On the Complexity of the (ℓ, k)-Median Problems

Luís Cunha
Instituto de Computação, Universidade Federal Fluminense, Niterói, Brazil

Thiago Nascimento
Instituto de Computação, Universidade Federal Fluminense, Niterói, Brazil

Marilia D. V. Braga
Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Germany

Jens Stoye
Faculty of Technology and Center for Biotechnology (CeBiTec), Bielefeld University, Germany

Abstract

The genome median problem is a central computational problem in comparative genomics, as it models the reconstruction of an ancestral genome from a set of related genomes. Given ℓ genomes and a distance measure, the problem asks for a genome that minimizes the sum of the distances to the input genomes. Two classical distances are the breakpoint distance and the double-cut-and-join (DCJ) distance. For multichromosomal circular genomes, the median problem is polynomial-time solvable under the breakpoint distance, whereas it is NP-hard under the DCJ distance.

For even integer $k \geq 2$, the σ_k distance interpolates between these two extremes: σ_2 corresponds to the breakpoint distance, while σ_∞ corresponds to the DCJ distance. A central open problem in this setting is the $(3, 4)$-Median problem, which asks for a median of three genomes under the σ_4 distance, the first intermediate distance after the breakpoint distance.

Motivated by this question, we study the more general (ℓ, k)-Median problem, in which ℓ is the number of input genomes and k determines the σ_k distance. We prove that $(\ell, 6)$-Median for every $\ell \geq 4$ and $(3, 12)$-Median are NP-complete. We then extend the hardness of $(\ell, 6)$-Median to (ℓ, k)-Median for all even $k \geq 6$, and the hardness of $(3, 12)$-Median to $(3, k)$-Median for all even $k \geq 12$. These results identify broad hardness regions in the (ℓ, k) parameter space and delimit the remaining open cases around the fundamental $(3, 4)$-Median problem.

2012 ACM Subject Classification Theory of computation → Proof complexity

Keywords and phrases Genome rearrangement, median problem, sigma-k distance

Digital Object Identifier 10.4230/LIPIcs.WABI.2026.19

Funding *Luís Cunha*: FAPERJ: JCNE - E-26/201.372/2022 and E-26/204.301/2025; CNPq: PQ - 307666/2025-5, Universal - 406173/2021-4, National Institute of Science and Technology for Applied Quantum Computing - 408884/2024-0.
Marilia D. V. Braga: CAPES PRINT: 88887.310247/2018-00.
Jens Stoye: CAPES PRINT: 88887.310247/2018-00.

1 Introduction

Phylogenetics studies evolutionary relationships among species using genetic information obtained, for example, through DNA sequencing. A central task in this context is to infer ancestral genomes from a set of related extant genomes. Comparative genomics, in which genomes of different species are compared, helps address fundamental questions at the intersection of genetics and evolution [9].

From a computational perspective, ancestral reconstruction is naturally modeled by the *median problem* [13]. Given a set of $\ell \geq 3$ genomes and a distance measure, the median problem asks for a genome that minimizes the sum of its distances to the input genomes.

26th International Conference on Algorithms for Bioinformatics (WABI 2026).
Editors: Nadia El-Mabrouk and Fabio Vandin; Article No. 19; pp. 19:1–19:17
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Although the case $\ell = 3$ is the classical formulation of the problem, the problem is naturally defined for any number of input genomes and has also been studied in this more general setting [8]. A median genome can be interpreted as a candidate ancestral genome for the input genomes.

The choice of the distance measure depends on the genome model. In this paper, a genome is represented as a set of *chromosomes*, and each chromosome is a circular sequence of oriented *genes* classified into gene *families*. Two genomes over the same set of genes form a *canonical pair* if they are *singular*, i.e., each genome contains exactly one gene from each family. One classical distance in this setting is the *breakpoint distance* [14], which counts the non-common adjacencies of two genomes, where an *adjacency* is the oriented neighborhood between two consecutive genes in a chromosome.

Genomes can also be restricted according to the structure of their chromosomes: they may be monochromosomal or multichromosomal, and their chromosomes may be linear, circular, or mixed. Different genomic distances have been studied for these scenarios, ranging from the breakpoint distance to the more general *double-cut-and-join* (DCJ) distance [16, 4].

The relationship between two genomes can be represented by the *breakpoint graph*. For canonical circular genomes, this graph is a collection of even cycles. We call a cycle with i edges an *i-cycle*, and an *even cycle* if i is even. Let c_i be the number of i-cycles in the breakpoint graph. If n is the number of genes in each genome, the breakpoint distance is $n - c_2$. Similarly, for rearrangements modeled by the DCJ operation, the DCJ distance is $n - c$, where c is the total number of cycles.

The computational complexity of the median problem depends on both the genome model and the chosen distance. For circular genomes, the median problem is polynomial-time solvable under the breakpoint distance, but NP-hard under the DCJ distance [13].

A natural way to study the gap between these two extremes is through the family of σ_k distances. For an even integer $k \geq 2$, let $\sigma_k = c_2 + c_4 + \cdots + c_k$ and define $\mathrm{d}_k = n - \sigma_k$ [6]. Thus, the distance defined by σ_2 corresponds to the breakpoint distance, while σ_∞ corresponds to the DCJ distance.

The σ_k family has also been studied in the context of the *double distance* problem, whose goal is to compute the minimum number of genomic rearrangements between a singular and a duplicated genome that – in addition to rearrangements – are separated by a whole genome duplication. For that problem, the cases $k = 2, 4, 6$ can be solved in linear time [5], whereas the problem is NP-complete for every $k \geq 8$ [7]. Together with the classical results for the median problem under the breakpoint and DCJ distances [13], these works motivate a finer investigation of median problems under intermediate σ_k distances.

The main motivation for this work is the complexity of the $(3, 4)$-Median problem. This problem asks for a median of three genomes under the σ_4 distance and represents the first nontrivial intermediate case between the polynomial-time breakpoint median problem and the NP-hard DCJ median problem. The relevance of this question has been highlighted in previous work on genome rearrangement distances and median problems [13, 12].

Motivated by this question, we study the more general (ℓ, k)-Median problem. This formulation allows us to investigate how the complexity depends on the number ℓ of input genomes and on the distance parameter k. Eventually, it may lead to better approximations to the (more relevant) DCJ median, while still staying at or close to the polynomial time complexity of the breakpoint median.

This paper is organized as follows. In Section 2, we introduce the necessary background on genomes, breakpoint graphs, and the σ_k distance, and we formally define the (ℓ, k)-Median problem. In Section 3 we introduce concepts and properties that are common to our reductions, including the construction of multiple breakpoint graphs from regular digraphs. In Section 4, starting from the NP-complete problem *Maximum-cardinality edge-disjoint directed triangle decomposition*, we present our results for the smallest values of k: We first prove the NP-completeness of the median problem for the cases $\ell \geq 4$ and $k = 6$ and for the case $\ell = 3$ and $k = 12$. In Section 5, we extend these results to larger values of $k \geq 8$ for $\ell \geq 4$ and of $k \geq 14$ for $\ell = 3$. Table 1 summarizes the complexity landscape for (ℓ, k)-Median and highlights the results proved in this paper.

Table 1 Complexity of the multichromosomal circular (ℓ, k)-Median problem. Each row represents the number of input genomes, and each column represents the parameter k used in the σ_k distance. Results proved in this paper are shown in blue.

# of genomes \ σ_k distance	2	4	6	8	10	12	$k \geq 14$	∞
3	P [13]	Open	Open	Open	Open	NP-hard	NP-hard	NP-hard (DCJ) [13]
4	P [13]	Open	NP-hard	NP-hard	NP-hard	NP-hard	NP-hard	NP-hard (DCJ) [13]
$\ell \geq 5$	P [13]	Open	NP-hard	NP-hard	NP-hard	NP-hard	NP-hard	NP-hard (DCJ) [13]

2 Preliminaries

A genome is a set of *chromosomes*. In this work, all chromosomes are circular, representing circular DNA molecules, and are modeled as oriented circular sequences of *genes*. Each gene belongs to a *gene family*. If all genes in a genome are from different gene families, the genome is called *singular*. A gene $\mathtt{X}$ has two extremities, denoted by $\mathtt{X}^t$ and $\mathtt{X}^h$, which correspond to its tail and its head, respectively. The gene is written as $\mathtt{X}$ when read in its forward orientation, from tail to head, and as $\bar{\mathtt{X}}$ when read in its reverse orientation, from head to tail. A chromosome is represented by a sequence of genes enclosed in parentheses, for example $(\bar{3}\ 2\ \bar{1})$.

An *adjacency* in a chromosome is an unordered pair of neighboring gene extremities. For instance, the chromosome $(\bar{3}\ 2\ \bar{1})$ has $\{3^t 2^t, 2^h 1^h, 1^t 3^h\}$ as its adjacencies. Since an adjacency is an unordered pair, $3^t 2^t$ and $2^t 3^t$ represent the same adjacency.

