Maximal Strip Recovery Problem with Gaps: Hardness and Approximation Algorithms

Laurent Bulteau, Guillaume Fertin, Irena Rusu

Maximal Strip Recovery Problem with Gaps: Hardness and Approximation Algorithms[☆]

Laurent Bulteau, Guillaume Fertin, Irena Rusu

Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR CNRS 6241
Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3 - France

Abstract

Given two comparative maps, that is two sequences of markers each representing a genome, the Maximal Strip Recovery problem (MSR) asks to extract a largest sequence of markers from each map such that the two extracted sequences are decomposable into non-intersecting strips (or synteny blocks). This aims at defining a robust set of synteny blocks between different species, which is a key to understand the evolution process since their last common ancestor. In this paper, we add a fundamental constraint to the initial problem, which expresses the biologically sustained need to bound the number of intermediate (non-selected) markers between two consecutive markers in a strip. We therefore introduce the problem δ-gap-MSR, where δ is a (usually small) non-negative integer that upper bounds the number of non-selected markers between two consecutive markers in a strip. We show that, if we restrict ourselves to comparative maps without duplicates, the problem is polynomial for $\delta = 0$, NP-complete for $\delta = 1$, and APX-hard for $\delta \geq 2$. For comparative maps with duplicates, the problem is APX-hard for all $\delta \geq 0$.

Keywords: algorithmic complexity, approximation algorithms, comparative maps, genome comparison, synteny blocks

1. Introduction

In comparative genomics, finding *synteny blocks* (that is, regions with similar content and gene order) of two genomes is a crucial task, as the decomposition of genomes into synteny blocks allows to estimate the nature of genome rearrangement events that took place during the evolution process since the last common ancestor of the genomes.

In addition to the difficulty to define a synteny block precisely, another difficulty is introduced by the quality of genome annotation. Zheng et al. [31] make a list of possible errors and ambiguities introduced by the mapping technology, which is used to obtain a representation of a genome as a sequence of *markers*, called a *genomic map*. Each marker represents a small, specific element which has been identified on the genome, at a specific position which is the

[☆] A preliminary version of this paper appeared in the proceedings of the 20th International Symposium on Algorithms and Computation, ISAAC 2009 [8]

Email addresses: Laurent.Bulteau@univ-nantes.fr (Laurent Bulteau), Guillaume.Fertin@univ-nantes.fr (Guillaume Fertin), Irena.Rusu@univ-nantes.fr (Irena Rusu)

marker's position. Comparing two genomes is then possible using their genomic maps, assuming that the pairs of identical markers on the two genomes are known (the maps are then called *comparative maps*). Comparative maps are less precise than genome sequences (either as DNA sequences or as sequences of genes), but still allow the identification of synteny blocks.

The problem that needs to be solved when no error occurs is the following: *Given two comparative maps, decompose them into non-intersecting synteny blocks*. In case of errors or ambiguities, Zheng et al. [31] propose to switch to the following problem: *Given two comparative maps, find a longest (possibly non-contiguous) subsequence of markers in each comparative map, such that the subsequences are decomposable into non-intersecting synteny blocks*. The idea behind this maximization problem is that true synteny is possibly interrupted by erroneous or ambiguous markers, which should be discarded before searching for synteny blocks.

The problem, called MAXIMAL STRIP RECOVERY (MSR), is obtained from this maximization problem using comparative maps with signed, but not duplicated, markers, and a specific definition of synteny blocks. Synteny blocks of two sequences are defined as *strips*, which are contiguous sequences of *at least two* markers that occur on each sequence either in the same order, or in reverse order and with a reversed sign.

Zheng et al. [31] and Choi et al. [13] propose two heuristics to solve the MSR problem. Chen et al. [10, 11] devise a 4-approximation algorithm for it, and propose several extensions of MSR, namely MSR-d, which compares an arbitrary number $d \geq 2$ of genomes, MSR-DU, which allows markers to be duplicated in the input maps, MSR-WT, where one takes into account the biological importance of the markers by giving a weight to each of them, and finally MSR-NB, which uses the number of non-breaking points (or adjacencies), instead of the length, as score function. NP-completeness results are obtained in [10] for a number of those extensions, and by Wang et al. [29] for MSR. A more precise hardness result is given by Jiang [19], who proves the APX-completeness of MSR. A more general review on related problems can be found in [28].

