# On the Matrix Median Problem

João Paulo Pereira Zanetti[1], Priscila Biller[1], and João Meidanis[1,2]

[1] Institute of Computing, University of Campinas, SP, Brazil
[2] Scylla Bioinformatics, Campinas, Brazil

**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 general form of the problem, the matrix median problem. It is known that, for any metric distance, at least one of the corners is a $\frac{4}{3}$-approximation of the median. Our results allow us to compute up to three additional matrix median candidates, all of them with approximation ratios at least as good as the best corner, when the input matrices come from genomes. From the application point of view, it is usually more interesting to locate medians farther from the corners. We also show a fourth median candidate that gives better results in cases we tried. However, we do not have proven bounds for this fourth candidate yet.

## 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 parsimonious 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 rearrangement models, with notable exceptions being the multichromosomal breakpoint distance [10] and the Single-Cut-or-Join (SCJ) model [3]. 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 [8], the more recent GRAPPA [7], and MGR [1]).

In this paper we will focus on the algebraic rearrangement model proposed by Meidanis and Dias [6], recently extended to allow linear chromosomes in a very natural way by Feijão and Meidanis [4]. This extended algebraic rearrangement model is similar to the well-known Double-Cut-and-Join (DCJ) model [11], 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. The median problem can be stated as follows: given three genomes $\pi_1$, $\pi_2$, and $\pi_3$, and a distance metric $d$, find a genome $\mu$ 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 paper 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.

It is known that, for any distance satisfying the axioms of a metric, at least one of the corners is a $\frac{4}{3}$-approximation of the median [9]. Our results allow us to compute up to three additional matrix median candidates, all of them with approximation ratios at least as good as the best corner, when the input matrices come from genomes. Also, application-wise, it is usually more interesting to locate medians farther from the corners [5].

The rest of this paper is organized as follows. In Section 2, we have basic definitions regarding the algebraic adjacency theory needed in this work. In Section 3, we show how genomes can also be seen as matrices, define the matrix median problem, show our results, and propose an algorithm. Finally, in Section 4, we present our conclusions.

## 2 Algebraic Rearrangement Theory

We will start this section showing some basic definitions of the algebraic theory of Feijão and Meidanis [4].

### 2.1 Permutations

Given a set $E$, a *permutation* $\alpha : E \to E$ is a bijective map from $E$ onto itself. 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 $b$, $b$ 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 $\alpha$, $\beta$ 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 $\alpha$ has an *inverse* $\alpha^{-1}$ 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 $\alpha$ is a representation of $\alpha$ as a product of 2-cycles, not necessarily disjoint. All permutations have a 2-cycle decomposition. The *norm* of a permutation $\alpha$, denoted by $\|\alpha\|$, is the minimum number of cycles in a 2-cycle decomposition of $\alpha$. For example, the permutation $\alpha = (a\ b\ c\ d)$ can be decomposed as $(a\ b)(b\ c)(c\ d)$, and $\|\alpha\| = 3$.

### 2.2 Modeling Genomes as Permutations

To model genomes with the Algebraic Theory, the formulation is similar to the set representation of a genome, used in several related works [10,3]. 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 (the telomeres may be omitted, when the gene set is given) where each extremity appears at most once.

According to algebraic rearrangement theory [4], 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 paper, the *distance* between two genomes or two permutations $\pi$ and $\sigma$ will be defined as $d(\pi, \sigma) = \|\sigma\pi^{-1}\|$. 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 paper.

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 of a genome and its representation as a permutation.

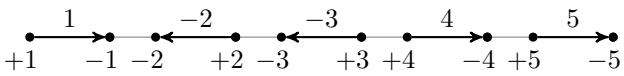


**Fig. 1.** A genome with one linear chromosome, represented by the permutation $\pi = (-1\ -2)(+2\ -3)(+3\ +4)(-4\ +5)$.

With these definitions, a distance between two genomes $\sigma$ and $\pi$ can be defined as $d(\pi, \sigma) = \|\sigma\pi^{-1}\|$, as mentioned in the introduction. The resulting distance agrees with DCJ in circular genomes, and in general these two distances are very close.

## 3 Results

Permutations can be seen as matrices, and, 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, it applies to all square matrices, not only those associated to permutations. Such a metric can be useful in the computation of genome medians, and we show here how to compute an approximate solution to the matrix median problem, by solving a system of linear equations.