A graphical representation of a singular genome $\mathbb{G}$ is the *genome graph* whose vertex set contains two vertices for each gene $\mathtt{X}$ of $\mathbb{G}$, one representing its tail $\mathtt{X}^t$ and one representing its head $\mathtt{X}^h$, and whose edge set contains the adjacencies of $\mathbb{G}$: for each adjacency AB, there is an edge (called $\mathbb{G}$-*edge*) connecting A and B. Every vertex is therefore incident to one $\mathbb{G}$-edge and the circular genome $\mathbb{G}$ represents a perfect matching.

While the genome graph is just an alternative representation of a genome, the relationships between two or more *canonical genomes*, i.e., singular genomes over the same set of genes, can be represented by a structure obtained by merging their genome graphs:

▶ **Definition 1** (Multiple breakpoint graph)**.** *Given ℓ canonical genomes $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_\ell$, the* multiple breakpoint graph *$MBG(\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_\ell) = (V, E)$ is constructed as follows: The vertex set V contains two vertices for each gene $\mathtt{X}$, one representing its tail $\mathtt{X}^t$ and one representing its head $\mathtt{X}^h$. The edge multiset E contains the adjacencies of $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_\ell$: For each adjacency AB in genome $\mathbb{G}_i$, there is a $\mathbb{G}_i$-edge connecting A and B.*

The multiple breakpoint graph of $\ell = 2$ genomes is also called just *breakpoint graph*. Since each circular genome corresponds to a perfect matching, the breakpoint graph decomposes into a set of cycles alternating between edges of the two genomes. We denote by c_i the number of such cycles of length i, where integer i is obviously even.

The breakpoint distance between two canonical genomes $\mathbb{G}_1$ and $\mathbb{G}_2$ can be derived from their breakpoint graph. Assuming that the genomes have n genes each, the corresponding breakpoint distance is $d_{BP}(\mathbb{G}_1, \mathbb{G}_2) = n - c_2$. Similarly, the DCJ distance is $d_{DCJ}(\mathbb{G}_1, \mathbb{G}_2) = n - c$, where c is the total number of cycles in $MBG(\mathbb{G}_1, \mathbb{G}_2)$. Equivalently, it can be written as $d_{DCJ}(\mathbb{G}_1, \mathbb{G}_2) = n - (c_2 + c_4 + \cdots + c_\infty)$.

The intermediate region between these two extremes can be studied through the family of σ_k *distances*. For any even integer $k \geq 2$, let $\sigma_k = c_2 + c_4 + \cdots + c_k$. Then, given two canonical genomes $\mathbb{G}_1$ and $\mathbb{G}_2$ over the same set of n genes $\mathcal{G}$ and an even integer $k \geq 2$, their σ_k distance, denoted by $\mathrm{d}_k(\mathbb{G}_1, \mathbb{G}_2)$, is defined as $\mathrm{d}_k(\mathbb{G}_1, \mathbb{G}_2) = n - \sigma_k$ [6]. Thus, the breakpoint distance corresponds to d_2, while the DCJ distance corresponds to d_∞.

In this work, we study the following problem:

▶ **Problem 2** ((ℓ, k)-Median). *Given a set of $\ell \geq 3$ canonical circular genomes $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_\ell$ over the same set of n genes $\mathcal{G}$ and an even integer $k \geq 2$, find a circular genome $\mathbb{M}$ over $\mathcal{G}$ that is canonical with the input genomes and minimizes the sum $\sum_{i=1}^{\ell} \mathrm{d}_k(\mathbb{G}_i, \mathbb{M})$.*

Observe that, like the input circular genomes, the circular median genome $\mathbb{M}$ also corresponds to a perfect matching in the multiple breakpoint graph. Since the number of genes is fixed at n for all canonical genomes in the instance, minimizing the sum of distances is equivalent to maximizing the *k-score* $\sum_{i=1}^{\ell} \sigma_k(\mathbb{G}_i, \mathbb{M})$. In other words, the goal is to maximize the total number of alternating cycles in the multiple breakpoint graph, counted between the median genome $\mathbb{M}$ and the input genomes $\mathbb{G}_i$, for $i \in \{1, \ldots, \ell\}$. A cycle between the median and genome $\mathbb{G}_i$ is called $\mathbb{G}_i$*-cycle*. In our illustrations of multiple breakpoint graphs, each genome is represented by a distinct color. Therefore, we may refer to the genome $\mathbb{G}_i$ or to the color $\mathbb{G}_i$ interchangeably.

We focus on the *decision version* of the problem, which is the one used for complexity analysis: Given an instance as above and an integer threshold T, the question is whether there exists a genome $\mathbb{M}$ such that $\sum_{i=1}^{\ell} \mathrm{d}_k(\mathbb{G}_i, \mathbb{M}) \leq T$, or equivalently, $\sum_{i=1}^{\ell} \sigma_k(\mathbb{G}_i, \mathbb{M}) \geq \ell n - T$.

3 Basic tools

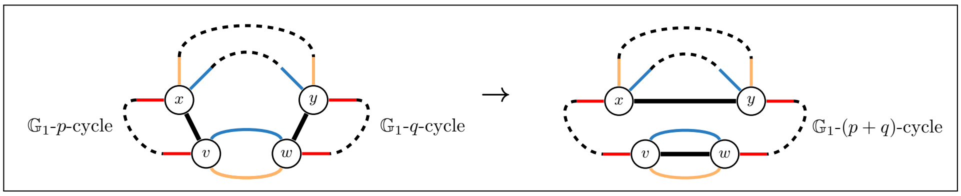
In this section we derive basic properties of multiple breakpoint graphs and median genomes, and introduce general gadgets that we will use in our reductions.

3.1 Subdivision gadget and its gates

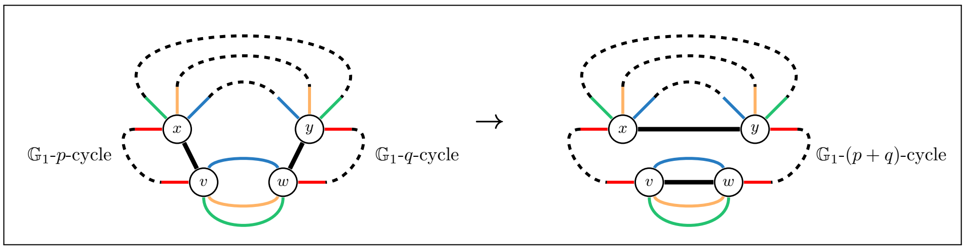
We first demonstrate an extension of a property that has been used in other works concerning the median problem [15, 10, 12]. Let a $\widetilde{\mathbb{G}}_i$-adjacency vw be an adjacency occurring in all genomes but $\mathbb{G}_i$.

▶ **Lemma 3.** *Let $\mathbb{G}_i$ be one genome of an instance of (ℓ, k)-median, for some $\ell \geq 3$ and some $k \geq 2$, and let vw be a $\widetilde{\mathbb{G}}_i$-adjacency. It holds that vw occurs in at least one median if $\ell = 3$ and in all medians if $\ell \geq 4$.*

Proof. Let the ℓ input genomes be $\mathbb{G}_1$, $\mathbb{G}_2$, ..., $\mathbb{G}_\ell$. We assume that vw occurs in genomes $\mathbb{G}_2, \ldots, \mathbb{G}_\ell$ and not in $\mathbb{G}_1$. Clearly, for the vertices v and w, either $vw \in \mathbb{M}$ or $vw \notin \mathbb{M}$. Assume that $vw \notin \mathbb{M}$ and, since $\mathbb{M}$ is a perfect matching, let x and y be vertices such that $vx, wy \in \mathbb{M}$; see Figure 1 (a) for $\ell = 3$ and Figure 1 (b) for $\ell = 4$. In this case $\mathbb{M}$ forms paths with genomes $\mathbb{G}_2$, ..., $\mathbb{G}_\ell$, and each of them is part of a ($\mathbb{G}_2$- or ... or $\mathbb{G}_\ell$-) cycle of length at least 4. Furthermore, $\mathbb{M}$ forms two paths with genome $\mathbb{G}_1$, that could in the best case be part of two $\mathbb{G}_1$-cycles, each one of length at least 2. By replacing the edges vx, wy with vw, xy in $\mathbb{M}$, we split each of the $\mathbb{G}_2$-, ..., $\mathbb{G}_\ell$-cycles into two smaller cycles, one being a 2-cycle, which would increase the k-score by $\ell - 1$. This move would merge the two possible $\mathbb{G}_1$-cycles, possibly decreasing the score by 2. Therefore, if $\ell = 3$, the k-score would certainly not decrease, and if $\ell \geq 4$, the k-score would even certainly increase. ◀



(a) $\ell = 3$.

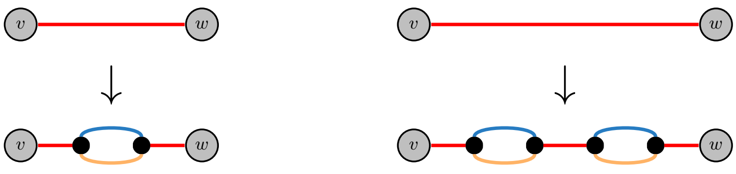


(b) $\ell = 4$.

