Ax = y \neq x$ and x is a telomere in M but y is not, $M - J(y, My)$ has a better score than M. If y is a telomere in M but x is not, use the cut $-J(x, Mx)$.

If $Ax = y \neq x$ and neither x nor y are telomeres in M, then $M + W(x, y, My, Mx)$ has a better score than M.

Finally, if $Ax = x$ and $Mx = y \neq x$, then $M - J(x, y)$ has a better score than M.

15. (Leonid Chindelevitch) Consider the following genomes:

$$\begin{aligned} A &= (a\ b)(c\ d)(e\ f)(g\ h) \\ B &= (a\ d)(c\ f)(e\ h)(b\ g) \\ C &= (a\ c)(b\ d)(f\ g). \end{aligned}$$

Show that it is impossible to construct a basis for

$$V_5 = \mathrm{im}(A - B) \cap \mathrm{im}(B - C) \cap \mathrm{im}(C - A)$$

composed solely of vectors with coefficients -1, 0, and $+1$ with respect to the canonical basis a, b, c, d, e, f, g, h.

Solution:

The first step is to realize that a basis of V_5 will have exactly two vectors, and they will have to be orthogonal to all of the following vectors:

$$\begin{array}{l} a+b+c+d+e+f+g+h \\ a+c+e+g \\ a+d \\ b+c \\ a+b+f \\ c+d+g \end{array},$$

because each one of these vectors satisfies $Av = Bv$, or $Bv = Cv$, or $Cv = Av$.

Let $a^t v, b^t v, \ldots, h^t v$ be the coefficients of $v \in V_5$ with respect to $a, b, \ldots, h$. Due to these orthogonality conditions, we can write all these coefficients in terms of the two first coefficients $a^t v$ and $b^t v$:

$$\begin{aligned} c^t v &= -b^t v \\ d^t v &= -a^t v \\ f^t v &= -a^t v - b^t v \\ g^t v &= a^t v + b^t v \\ e^t v &= -2a^t v \\ h^t v &= 2a^t v. \end{aligned}$$

Since $a^t v$ has to be -1, 0, or $+1$, if we have $a^t v = -1$ or $+1$ this implies that $e^t v$ and $h^t v$ will not satisfy this. However, if $a^t v = 0$ we have only one direction in V_5, namely, the one generated by $b - c - f + g$. Therefore, in a basis of V_5 we will have at most one vector with coefficients -1, 0, and $+1$.

16. True or false: as we move B towards the median in the orthogonal algorithm, V_5 does not change.

17. True or false: M is a median of A, B, and C if and only if the following three conditions hold:

$$\begin{aligned} V_*(.A.B.M.) &= \{0\}, \\ V_*(.B.C.M.) &= \{0\}, \\ V_*(.C.A.M.) &= \{0\}. \end{aligned}$$

18. True or false: a matrix median M of three genomes satisfies $M^k = I$ for some integer k, even when M is not a genome.

19. Consider the following genome matrices:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Find all medians M of A, B, and C. Show that they lie in a circle defined by

$$(M - D)^t (M - D) = R,$$

for certain matrices D and R.

Solution:

Notice that B is just one step away from any median M, because

$$d(B, M) = \frac{1}{2}(d(A, B) + d(B, C) - d(C, A)) = \frac{1}{2}(2 + 3 - 3) = 1.$$

All medians are therefore of the form $M = B - 2uu^t B/u^t u$, for some nonzero vector $u \in \mathrm{im}(A - B) \cap \mathrm{im}(C - B)$. However, since $\mathrm{im}(A - B) \subseteq \mathrm{im}(C - B)$, it suffices to look for vectors u in $\mathrm{im}(A - B)$.

