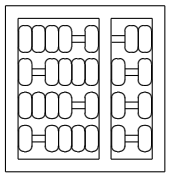

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

On the algebraic genome median

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

Technical Report - IC-12-27 - Relatório Técnico

December - 2012 - Dezembro

On the algebraic genome median

P. Biller*† J. P. P. Zanetti*† P. Feijão† J. Meidanis‡

Abstract

The Genome Median Problem is an important problem in phylogenetic reconstruction under rearrangement models. It can be stated as follows: given three genomes, find a fourth that minimizes the sum of the pairwise rearrangement distances between it and the three input genomes. Recently, Feijão and Meidanis extended the algebraic theory for genome rearrangement to allow for linear chromosomes, thus yielding a new rearrangement model (the algebraic model), very close to the celebrated DCJ model. In this paper, we study the genome median problem under the algebraic model, whose complexity is currently open, proposing a more generalized form of the problem, the matrix median problem, that can be approximated in linear time to a factor of $\frac{4}{3}$ of the optimum. The study of the matrix median might help in the solution of the algebraic median problem.

1 Introduction

Genome rearrangements are evolutionary events where large, continuous pieces of the genome shuffle around, changing the order of genes in the genome of a species. Gene order data can be very useful in estimating the evolutionary distance between genomes, and also in reconstructing the gene order of ancestral genomes. The simplest form of inference of evolutionary scenarios based on gene order is the pairwise genome rearrangement problem: given two genomes, find a plausible rearrangement scenario between them, that is, the smallest sequence of rearrangement events that transforms one genome into the other.

For most rearrangement events proposed, this problem has already been solved, usually with linear or subquadratic algorithms. However, when more than two genomes are considered, inferring evolutionary scenarios becomes much more difficult. This problem is known as the *multiple genome rearrangement problem* (MGRP): given a set of genomes, find a tree with the given genomes as leaves and an assignment of genomes to the internal nodes such that the sum of all edge lengths (the pairwise distance between adjacent genomes) is minimal.

The MGRP may still be hard even when only three genomes are considered, the so called *genome median problem* (GMP): given three genomes, find a fourth genome that minimizes the sum of its pairwise distances to the other three. The GMP is NP-complete in most

*These authors contributed equally to this work.
†Institute of Computing, University of Campinas, SP, Brazil.
‡Institute of Computing, University of Campinas, SP, Brazil; Scylla Bioinformatics, Campinas, Brazil.

rearrangement models, with notable exceptions being the multichromosomal breakpoint distance [8] and the Single-Cut-or-Join (SCJ) model [2]. The GMP is a particularly interesting problem, because several algorithms for the MGRP are based on repeatedly solving GMP instances, until convergence is reached (for instance, the pioneering BPAnalysis [7], the more recent GRAPPA [6], and MGR [1]).

In this paper we will focus on the algebraic rearrangement model proposed by Meidanis and Dias [5], recently extended to allow linear chromosomes in a very natural way by Feijão and Meidanis [3]. This extended algebraic rearrangement model is similar to the well-known Double-Cut-and-Join (DCJ) model [9], with a slight difference in the weight of single cut/join operations, where the weight for this operations is 1 in the DCJ model, but 1/2 in the algebraic model. The algebraic pairwise distance problem can be solved in linear time, but the median problem remains open.

The main goal of this paper is to investigate the problem of computing the algebraic median of three genomes. According to algebraic rearrangement theory [3], a genome can be seen as a permutation $\pi : E \mapsto E$, where E is the set of gene extremities, with the added property that $\pi^2 = \mathbf{1}$, the identity permutation. In this report, the *distance* between two genomes or two permutations π and σ will be defined as $||\sigma\pi^{-1}||$, where $||\alpha||$ designates the *norm* of a permutation, which will be defined in the Methods section. It is important to note that, in the original paper, the algebraic distance is defined as $\frac{||\sigma\pi^{-1}||}{2}$. However, to avoid dealing with fractional numbers and to simplify the calculations, we will multiply the distances by 2 in this report.

The median problem can be stated as follows: given three genomes π_1, π_2, and π_3, find a genome μ that minimizes $d(\mu; \pi_1, \pi_2, \pi_3)$, defined as the *total score*

$$d(\mu; \pi_1, \pi_2, \pi_3) = d(\mu, \pi_1) + d(\mu, \pi_2) + d(\mu, \pi_3).$$

We do not know yet the status of the median problem for the algebraic distance, but suspect it may be NP-hard as well. However, in this note we show that viewing genomes (or even general permutations) as matrices, the median can be approximated quickly, although the matrix solution may not be always translated back into permutations or genomes. Nevertheless, this positive result can help shed more light into the problem, by leading to approximation solutions, or to special cases that can be solved polynomially in the genome setting.