Figure 1 Illustration of the proof of Lemma 3, with median edges drawn in black. It shows a local part of the multiple breakpoint graph assuming $vw \notin \mathbb{M}$, followed by the same local part after replacing vx, wy with vw, xy in $\mathbb{M}$ for **(a)** $\ell = 3$ and **(b)** $\ell = 4$.

Let v and w be vertices of graph $H = MBG(\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_\ell)$ and let i be an integer with $1 \leq i \leq \ell$. If vw is a $\widetilde{\mathbb{G}}_i$-adjacency in H, then we say that vertices v and w form a *Z-pair*.

We build a general gadget that can be defined for a set of ℓ colors as follows. The *ℓ-subdivision* gadget, denoted by Z_ℓ, splits a $\mathbb{G}_i$-edge vw and inserts two new vertices v' and w', such that vv' and ww' are $\mathbb{G}_i$-edges and $v'w'$ is a $\widetilde{\mathbb{G}}_i$-adjacency. We then extend it to Z_ℓ^q, the *ℓ^q-subdivision* of a $\mathbb{G}_i$-edge uv for an integer $q \geq 1$: This operation inserts q $\widetilde{\mathbb{G}}_i$-adjacencies in vw. Since vw is a $\mathbb{G}_i$-edge, we can say that its corresponding Z_ℓ^q-gadget is of color $\mathbb{G}_i$. The outermost vertices of Z_ℓ^q are still vertices v and w, and are now called *gates*, while the (new) inner vertices of Z_ℓ^q are distributed into Z-pairs. Note that $Z_\ell^1 = Z_\ell$. Examples of Z_3 and Z_3^2 are shown in Figure 2.



(a) **(b)**

Figure 2 Let i be an integer between 1 and 3. **(a)** A Z_3^1- or simply Z_3-gadget splits a $\mathbb{G}_i$-edge vw and inserts a $\widetilde{\mathbb{G}}_i$-adjacency (a pair of vertices connected by two parallel edges in the two other colors). **(b)** A Z_3^2-gadget inserts two $\widetilde{\mathbb{G}}_i$-adjacencies into a $\mathbb{G}_i$-edge vw. In both (a) and (b), the two endpoints of each $\mathbb{G}_i$-adjacency are Z-pairs, and vertices v and w are the gates of the gadget.

3.2 The α-color and Θ-transformation of p-regular digraphs

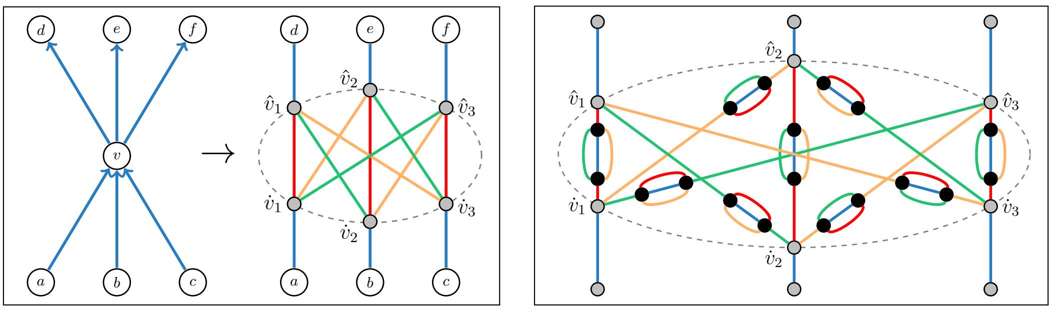
In our reductions, we construct edge ℓ-colored graphs in which $\ell-1$ colors, named $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_{\ell-1}$, are symmetrically distributed, while the ℓth color has a special role. Therefore, instead of calling it $\mathbb{G}_\ell$ we call the color $\mathbb{G}_\alpha$ or simply α. We denote the edges of this color $\mathbb{G}_\alpha$-edges or simply α-edges, and its median cycles $\mathbb{G}_\alpha$-cycles or simply α-cycles.

In the following parts of this manuscript, a color $\mathbb{G}_i$ refers to some color that is not α, that is, $\mathbb{G}_i \neq \alpha$, therefore any set $\widetilde{\mathbb{G}}_i$ (of all colors but $\mathbb{G}_i$) always includes α. We denote by $\mathbb{G}_{\hat{\alpha}}$ any unspecified color that is not α.

We start from directed graphs (*digraphs*). A digraph is *p-regular* if every vertex has both in-degree and out-degree equal to p [3]. We will now explain how to construct, from a p-regular digraph D and an integer $q \geq 1$, an undirected $(p+1)$-colored graph $\Theta(D, q)$ that includes $(p+1)^q$-subdivision gadgets.

▶ **Construction 4** (Θ-transformation)**.** *Given a p-regular digraph $D = (V, E)$ (for $p \geq 2$) and an integer $q \geq 1$, we construct the undirected edge $(p+1)$-colored graph $\Theta(D, q)$ as follows:*

1. *(Vertex $2p$-expansions with $\mathbb{G}_{\hat{\alpha}}$-edges) For each vertex $v \in V$, create a p-colored undirected complete bipartite graph $B^p(u)$ whose partitions are $\{\dot{v}_1, \dot{v}_2, \ldots, \dot{v}_p\}$ (*in-plugs*) and $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_p\}$ (*out-plugs*). The edges of $B^p(v)$ are distributed into p disjoint perfect matchings, and each matching has one of the p $\mathbb{G}_{\hat{\alpha}}$-colors $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_p$ as shown in Figure 3(a) for $p = 3$.*
2. *(q-Subdivisions of $\mathbb{G}_{\hat{\alpha}}$-edges) For each vertex $v \in V$, by a transformation of $B^p(v)$, where each of its p^2 $\mathbb{G}_{\hat{\alpha}}$-edges is subdivided by the insertion of a Z^q_{p+1}-gadget whose additional color is α, we obtain the gadget $X^{p,q}(v)$ (see Figure 3(b) for $p = 3$ and $q = 1$). Note that each in-plug $\dot{v}$ and each out-plug $\hat{v}$ of $B^p(v)$ becomes a gate of p Z^q_{p+1}-gadgets in $X^{p,q}(v)$ (of all p distinct $\mathbb{G}_{\hat{\alpha}}$-colors).*



(a) From a vertex v of D to $B^3(v)$. **(b)** Gadget $X^{3,1}(v)$.

Figure 3 Construction of gadget $X^{3,1}(v)$. **(a)** Vertex expansion: Each vertex v of 3-regular digraph D is replaced by a complete undirected bipartite graph $B^3(v)$. The edges are split into 3 perfect matchings, each one receives one color $\mathbb{G}_1$, $\mathbb{G}_2$ or $\mathbb{G}_3$ (here red, green and orange). The original (directed) edges become undirected α-edges (here blue). **(b)** Subdivision: The gadget $X^{3,1}(v)$ is obtained from $B^3(v)$ by replacing each of its internal edges by the corresponding Z_4-gadget. From the six vertices a, b, c, d, e, f adjacent to v we only show the respective plugs connected by α-edges to the plugs of $X^{3,1}(v)$.

3. *(Connection between gadgets with α-edges) For each vertex $v \in V$, for each out-edge $(v \to w)$, create in $\Theta(D, q)$ an undirected α-edge connecting one of the out-plugs of $X^{p,q}(v)$ to one of the in-plugs of $X^{p,q}(w)$. Similarly, for each in-edge $(u \to v)$, create an undirected α-edge connecting one of the out-plugs of $X^{p,q}(u)$ to one of the in-plugs of $X^{p,q}(v)$ in $\Theta(D, q)$. These connections should be done such that exactly one α-edge is incident to each plug (Figure 4).*

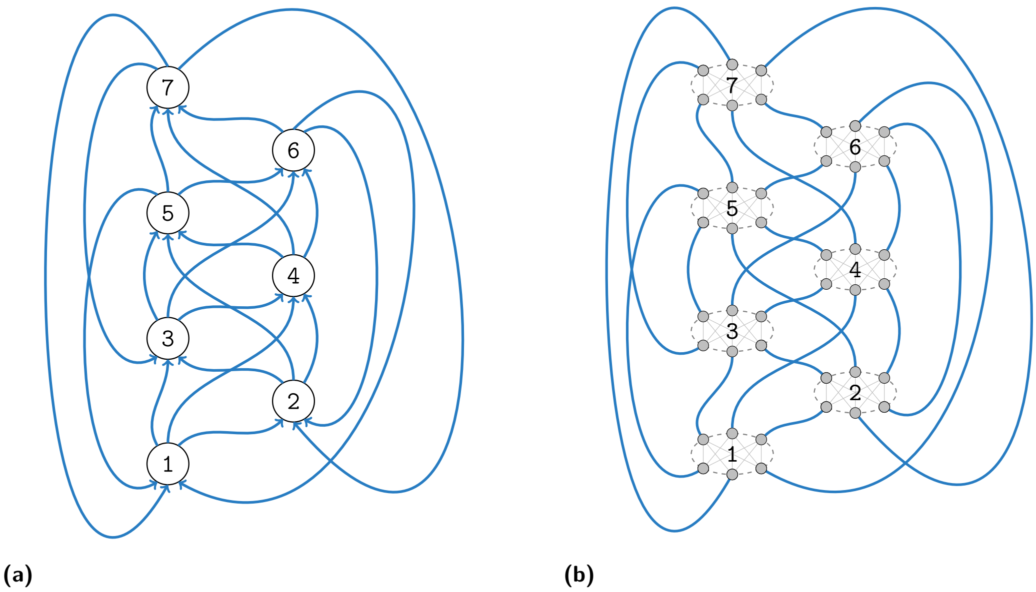


Figure 4 Creation of the undirected, 4-colored graph $\Theta(D, q)$ from a 3-regular digraph D and $q \geq 1$. **(a)** A 3-regular digraph D. **(b)** The graph $\Theta(D, q)$ obtained from D via Construction 4, where each $X^{3,q}$-gadget is represented by a dashed ellipse. Colors $\mathbb{G}_1, \mathbb{G}_2$, and $\mathbb{G}_3$ are present only in edges inside the $X^{3,q}$-gadgets, masked by the ellipses. Color α (here blue) is a union of the α-edges between ellipses and the α-edges present in all Z_4^q-gadgets, the latter being also masked by the ellipses.

