# On the matrix median problem

João Paulo Pereira Zanetti, Priscila Biller, and João Meidanis
jppzanetti@gmail.com

## Genome median

The *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-Hard in most rearrangement models.

## Algebraic rearrangement model

Our ultimate goal is to investigate the problem of computing the algebraic median of three genomes. According to algebraic rearrangement theory [1], 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. The *distance* between two genomes $\pi$ and $\sigma$ is defined as $||\sigma\pi^{-1}||$, where $||\alpha||$ designates the *norm* of a permutation.

## Why use matrices?

Since we believe the GMP is NP-Hard under the algebraic adjacency model, we look for answers by relaxing the restrictions of the problem.

One way to do that is to accept any permutations, not just genomic ones. Moreover, we can generalize even further the problem, accepting matrices in general.

First, note that permutations can be seen as matrices. Given a permutation $\alpha$, its corresponding matrix can be obtained by permuting the columns or rows of the identity matrix according to $\alpha$. For example,

$$(a\ b)(c\ d) \longrightarrow \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

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

$$d(A, B) = \operatorname{rank}(B - A).$$

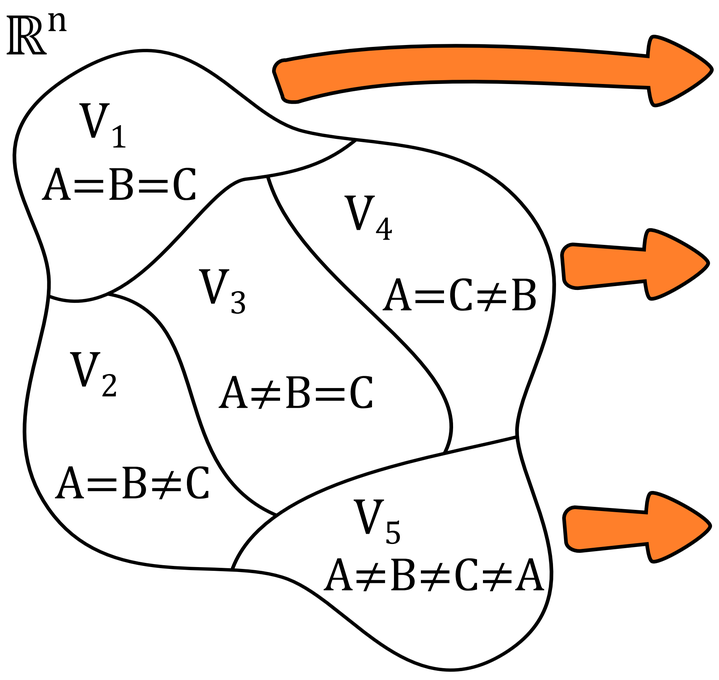
Since the correspondence between permutations and matrices preserves distances, it makes sense to study the matrix median problem as a way of sheding light into the permutation median problem, which in turn is related to the genome median problem.

Let $A$, $B$, and $C$ be three $n \times n$ matrices. We want to find a matrix $M$ that minimizes the total score

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

## Partitioning $\mathbb{R}^n$ to define candidates

When the input matrices $A$, $B$ and $C$ are permutation matrices, we can decompose $\mathbb{R}^n$ into a direct sum of five subspaces, according to the behavior of the three matrices, as illustrated in the figure below.




## Approximation algorithm

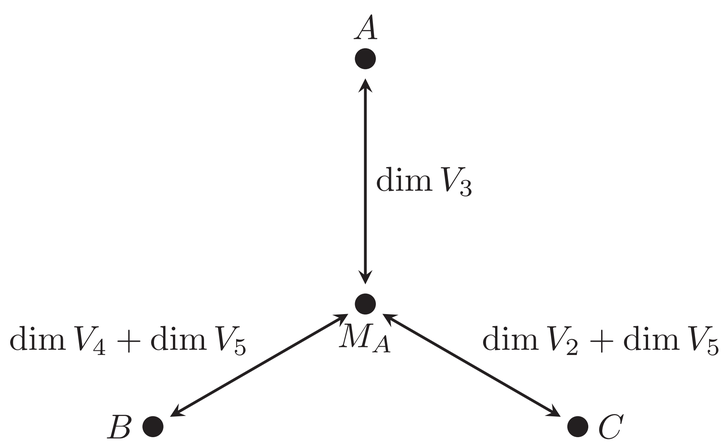
We then have the following algorithm for permutation matrices, in three steps:

1. Determine the $V_i$ subspaces
2. For each $V_i$ subspace, compute a projection matrix $P_i$
3. $M_A \leftarrow A + (B - A)P_3$
   $M_B \leftarrow B + (A - B)P_4$
   $M_C \leftarrow C + (B - C)P_2$

Notice that the three candidate matrices have the same total score:

$$\dim V_2 + \dim V_3 + \dim V_4 + 2 \dim V_5.$$

The following figure shows how each subspace contibutes to the distances between each matrix, using $M_A$ as en example.




## Approximation factor

Preliminary tests suggested an approximation factor of $\frac{4}{3}$. We prove that it is indeed the case, using the ratio between the candidate's total score and the trivial lower bound.

That means that, at worst, the matrices $M_A$, $M_B$, and $M_C$ have a total score $\frac{4}{3}$ times larger than the median.

## Acknowledgments

The authors thank the funding agencies CAPES and FAPESP for the financial support, and the University of Campinas for their infrastructure.

## Further information

An article detailing the problem and the approximation algorithm will be presented at the Workshop on Algorithms in Bioinformatics (WABI) 2013, in Sophia Antipolis, France [2].

## References

[1] P. Feijão and J. Meidanis. Extending the Algebraic Formalism for Genome Rearrangements to Include Linear Chromosomes. In *BSB 2012, Brazilian Symposium on Bioinformatics.*

[2] J. P. P. Zanetti, P. Biller and J. Meidanis. On the matrix median problem. In *WABI 2013, Workshop on Algorithms in Bioinformatics.*