2 Algebraic rearrangement theory

We will start this section showing some basic definitions of the algebraic theory of Feijão and Meidanis [3].

2.1 Basic Concepts

Given a set E, a *permutation* $\alpha : E \to E$ is a map from E onto itself, which is therefore also injective. Permutations are represented as parenthesized lists, with each element followed by its image. For instance, on $E = \{a, b, c\}$, $\alpha = (a\ b\ c)$ is the permutation that maps a to

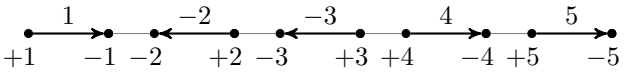

Figure 1: A genome with one linear chromosome, represented by the permutation $\pi = (-1\ -2)(+2\ -3)(+3\ +4)(-4\ +5)$.

b, b to c, and maps c back to a. This representation is not unique; $(b\ c\ a)$ and $(c\ a\ b)$ are equivalent. Permutations are composed of one or more *cycles*. For instance, the permutation $\alpha = (a\ b\ c)(d\ e)(f)$ has three cycles. A cycle with k elements is called a *k-cycle*. An 1-cycle represents a fixed element in the permutation and is usually omitted.

The *product* or *composition* of two permutations α, β is denoted by $\alpha\beta$. The product $\alpha\beta$ is defined as $\alpha\beta(x) = \alpha(\beta(x))$ for $x \in E$. For instance, with $E = \{a, b, c, d, e, f\}$, $\alpha = (b\ d\ e)$ and $\beta = (c\ a\ e\ b\ f\ d)$, we have $\alpha\beta = (c\ a\ b\ f\ e\ d)$.

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

A *2-cycle decomposition* of a permutation α is a representation of α as a product of 2-cycles, not necessarily disjoint. All permutations have a 2-cycle decomposition. The *norm* of a permutation α, denoted by $\|\alpha\|$, is the minimum number of cycles in a 2-cycle decomposition of α.

2.2 Modeling Genomes with Adjacency Algebraic Theory

To model genomes with the Algebraic Theory, the formulation is similar to the set representation of a genome, used in several related works [8, 2]. In this representation, each gene a has two *extremities*, called *tail* and *head*, respectively denoted by a_t and a_h, or alternatively using signs, where $-a = a_h$ and $+a = a_t$. An *adjacency* is an unordered pair of extremities indicating a linkage between two consecutive genes in a chromosome. An extremity not adjacent to any other extremity in a genome is called a *telomere*. A genome is represented as a set of adjacencies and telomeres (possibly omitted, when the gene set is given) where each extremity appears at most once.

In the algebraic theory, genomes are represented by permutations, with a *genome* being a product of 2-cycles, where each 2-cycle corresponds to an adjacency. Figure 1 shows an example genome and its representation as a permutation.

3 Results

In this section we define a matrix distance that corresponds to the algebraic distance and prove that it is indeed a valid metric in general, that is, even for metrics that do not represent genomes. An approximate solution to the corresponding matrix median problem can be found in polynomial time, by solving a system of linear equations. We show a linear time algorithm that computes a $\frac{4}{3}$–approximation to the matrix median distance.

3.1 Matrix distance

Permutations can be seen as matrices, and there is a matrix metric that corresponds to algebraic distance. However, this metric is more general, in the sense that it applies to all square matrices, not only those associated to permutations. We show here how to compute an approximate solution to the matrix median problem, which can be useful in the computation of genome medians.

Given two $n \times n$ matrices A and B, we define the *distance* between them as:

$$d(A, B) = n - \dim \ker(B - A),$$

where dim denotes the dimension of a vector space and $\ker X$ denotes the kernel of the matrix X, representing a subspace of vectors v such that $Xv = 0$.

This distance can be shown to satisfy the conditions of a metric, that is, it is symmetric, obeys the triangle inequality, and $d(A, B) = 0$ if and only if $A = B$. The hardest part is the triangle inequality, which we show here.