The MSR problem takes into account the need to keep as much of the data as possible from the initial comparative maps and the need to have conflict-free synteny blocks. However, it is too permissive as it allows two consecutive elements from one strip to be separated by an arbitrary long gap (in terms of intermediate markers) on the initial comparative maps, and possibly to be very close on one map and very far from each other on the other. As the discarded elements are supposed to be errors and ambiguities (which are rather the exception than the rule), and the elements kept in the subsequences are supposed to be the safe information (which is the major part of the comparative information), it follows that a safe synteny block should not allow arbitrarily long gaps.

We therefore introduce and study in this paper the δ-gap-MSR problem, a restriction of the MSR problem where the allowed gaps along the comparative maps between two consecutive elements in a strip are upper bounded by parameter δ, where δ is a given (usually small) non-negative integer. We investigate the algorithmic complexity of δ-gap-MSR depending on the allowed multiplicity for a marker and prove the results given in Table 1 (corresponding section numbers are given in brackets). For the NP-complete or APX-hard cases, we provide three approximation algorithms, whose approximation ratios are also given in Table 1.

The organization of the paper is as follows. In Section 2, we introduce some notations, and we formally define MSR, MSR-DU, δ-gap-MSR and δ-gap-MSR-DU. We prove in Section 3 the hardness results: after a preliminary result in Section 3.1, we prove the NP-completeness of 1-gap-MSR in Section 3.2; the APX-completeness of δ-gap-MSR, $\delta \geq 2$, in Section 3.3; and the APX-completeness of δ-gap-MSR-DU, $\delta \geq 0$, in Section 3.4. We then give polynomial-time algorithms in Section 4: an exact algorithm for 0-gap-MSR and a general 4-approximation in

Table 1: Hardness and approximability of variants of MSR.

Problem	Complexity		Approximation ratio	
0-gap-MSR	P	(4.1)	-	
1-gap-MSR	NP-hard	(3.2)	1.8	(4.2)
δ-gap-MSR ($\delta \geq 2$)	APX-hard	[21],(3.3)	4	(4.1)
MSR	APX-hard	[21]	4	[11]
0-gap-MSR-DU	APX-hard	(3.4)	2.25	(4.3)
δ-gap-MSR-DU ($\delta \geq 1$)	APX-hard	(3.4)	4	(4.1)
MSR-DU	APX-hard	[21]	4	[11]

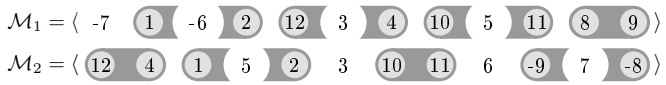
Section 4.1; a 1.8-approximation for 1-gap-MSR in Section 4.2; and a 2.25-approximation for 0-gap-MSR-DU in Section 4.3.

2. Notations and Definitions

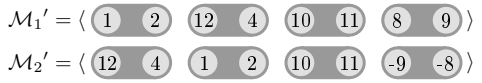
A *comparative map* $\mathcal{M}$ is a sequence of signed integers, where the absolute value of each integer represents a specific marker, and the sign represents the orientation of the marker on the chromosome, see for example Figure 1a. A marker may appear several times in a comparative map, possibly with different orientations: in this case, we say that the comparative map $\mathcal{M}$ contains *duplicates* (the presence of duplicates is useful if the markers represent genes possibly having paralogs in the comparative map). Note that a comparative map is suited to represent uni-chromosomal genomes. However, the algorithms we present can easily be adapted to handle multi-chromosomal instances. A *sequence* $\mathcal{M}$ is denoted $\mathcal{M}$=$\langle m_1, m_2, \ldots, m_l \rangle$, and its i^{th} element m_i is (also) denoted $\mathcal{M}[i]$.

A *subsequence* σ of $\mathcal{M}$ is a sequence $\langle \sigma_1, \ldots, \sigma_h \rangle$ of markers from $\mathcal{M}$ with $h \geq 2$ and positions $i_1 < i_2 < \ldots < i_h$ respectively on $\mathcal{M}$. The vector $(i_1, \ldots, i_h)$ is denoted $idx(\sigma, \mathcal{M})$. The *gap* of σ in $\mathcal{M}$ is $\max\{i_{k+1} - i_k - 1 \mid 1 \leq k < h\}$. The *length* $|\sigma|$ of σ is h. Two subsequences σ and τ are *non-overlapping* in $\mathcal{M}$ if one appears strictly before the other, that is, if the last element of $idx(\sigma, \mathcal{M})$ (resp. of $idx(\tau, \mathcal{M})$) is strictly smaller than the first element of $idx(\tau, \mathcal{M})$ (resp. of $idx(\sigma, \mathcal{M})$). The *reversed opposite* of $\langle \sigma_1, \ldots, \sigma_h \rangle$ is $\langle -\sigma_h, -\sigma_{h-1}, \ldots, -\sigma_1 \rangle$.

