Center Genome With Respect to the Rank Distance

P. Biller[1,2][0000−0002−2937−5397], J. P. P. Zanetti[4][0000−0002−9955−7751], and J. Meidanis[1,3][0000−0001−7878−4990]

[1] Institute of Computing, University of Campinas, Campinas, Brazil
[2] Department of Mathematics, Simon Fraser University, Burnaby, Canada
[3] Scylla Bioinformatics, Campinas, Brazil
[4] R&D Seismology and Acoustics, Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands

Abstract. The rank distance between matrices has been applied to genome evolution, specifically in the area of genome rearrangements. It corresponds to looking for the optimal way of transforming one genome into another by cuts and joins with weight 1 and double-swaps with weight 2. In this context, the genome median problem, which takes three genomes A , B, and C and aims to find a genome M such that $d(A, M) + d(B, M) + d(C, M)$ is minimized, is relevant. This problem can be stated for any genomic distance, not just the rank distance. In many cases, the genome median problem is NP-hard, but a number of approximate methods have been developed.
Here we examine a related problem, the so-called center genome problem, where we aim to minimize the maximum (instead of the sum) of pairwise distances between the center genome and the inputs. We show that, for the rank distance, and for two genomic inputs A and B, it is not possible to always attain the well-known lower bound $\lceil d(A, B)/2 \rceil$. The issue arises when A and B are co-tailed genomes (i.e., genomes with the same telomeres) with $d(A, B)$ equal to twice an odd number, when the optimal attainable score is 1 unit larger than the lower bound. In all other cases, we show that the lower bound is attained.

Keywords: Genome rearrangements, genome matrices

1 Introduction

The rank distance between matrices has been very successfully used in coding theory since at least 1985, when Gabidulin published his discoveries in matrix codes [5]. Recently, applications of the rank distance to genome evolution, specifically in the area of genome rearrangements, started to emerge [9]. In this context, the genome median problem, which takes a number of genomes A_1 , A_2, ..., A_k and aims to find a genome M such that $d(A_1, M) + d(A_2, M) + \ldots + d(A_k, M)$ is minimized, is relevant. This problem can be stated for any genomic distance, not just the rank distance. In many cases, the genome median problem is NP-hard, but a number of approximate methods have been developed.

With regard to genome medians, much work has been published, especially in the case of exactly three inputs. This is one of the seminal steps in building phylogenetic trees. Finding a genome median is NP-hard for several genome distances, with the exception of SCJ and breakpoint for multichromosomal genomes. [8, 4].

Center genomes, also called closest genomes or minimax genomes, are also aimed at somehow representing all the inputs, as a sort of average genome. The center genome problem takes genome inputs A_1 , A_2, ..., A_k and looks for a genome M minimizing $\max(d(A_1, M), d(A_2, M), \ldots, d(A_k, M))$. There is an important difference between using central genomes and median genomes as subroutines for ancestral reconstruction methods: when just two inputs are used for the median, the solution will probably be not very relevant, because many solutions exist, including both input genomes and anything in an optimal path from one to the other; on the other hand, the center genome, even with just two inputs, is already restricted enough to be relevant with respect to ancestral genomes.

For any distance defined as the minimum number of operations, when all operations have the same weight, clearly the theoretical lower bound for two genomes is readily achievable: it suffices to start at one of the genomes and walk towards the other, stopping when the right number of steps have been performed. However, if an arbitrary number of inputs is allowed, the problem becomes NP-hard, even for very simple distances such as the SCJ [2].

In contrast, distance measures where operations have distinct weights may not be able to always attain the lower bound. Here we concentrate on two inputs and examine the rank distance, which can be defined as the rank of $A - B$ for genomes (matrices) A and B, but also as the minimum number of cuts, joins, and double swaps, with weights 1, 1, and 2, respectively, that bring one genome to the other. Since we have different weights, it is not obvious the lower bound can be achieved. In fact, we show that it cannot in the case where $d(A, B) = 2n$ with n odd. In all other cases, the lower bound is achieved.

The rest of this paper is organized as follows. Section 2 contains the definitions used throughout the text. Section 3 presents the results. Finally, Section 4 summarizes our work and points to possible continuation of this research.

2 Definitions

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

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

For genomes with just one gene, we have just two extremities. There are only two genomes: one with an adjacency linking these two extremities, and the other

with just telomeres. 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 therefore square matrices of size $(2n) \times (2n)$ 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 $A^\top = A$.
- They are **orthogonal** matrices, that is, they satisfy $A^\top = A^{-1}$.
- 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.

Extremities x such that $Ax = x$ are called **telomeres** of A. A genome with no telomeres is called **circular**. Two genomes with exactly the same set of telomeres are called **co-tailed**.

3 Results

We recall a lower bound for the score relative to two genomes, 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.

We start by recalling the notion of **intermediate genomes**, defined as genomes that appear in an optimal scenario between two genomes A and B. The definition depends on A and B, so sometimes we will call them AB-intermediates for improved clarity. Although initially defined for DCJ [3], the definition works for any distance.

In addition to being optimal scenario members, intermediate genomes can be characterized as those for which the triangle inequality becomes an equality. They are also the medians of two genomes.