A generic vector in $\mathrm{im}(A-B)$ can be written as $u = [x\ y\ -y\ -x]^t$. Then $u^t u = x^2+y^2+y^2+x^2 = 2(x^2 + y^2)$. And

$$uu^t = \begin{bmatrix} x^2 & xy & -xy & -x^2 \\ xy & y^2 & -y^2 & -xy \\ -xy & -y^2 & y^2 & xy \\ -x^2 & -xy & xy & x^2 \end{bmatrix},$$

leading to

$$uu^t B = \begin{bmatrix} -xy & -x^2 & x^2 & xy \\ -y^2 & -xy & xy & y^2 \\ y^2 & xy & -xy & -y^2 \\ xy & x^2 & -x^2 & -xy \end{bmatrix}.$$

With this, we can compute $M = B - 2uu^tB/u^tu$ as follows:

$$\begin{aligned} M &= B + \frac{1}{x^2+y^2}\begin{bmatrix} xy & x^2 & -x^2 & -xy \\ y^2 & xy & -xy & -y^2 \\ -y^2 & -xy & xy & y^2 \\ -xy & -x^2 & x^2 & xy \end{bmatrix} \\ &= \frac{1}{x^2+y^2}\begin{bmatrix} xy & x^2 & y^2 & -xy \\ y^2 & xy & -xy & x^2 \\ x^2 & -xy & xy & y^2 \\ -xy & y^2 & x^2 & xy \end{bmatrix}. \end{aligned}$$

This is, therefore, the general form of a median. The formula actually specifies a one-parameter family, because multiplying u by any nonzero scalar results in the same M.

If all medians lie on a circle, one way to find the center is to look for antipodals. Indeed, without much effort we find that

$$M(x,y) + M(-y,x) = A + B.$$

This means that, if there is a center, $D = (A+B)/2$ is a good candidate for it. Developing the product $(M-D)^t(M-D)$, we find:

$$(M-D)^t(M-D) = I - AB.$$

The circle formula then becomes:

$$\left(M - \frac{A+B}{2}\right)^t \left(M - \frac{A+B}{2}\right) = I - AB.$$

20. True or false: the orthogonal algorithm can generate all medians of A, B, and C by suitable choice of u's.

Solution:
The following result can help.

Theorem 18. *Let A, B, C be three orthogonal matrices. If M is any median of A, B, and C, then there is a vector u in $\mathrm{im}(A-B) \cap \mathrm{im}(C-B)$ that makes the orthogonal algorithm walk towards M.*

Proof. Since M is a median, it is in a shortest path between A and B. Therefore, $\mathrm{im}(M-B) \subseteq \mathrm{im}(A-B)$. Analogously, $\mathrm{im}(M-B) \subseteq \mathrm{im}(C-B)$. Take $u \in \mathrm{im}(M-B)$. It then belongs to the intersection $\mathrm{im}(A-B) \cap \mathrm{im}(C-B)$ and can be used by the algorithm. But using this u will approximate B to M, as in the proof of one of our theorems. □

This result seems to tip the scale in favor of "true".

21. Let M be an $n \times n$ matrix and P a projection onto a subspace $V \subseteq \mathbb{R}^n$. Show that $MV \subseteq V$ if and only if $PMP = MP$.

Solution:
If $MV \subseteq V$ then $Mv \in V$ for any vector $v \in V$. In particular, MPu is in V for any vector $u \in \mathbb{R}^n$. It follows that $PMPu = MPu$ for any $u \in \mathbb{R}^n$. We conclude that $PMP = MP$.

Conversely, if $PMP = MP$, consider any vector $v \in V$. Since P is onto V, we have $v = Pv$. Then $Mv = MPv = PMPv \in V$.

22. Show that if A, B, and C are genomes, then M_I is symmetric.

Solution 1:
There is a proof in the paper "On the rank-median of 3 permutations", by Chindelevitch and Meidanis.

Solution 2:
This is a point-free solution. We have $M_I = AP_1 + AP_2 + BP_3 + AP_4 + P_5$. The strategy is to show that $P_1^t M_I$ is symmetric, then that $P_S^t M_I$ is symmetric, where $P_S = P_2 + P_3 + P_4$, and finally that $P_5^t M_I$ is symmetric. If we manage to show all that, it follows that

$$M_I = I^t M_I = (P_1 + P_2 + P_3 + P_4 + P_5)^t M_I = P_1^t M_I + P_S^t M_I + P_5^t M_I,$$

so M_I is symmetric as a sum of symmetric matrices.

