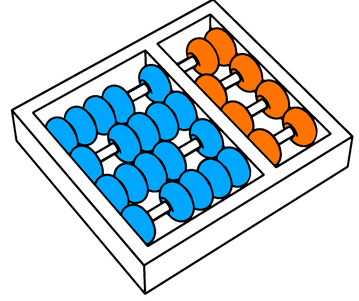
On the algebraic genome median

João Meidanis, Pedro Feijão, Priscila Biller and João P. P. Zanetti
jppzanetti@gmail.com

UNICAMP



Problem

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 main 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 π and σ is defined as $||\sigma\pi^{-1}||$, where $||\alpha||$ designates the *norm* of a permutation.

Why use matrices?

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 $Xv = 0$.

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

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

Approximation algorithm

The definition of the matrix distance suggests the following strategy. Decompose $\mathbb{R}^n$ into a direct sum of the subspaces below:

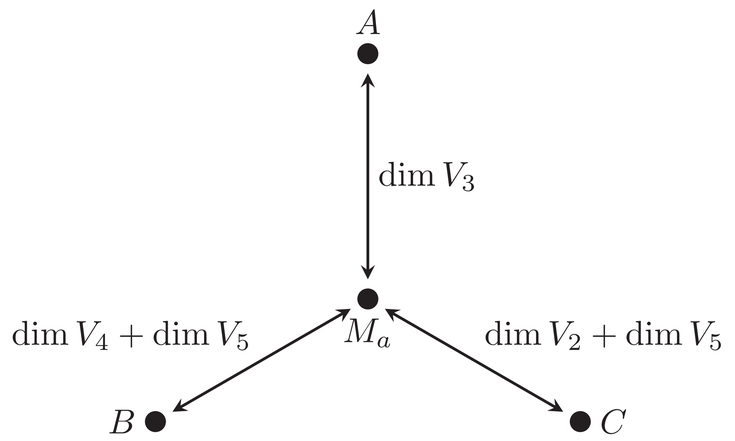
- V_1, where $A = B = C$;
- V_2, where $A = B \neq C$;
- V_3, where $A \neq B = C$;
- V_4, where $A = C \neq B$;
- V_5, where $A \neq B \neq C \neq A$.

We then make our median candidate M_a equal to A, except in the third subspace, where it should be equal to B and C. That procedure tries to minimize the pairwise distances and should yield a good approximation of a median, if not a median.

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

$$d(M_a; A, B, C) = \dim V_2 + \dim V_3 + \dim V_4 + 2 \dim V_5.$$

The following figure shows how each subspace contibutes to the distances between each matrix.



Example

Here's an example of a computation using our procedure that leads to a median.

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

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

$$d(M_a; A, B, C) = 1 + 0 + 1 + 0 = 2$$

From the triangle inequality, we get a lower bound for a median M that is

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

Therefore, M_a is a median for A, B, and C.

Counter-Example

Here we have an instance for the matrix median problem for which M_a isn't a median.

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

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

$$d(M_a; A, B, C) = 0 + 0 + 0 + 4 = 4$$

But the identity matrix I has a better score than M_a:

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

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

Furthermore, looking at the lower bound, I is actually a median in this case.

Approximation factor

The counter-example suggests an approximation factor of $\frac{4}{3}$. We prove that it is indeed the case, using the ratio between $d(M_a; A, B, C)$ and the trivial lower bound.

That means that, at worst, the matrix M_a has a 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

A detailed report of this work will be made available soon in a technical report of the IC-Unicamp, in http://www.ic.unicamp.br/~reltech/.

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.*