Lemma 1. *For any three $n \times n$ matrices A, B, and C, we have*

$$d(A, C) \leq d(A, B) + d(B, C).$$

Proof. It is based on the fact that

$$\ker(B - A) \cap \ker(C - B) \subseteq \ker(C - A),$$

since when $Bv = Av$ and $Cv = Bv$ then $Cv = Av$ by transitivity. Using the know formula for the dimension of a space sum [4]

$$\begin{aligned} \dim(\ker(B - A) + \ker(C - B)) &= \dim \ker(B - A) + \\ &\quad + \dim \ker(C - B) - \\ &\quad - \dim(\ker(B - A) \cap \ker(C - B)) \end{aligned}$$

we have

$$\begin{aligned} d(A, C) &= n - \dim \ker(C - A) \\ &\leq n - \dim(\ker(B - A) \cap \ker(C - B)) \\ &= n - \dim \ker(B - A) - \dim \ker(C - B) + \\ &\quad + \dim(\ker(B - A) + \ker(C - B)) \\ &\leq d(A, B) + d(B, C), \end{aligned}$$

since

$$\dim(\ker(B - A) + \ker(C - B)) \leq n,$$

because no subspace can have higher dimension than R^n itself. □

Given this metric, the first interesting observation is that permutations (including genomes) can be mapped to matrices in a distance-preserving way. Given a permutation $\alpha : E \mapsto E$, with $|E| = n$, we first identify each element of E with a unit vector of $\mathbb{R}^n$, and then define A, the matrix counterpart of α, so that

$$Av = \alpha v.$$

An example will help. Let α be the permutation $(a\ b)(c\ d)$. Identify $a = \begin{bmatrix}1\ 0\ 0\ 0\end{bmatrix}^t$, $b = \begin{bmatrix}0\ 1\ 0\ 0\end{bmatrix}^t$, $c = \begin{bmatrix}0\ 0\ 1\ 0\end{bmatrix}^t$, and $d = \begin{bmatrix}0\ 0\ 0\ 1\end{bmatrix}^t$, where v^t is the transpose of vector v. We then have

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

It is well known that this mapping produces matrices A that are invertible, and that satisfy $A^{-1} = A^t$, where A^t denotes the transpose of matrix A. Also, the identity permutation corresponds to the identity matrix I, and the product $\alpha\beta$ corresponds to matrix AB, where A is the matrix corresponding to α, and B is the matrix corresponding to β. If α happens to be a genome, that is, if $\alpha^2 = \mathbf{1}$, then A is a symmetric matrix and vice-versa.

To see why this mapping is distance-preserving, one can reason as follows. First, notice that it suffices to show that

$$||\alpha|| = n - \dim\ker(A - I), \tag{1}$$

for any permutation α and associated matrix A. Indeed, if P is invertible then

$$\dim\ker(S - P) = \dim\ker(SP^{-1} - I),$$

and then Equation (1) will relate $d(\sigma, \pi)$ to the distances of the corresponding matrices S and P.

Then proceed to show that, for a k-cycle, Equation (1) is true since both sides are equal to $k-1$. Finally, for a general permutation, decompose it in disjoint cycles, and use the fact that, for *disjoint* permutations α and β, with associated matrices A and B, respectively, we have

$$\ker(A - I) \cap \ker(B - I) = \ker(AB - I),$$

and

$$\ker(A - I) + \ker(B - I) = \mathbb{R}^n,$$

which guarantee that

$$n - \dim\ker(AB - I) = n - \dim\ker(A - I) + n - \dim\ker(B - I).$$

Therefore, if Equation (1) is valid for α and β, and if α and β are disjoint, then Equation (1) is valid for the product $\alpha\beta$. Since any permutation can be written as a product of disjoint cycles, Equation (1) is valid in general.

3.2 Matrix median

Since the correspondence between permutations and matrices preserves distances, it makes sense to study the matrix median problem as a way of sheding light into the permutation median problem, which in turn is related to the genome median problem. As we will see in the next section, the matrix median problem can be approximated within a $\frac{4}{3}$ factor by a polynomial time algorithm.

Let A, B, and C be three $n \times n$ matrices. Suppose we want to find a matrix M such that

$$d(M; A, B, C) = d(M, A) + d(M, B) + d(M, C)$$

is minimized. In order to have small $d(M, A)$, M must be equal to A in a large subspace, so that $\ker(A - M)$ is large. Similarly with B and C.

This suggests the following strategy. Decompose $\mathbb{R}^n$ into a direct sum of the subspaces below. In these subspaces, the following relations are true: (i) $A = B = C$, (ii) $A = B \neq C$, (iii) $A \neq B = C$, (vi) $A = C \neq B$, and (v) $A \neq B \neq C \neq A$.