▶ **Proposition 5.** *If graph $\Theta(D, q)$ is obtained via Construction 4 on p-regular digraph D and an integer $q \geq 1$, then its $p+1$ colors define $p+1$ circular genomes $\mathbb{G}_1$, $\mathbb{G}_2$, …, $\mathbb{G}_p$ and $\mathbb{G}_\alpha$ such that $\Theta(D, q) = MBG(\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_p, \mathbb{G}_\alpha)$.*

Proof. Note that $\Theta(D, q)$ has an even number of vertices: Each $X^{p,q}$-gadget has an even number of plugs and each Z_{p+1}^q gadget adds q Z-pairs. Furthermore, it is easy to see that each of the vertices of $\Theta(D, q)$ is incident to exactly $p+1$ edges, each one of a distinct color $\mathbb{G}_1$, $\mathbb{G}_2$, …, $\mathbb{G}_p$ or $\mathbb{G}_\alpha$. This guarantees that each of the $p+1$ colors defines a perfect matching of the vertices of $\Theta(D, q)$, therefore we can say that each color defines the adjacencies of one circular genome. The $p+1$ genomes are then named $\mathbb{G}_1$, $\mathbb{G}_2$, …, $\mathbb{G}_p$ and $\mathbb{G}_\alpha$. ◀

3.3 Median partitions $\mathbb{M}_Z$ and $\mathbb{M}_*$, $\mathbb{G}_{\hat{\alpha}}$-bridges and α-shortcuts

Due to Lemma 3, we consider only medians for $\Theta(D, q)$ that include an edge connecting the two vertices of each of its Z-pairs.

▶ **Proposition 6.** *Let $k \geq 2$ and $\Theta(D, q)$ be a graph obtained via Construction 4 on a p-regular digraph $D = (V, E)$ and an integer $q \geq 1$. There is a median $\mathbb{M}$ for $\Theta(D, q)$ that can be partitioned into sets $\mathbb{M}_Z$ and $\mathbb{M}_*$, where $\mathbb{M}_Z$ has one edge connecting each Z-pair and $\mathbb{M}_*$ is a perfect matching of the plugs of $\Theta(D, q)$. Furthermore, the k-score of $\mathbb{M}$ is $s_Z + s_*$, where the first term, also denoted Z-score, is a constant $s_Z = qp^3|V|$, while the second term, s_*, is the maximum k-score attainable by connecting the plugs of $\Theta(D, q)$.*

Proof. The fact that there exists a median $\mathbb{M}$ for $\Theta(D, q)$ that includes an edge for each Z-pair in $\Theta(D, q)$ is a direct consequence of Lemma 3. Each of the Z-pairs induces p 2-cycles, that contribute to any k-score for $k \geq 2$. After fixing $\mathbb{M}_Z$, there is no "free" vertex in $\Theta(D, q)$

that is not a plug, therefore $\mathbb{M}_*$ must be a perfect matching of the plugs. Since there are p^2 Z^q_{p+1}-gadgets for each vertex of D, and each one has pq Z-pairs, they yield the overall score $s_Z = qp^3|V|$, which is part of the k-score of any optimal solution that includes set $\mathbb{M}_Z$. ◀

All edges of $\Theta(D, q)$ (from all defined $p+1$ colors) that form 2-cycles with $\mathbb{M}_Z$ cannot be part of any cycle induced by $\mathbb{M}_*$ and are therefore said to be *Z-saturated*.

Note that each $\mathbb{G}_i$-adjacency in $\Theta(D, q)$ is either Z-saturated or part of a Z^q_ℓ-gadget Y of color $\mathbb{G}_i$. Since Y is a chain of $\mathbb{G}_i$-edges that alternate with $\widetilde{\mathbb{G}}_i$-adjacencies, together with edges from $\mathbb{M}_Z$, it induces a fixed $\mathbb{G}_i$-path of length $2q+1$. Such a path is called *$\mathbb{G}^q_i$-bridge*, whose endpoints (gates) are necessarily plugs of $\Theta(D, q)$. If $q = 1$ we can use the simpler notation $\mathbb{G}_i$-bridge. An example is given in Figure 5 (a). Since any $\mathbb{G}_{\hat{\alpha}}$-edge that is not Z-saturated is part of a bridge, we have:

▶ **Observation 7.** *Let $\mathbb{G}_i$ be a $\mathbb{G}_{\hat{\alpha}}$-color and recall that each $\mathbb{G}^q_i$-bridge is a $(2q+1)$-path. Any $\mathbb{G}_i$-cycle induced by $\mathbb{M}_*$ alternates median edges (from $\mathbb{M}_*$) with $\mathbb{G}^q_i$-bridges and has therefore length $b(2q+1+1) = 2b(q+1)$, where $b \geq 1$ is the number of involved $\mathbb{G}^q_i$-bridges.*

The idea is then to set the value of q such that a median cycle involving a single bridge has length at most k, while a cycle involving two or more bridges has length greater than k. For any $k \geq 4$, the smallest q that fulfills this requirement is $q = \lfloor \frac{k}{4} \rfloor$: if $b = 1$, then $2b\left(\lfloor \frac{k}{4} \rfloor + 1\right) = 2\left(\lfloor \frac{k}{4} \rfloor + 1\right) \leq k$. If $b \geq 2$, then $2b\left(\lfloor \frac{k}{4} \rfloor + 1\right) \geq 4\left(\lfloor \frac{k}{4} \rfloor + 1\right) > k$. Table 2 displays the behavior of q as k grows. In this setting, only $\mathbb{G}_{\hat{\alpha}}$-cycles involving a single bridge contribute to the k-score.

Table 2 For each possible $k \geq 4$ we show: the value of q, the length of $\mathbb{G}^q_{\hat{\alpha}}$-bridges induced by $\mathbb{M}_Z$ and the lengths of $\mathbb{G}_{\hat{\alpha}}$-cycles induced by $\mathbb{M}_*$.

k	4 or 6	8 or 10	12 or 14	16 or 18	20 or 22	$k \geq 24$
q	1	2	3	4	5	$q = \lfloor \frac{k}{4} \rfloor$
length of all $\mathbb{G}^q_{\hat{\alpha}}$-bridges	3	5	7	9	11	$2q+1$
length of a $\mathbb{G}_i$-cycle involving b $\mathbb{G}^q_i$-bridges	$4b$	$6b$	$8b$	$10b$	$12b$	$2b(q+1)$

Such a cycle is obtained when an edge in $\mathbb{M}_*$, which is then called *cross edge*, connects plugs that are the two endpoints (gates) of the same $\mathbb{G}^q_i$-bridge. In each $X^{p,q}$-gadget at most p $\mathbb{G}_{\hat{\alpha}}$-cycles of length $2(q+1) \leq k$ can be built with cross edges. (Figure 5 (b) illustrates this scenario for a $X^{3,1}$-gadget.)

If an edge of $\Theta(D, q)$ is not Z-saturated nor part of a bridge, it must be an α-edge connecting plugs and is then called *α-shortcut*. Observe that $\Theta(D, q)$ has $2p|V|$ plugs, thus any perfect matching of the plugs has $p|V|$ edges. The α-shortcuts form a perfect matching of the plugs such that each α-shortcut connects an in- and an out-plug of distinct $X^{p,q}$-gadgets.

An edge in $\mathbb{M}_*$ is called *loop edge* if it connects plugs that are the two endpoints of the same α-shortcut. A median edge that is not a loop edge nor a cross edge is called *jump edge*. Examples are given in Figure 6 (a).