### 3.1 Matrix Distance

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

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

where $r(X)$ denotes the **rank** of matrix $X$. It is well-known that

$$r(X) = \dim \mathrm{im}(X), \tag{1}$$

where dim denotes the dimension of a vector space and $\mathrm{im}(X)$ is the **image** of $X$, namely, the space of all vectors that can be written as $Xv$ for some $v \in \mathbb{R}^n$. Here we treat vectors as column matrices, that is, $n \times 1$ matrices. Therefore, the distance satisfies

$$d(A, B) = \dim \mathrm{im}(B - A).$$

for every pair of matrices $A$ and $B$. It is also well-known that

$$r(X) = n - \dim \ker(X), \tag{2}$$

for every $x \times n$ matrix $X$, where $\ker(X)$ is the **kernel** of $X$, defined as the space of all vectors $v \in \mathbb{R}^n$ such that $Xv = 0$, so the distance satisfies yet another formula:

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

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 $A = B$ [2].

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 $v$ of $E$ with a unit vector of $\mathbb{R}^n$, and then define $A$, the matrix counterpart of $\alpha$, so that

$$Av = \alpha v. \tag{3}$$

In Equation 3, we use $v$ both as a unit vector of $\mathbb{R}^n$ in the left side, and as an element of $E$ in the right side.

An example will help. Let $\alpha$ 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 $\alpha$, and $B$ is the matrix corresponding to $\beta$. If $\alpha$ happens to be a genome, that is, if $\alpha^2 = \mathbf{1}$, then $A$ is a symmetric matrix and vice-versa.

We now show that this mapping is distance-preserving.

**Lemma 1.** *For any permutation $\sigma$ and $\pi$, and their respective associated matrices $S$ and $P$, we have:*

$$d(\sigma, \pi) = d(S, P).$$

*Proof.* First, notice that it suffices to show that

$$\|\alpha\| = r(A - I), \tag{4}$$

for any permutation $\alpha$ and associated matrix $A$. Indeed, if $P$ is invertible then

$$r(S - P) = r(SP^{-1} - I),$$

and then Equation (4) 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 (4) 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 $\alpha$ and $\beta$, 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

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

or

$$r(AB - I) = r(A - I) + r(B - I),$$

because of Equation 2.

Therefore, if Equation (4) is valid for $\alpha$ and $\beta$, and if $\alpha$ and $\beta$ are disjoint, then Equation (4) is valid for the product $\alpha\beta$. Since any permutation can be written as a product of disjoint cycles, Equation (4) is valid in general. □

### 3.2 Matrix Median

Because the correspondence between permutations and matrices preserves distances, it makes sense to study the matrix median problem as a way of shedding light into the permutation median problem, which in turn is related to the genome median problem.

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 $\mathbb{R}^n$ as a direct sum of five subspaces, where the following relations are true: (1) $A = B = C$, (2) $A = B \neq C$, (3) $A \neq B = C$, (4) $A = C \neq B$, and (5) $A \neq B \neq C \neq A$.

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 $M$ would be to mimic one of $A$, $B$, or $C$.

Therefore, making $M$ equal to $A$, except in the third subspace, where it should be equal to $B$ (and $C$) 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.

### 3.3 Partitioning $\mathbb{R}^n$

We begin by introducing notation aimed at formalizing subspaces such as $A = B \neq C$. 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:

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

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 distinct classes actually disagree, that is, subspaces where vectors have different images under each class. There can be more than one subspace satisfying this property, but we can use orthogonality to define a unique subspace. For a partition $p$, we define $V_*(p)$ as the orthogonal complement of the sum of the partitions strictly refined by $p$ with respect to $V(p)$:

$$V_*(p) = V(p) \cap (\sum_{p<q} V(q))^{\perp}$$

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 orthogonal to the sum of the subspaces corresponding to coarser partitions. Notice that $p < q$ implies $V(q) \subseteq V(p)$.

It is easy to see that the $V_*$ subspaces are pairwise disjoint, but this is not enough to prove that their direct sum is $\mathbb{R}^n$. So, the proof will start from the basic sum $V_*(.ABC.) \oplus V_*(.AB.C.)$, where we already know that $V_*(.ABC.) \cap V_*(.AB.C.) = \{0\}$, and it will add one subspace in the sum at a time, ensuring that the intersection between the new subspace and the sum of the subspaces previously included contains the zero vector only.

**Lemma 2.** *If $A$, $B$, and $C$ are permutation matrices, then*

$$(V_*(.ABC.) + V_*(.AB.C.)) \cap V_*(.BC.A.) = \{0\}.$$

*Proof.* Let $v$ be a vector such that

$$v \in (V_*(.ABC.) + V_*(.AB.C.)) \cap V_*(.BC.A.).$$

We have that $Av = Bv$ and also $Av = Cv$. Therefore $Av = Bv = Cv$ and, consequently, $v = 0$, since $v \in V_*(.ABC.)$. □

Before we add more subspaces to the sum, we will need the result of the following lemma.