In the first subspace, since A, B, and C all have the same behaviour, M should also do the same thing. In the second subspace, since $A = B$ but C is different, it is better for M to go with A and B. Likewise, in the third subspace M should concur with B and C, and with A and C in the fourth. Finally, in the final subspace it seems hard to gain points in two different distances, so the best course for M is to mimic one of A, B, or C.

Notice that making M equal to A, except in the third subspace, where it should be equal to B and C, satisfies what we said in the last paragraph, and should yield a good approximation of a median, if not a median. The rest of this section will be devoted to showing the details on this construction.

3.2.1 An example

Let us exemplify with a concrete case before generalizing the result. Consider $\alpha = \mathbf{1}$, $\beta = (a\ b\ c\ d\ e\ f)$, $\gamma = (a\ f\ e\ d\ c\ b)$, and let A, B, C be the corresponding matrices, respectively.

We need to decompose the whole space $\mathbb{R}^n$ into the parts described earlier. In addition, we need to make the median M equal to A, except in the subspace $A \neq B = C$. To implement this idea, we need a base for each relevant subspace. We will perform a detailed procedure for the subspace $A \neq B = C$. The others can be dealt with in a similar way.

To find a base for subspace $A \neq B = C$, we begin by considering $\ker(B - C)$. If $v = (v_a, v_b, v_c, v_d, v_e, v_f)$ is a vector in $\ker(B - C)$, then

$$(B - C)v = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ -1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} v_a \\ v_b \\ v_c \\ v_d \\ v_e \\ v_f \end{bmatrix} = 0$$

which implies that $v_a = v_c = v_e$ and $v_b = v_d = v_f$. We conclude that the vectors in this kernel have the form $v = x(a+c+e) + y(b+d+f)$, for scalars x and y, where we use a, b, etc. meaning the corresponding unit vectors. Hence, vectors

$$a + c + e, \text{ and } b + d + f \tag{2}$$

form a base of the subspace $B = C$.

To go on in our quest for a base of $A \neq B = C$, we must first realize that this subspace is not uniquely defined. Instead, several subspaces can play this role, namely, any subspace whose direct sum with the subspace $A = B = C$ yields the subspace $B = C$. In contrast, notice that both $B = C$ and $A = B = C$ are unique, well-defined subspaces, since they are characterized by equalities and therefore closed under sum and scalar multiplication, which are the requirements for a bona fide vector subspace. However, $A = B = C$ can be completed in a number of ways to form $B = C$. The idea here is to find a base of $A = B = C$ and then add linearly independent vectors from $B = C$ until we generate the entire subspace $B = C$. The extra vectors added will constitute a base for one of the many possible $A \neq B = C$ subspaces.

Let us find a base for $A = B = C$. We can write $A = B = C$ as $\ker(B-C) \cap \ker(A-B)$, for instance. We already know the format of a $\ker(B - C)$ vector. If it also belongs to $\ker(A - B)$, a vector $v = (v_a, v_b, v_c, v_d, v_e, v_f)$ must satisfy

$$v_a = v_b = v_c = v_d = v_e = v_f,$$

that is, the subspace $A = B = C$ has dimension 1 and a base for it is

$$a + b + c + d + e + f = \sum E. \tag{3}$$

We may choose any vector from the list (2) to form a base for $A \neq B = C$, because none of them is a linear combination of the base (3) for $A = B = C$. According to which vector we choose, we get a different subspace where $A \neq B = C$, but the dimension of this complementary subspace is 1 in any case.

Let us take $a + c + e$ as a base for $A \neq B = C$, and $\sum E$ as a base for $A = B = C$. Proceeding analogously in the remaining subspaces, we can find bases for all subspaces, as shown in Table 1.

A median candidate M can then be defined as follows. We shall make M equal to A everywhere, except in the subspace $A \neq B = C$, where it is better to make it equal to B (and C). Given the dimensions shown in Table 1, we expect the distances from M to the input matrices to be as follows.

$$\begin{aligned} d(M, A) &= 1 \text{ (due to subspace } A \neq B = C) \\ d(M, B) &= 4 \text{ (due to subspaces } A \neq B \neq C \neq A \text{ and } A = C \neq B) \\ d(M, C) &= 4 \text{ (due to subspaces } A \neq B \neq C \neq A \text{ and } A = B \neq C) \end{aligned}$$

Making M agree with B in $A \neq B = C$ and with A everywhere else leads to the following

Table 1: Space partition for the matrix median — Partition of the entire $\mathbb{R}^n$ space into relevant pieces to find a median candidate for A, B, and C, with a possible choice of bases.