▶ **Observation 8.** *An oriented t-cycle C in p-regular digraph D implies that, for each edge $(v \to w)$ in C, the graph $\Theta(D, q)$ has an α-shortcut connecting an out-plug $\hat{v}$ of $X^{p,q}(v)$ to an in-plug $\dot{w}$ of $X^{p,q}(w)$. When $\mathbb{M}_*$ includes the cross edges chaining the endpoints of all α-shortcuts derived from t-cycle C, then $\mathbb{M}_*$ induces an α-$2t$-cycle in $\Theta(D, q)$.*

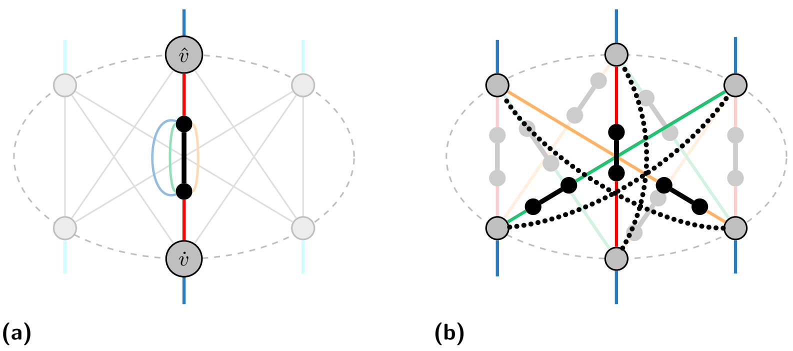


Figure 5 The (fixed) median edges from partition $\mathbb{M}_Z$ are represented as solid black, while (candidate) median edges from partition $\mathbb{M}_*$ are represented as dotted black. **(a)** Plugs $\dot{v}$ and $\hat{v}$ are the gates of a Z_4^1 gadget of color $\mathbb{G}_1$ (red). Median partition $\mathbb{M}_Z$ includes the (black) median edge connecting the pair of black vertices (called Z-pair). The shaded edges of colors $\mathbb{G}_2$, $\mathbb{G}_3$ and α (here green, orange and blue) connecting the Z-pair of black vertices are Z-saturated: They form 2-cycles with the median edge. Together with the (red) $\mathbb{G}_1$-edges, the median edge also forms a $\mathbb{G}_1$-bridge, that is the alternating path of length 3 connecting $\dot{v}$ to $\hat{v}$. **(b)** Here we omit Z-saturated edges, but all inner connections are displayed: Between each in-plug and each out-plug there is a Z_4^1-gadget of some color $\mathbb{G}_i$ that gives a $\mathbb{G}_i$-brigde of length 3. We can then have up to $p = 3$ $\mathbb{G}_{\hat{\alpha}}$-cycles, each of length 4, with (dotted black) median cross edges from partition $\mathbb{M}_*$. Here each of these cycles is with a distinct $\mathbb{G}_i$, but in general they can have other distributions among the p $\mathbb{G}_i$s.

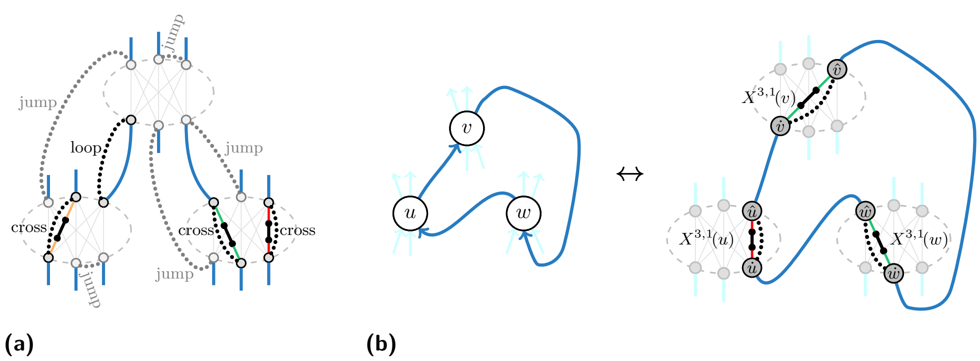


Figure 6 In these pictures, Z-saturated edges are omitted. The (fixed) median edges from partition $\mathbb{M}_Z$ are represented as solid black, while (candidate) median edges from partition $\mathbb{M}_*$ are represented as dotted black or gray. **(a)** Section of a graph showing three $X^{3,1}$-gadgets. One of the black dotted edges is a median loop edge, parallel to a (blue) α-shortcut. The other three black dotted edges are median cross edges, that are parallel to $\mathbb{G}_{\hat{\alpha}}$-bridges. All gray dotted edges are median jump edges: either they connect two (bottom) in- or two (top) out-plugs of the same or of distinct $X^{3,1}$-gadgets. Another type of jump edge connects a (bottom) in-plug and a (top) out-plug of distinct $X^{3,1}$-gadgets but is not parallel to an α-shortcut. **(b)** A directed 3-cycle (uvw) in a 3-regular digraph D and the potential corresponding 6-cycle in $\Theta(D, 1)$: The directed edge $(u \to v)$ in D implies that there is an α-shortcut $\hat{u}\dot{v} \in \Theta(D, 1)$ for some out-plug $\hat{u}$ of $X^{3,1}(u)$ and some in-plug $\dot{v}$ of $X^{3,1}(v)$. Similarly, the directed edges $(v \to w)$ and $(w \to u)$ imply that there are α-shortcuts $\hat{v}\dot{w}$ and $\hat{w}\dot{u} \in \Theta(D, 1)$. The α-shortcuts are colored in blue. If (dotted black) cross edges $\dot{u}\hat{u}$, $\dot{v}\hat{v}$ and $\dot{w}\hat{w}$ are in $\mathbb{M}_*$, we have an α-6-cycle $(\dot{u}\hat{u}\dot{v}\hat{v}\dot{w}\hat{w})$. Each of the three median cross edges shown here also induces a $\mathbb{G}_{\hat{\alpha}}$-cycle of length 4.

Therefore, only α-cycles that exclusively alternate α-shortcuts with median cross edges in $\Theta(D, q)$ can correspond to oriented cycles in D. An example of a 3-cycle in a 3-regular graph D corresponding to an α-6-cycle in $\Theta(D, 1)$ is given in Figure 6 (b).

4 NP-completeness results for the minimum k

In this section we will show our results with smallest k, which are the NP-completeness of the $(\ell, 6)$-Median problem for any $\ell \geq 4$ and the NP-completeness of the $(3, 12)$-Median problem.

4.1 NP-completeness of the $(\ell, 6)$-Median problem for $\ell \geq 4$

We will consider a restricted version of the NP-complete problem *maximum-cardinality edge-disjoint directed triangle decomposition*, described as follows: For an integer $p \geq 3$, the input is a p-regular digraph D with no 2-cycles, and the goal is to find a maximum-cardinality collection of edge-disjoint directed 3-cycles in D. This restricted version is denoted MAX3CDEC and remains NP-complete [1].

Here we have $k = 6$, therefore $q = \lfloor \frac{6}{4} \rfloor = 1$. Let $\ell \geq 4$ and $p = \ell - 1$, let $D = (V, E)$ be a p-regular digraph, and let graph $\Theta(D, 1)$ be obtained via Construction 4. Observation 8 guarantees that an oriented 3-cycle in graph D corresponds to a potential α-6-cycle induced by set $\mathbb{M}_*$ in $\Theta(D, 1)$. Recall that, since for $q = 1$ the length of any $\mathbb{G}_i$-cycle is a multiple of 4 (Table 2), a $\mathbb{G}_i$-cycle can only contribute to the 6-score if it involves a single $\mathbb{G}_i$-bridge.

▶ **Proposition 9.** *There is a median $\mathbb{M}$ of $\Theta(D, 1)$ that is of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$ and is free of jump edges.*

Proof. The choice of including a jump edge e in $\mathbb{M}_*$ could always be replaced by another one using the same plugs, but with at least the same 6-score: Let us suppose that e is part of an α-4-cycle c_4^α, whose second median edge is e'. If e' is a cross edge, the cycle c_4^α coexists with a $\mathbb{G}_{\hat{\alpha}}$-4-cycle. Otherwise e' is also a jump edge and c_4^α cannot co-exist with any $\mathbb{G}_{\hat{\alpha}}$-4-cycle. Replacing e and e' by two loop edges that are parallel to the α-shortcuts of c_4^α would produce two α-2-cycles, occupying the same four plugs. Although this new arrangement certainly prevents the incident bridges to form $\mathbb{G}_{\hat{\alpha}}$-4-cycles, it still has at least the same 6-score. Now let e be part of an α-cycle c_m^α with length $m \geq 6$. Note that c_m^α co-exists with at most $\frac{m}{2} - 1$ $\mathbb{G}_{\hat{\alpha}}$-4-cycles. Then it is clear that c_m^α could be replaced by $\frac{m}{2}$ α-2-cycles using the same plugs, achieving at least the same 6-score if $m = 6$ and a higher 6-score if $m \geq 8$. ◀

We will now relate solutions of MAX3CDEC to median solutions of the instance represented by $\Theta(D, 1) = MBG(\mathbb{G}_1, \mathbb{G}_2, ..., \mathbb{G}_p, \mathbb{G}_\alpha)$.