Starting with $P_1^t M_I$, we have:

$$P_1^t M_I = P_1^t (AP_1 + AP_2 + BP_3 + AP_4 + P_5) = P_1^t AP_1,$$

because all terms except the first will be zero. To see that, notice that $P_1^t AP_2 = P_1 P_2 AP_2 = 0$, since P_1 is symmetric as an orthogonal projection, $AP_2 = P_2 AP_2$ from Problem 21, and $P_1 P_2 = 0$ since P_2 projects onto a subspace orthogonal to P_1's image. The next two terms can be treated analogously. The last term is easier because we don't need Problem 21.

Now $P_1^t AP_1$ is symmetric by its very formula, since transposing it and applying the rules for transposition of a product we end up with the same expression, given that $A^t = A$ as a genome.

Now

$$\begin{aligned} P_2^t M_I &= P_2^t AP_2 + P_2^t BP_3 + P_2^t AP_4 = P_2^t AP_S, \\ P_3^t M_I &= P_3^t AP_2 + P_3^t BP_3 + P_3^t AP_4 = P_3^t BP_S, \\ P_2^t M_I &= P_4^t AP_2 + P_4^t BP_3 + P_4^t AP_4 = P_4^t CP_S, \end{aligned}$$

because $P_2^t B = P_2^t A$ comes from $BP_2 = AP_2$, the equality $P_3^t AP_4 = P_3^t BP_4$ comes in two steps from $AP_4 = CP_4$ and $P_3^t C = P_3^t B$, and the last one comes from $P_4^t A = P_4^t C$, $BP_3 = CP_3$ and some of the previous equalities.

Adding these three equalities we end up with

$$P_s^t M_I = (P_2^t A + P_3^t B + P_4^t C) P_s.$$

However, it is possible to show that $M_I^t P_S$ is equal to the right hand side of this last equation, because terms with P_1 or P_5 in M_I vanish in this product. We end up with

$$P_S^t M_I = M_I^t P_S,$$

showing that $P_s^t M_I$ is symmetric.

Finally, $P_5^t M_I = P_5^t P_5$, which is symmetric by its very formula also. We conclude that M_I is symmetric as the sum of the three symmetric matrices $P_1^t M_I$, $P_S^t M_I$, and $P_5^t M_I$.

23. Let M be an orthogonal matrix such that $Me = e$, where $e = [1\ 1\ 1 \ldots 1]^t$. Show that M cannot have an entry with value -1 or lower. Can it have entires arbitrarily close to -1?

Solution:
Hint: consider lines of the form $[1/n] * (2 * n - 1) + [-1 + 1/n]$.

24. True or false: a median genome of three linear genomes A, B, and C is always linear.

Solution:
We think that $A = [1, 2, 3]$, $B = [2, 3, 1]$, and $C = [3, 1, 2]$ provide a counter-example. The unique median is circular genome $M = (1, 2, 3)$.

25. True or false: a median genome of three circular genomes A, B, and C is always circular.

Note: if this is true, we have a proof that computing (rank) genome medians of genomes is NP-hard.

8.4 Divisibility

26. True or false: A join $J(x, y)$ applicable to A gets it closer to C if and only if x and y are in the same AC-orbit.

27. True or false: A cut $-J(x, y)$ applicable to A gets it closer to C if and only if x and y are in the same AC-orbit.

28. True or false: A double swap $W(x, y, z, w)$ applicable to A gets it closer to C if and only if x and y are in the same AC-orbit and z is not in this same AC-orbit.

29. True or false: for any two genome matrices A, and B, if $|A|+|B| > |AB|$ then $|\gcd(A, B)| \geq 1$.

30. Show that, if A is an orthogonal matrix, for any non-null vector $v \in \mathrm{im}(A - I)$ the Householder transformation $H(v) = I - vv^t/v^tv$ divides A.