**Lemma 3.** *If $A$, $B$, and $C$ are permutation matrices, and if*

$$2Bv = Av + Cv$$

*for a given vector $v$, then $Av = Cv$.*

*Proof.* Denote by $|x|$ the norm of a vector $x$. Note that $A$, $B$, and $C$ preserve norms, thus

$$|Av + Cv| = |2Bv| = 2|v| = |v| + |v| = |Av| + |Cv|.$$

But, if the norm of the sum is equal to the sum of the norms, then the two vectors are parallel, and have the same orientation. In other words, there is a positive scalar $c$ such that $cAv = Cv$. But we already have that $|Av| = |v| = |Cv|$. Therefore, $c = 1$. □

**Lemma 4.** *If $A$, $B$, and $C$ are permutation matrices, then*

$$(V_*(.ABC.) + V_*(.AB.C.) + V_*(.BC.A.)) \cap V_*(.AC.B.) = \{0\}.$$

*Proof.* Suppose that $u + v + w \in V_*(.AC.B.)$, where $u \in V_*(.ABC.)$, $v \in V_*(.AB.C.)$, and $w \in V_*(.BC.A.)$. We have that $A(u + v + w) = C(u + v + w)$, which implies $A(v + w) = C(v + w)$, since $Au = Cu$. Thus, we have

$$\begin{aligned} Av + Aw &= Cv + Cw \\ Bv + Aw &= Cv + Bw \\ B(v - w) &= Cv - Aw. \end{aligned}$$

Now we will apply $A$ and $C$ to $v - w$, and sum the results, obtaining:

$$\begin{aligned} A(v - w) &= Bv - Aw \\ C(v - w) &= Cv - Bw \\ A(v - w) + C(v - w) &= B(v - w) + Cv - Aw = 2B(v - w) \end{aligned}$$

By Lemma 3, we conclude that $A(v - w) = C(v - w)$. But we also have that $A(v + w) = C(v + w)$, which implies $Av = Cv$ and $Aw = Cw$. In other words, $u + v + w \in V_*(.ABC.)$. But $u + v + w$ also belongs to $V_*(.AC.B.)$. It follows that $u + v + w = 0$. □

**Lemma 5.** *If $A$, $B$, and $C$ are permutation matrices, then*

$$(V_*(.ABC.) + V_*(.AB.C.) + V_*(.BC.A.) + V_*(.AC.B.)) \cap V_*(.A.B.C.) = \{0\}.$$

*Proof.* By definition, the two subspaces whose intersection is indicated in the left-hand side are orthogonal to one another. Therefore, they have a zero intersection. □

With Lemmas 2 through 5 we get to the following theorem:

**Theorem 1.** *If $A$, $B$, and $C$ are permutation matrices, then*

$$\mathbb{R}^n = V_*(.ABC.) \oplus V_*(.AB.C.) \oplus V_*(.BC.A.) \oplus V_*(.AC.B.) \oplus V_*(.A.B.C.).$$

It is important to observe that Theorem 1 does not apply to general matrices, for instance:

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

With these three matrices, we have

$$V_*(.ABC.) = \{0\},$$
$$V_*(.AB.C.) = \left\langle [1\ 0]^t \right\rangle,$$
$$V_*(.BC.A.) = \left\langle [1\ 0]^t, [0\ 1]^t \right\rangle,$$
$$V_*(.AC.B.) = \left\langle [0\ 1]^t \right\rangle,$$

where $\langle X \rangle$ denotes the space spanned by the set $X$.

We can see that, in this case,

$$(V_*(.ABC.) + V_*(.AB.C.) + V_*(.BC.A.)) \cap V_*(.AC.B.) \neq \{0\}.$$

### 3.4 Computing Median Candidates

We now get into further detail in the procedure to compute these matrices. We saw that, when $A$, $B$, and $C$ are permutation matrices, $\mathbb{R}^n$ can be decomposed into a direct sum of $V_*$ subspaces. We will now implement the procedure outlined in Section 3.2, summarized in Table 1, to obtain a median candidate $M_A$.

**Table 1.** Distance contribution — Given three permutation matrices $A$, $B$ and $C$, this table shows the distance contribution of each of the five subspaces partitioning $\mathbb{R}^n$ to the distances $d(M_A, A)$, $d(M_A, B)$, and $d(M_A, C)$, for a candidate median matrix $M_A$.