▶ **Theorem 10.** *The problem $(\ell, 6)$-Median is* NP*-complete for any $\ell \geq 4$.*

Proof. Take the subdivision factor $q = \lfloor \frac{6}{4} \rfloor = 1$. For $p = \ell - 1$, let D be a p-regular digraph free of 2-cycles and let $\Theta(D, 1)$ be a graph obtained via Construction 4. Proposition 6 guarantees that there is a median $\mathbb{M}$ for $\Theta(D, 1)$ of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$, where $\mathbb{M}_Z$ is fixed. Furthermore, the 6-score of $\mathbb{M}$ is $s_Z + s_*$, and $s_Z = qp^3|V|$, that gives $s_Z = p^3|V|$ for $q = 1$. We prove the following equivalence: An optimal solution of MAX3CDEC for a digraph $D = (V, E)$ contains f directed 3-cycles if and only if $s_* = f + p|V|$.

First we show that an optimal solution $\mathcal{S}$ of MAX3CDEC in D with f directed 3-cycles yields a set $\mathbb{M}_*$ for which $s_* = f + p|V|$.

For each oriented 3-cycle $C = (uvw)$ in the solution $\mathcal{S}$, we can add the median cross edges to $\mathbb{M}_*$ so that we build an α-6-cycle as described in Observation 8. After doing that for each of the f 3-cycles of $\mathcal{S}$, looking at each gadget $X(v)$, either each of its in-plugs is

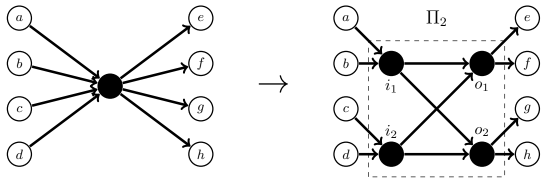


Figure 7 Transformation of a vertex with in-degree and out-degree 4 into Π_2.

already connected to one of its out-plugs in $\mathbb{M}$ to form an α-6-cycle (as described in the previous step), or there are the same number of free in- and out-plugs. Simply add to $\mathbb{M}_*$ one cross edge connecting each free in-plug $\dot{v}$ to a free out-plug $\hat{v}$. In the end of this procedure, $\mathbb{M}_*$ contains p cross edges forming a perfect matching of in- and out-plugs of each $X(v)$. Each of these cross edges creates a $\mathbb{G}_i$-4-cycle. Therefore, in total there are $p|V|$ 4-cycles between $\mathbb{M}$ and the genomes $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_p$. Since D has no 2-cycles, any cycle of D that is not a directed 3-cycle has length greater than 3. This corresponds to a cycle of length greater than 6 in $\Theta(D, 1)$ and therefore does not affect the 6-score. Thus, the resulting score is $s_* = f + p|V|$.

Conversely, we will now show that if $s_* = f + p|V|$, we can derive an optimal solution $\mathcal{S}$ of MAX3CDEC in D with f directed 3-cycles.

Proposition 9 guarantees the existence of an optimal set $\mathbb{M}_*$, such that each edge in $\mathbb{M}_*$ is either a loop or a cross edge. If $\mathbb{M}_*$ is exclusively composed of loop edges, it induces $p|V|$ α-2-cycles. In the other extreme, if $\mathbb{M}_*$ is exclusively composed of cross edges, it induces $p|V|$ $\mathbb{G}_{\hat{\alpha}}$-4-cycles. Note, however, that loop edges are incompatible, while cross edges are compatible with α-6-cycles. Therefore, $\mathbb{M}_*$ will include all cross edges that also produce α-6-cycles. Each α-6-cycle with 3 cross edges corresponds to a directed 3-cycle in D (Observation 8). If f is the number of α-6-cycles, then $\mathbb{M}_*$ will have at least the $3f$ cross edges that build these cycles, and each of these $3f$ cross edges also builds a $\mathbb{G}_{\hat{\alpha}}$-4-cycle. For the remaining $3|V| - 3f$ edges of $\mathbb{M}_*$, they are freely distributed among (compatible) cross and loop edges, and each one adds one cycle to the 6-score. Therefore $s_* = f + 3f + 3|V| - 3f = f + p|V|$ and there are f directed 3-cycles in an optimal solution $\mathcal{S}$ of D. ◀

4.2 NP-completeness of the $(3, 12)$-Median problem

The proof of NP-completeness for three input genomes starts from MAX3CDEC using 4-regular digraphs. Given a 4-regular digraph D, we first apply a polynomial transformation that produces a 2-regular digraph $D_\Pi = (V', E')$ in which directed 3-cycles of D correspond to directed 6-cycles of D_Π.

A *permutation network* is a directed graph Π_d with d input nodes $i_1, \ldots, i_d$ and d output nodes $o_1, \ldots, o_d$. For every permutation π of $\{1, \ldots, d\}$, there exist edge-disjoint paths $P_1, \ldots, P_d$ in Π_d such that P_j goes from i_j to $o_{\pi(j)}$ [2]. Figure 7 illustrates the transformation of a vertex into Π_2. This transformation replaces one vertex with in- and out-degree 4 with four vertices with in- and out-degree 2 each.

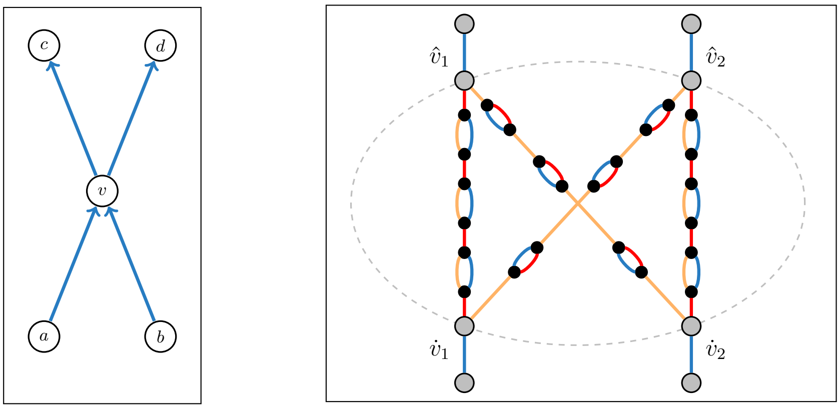
Let D be a 4-regular digraph with no 2-cycles. A 2-regular digraph D_Π is obtained by replacing each vertex of D with a Π_2. Since the digraph Π_2 has one edge between any input and output nodes, any path of length m in D has a corresponding path in D_Π of length $2m$. Therefore, finding an optimal solution of MAX3CDEC for D is equivalent to finding a maximum-cardinality collection of edge-disjoint directed 6-cycles in D_Π, and the latter problem is denoted MAX6CDEC. Note that, since D is free of 2-cycles, the shortest cycle in D_Π has length at least 6.

Here we have $k = 12$, therefore $q = \lfloor \frac{12}{4} \rfloor = 3$. Now let $\Theta(D_\Pi, 3)$ be obtained via Construction 4, as illustrated in Figure 8. Observation 8 guarantees that an oriented 6-cycle in D_Π corresponds to a potential α-12-cycle induced by set $\mathbb{M}_*$ in $\Theta(D_\Pi, 3)$. The reduction of $(3, 12)$-Median is analogous to the one of $(4, 6)$-Median (described in the previous subsection). Recall that, since for $q = 3$ the length of any $\mathbb{G}_i$-cycle is a multiple of 8 (Table 2), a $\mathbb{G}_i$-cycle can only contribute to the 6-score if it involves a single $\mathbb{G}_i$-bridge.

Here again we can focus on medians free of jump edges:

▶ **Proposition 11.** *There is a median $\mathbb{M}$ of $\Theta(D_\Pi, 3)$ that is of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$ and is free of jump edges.*

Proof. Analogous to the proof of Proposition 9. ◀



(a) Vertex v of D_Π. **(b)** Gadget $X^{2,3}(v)$ of $\Theta(D_\Pi, 3)$.

Figure 8 (a) A vertex v of D_Π and its four adjacent vertices. **(b)** The gadget $X^{2,3}(v)$ of $\Theta(D_\Pi, 3)$ obtained from v. From the four vertices adjacent to v in (a) we only show the respective plugs connected by α-shortcuts to the plugs of $X^{2,3}(v)$.

These observations are sufficient to support our arguments. Therefore we have:

▶ **Theorem 12.** *The problem $(3, 12)$-Median is* NP*-complete.*

Proof. Take the subdivision factor $q = \lfloor \frac{12}{4} \rfloor = 3$. Note that here we use $p = 2$: Our $\Theta(D_\Pi, 3)$ is a graph obtained via Construction 4 on 2-regular graph D_Π whose cycles have length at least 6. Proposition 6 guarantees that there is a median $\mathbb{M}$ for $\Theta(D_\Pi, 3)$ of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$, where $\mathbb{M}_Z$ is fixed. Furthermore, the 12-score of $\mathbb{M}$ is $s_Z + s_*$, where $s_Z = qp^3|V'| = 24|V'|$.