31. Show that, if A is an orthogonal matrix, any Householder transformation H that divides A also divides A^{-1}.

32. Show that $\ker(A - I) = \ker(A^t - I)$ and $\mathrm{im}(A - I) = \mathrm{im}(A^t - I)$ if A is an orthogonal matrix.

33. Compute the norm of a GCD between matrices A and B corresponding to permutations (2 4 5)(3 6 7) and (1 3 4)(2 9 7)(5 8 6).

Solution:
We have that the norm of a GCD is $(|A|+|B|-|AB^{-1}|)/2$. Since $|A| = 4$, $|B| = 6$, and $|AB^{-1}| = 8$, the norm of a GCD is just 1.

34. Compute the norm of a GCD between matrices A and B corresponding to permutations (1 8 6)(2 9 4)(3 7 5) and (1 2 3)(4 5 6)(7 8 9).

Solution:
We have that the norm of a GCD is $(|A|+|B|-|AB^{-1}|)/2$. Since $|A| = 6$, $|B| = 6$, and $|AB^{-1}| = 6$, the norm of a GCD is 3.

35. Let A, B, and C be orthogonal matrices and let u and v be two linearly independent vectors in $\mathrm{im}(A - B) \cap \mathrm{im}(C - B)$. We know that $H(u)$ and $H(v)$ are transformations that divide both AB^{-1} and CB^{-1}. Under which conditions we have that $H(u)H(v)$ also divides these two quotients?

36. Show that, for an orthogonal matrix A, if a nonzero vector v is in $\mathrm{im}(A - I)$ then the Householder transformation $H(v) = I - 2vv^t/v^tv$ divides A. Moreover, show that u does not belong to $\ker(A - I)$ but it belongs to $\ker(AH(v) - I)$, and Au does not belong to $\ker(A - I)$ but it belongs to $\ker(H(v)A - I)$. Therefore, $\ker(A - I)$ grows by u or by Au, depending on the side you multiply $H(v)$.

37. Let $A = (b\ d\ e)(c\ f\ g)$ and $B = (a\ c\ d)(b\ i\ g)(e\ h\ f)$. Find a vector u_A such that $(A - I)u_A = v$, where and $v = b - d + c - g$. Then find a vector u_B such that $(B - I)u_B = v$ for this same vector v.

Solution:
One way of doing this is step-by-step. Since $v = b - \ldots$ we may try $u_A = e + x$. Then $Ax - x = A(u_A - e) - (u_A - e) = v - Ae + e = e - d + c - g$. Again, we can try $x = d + y$, leading to $Ay - y = A(x - d) - (x - d) = c - g$. Then $y = g$ is the solution, so $u_A = e + d + g$ is one possible solution.

Likewise, from $Bu_B - u_B = b - d + c - g$ we guess $u_B = g + x$, where x satisfies $Bx - x = c - d$, and $x = -c$ is a solution, leading to $u_B = g - c$.

38. Let A and B be two orthogonal matrices such that both $\ker(A-I)$ and $\ker(B-I)$ have bases composed of 0-1 vectors, and suppose I is not intermediate between A and B. Can we always find a nontrivial common divisor D such that $\ker(AD - I)$ and $\ker(BD - I)$ still have bases composed solely of 0-1 vectors?

39. Let A and B correspond to constructive permutations. Then $\ker(AB-1) = \ker(A-I) \cap \ker(B-I)$.

8.5 Permutations

This section contains exercises on permutations in the algebraic sense. The goal is to translate them all to matrix notation, with algebraic permutations being replaced by permutation matrices, which have exactly the same properties, and more. For now, we are keeping them in this format so that we don't lose them, while translation is in progress.

8.5.1 Medians

Definition 13. *A permutation μ is a **better median candidate** then σ for permutations α, β, and γ when* $\mathrm{sc}(\mu;\alpha,\beta,\gamma) < \mathrm{sc}(\sigma;\alpha,\beta,\gamma)$.

40. Show that $\alpha\pi$ is a better median candidate for π, σ, and τ than π if $\alpha|\sigma\pi^{-1}$ and $\alpha|\tau\pi^{-1}$.

Note: score improves by $|\alpha|$.