Given two genomes A, and B, a **center** genome for them is a genome M that minimizes the **score** $\mathrm{sc}(M; A, B)$, defined as:

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

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

Lemma 1. *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:

$$\mathrm{sc}(M;A,B) \geq \left\lceil \frac{d(A,B)}{2} \right\rceil. \tag{1}$$

It would be tempting to state the following conjecture:

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 center genome, since it would attain the lower bound established in Equation 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 center conjecture is **not** true.

3.1 Co-Tailed Genomes

When A and C are co-tailed, we do not always get a center genome satisfying the lower bound, but we can get within 1 unit of it. This case opens up the possibility of center genomes that are not intermediates, e.g., in the example of Section 3, the identity matrix is a center genome, but not an intermediate. Let's begin by studying properties of intermediate genomes between two co-tailed ones.

Lemma 2. *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 1 of a paper by Chindelevitch and Meidanis [1], 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 AC-intermediate genomes.

So let's assume that B has a telomere x not shared by A. In this case, at some point in an optimal operation series going from A to B, there must be a cut. However, any optimal such series can be extended to a sorting series going from A to C, since B is intermediate between A and C. However, no cuts can be present in an optimal scenario linking co-tailed genomes [6]. This 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 3. *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 2, 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

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

It is easy to find center genomes for co-tailed genomes whose distance is a multiple of 4. However, if their distance is not divisible by 4, we are forced to take the second best, which is 1 unit off the lower bound.

Lemma 4. *If A and C are co-tailed genomes and $d(A, C)/2$ is even, then there is a genome B satisfying the center 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 3. 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 AC-intermediate genome satisfying the center lower bound.

Lemma 5. *If A and C are co-tailed genomes and $d(A, C)/2$ is odd, then there is no genome B satisfying the center 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 2, B would be co-tailed with A. But then, by Lemma 3, $d(A, B) = d(A, C)/2$ would have to be even, contradicting the hypothesis.

Lemma 6. *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 $B = A$. 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 of 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.

3.2 Genomes Not Co-Tailed

If A and C are not co-tailed, then there are AC-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 7. *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 8. *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, destroying the adjacency of x in A 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 $d(A, C) \geq 2$. Co-tailed genomes can be sorted by double swaps, so 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 7, 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.

3.3 Main Result

Theorem 1. *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 8, since $0 \leq \lceil d(A, C)/2 \rceil \leq d(A, C)$. Part 2 is a consequence of Lemma 4. Part 3 is a consequence of Lemmas 5 and 6.

4 Conclusions

In this paper we showed that center genomes do not always attain the theoretical lower bound in the case of two genomes, with respect to the rank distance. In spite of that, their are easy to calculate, and provide an attractive alternative to the median in ancestral genome reconstruction, even in the two-input version, which is already more restrictive than its median counterpart. Given that computing a median is NP-hard for the majority of relevant distances, its replacement by a center solution would bring a significant gain.

Nevertheless, it would be interesting to extend this analysis to three inputs, and determine what happens there. Probably the arbitrary input version is NP-hard, as similar problems with simpler distances have already been proved NP-hard [7, 2]. In addition, considering genomes with unequal gene content would also be worthwhile.

Acknowledgements

We thank funding agency FAPESP (Brazil) for financial support (Grant numbers 2012/13865-7, 2012/14104-0, and 2018/00031-7). PB would also like to acknowledge the Canada 150 Research Chair program.

References

1. Chindelevitch, L., Zanetti, J.P.P., Meidanis, J.: On the rank-distance median of 3 permutations. BMC Bioinformatics **19**(Suppl 6), 142 (2018)
2. Cunha, L.F.I., Feijão, P., dos Santos, V.F., Kowada, L.A., de Figueiredo, C.M.H.: On the computational complexity of closest genome problems. Discrete Applied Mathematics **274**, 26–34 (2020)
3. Feijão, P.: Reconstruction of ancestral gene orders using intermediate genomes. BMC Bioinformatics **16**(Suppl 14), S3 (2015)
4. Feijão, P., Meidanis, J.: SCJ: a breakpoint-like distance that simplifies several rearrangement problems. Trans Comput Biol Bioinform **8**, 1318–1329 (2011)
5. Gabidulin, E.M.: Theory of codes with maximum rank distance. Probl. Peredachi Inf. **21**(1), 3–16 (1985)
6. Meidanis, J., Biller, P., Zanetti, J.P.P.: A matrix-based theory for genome rearrangements. Tech. Rep. IC-18-10, Institute of Computing, University of Campinas (Aug 2018)
7. Popov, V.: Multiple genome rearrangement by swaps and by element duplications. Theoretical Computer Science **385**(1–3), 115–126 (2007)
8. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. BMC Bioinform **10**(1), 120 (2009)
9. Zanetti, J.P.P., Biller, P., Meidanis, J.: Median approximations for genomes modeled as matrices. Bulletin of Mathematical Biology **78**(4), 786–814 (2016), a preliminary version appeared on the Proceedings of the Workshop on Algorithms for Bioinformatics (WABI) 2013