The cross edges form two $\mathbb{G}_i$-8-cycles with $\mathbb{G}_i^3$-bridges in each $X^{2,3}$-gadget. Therefore we need to show the following equivalence: An optimal solution $\mathcal{S}$ of MAX6CDEC for digraph $D_\Pi = (V', E')$ has a partition into f 6-cycles if and only if $s_* = f + 2|V'|$. The proof is analogous to the proof of Theorem 10 and shows that a set $\mathbb{M}_*$ including the $6f$ cross edges that build f α-12-cycles is optimal and allows the correspondence between the f α-12-cycles induced by $\mathbb{M}_*$ to the f 6-cycles of $\mathcal{S}$. The remaining edges of $\mathbb{M}_*$ are freely distributed into (compatible) loop and cross edges. ◀

5 Larger values of k with Θ'-transformation and α-brigdes

For generalizing our previous results to larger values of k, we follow the strategy used in the proof of NP-hardness for the marked directed Eulerian cycle decomposition problem [2]. In that construction, the edges are partitioned into three groups so that each directed 3-cycle contains exactly one edge from each group; then one group is marked. We use the analogous marked version of MAX3CDEC, in which every directed triangle contains exactly one marked edge and every cycle of length greater than 3 contains at least one marked edge.

The *marked maximum-cardinality edge-disjoint directed triangle decomposition*, denoted MARKED-MAX3CDEC, is described as follows. For $p \geq 3$, given a p-regular digraph with no 2-cycles and a set of marked edges, find a maximum-cardinality collection of edge-disjoint directed 3-cycles such that every selected triangle contains exactly one marked edge.

We adapt Construction 4 into a variation using marked edges:

▶ **Construction 13** (Θ'-transformation)**.** *Let $D = (V, E)$ be a p-regular digraph with x marked edges, and let two integers q and r be such that $q \geq 1$ and $r \geq 1$. We first construct the graph $\Theta(D, q)$ and then we perform a ℓ^r-subdivision of α-shortcuts that correspond to marked edges of D, obtaining the $(p+1)$-colored graph $\Theta'(D, q, r)$. The detailed subdivision, after steps 1–3 from Construction 4, is as follows:*

4. *(r-Subdivision of marked α-shortcuts) For each α-shortcut $\hat{v}\dot{w} \in \Theta(D, q)$, if the corresponding edge vw is marked in D, subdivide $\hat{v}\dot{w}$ by the insertion of a Z^r_{p+1}-gadget. Once this is done for all x marked edges of D, the resulting graph is $\Theta'(D, q, r)$.*

▶ **Proposition 14.** *If graph $\Theta'(D, q, r)$ is obtained via Construction 13 on p-regular digraph D and integers $q \geq 1$ and $r \geq 1$, its $p+1$ colors define $p+1$ circular genomes $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_p$ and $\mathbb{G}_\alpha$ such that $\Theta'(D, q, r) = MBG(\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_p, \mathbb{G}_\alpha)$.*

Proof. This was already true for $\Theta(D, q)$, and here we have only additional Z-pairs, and each vertex of such a pair is incident to exactly $p+1$ edges, each one of a distinct color $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_p$ or $\mathbb{G}_\alpha$. Therefore each of the $p+1$ colors still defines a perfect matching of the vertices of $\Theta'(D, q, r)$. ◀

We can still focus on a median $\mathbb{M}$ whose partitions are $\mathbb{M}_Z$ and $\mathbb{M}_*$, but note that here the partition $\mathbb{M}_Z$ also induces an α^r-bridge for each of the x ℓ^r-subdivided α-shortcuts in $\Theta'(D, q, r)$. Furthermore, the Z-score induced by $\mathbb{M}_Z$ needs to be updated accordingly.

▶ **Proposition 15.** *Let $k \geq 2$ and $\Theta'(D, q, r)$ be a graph obtained via Construction 13 on a p-regular digraph $D = (V, E)$ with x marked edges and integers $q \geq 1$ and $r \geq 1$. There is a median $\mathbb{M}$ for $\Theta'(D, q, r)$ that can be partitioned into sets $\mathbb{M}_Z$ and $\mathbb{M}_*$, where $\mathbb{M}_Z$ has one edge connecting each Z-pair and $\mathbb{M}_*$ is a perfect matching of the plugs of $\Theta'(D, q, r)$. Furthermore, the k-score of $\mathbb{M}$ is $s_Z + s_*$, where the first term, also denoted Z-score, is a constant $s_Z = qp^3|V| + prx$, while the second term, s_*, is the maximum k-score attainable by connecting the plugs of $\Theta'(D, q, r)$.*

Proof. The Z-score in $\Theta(D, q)$ was $s_Z = qp^3|V|$. In $\Theta'(D, q, r)$, each of the x ℓ^r-subdivisions creates r additional Z-pairs, each of which includes the endpoints of an $\widetilde{\mathbb{G}}_\alpha$-adjacency. Since $\mathbb{M}_Z$ induces p 2-cycles with each of these new $\widetilde{\mathbb{G}}_\alpha$-adjacencies, the new Z-score is $s_Z = qp^3|V| + prx$. ◀

The α^r-bridges have the goal of increasing the lengths of the α-cycles with cross edges induced by $\mathbb{M}_*$.

▶ **Proposition 16.** *An oriented t-cycle C in p-regular digraph D with $s \geq 1$ marked edges implies that, for each edge $(v \to w)$ in C, if $(v \to w)$ is not marked, the graph $\Theta'(D, q, r)$ has an α-shortcut connecting an out-plug $\hat{v}$ of $X^{p,q}(v)$ to an in-plug $\dot{w}$ of $X^{p,q}(w)$. Otherwise, if $(v \to w)$ is one of the s marked edges, the graph $\Theta'(D, q, r)$ has an α^r-bridge connecting $\hat{v}$ to $\dot{w}$. When $\mathbb{M}_*$ includes the cross edges chaining the endpoints of all α-shortcuts and α^r-bridges derived from the t-cycle C, then $\mathbb{M}_*$ induces an α-cycle of length $2(t + rs)$ in $\Theta'(D, q, r)$.*

Proof. Each marked edge in D corresponds to an α^r-bridge in $\Theta'(D, q, r)$. Since an α^r-bridge has length $2r + 1$, a t-cycle in D_Π containing s marked edges gives in $\Theta'(D, q, r)$ an α-cycle whose length is $2t - s + 2rs + s = 2(t + rs)$ (assuming the corresponding median cross edges are in set $\mathbb{M}_*$). ◀

The idea is to set r to a value such that the smallest α-cycle with only median cross edges has length k. This can be achieved by increasing r by one as k increases by two, as shown in Table 3.

Table 3 The values of the subdivision factor r as k grows and the corresponding length increases produced on α-cycles.

r	1	2	3	4	$r \geq 5$
length of α^r-bridges	3	5	7	9	$2r + 1$
increase on the length of α-cycles produced by each α^r-bridge	+2	+4	+6	+8	$+2r$
k for $\ell \geq 4$ and for $\ell = 3$	8 and 14	10 and 16	12 and 18	14 and 20	$6 + 2r$ and $12 + 2r$

Observe that, in graph $\Theta'(D, q, r)$, a median loop edge can be parallel to an α-shortcut, making an α-2-cycle, or to an α^r-bridge, making an α-cycle of length $2(r + 1)$. By following the values of r given in Table 3 it is clear that we always have $2(r + 1) < k$, therefore in this setting any α-cycle with a loop edge still contributes to the k-score.

We are now ready to generalize the NP-completeness results from Subsection 4.1 and Subsection 4.2 to larger values of k.

5.1 NP-completeness of (ℓ, k)-Median for $\ell \geq 4$ and even $k \geq 8$

For $p \geq 3$, let $D = (V, E)$ be a p-regular digraph with x marked edges, such that each 3-cycle in D has exactly one marked edge, and each larger cycle has at least one marked edge. Now take $q = \lfloor \frac{k}{4} \rfloor$ and $r = \frac{k-6}{2}$ for constructing the graph $\Theta'(D, q, r)$. We have $k = 6 + 2r$, and from Table 3 it is easy to see that each 3-cycle C in D (with a single marked edge) corresponds to a (potential) α-k-cycle in $\Theta'(D, q, r)$. Moreover, from Table 2 it can be verified that only $\mathbb{G}_i$-cycles involving a single bridge contribute to the k-score.

▶ **Proposition 17.** *There is a median $\mathbb{M}$ of $\Theta'(D, q, r)$ that is of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$ and is free of jump edges.*

Proof. Analogous to the proof of Proposition 9. ◀

With this framework, it is now straightforward to prove the following:

▶ **Theorem 18.** *Problem (ℓ, k)-Median is* NP*-complete for $\ell \geq 4$ and even $k \geq 8$.*

Proof. For $p = \ell - 1$, let D be a p-regular digraph free of 2-cycles and with x marked edges such that each 3-cycle has a single marked edge and each larger cycle has at least one marked edge. Now take the subdivision factors $r = \frac{k-6}{2}$ and $q = \lfloor \frac{k}{4} \rfloor$, and let $\Theta'(D, q, r)$ be the graph obtained via Construction 13.

Proposition 15 guarantees that there is a median $\mathbb{M}$ for $\Theta'(D, q, r)$ of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$, where $\mathbb{M}_Z$ is fixed. Furthermore, the k-score of $\mathbb{M}$ is $s_Z + s_*$, where $s_Z = qp^3|V'| + prx$.