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

One way to implement this strategy is to compute orthogonal projection matrices $P_1$, $P_2$, $P_3$, $P_4$, and $P_5$ for each of the subspaces and then compute $M_A$ as follows:

$$\begin{aligned} M_A &= AP_1 + AP_2 + BP_3 + AP_4 + AP_5 \\ &= A + (B - A)P_3\,, \end{aligned}$$

since $P_1 + P_2 + P_3 + P_4 + P_5 = I$. To obtain each matrix $P_i$ all we need is a $n \times k_i$ matrix $B_i$ whose columns form an orthonormal basis of the corresponding subspace, because $P_i$ can then be written as $B_i B_i^t$.

To build the $B_i$s, one possibility is to use Function Add below, a basic routine to expand an orthonormal basis so that it can generate a given extra vector $v$. It projects the new vector in the orthogonal complement of the subspace generated by the original basis and then adds the normalized projection to form the new basis. Function Add's complexity is $O(kn)$ arithmetic operations (additions, subtractions, multiplications, and divisions).

**Function** Add($B$, $v$) Augments an orthonormal basis $B$ so that it can generate the vector $v$.

**Data**: An orthonormal basis $B = \{v_1, \ldots, v_k\}$ and a vector $v$.
**Result**: An augmented basis.

```
1 w ← v − Σ_{i=1}^{k} v_i v^t v_i
2 if w = 0 then
3 |   Return B
4 else
5 |   Normalize w
6 |   Return B ∪ {w}
```

Algorithm 1 uses Function Add and a linear system solving method Solve, that returns a linearly independent set of vectors spanning the solution space, to determine orthonormal bases for each of the five $V_*$ subspaces. Assuming that we can solve a system of $n$ linear equations in time $O(n^3)$, the total complexity for Algorithm 1 is $O(n^3)$ as well.

Once we have determined orthonormal bases for the $V_*$ subspaces, it is easy to find projection matrices for them. For each $B_i$, its projection matrix is $P_i = B_i B_i^t$, as previously said.

Knowing the projection matrices, we can then finally compute the median candidates, in Algorithm 2. Previously, we saw how to compute $M_A$. It is also possible to define $M_B$ and $M_C$ in an analogous way. The matrix $M_B$ follows $B$ in $V_*(.A.B.C.)$ instead of $A$, and $M_C$ follows $C$. The entire computation takes $O(n^3)$ arithmetic operations.

We can also define a median candidate $M_I$, that follows the identity in the subspace $V_*(.A.B.C.)$. That is,

$$M_I = A(P_1 + P_2 + P_4) + BP_3 + P_5 = A + (B - A)P_3 + (I - A)P_5 \,.$$

We still have no proven results on its total median score, but we conjecture that, for genomic matrices, $M_I$ is better than $M_A$, $M_B$ and $M_C$, or it might even be a median. This is due to the symmetric nature of the genomic matrices.

**Algorithm 1:** Computation of orthonormal bases for each of the $V_*$ subspaces.

**Data**: Three $n \times n$ permutation matrices $A$, $B$ and $C$.
**Result**: The bases for the $V_*$ subspaces.

$B \leftarrow \emptyset$
$L \leftarrow$ Solve $\{(A - B)v = 0,\ (A - C)v = 0\}$
**foreach** $v \in L$ **do**
    $B \leftarrow$ `Add(`$B$`,`$v$`)`
$B_1 \leftarrow B$ `// Basis for` $V_*(.ABC.)$
$L \leftarrow$ Solve $\{(A - B)v = 0\}$
**foreach** $v \in L$ **do**
    $B \leftarrow$ `Add(`$B$`,`$v$`)`
$B_2 \leftarrow B - B_1$ `// Basis for` $V_*(.AB.C.)$
$L \leftarrow$ Solve $\{(B - C)v = 0\}$
**foreach** $v \in L$ **do**
    $B \leftarrow$ `Add(`$B$`,`$v$`)`
$B_3 \leftarrow B - B_1 - B_2$ `// Basis for` $V_*(.BC.A.)$
$L \leftarrow$ Solve $\{(A - C)v = 0\}$
**foreach** $v \in L$ **do**
    $B \leftarrow$ `Add(`$B$`,`$v$`)`
$B_4 \leftarrow B - B_1 - B_2 - B_3$ `// Basis for` $V_*(.AC.B.)$
$L \leftarrow$ canonical basis for $\mathbb{R}^n$
**foreach** $v \in L$ **do**
    $B \leftarrow$ `Add(`$B$`,`$v$`)`
$B_5 \leftarrow B - B_1 - B_2 - B_3 - B_4$ `// Basis for` $V_*(.A.B.C.)$

**Algorithm 2:** Computation of median candidates.

**Data**: Three $n \times n$ permutation matrices $A$, $B$ and $C$.
**Result**: Three median candidates $M_A$, $M_B$, and $M_C$.