Solution:

If $\alpha|\sigma\pi^{-1}$, then $|\sigma\pi^{-1}\alpha^{-1}| = |\sigma\pi^{-1}| - |alpha|$. Likewise, $|\tau\pi^{-1}\alpha^{-1}| = |\tau\pi^{-1}| - |alpha|$. But

$$\begin{aligned} \mathrm{sc}(\alpha\pi;\pi,\sigma,\tau) &= |\pi\pi^{-1}\alpha^{-1}| + |\sigma\pi^{-1}\alpha^{-1}| + |\tau\pi^{-1}\alpha^{-1}| \\ &= |\alpha^{-1}| + |\sigma\pi^{-1}| - |\alpha^{-1}| + |\tau\pi^{-1}| - |\alpha^{-1}| \\ &= \mathrm{sc}(\pi;\pi,\sigma,\tau) - |\alpha|. \end{aligned}$$

41. If there is a permutation β with $|\beta| = 2$ such that $\beta\pi$ is a better median candidate for π, σ, and τ than π, then for any divisor α of β with $|\alpha| = 1$ we have that $\alpha\pi$ is also a better median candidate for π, σ, and τ than π.

Note: because the converse of Exercise 40 holds here.

42. The conclusion of Exercise 41 is not true for $|\beta| = 3$: take $\pi = 1$, $\sigma = (a\ e\ c)(b\ d\ f)$, $\tau = (a\ b)(c\ d)(e\ f)$, and $\beta = \tau\pi^{-1} = \tau$.

43. Perhaps Exercise 41 is true for $|\beta| = 3$ if π is opposite a maximum edge?

Note: hard.

44. Among π, σ, and τ, the best median candidates are opposite maximum edges.

Solution:

It follows from the fact that adding the score of $\mu = \pi$, σ, or τ to the size of its opposite edge results in the same value, which is the perimeter of the triangle π, σ, τ. Therefore, smaller scores correspond to larger opposite edges and vice-versa.

45. Sometimes, there is no genome satisfying the lower bound, but there is a permutation satisfying the lower bound: $\pi = (a\ b)(c\ d)$, $\sigma = (b\ c)(d\ e)$, $\tau = (a\ c)(b\ d)$, $\mu = (a\ b\ c\ d)$.

Solution:

It is straightforward to verify that $d(\mu,\pi) = 1$, $d(\mu,\sigma) = 3$, and $d(\mu,\tau) = 1$. The lower bound is 5, so $\mathrm{sc}(\mu;\pi,\sigma,\tau)$ is equal to the lower bound.

A genome satisfying the lower bound must have odd norm, because and even-normed permutation would produce even distances from each of the (even-normed) corners, leading to an even score by the parity lemma, which couldn't possibly be equal to 5.

Therefore, the distances to the corners would have to be 1,1,3. This means the genome would have to be at distance 1 from two of the corners. This is only possible for π and τ, which are 2

units apart. But every permutation in a shortest path between π and τ is of the form $\pi\delta$, where δ is a divisor of $\pi\tau = (a\ d)(b\ c)$. The only nontrivial possibilities are μ and μ^{-1}; neither is a genome.

46. Sometimes, there is no permutation satisfying the lower bound, but there is a matrix satisfying the lower bound: $\pi = (a\ b)(c\ d)$, $\sigma = (b\ c)(d\ a)$, $\tau = (a\ c)(b\ d)$, $\mu = (\pi + \sigma + \tau - 1)/2$.

47. True or false: if there is no genome between genomes π and σ, except the inputs, then the genome median of π, σ, and arbitrary τ is one of π, σ.