Subspace	Base	Dimension
$A \neq B \neq C \neq A$	a, b, c, d	4
$A = B \neq C$	(empty)	0
$A \neq B = C$	$a + c + e$	1
$A = C \neq B$	(empty)	0
$A = B = C$	$\sum E$	1

equations:

$$\begin{aligned} M(a) &= a \\ M(b) &= b \\ M(c) &= c \\ M(d) &= d \\ M(a+c+e) &= b+d+f \\ M(\sum E) &= \sum E \end{aligned}$$

These equations admit a unique solution M, shown below.

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Computing the distances, we arrive at $d(M; A, B, C) = 1 + 4 + 4 = 9$, as expected.

3.2.2 Generalization of the example

In a previous section we saw an example of a matrix computation that leads to a median candidate. In this section, our goal is to generalize this procedure and find an exact formula for its median score.

We begin by introducing notation aimed at formalizing subspaces such as $A = B \neq C$ employed in the example. Given $n \times n$ matrices A, B, and C, we will use a dotted notation to indicate a partition, e.g., $.AB.C.$ means a partition where A and B are in one class, and C is in another class by itself. To each such partition, we associate a vector subspace of $\mathbb{R}^n$ formed by those vectors having the same image in each class:

$$V(.AB.C.) = \{v \in \mathbb{R}^n | Av = Bv\}.$$

Table 2: Distance contribution — Given three matrices A, B and C, this table shows the distance contribution of each of the five subspaces partitioning $\mathbb{R}^n$ to the distances $d(M, A)$, $d(M, A)$, and $d(M, A)$, for a candidate median matrix M.

		Contributes to ...		
Subspace	$M = \ldots$	$d(M, A)$	$d(M, B)$	$d(M, C)$
$V_*(.A.B.C.)$	A	no	yes	yes
$V_*(.AB.C.)$	A	no	no	yes
$V_*(.A.BC.)$	B	yes	no	no
$V_*(.AC.B.)$	A	no	yes	no
$V_*(.ABC.)$	A	no	no	no

Notice that singleton classes do not impose additional restrictions. With this notation, the subspace we used to call $A = B = C$ can be written as $V(.ABC.)$. Notice also that $V(.A.B.C.) = \mathbb{R}^n$.

We need also a notation for strict subspaces, such as $A = B \neq C$, where different classes actually disagree. As we have learned, these cannot be defined uniquely, but they can be specified unambiguously if we have bases for the unrestricted subspaces. Suppose we are given a mapping $\mathcal{B}$ such that $\mathcal{B}(p)$ is an ordered base for $V(p)$, where p runs over all partitions involving A, B, and C. We define a strict version $V_*(p, \mathcal{B})$ with respect to $\mathcal{B}$ in two steps. First we specify its dimension as

$$\dim V_*(p, \mathcal{B}) = \dim V(p) - \dim \sum_{p<q} V(q),$$

where $p < q$ means that partition p strictly refines partition q. In other words, we want to capture the part of the subspace $V(p)$ that is not covered by subspaces corresponding to coarser partitions. Notice that $p < q$ implies $V(q) \subseteq V(p)$.

Then, given its dimension k, we define $V_*(p, \mathcal{B})$ as the subspace generated by the k first vectors of $\mathcal{B}(p)$ that are not in $\sum_{p<q} V(q)$. Since the vectors of a base are linearly independent, this definition guarantees that $V_*(p, \mathcal{B})$ will have dimension k, and that

$$V(p) = V_*(p, \mathcal{B}) \oplus \sum_{p<q} V(q).$$

We will now assume that a base collection $\mathcal{B}(p)$ has been chosen and will drop the reference to $\mathcal{B}$ from here on. Section 3.5.1 will show a standard way of obtaining such bases.

For three matrices A, B, and C, we partition $\mathbb{R}^n$ into five subspaces, and define a median candidate M as shown in Table 2.

The matrix M will have a total distance to A, B, and C equal to:

$$\begin{aligned} d(M;A,B,C) &= d(M,A) + d(M,B) + d(M,C) \\ &= \dim V_*(.A.BC.) + \\ &\quad \dim V_*(.A.B.C.) + \dim V_*(.AC.B.) + \\ &\quad \dim V_*(.A.B.C.) + \dim V_*(.AB.C.) \\ &= 2\dim V_*(.A.B.C.) + \\ &\quad \dim V_*(.AB.C.) + \dim V_*(.AC.B.) + \dim V_*(.A.BC.). \end{aligned} \tag{4}$$

3.3 When it does not work

Unfortunately, the procedure described in the previous section does not always yield a median, as witnessed by the small instance presented in this section.