Compute the bases $B_1$, $B_2$, $B_3$, $B_4$, and $B_5$ with Algorithm 1
**foreach** $i \in \{1, 2, 3, 4, 5\}$ **do**
    $P_i \leftarrow B_i B_i^t$
$M_A \leftarrow A + (B - A)P_3$
$M_B \leftarrow B + (A - B)P_4$
$M_C \leftarrow C + (B - C)P_2$

### 3.5 Approximation Factor

We already know that the direct sum of the $V_*$ subspaces is $\mathbb{R}^n$, when $A$, $B$, and $C$ are permutation matrices. Now we can use this property to express the distance in terms of these subspaces dimensions. The matrix $M_A$ will have a total distance to $A$, $B$, and $C$ equal to:

$$\begin{aligned} d(M_A; A, B, C) &= d(M, A) + d(M, B) + d(M, C) \\ &= \dim V_*(.BC.A.) + \\ &\quad \dim V_*(.A.B.C.) + \dim V_*(.AC.B.) + \\ &\quad \dim V_*(.A.B.C.) + \dim V_*(.AB.C.) \\ &= 2 \dim V_*(.A.B.C.) + \\ &\quad \dim V_*(.AB.C.) + \dim V_*(.AC.B.) + \dim V_*(.BC.A.). \quad (5) \end{aligned}$$

There are cases where $M_A$ is not a median, even if the input matrices are genomic. Take, for example:

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

By Equation (5), the matrix $M_A$ has the following total score:

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

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

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

Thus, given that the procedure described in the Section 3.4 does not guarantee a matrix median, it is interesting to know whether it is an approximation algorithm, namely, whether there is a constant $\rho$ such that the candidate's total score is at most $\rho$ times the score of a median.

In general, consider permutation 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 Equation (5), the median score of the approximate solution $M_A$ constructed in Section 3.4 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:

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

Then, to prove that the matrix $M_A$ is indeed an approximate solution, it suffices to show that there is a constant $\rho$ 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 $d(M; A, B, C)$. The same result holds for $M_B$ and $M_C$.

## 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. For the case where the input matrices represent genomes, we have shown how to compute matrices that approximate the median with factors at least as good as the best corner, that is, they are approximations to the median within a factor of $\frac{4}{3}$. In addition, we showed a construction, in the form of matrix $M_I$, that in several examples we worked out is even better than the other median candidates. We conjecture that this matrix might be a median if $A$, $B$, and $C$ come from genomes. 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, the Research Funding Agency of the State of Sao Paulo (FAPESP), for grants 2012/13865-7 and 2012/14104-0, and the anonymous referees for helpful comments that improved the paper significantly.

## References

1. Bourque, G., Pevzner, P.A.: Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research 12(1), 26–36 (2002)
2. Delsarte, P.: Bilinear forms over a finite field, with applications to coding theory. Journal of Combinatorial Theory, Series A 25(3), 226–241 (1978)
3. Feijao, P., Meidanis, J.: SCJ: a breakpoint-like distance that simplifies several rearrangement problems. IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics 8, 1318–1329 (2011)
4. Feijao, P., Meidanis, J.: Extending the Algebraic Formalism for Genome Rearrangements to Include Linear Chromosomes. In: de Souto, M., Kann, M. (eds.) BSB 2012, LNBI 7409. pp. 13–24. Springer-Verlag Berlin, Heidelberg (2012)
5. Haghighi, M., Sankoff, D.: Medians seek the corners, and other conjectures. BMC bioinformatics 13(Suppl 19), S5 (2012)
6. Meidanis, J., Dias, Z.: An alternative algebraic formalism for genome rearrangements. In: Sankoff, D., Nadeau, J. (eds.) Comparative Genomics: Empirical and Analyitical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families, pp. 213–223. Kluwer Academic Publishers (2000)
7. Moret, B.M., Wang, L.S., Warnow, T., Wyman, S.K.: New approaches for reconstructing phylogenies from gene order data. Bioinformatics 17 Suppl 1(suppl 1), S165–S173 (2001)
8. Sankoff, D., Blanchette, M.: Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology 5(3), 555–570 (1998)
9. Sankoff, D., Sundaram, G., Kececioglu, J.: Steiner points in the space of genome rearrangements. International Journal of Foundations of Computer Science 7(01), 1–9 (1996)
10. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics 10, 120 (2009)
11. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21(16), 3340–3346 (2005)