48. Does the Algorithm 1 always return a permutation median of π, σ, and τ?

Algorithm 1: Walking from π

`Median(`π, σ, τ`)`
while *there is a 2-cycle* α *dividing* $\sigma\pi^{-1}$ *and* $\tau\pi^{-1}$ **do**
  $\pi \leftarrow \alpha\pi$
return π

49. How about Algorithm 2?

Algorithm 2: Walking from an arbitrary permutation

`Median(`π, σ, τ`)`
start with any permutation β
while *there is a 2-cycle* α *dividing at least two of* $\pi\beta^{-1}$, $\sigma\beta^{-1}$ *and* $\tau\beta^{-1}$ **do**
  $\beta \leftarrow \alpha\beta$
return β

50. Let π, σ, τ, π', σ', τ' be permutations such that

$$(\mathrm{supp}(\pi) \cup \mathrm{supp}(\sigma) \cup \mathrm{supp}(\tau)) \cap (\mathrm{supp}(\pi') \cup \mathrm{supp}(\sigma') \cup \mathrm{supp}(\tau')) = \emptyset.$$

True or false: every median of $\pi\pi'$, $\sigma\sigma'$, $\tau\tau'$ is of the form $\mu\mu'$, where μ is a median of π, σ, τ and μ' is a median of π', σ', τ'.

51. True or false: if one cannot walk from either π, σ, or τ, then one of them is a permutation median.

52. Same as Exercise 51, but for genomes and steps of size 1 or 2.

53. True or false: a permutation median μ of α, β, and γ satisfies:

$$\mathrm{supp}(\mu) \subseteq \mathrm{supp}(\alpha) \cup \mathrm{supp}(\beta) \cup \mathrm{supp}(\gamma).$$

8.5.2 Divisibility

54. True or false: if $|\alpha| + |\beta| > |\alpha\beta|$, then there is a 2-cycle dividing α and β simultaneously.

Solution:
False. Take $\alpha = (a\ b)(c\ d)$, $\beta = (a\ c)(b\ d)$. Then $|\alpha| + |\beta| = 4 > 2 = |\alpha\beta|$, but there is no common divisor of α and β.

55. True or false: if α and β divide π and are disjoint, then the product $\alpha\beta$ divides π.

56. True or false: if α and β are constructive and π divides α, then π divides the product $\alpha\beta$.

57. True or false: if $\alpha|\gamma$ and $\beta\delta$ and γ, δ are constructive, then $\alpha\beta$ divides $\gamma\delta$.

58. True or false: if $|\alpha| + |\beta| + |\gamma| = |\alpha\beta\gamma|$, then β divides $\alpha\beta\gamma$.

59. True or false: if α is a GCD of β and γ, then $|\beta| + |\gamma| = |\beta\gamma| + |\alpha|$.

60. True or false: if for all τ with $|\tau| \leq 2$ we have $\tau|\rho \rightarrow \tau|\sigma$, then for all τ with $|\tau| = 3$ we have $\tau|\rho \rightarrow \tau|\sigma$.

61. True or false: if α and β are constructive and $\gamma|\beta$, then α and γ are constructive.

62. True or false: if α and β are cycles, then $\alpha|\beta$ if and only if there is a set $\mathcal{S}$ such that $\alpha = \beta - \mathcal{S}$.

63. True or false: if $\alpha|\beta$ then $orb(\alpha)$ refines $orb(\beta)$.

64. True or false: if σ is a cycle, α and β are disjoint permutations that divide σ, then $\alpha\beta|\sigma$ if and only if $orb(\sigma\alpha^{-1}) \perp orb(\sigma\beta^{-1})$, that is, for every orbit $\mathcal{S}$ of $\sigma\alpha^{-1}$ and every orbit $\mathcal{X}$ of $\sigma\beta^{-1}$ we have $\mathcal{S} \perp \mathcal{X}$, meaning that either $\mathcal{S} \subseteq \mathcal{X}$, or $\mathcal{X} \subseteq \mathcal{S}$, or $\mathcal{X} \cap \mathcal{S} = \emptyset$.

65. True or false: if for all τ with $|\tau| \leq 2$ we have $\tau|\rho \rightarrow \tau|\sigma$, then $\rho|\sigma$.

66. True or false: for $k \geq 2$, we have: if for all α with $|\alpha| \leq k$ we have $\alpha|\sigma \rightarrow \alpha|\tau$, then for all α with $|\alpha| \leq k + 1$ we have $\alpha|\sigma \rightarrow \alpha|\tau$.