Consider the following three input matrices:

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

To find a median candidate M_a and its median score $d(M_a; A, B, C)$, we have to compute the relevant subspaces. As we did in the first example, we first compute the kernels of $B-C$, $A-C$, and $A-B$, to find the bases for $B=C$, $A=C$, and $A=B$, respectively.

Lets start with the subspace $B=C$. If $v=(v_a, v_b, v_c)$ is a vector in $\ker(B-C)$, then

$$(B-C)v = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} = 0$$

which implies that $v_a = v_b = v_c$. Therefore, $\sum E$ is a base for the subspace $B=C$. The same thing happens for subspaces $A=C$ and $A=B$. So, we have that $\sum E$ is also a base for the subspace $A=B=C$.

The next step is to find a base for the subspace $A \neq B = C$ by adding linearly independent vectors to the base of $A=B=C$ until it generates $B=C$. But, in this example, the subspaces $A=B=C$ and $B=C$ are equal. Thus, $A \neq B = C$ is empty. In the same way, the subspaces where $A=C \neq B$ and $A=B \neq C$ are empty.

The last subspace is the one where $A \neq B \neq C \neq A$. This subspace is $\mathbb{R}^n$ minus the sum of the subspaces we already computed. So, we just need to add linearly independent vectors to the one we already have until the whole $\mathbb{R}^n$ is generated. Choosing a and b in this capacity, we end up with the space partition for this instance is summarized in Table 3.

So, by Equation (4), our algorithm gives the following total score for this instance:

$$d(M_a; A, B, C) = 2 \times 2 + 0 + 0 + 0 = 4.$$

However, the identity matrix I has a better total score than M_a, and is actually a median in this case:

$$d(I; A, B, C) = 1 + 1 + 1 = 3.$$

Table 3: Space partition for the matrix median — Partition of the entire $\mathbb{R}^n$ space into relevant pieces to find a median candidate for A, B, and C, with a possible choice of bases.

Subspace	Base	Dimension
$A \neq B \neq C \neq A$	a, b	2
$A = B \neq C$	(empty)	0
$A \neq B = C$	(empty)	0
$A = C \neq B$	(empty)	0
$A = B = C$	$\sum E$	1

Therefore, our procedure does not always produce a matrix median, but we can still show that it is an approximation algorithm for the problem.

3.4 Approximation factor

Given that the procedure described in the Section 3.2.2 does not guarantee a matrix median, it is interesting to know whether it is an approximation algorithm, namely, whether there is a constant ρ such that the candidate's total score is at most ρ times the score of a median.

Take matrices A, B, and C and a matrix M such that $d(M; A, B, C)$ is minimum, that is, M is a median. There is a trivial lower bound for the median score of a matrix, easily obtained with the help of the triangle inequality, namely

$$d(M; A, B, C) \geq \frac{1}{2}(d(A, B) + d(B, C) + d(C, A)).$$

According to Eq. (4), the median score of the approximate solution M_a constructed in Section 3.2.2 is given by:

$$\begin{aligned} d(M_a; A, B, C) &= 2 \dim V_*(.A.B.C.) + \\ &\quad \dim V_*(.AB.C.) + \dim V_*(.AC.B.) + \dim V_*(.A.BC.), \end{aligned}$$

For comparison, we can write the trivial lower bound in terms of subspace dimensions. It suffices to write each distance as a dimension sum of the subspaces where they differ. The result is:

$$\begin{aligned} \frac{1}{2}(d(A, B) + d(B, C) + d(C, A)) &= \frac{3}{2} \dim V_*(.A.B.C.) + \dim V_*(.AB.C.) + \\ &\quad \dim V_*(.AC.B.) + \dim V_*(.A.BC.). \end{aligned}$$

Then, to prove that the matrix M_a is indeed an approximate solution, it suffices to show that there is a constant ρ such that

$$d(M_a; A, B, C) \leq \rho d(M; A, B, C),$$

for any given matrices A, B, and C.

It is possible to demonstrate that $\frac{4}{3}$ is an approximate factor for our solution, as follows:

$$\begin{aligned} d(M_a; A, B, C) &= 2 \dim V_*(.A.B.C.) + \dim V_*(.AB.C.) + \dim V_*(.AC.B.) + \\ &\quad + \dim V_*(.A.BC.) \\ &\leq \frac{4}{3}[\frac{3}{2} \dim V_*(.A.B.C.) + \dim V_*(.AB.C.) + \dim V_*(.AC.B.) + \\ &\quad + \dim V_*(.A.BC.)] \\ &\leq \frac{4}{3}[\frac{1}{2}(d(A, B) + d(B, C) + d(C, A))] \\ &\leq \frac{4}{3} d(M; A, B, C). \end{aligned}$$

Thus, we proved that $d(M_a; A, B, C)$ is at most $\frac{4}{3}$ times $d(M; A, B, C)$.