The cross edges form p $\mathbb{G}_i$-cycles of length $2(q+1) < k$ with $\mathbb{G}_i^q$-bridges in each $X^{p,q}$-gadget. Therefore we need to show the following equivalence: An optimal solution $\mathcal{S}$ of MARKED-MAX3CDEC for digraph $D = (V, E)$ has a partition into f 6-cycles if and only if $s_* = f + p|V|$. The proof is analogous to the proof of Theorem 10 and shows that a set $\mathbb{M}_*$ including the $3f$ cross edges that build f α-k-cycles is optimal and allows the correspondence between the f α-k-cycles induced by $\mathbb{M}_*$ to the f 6-cycles of $\mathcal{S}$. The remaining edges of $\mathbb{M}_*$ are freely distributed into (compatible) loop and cross edges. ◀

5.2 NP-completeness of $(3, k)$-Median for even $k \geq 14$

Since the problem used for the proof of the NP-completeness of $(3, 12)$-Median is the same as the problem used for $(\ell, 6)$-Median, a similar adaptation of marked edges can be used.

Let $D = (V, E)$ be a 4-regular digraph with x marked edges and no 2-cycles and let $D_\Pi = (V', E')$ be obtained by replacing each vertex of D with a permutation network Π_2. Note that D_Π has the "same" marked edges as D, therefore each 6-cycle of D_Π has only one marked edge. Furthermore, each cycle of length greater than 6 has at least one marked edge.

Now let $q = \lfloor \frac{k}{4} \rfloor$, let $r = \frac{k-12}{2}$, and let $\Theta'(D_\Pi, q, r)$ be obtained via Construction 13. We have $k = 12 + 2r$, and from Table 3 it is easy to see that each 6-cycle C in D_Π (with a single marked edge) corresponds to a (potential) α-k-cycle in $\Theta'(D_\Pi, q, r)$. Furthermore, only $\mathbb{G}_i$-cycles involving a single bridge contribute to the k-score (Table 2).

▶ **Proposition 19.** *There is a median $\mathbb{M}$ of $\Theta'(D_\Pi, q, r)$ that is of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$ and is free of jump edges.*

Proof. Analogous to the proof of Proposition 9. ◀

These are all the relevant properties to derive our last result:

▶ **Theorem 20.** *Problem $(3, k)$-Median is NP-complete for $k \geq 14$.*

Proof. Let D be a 4-regular digraph free of 2-cycles and with x marked edges such that each 3-cycle has a single marked edge and each larger cycle has at least one marked edge. Then let $D_\Pi = (V', E')$ be the 2-regular digraph obtained by replacing each vertex of D by a Π_2. Recall that each m-cycle in D corresponds to a $2m$-cycle in D_Π and note that D_Π has the "same" marked edges as D, therefore each 6-cycle of D_Π has a single marked edge and each larger cycle has at least one marked edge. Now take the subdivision factors $q = \lfloor \frac{k}{4} \rfloor$, and $r = \frac{k-12}{2}$, and let $\Theta'(D_\Pi, q, r)$ be the graph obtained via Construction 13.

Proposition 15 guarantees that there is a median $\mathbb{M}$ for $\Theta'(D_\Pi, q, r)$ of the form $\mathbb{M} = \mathbb{M}_Z \cup \mathbb{M}_*$, where $\mathbb{M}_Z$ is fixed. Furthermore, the k-score of $\mathbb{M}$ is $s_Z + s_*$, where $s_Z = qp^3|V'| + prx = 8q|V'| + 2rx$.

The cross edges form two $\mathbb{G}_i$-cycles of length $2(q+1) < k$ with $\mathbb{G}_i^q$-bridges in each $X^{2,q}$-gadget. Therefore we need to show the following equivalence: an optimal solution $\mathcal{S}$ of MARKED-MAX6CDEC for digraph $D_\Pi = (V', E')$ has a partition into f 6-cycles if and only if $s_* = f + 2|V'|$. The proof is analogous to the proof of Theorem 10 and shows that a set $\mathbb{M}_*$ including the $6f$ cross edges that build f α-k-cycles is optimal and allows the correspondence between the f α-k-cycles induced by $\mathbb{M}_*$ and the f 6-cycles of $\mathcal{S}$. The remaining edges of $\mathbb{M}_*$ are freely distributed into (compatible) loop and cross edges. ◀

6 Conclusion

The study of median problems under intermediate genome distances provides a natural way to refine the complexity landscape between two classical endpoints: the breakpoint median problem, corresponding to σ_2, which is polynomial-time solvable for multichromosomal circular genomes, and the DCJ median problem, corresponding to σ_∞, which is NP-hard in the same setting. Within this landscape, the $(3, 4)$-Median problem plays a central role: it asks for a median of three genomes under the first intermediate distance after the breakpoint distance and has been highlighted in previous work as a fundamental case for understanding the transition from tractability to hardness.

Motivated by this question, we study the more general (ℓ, k)-Median problem. Rather than addressing only one isolated case, this formulation allows us to map broader regions of the parameter space and to identify where NP-hardness already arises. We prove that $(4, 6)$-Median and $(3, 12)$-Median are NP-complete for circular genomes and extend these results to larger values of ℓ and k. Summarized in Table 1, our results provide a clearer picture of the transition from tractable to intractable cases for median problems with σ_k distances. In particular, they delimit broad hardness regions around the central open case $(3, 4)$-Median and show that increasing either the number of input genomes or the distance parameter suffices to establish NP-completeness across large portions of the (ℓ, k) parameter space.

From a biological perspective, the restriction to short cycles can be viewed as a conservative way of emphasizing local, more reliable rearrangement signals, since long components in breakpoint graphs are often harder to interpret and may reflect breakpoint reuse, noise, or loss of historical signal [11]. Therefore, our work may lead to better rearrangement medians, computable similarly fast as the breakpoint median, but not as restricted.

References

1 Amihood Amir, Tzvika Hartman, Oren Kapah, Avivit Levy, and Ely Porat. On the cost of interchange rearrangement in strings. In *Proceedings of ESA 2007*, volume 4698 of *LNCS*, pages 99–110, 2007. doi:10.1007/978-3-540-75520-3_11.

2 Martin Bader. The transposition median problem is NP-complete. *Theoretical Computer Science*, 412:1099–1110, 2011. doi:10.1016/j.tcs.2010.12.009.

3 Jørgen Bang-Jensen and Gregory Z. Gutin. *Digraphs: theory, algorithms and applications*. Springer London, 2009. doi:10.1007/978-1-84800-998-1.

4 Anne Bergeron, Julia Mixtacki, and Jens Stoye. A unifying view of genome rearrangements. In *Proceedings of WABI 2006*, volume 4175 of *LNBI*, pages 163–173, 2006. doi:10.1007/11851561_16.

5 Marilia D. V. Braga, Leonie R. Brockmann, Katharina Klerx, and Jens Stoye. Investigating the complexity of the double distance problems. *Algorithms for Molecular Biology*, 19:1, 2024. doi:10.1186/s13015-023-00246-y.

6 Cedric Chauve. Personal communication at Dagstuhl Seminar no. 18451 (Genomics, Pattern Avoidance, and Statistical Mechanics), 2018.

7 Luís Cunha, Thiago Lopes, Uéverton Souza, Leonard Bohnenkämper, Marilia D. V. Braga, and Jens Stoye. Closing the complexity gap of the double distance problem. *IEEE Transactions on Computational Biology and Bioinformatics*, 22:2194–2205, 2025. doi:10.1109/TCBBIO.2025.3590588.

8 Luís Cunha, Ignasi Sau, and Uéverton Souza. On the parameterized complexity of the median and closest problems under some permutation metrics. *Algorithms for Molecular Biology*, 19:24, 2024. doi:10.1186/s13015-024-00269-z.

9 Anna E. Dewar, Laurence J. Belcher, and Stuart A. West. A phylogenetic approach to comparative genomics. *Nature Reviews Genetics*, 26:395–405, 2025. doi:10.1038/s41576-024-00803-0.

10 Jakub Kováč. On the complexity of rearrangement problems under the breakpoint distance. *Journal of Computational Biology*, 21:1–15, 2014. doi:10.1089/cmb.2013.0004.

11 David Sankoff. The signal in the genomes. *PLOS Computational Biology*, 2:e35, 2006. doi:10.1371/journal.pcbi.0020035.

12 Helmuth O. M. Silva, Diego P. Rubert, Eloi Araujo, Eckhard Steffen, Daniel Doerr, and Fábio V. Martinez. Algorithms for the genome median under a restricted measure of rearrangement. *RAIRO Operations Research*, 57:1045–1058, 2023. doi:10.1051/ro/2023052.

13 Eric Tannier, Chunfang Zheng, and David Sankoff. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10:120, 2009. doi:10.1186/1471-2105-10-120.

14 Geoffrey A. Watterson, Warren J. Ewens, Thomas E. Hall, and Alexander Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99:1–7, 1982. doi:10.1016/0022-5193(82)90384-8.

15 Andrew W. Xu. A fast and exact algorithm for the median of three problem: a graph decomposition approach. *Journal of Computational Biology*, 16:1369–1381, 2009. doi:10.1089/cmb.2009.0087.

16 Sophia Yancopoulos, Oliver Attie, and Richard Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21:3340–3346, 2005. doi:10.1093/bioinformatics/bti535.