67. Suppose π_1, π_2, π_3 are genomes and μ is a genome median of the three. If a two-cycle τ divides at least two of π_1, π_2, π_3, then τ divides μ.

68. True or false: if π is a cycle of length at least 3 and for all 3-cycles τ we have $\tau|\pi \rightarrow \tau|\sigma$, then $\pi|\sigma$.

69. True or false: if a partition $\mathcal{C}$ refines $orb(\alpha)$, then there is a unique permutation β such that $orb(\beta) = \mathcal{C}$ and $\beta|\alpha$.

70. True or false: if a partition $\mathcal{C}$ refines $orb(\alpha)$, then there is a unique permutation β of maximum norm such that $orb(\beta) = \mathcal{C}$ and $\beta|\alpha$.

71. True or false: for any permutations α, β, γ, δ we have $|\alpha\beta\gamma\delta| = |\alpha\gamma\beta\delta|$.

72. Does 3-way constructiveness depend on the order?

73. True or false: if α ans β are constructive and τ is a nontrivial cycle that divides $\alpha\beta$, then either $\tau|\alpha$ or $\tau|\beta$, but not both.

Bibliography

[1] P. Biller. The minimax genome problem. PhD Qualifying Exam Text, In Portuguese. URL: http://www.ic.unicamp.br/~meidanis/PUB/Doutorado/2012-Biller/eqe.pdf, July 2014.

[2] P. Feijao and J. Meidanis. SCJ: A breakpoint-like distance that simplifies several rearrangement problems. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 8(5):1318–1329, September 2011.

[3] P. Feijao and J. Meidanis. Extending the algebraic formalism for genome rearrangements to include linear chromosomes. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 10(4), July 2013. A preliminary version appeared in: BSB '12 — Proceedings of the 7th Brazilian Symposium on Bioinformatics, 2012.

[4] Jean H. Gallier. *Geometric Methods and Applications*. Texts in Applied Mathematics. Springer-Verlag, 2001.

[5] M. Isabel Garca-Planas and M. Dolors Magret. Eigenvectors of permutation matrices. *Advances in Pure Mathematics*, 5:390–394, 2015.

[6] S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, 46(1):1–27, January 1999. A preliminary version appeared in: STOC '95 — Proceedings of the Twenty-seventh Annual ACM Symposium on Theory of Computing, 1995.

[7] Zackary I. Johnson and Sallie W. Chisholm. Properties of overlapping genes are conserved across microbial genomes. *Genome Research*, 14(11):2268–2272, Nov 2004.

[8] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two chromosomes. *Algorithmica*, 13:180–210, 1995. A preliminary version appeared in: CPM '93 — Proceedings of the Fourth Annual Symposium on Combinatorial Pattern Matching, 1993.

[9] Izabela Makalowska, Chiao-Feng Lin, and Wojciech Makalowski. Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry*, 29:1–12, 2005.

[10] J. Meidanis and Z. Dias. An alternative algebraic formalism for genome rearrangements. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, pages 213–223. Springer Netherlands, 2000.

[11] J Meidanis and S. Yancopoulos. The emperor has no caps! a comparison of DCJ and algebraic distances. In Cedric Chauve, Nadia El-Mabrouk, and Eric Tannier, editors, *Models and Algorithms for Genome Evolution*, pages 207–243. Springer London, London, 2013.

[12] Tomohiro Nakayama, Satoshi Asai, Yasuo Takahashi, Oto Maekawa, and Yasuji Kasama. Overlapping of genes in the human genome. *International Journal of Biomedical Science*, 3(1):14–19, Mar 2007.

[13] Douglas West. *Introduction to Graph Theory*. Prentice-Hall, 2001.

[14] K. Wieand. Eigenvalue distributions of random permutation matrices. *The Annals of Probability*, 28(4):1563–1587, Oct 2000.

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

[16] J. P. P. Zanetti, P. Biller, and J. Meidanis. Median approximations for genomes modeled as matrices. *Bulletin of Mathematical Biology*, 78(4):786–814, 2016.