3.5 Linear time 4/3-approximation algorithm for the median distance

We saw that it is possible to compute a matrix median by solving a system of linear equations. However, if we are interested in approximating the median total score only, then a linear time algorithm can be used, as described in this section.

We begin by defining canonical bases for the partition subspaces introduced in Section 3.2.2. As we saw, a collection of bases for each V-subspace is necessary to remove the ambiguity on the V_*-subspaces. Then we relate these bases to orbits of permutations obtained from the input genomes, and show that the dimensions involved in computing the median distance are the result of counting orbits and their combinations, all doable in linear time.

3.5.1 Canonical bases

Suppose we need a base for the subspace $V(.AB.)$, where matrices A and B correspond to permutations α and β, respectively. The next results show us how to obtain a base for this subspace.

Theorem 2. *If $n \times n$ matrix P corresponds to permutation π, then the following set is a base for the subspace* $\ker(P - I)$*:*

$$B = \{\sum L \mid L \text{ is an orbit of } P\},$$

where the orbits of P are the orbits of π with each element replaced by the corresponding unit vector.

Proof. Orbits are minimal sets $L \subseteq E$ such that $PL = L$. We will show that (i) $B \subseteq \ker(P-I)$, (ii) the vectors in B are linearly independent, and (iii) every vector in $\ker(P-I)$ is a linear combination of vectors in B.

Let's start with (i). If L is an orbit of P, we have

$$P \sum L = P \sum_{a \in L} a = \sum_{a \in L} Pa = \sum_{P^{-1}b \in L} b = \sum_{b \in PL} b = \sum_{b \in L} b = \sum L.$$

Therefore, $\sum L \in \ker(P - I)$.

We need to prove next that the vectors in B are linearly independent. If $L_1, L_2, \ldots, L_k$ are orbits of P, and

$$c_1 \sum L_1 + c_2 \sum L_2 + \ldots + c_k \sum L_k = 0,$$

then each c_i must be zero, because there is at least one element in each L_i and this element does not appear in any other L_j with $i \neq j$.

Finally, we need to show that B generates $\ker(P - I)$. Consider a vector v such that $(P - I)v = 0$. We begin by showing that the coefficient in v of the unit vector of any element $a \in E$ is the same as the coefficient of any other element in a's orbit. To retrieve the coefficient of an element $a \in E$ in v we can perform a scalar product

$$a^t v,$$

where a^t is the row vector obtained by transposing the column vector a.

Recall that permutation matrices such as P satisfy $P^t = P^{-1}$. The coefficient of Pa in v is therefore

$$(Pa)^t v = a^t P^t v = a^t P^{-1} v = a^t v,$$

that is, it is equal to the coefficient of a in v. Using a similar argument, we see that the same happens with $P^2 a$, $P^3 a$, and so on. But these are exactly the elements that form a's orbit in P.

We conclude that, for each orbit L of P, there is a common coefficient c_L in v for unit vectors of this orbit. Therefore, v can be written as

$$v = \sum_{L \text{ orbit of } P} c_L \sum L,$$

showing that, in fact, v is a linear combination of vectors in B.

□

We denote by $\text{orb}(\pi)$ the set of orbits of a permutation π. This is a partition of the set E. The following corollary provides an alternative proof for a result in Section 3.1. Recall that $||\pi|| = n - |\text{orb}(\pi)|$.

Corollary 3. $\dim \ker(P - I) = |orb(\pi)|$.

The base B defined in Theorem 2 will be called the **canonical base** of $\ker(P - I)$. A base for $V(.AB.)$ can be obtained from canonical bases as follows. Recall that $v \in V(.AB.)$ when

$$Av = Bv,$$

which is equivalent to

$$(A^{-1}B - I)v = 0$$

when A is invertible, which is the case if it corresponds to a permutation.

We conclude that $V(.AB.)$ is the same as $\ker(A^{-1}B - I)$, for which we have a canonical base as seen in Theorem 2. Therefore, $\dim V(.AB.) = |\text{orb}(\alpha^{-1}\beta)|$. Notice also that

switching A with B we get $\beta^{-1}\alpha$, which is the inverse of $\alpha^{-1}\beta$ and therefore has the exact same orbits.

To define a canonical base for the subspace $V(.ABC.)$, we need to combine orbits from, say, $\alpha^{-1}\beta$ and $\beta^{-1}\gamma$. If a vector v belongs to $V(.ABC.)$, then $Av = Bv$, but also $Bv = Cv$. If L and K are orbits of $\alpha^{-1}\beta$ and $\alpha^{-1}\gamma$, respectively, such that $L \cap K \neq \emptyset$, then the coefficients of all elements of L and K have to be equal in v. This extends to all orbits that intersect. It follows that we need to replace orbits that intersect by their union, until no further intersections remain. This is the process of obtaining the finest partition that is refined by both $\mathrm{orb}(\alpha^{-1}\beta)$ and $\mathrm{orb}(\alpha^{-1}\gamma)$. We denote the resulting partition by

$$\mathrm{orb}(\alpha^{-1}\beta) \vee \mathrm{orb}(\alpha^{-1}\gamma).$$

A linear time algorithm to build this partition could be as follows. Build a graph with elements of E as vertices. For each orbit of $\alpha^{-1}\beta$, build a path linking its vertices. Do the same for $\alpha^{-1}\gamma$. The connected components of the resulting graph are the sets in $\mathrm{orb}(\alpha^{-1}\beta) \vee \mathrm{orb}(\alpha^{-1}\gamma)$.

Each set $L \in \mathrm{orb}(\alpha^{-1}\beta) \vee \mathrm{orb}(\alpha^{-1}\gamma)$ will yield a vector $\sum L$ for the canonical base of $V(.ABC.)$. Then

$$\begin{aligned}
\dim V(.ABC.) &= |\mathrm{orb}(\alpha^{-1}\beta) \vee \mathrm{orb}(\alpha^{-1}\gamma)|, \\
\dim V(.AB.C.) &= |\mathrm{orb}(\alpha^{-1}\beta)|, \\
\dim V(.AC.B.) &= |\mathrm{orb}(\alpha^{-1}\gamma)|, \\
\dim V(.A.BC.) &= |\mathrm{orb}(\beta^{-1}\gamma)|,
\end{aligned}$$

and from these values we can also write equations for the dimensions of the corresponding V_*-subspaces.

We need yet another value to compute the median distance: $\dim V_*(.A.B.C.)$. This dimension can be computed as follows:

$$\dim V_*(.A.B.C.) = n - \dim(V(.AB.C.) + V(.AC.B.) + V(.A.BC.)).$$

Putting everything together we get that the median distance is given by:

$$\begin{aligned}
d(M; A, B, C) = 2n- \\
|\mathrm{orb}(\alpha^{-1}\beta)| - |\mathrm{orb}(\alpha^{-1}\gamma)| - |\mathrm{orb}(\beta^{-1}\gamma)| + |\mathrm{orb}(\alpha^{-1}\beta) \vee \mathrm{orb}(\alpha^{-1}\gamma)|.
\end{aligned}$$

This value can be computed in linear time. Computation of inverses, products, and orbits of permutations can all be done in time linear in n, the number of elements of E. The computation of the finest refined partition can also be done in linear time by the algorithm described earlier in this section.

4 Conclusions

We showed in this paper that it is possible to define a distance on matrices in a way that yields exactly the algebraic distance when restricted to permutation matrices. And, more

importantly, that the median problem in matrices can be approximated to a factor of $\frac{4}{3}$ using standard linear algebra tools, which lead to polynomial time algorithms. The implications to computing algebraic genome medians can be significant.

5 Acknowledgements

We thank Luiz Antonio Barrera San Martin who suggested linear representations in a discussion on permutations.

References

[1] G. Bourque and P. A. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12(1):26–36, 2002.

[2] P. Feijao and J. Meidanis. SCJ: a breakpoint-like distance that simplifies several rearrangement problems. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:1318–1329, 2011.

[3] P. Feijao and J. Meidanis. Extending the Algebraic Formalism for Genome Rearrangements to Include Linear Chromosomes. In M. de Souto and M. Kann, editors, *BSB 2012, LNBI 7409*, pages 13–24, Heidelberg, 2012. Springer-Verlag Berlin.

[4] K. M. Hoffman and R. Kunze. *Linear Algebra*. Prentice Hall, 2nd edition, 1971.

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

[6] B. M. Moret, L. S. Wang, T. Warnow, and S. K. Wyman. New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17 Suppl 1(suppl 1):S165–S173, 2001.

[7] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5(3):555–570, 1998.

[8] E. Tannier, C. Zheng, and D. Sankoff. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10:120, 2009.

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