Given two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$, a *prestrip* is a subsequence σ of $\mathcal{M}_1$ of length at least 2, such that either σ or its reversed opposite is a subsequence of $\mathcal{M}_2$, and such that the markers in σ are pairwise distinct. A *sub-prestrip* σ' of a prestrip σ is a prestrip such that σ' is a subsequence of σ. The *gap* of a prestrip is the maximum of the gaps of the two corresponding subsequences in $\mathcal{M}_1$ and $\mathcal{M}_2$. Two prestrips are *non-overlapping* if the corresponding subsequences are non-overlapping, both in $\mathcal{M}_1$ and $\mathcal{M}_2$. A *strip* is a prestrip with gap 0. Strips represent synteny blocks between two comparative maps. A prestrip can also be seen as a synteny bock, but only if we consider that there is noise in the comparative maps (false markers appear between two consecutive markers of the "true" synteny block). A set of prestrips $\mathcal{S}$ is said to be *feasible* if it contains pairwise non-overlapping prestrips, and we write $\|\mathcal{S}\|$ for its *total size*: $\|\mathcal{S}\| = \sum_{\sigma \in \mathcal{S}} |\sigma|$. We call *peg marker*, and we write $\times$, a marker appearing in only one map (a peg marker never belongs to any prestrip, but it affects the gap of prestrips). A sequence of h consecutive peg markers is written $\times^h$.



(a) Two sequences $\mathcal{M}_1$, $\mathcal{M}_2$ without duplicates, and a feasible set of gap-1 prestrips of total length 8 $\{\langle 1, 2\rangle, \langle 12, 4\rangle, \langle 10, 11\rangle, \langle 8, 9\rangle\}$.



(b) Two subsequences $\mathcal{M}_1'$ and $\mathcal{M}_2'$ of $\mathcal{M}_1$ and $\mathcal{M}_2$ obtained by deleting markers $3, 5, 6, 7$, and partitioned into a set of strips.

Figure 1

Finally, we define some notions of graph theory: a graph $G = (V, E)$ is *cubic* if every vertex $u \in V$ has degree exactly 3. A set $X \subset V$ is said to be *independent* if for every edge $(u, v) \in E$, $u \notin X$ or $v \notin X$. The cardinality of a maximum independent set of G is written $\alpha(G)$.

The problems MSR (for MAXIMAL STRIP RECOVERY, see [31]) and MSR-DU [11] are defined, in their decision formulation, as follows:

Problem: MSR
Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$ without duplicates, $\ell \in \mathbb{N}$.
Question: Is there a feasible set $\mathcal{S}$ of prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $\|\mathcal{S}\| \geq \ell$?

Problem: MSR-DU
Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$ (possibly with duplicates), $\ell \in \mathbb{N}$.
Question: Is there a feasible set $\mathcal{S}$ of prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $\|\mathcal{S}\| \geq \ell$?

The idea behind both those problems is that, if we find a feasible set of prestrips with maximum total size, the elements appearing in no prestrip are considered as noise: we can remove them to "clean" the data. Indeed, once those elements are removed, the resulting comparative maps can be partitioned into common strips, i.e. the resulting genomes are decomposed into synteny blocks with the same set of blocks in both genomes, see Figure 1b. Heuristics for the first problem have been given in [31]. They have been improved in [11] into a 4-approximation algorithm. Finally, MSR (and thus MSR-DU) has been independently proved NP-hard in [32] and APX-hard in [19, 21]. The complementary problem, called CMSR, is the equivalent problem where one aims at minimizing the number k of deleted markers instead of maximizing the number ℓ of selected markers. Problem CMSR is also APX-hard [20, 21], and several approximation algorithms are known for it (with ratio 3 [33, 17], 3.5 [7] and 7/3 [22, 23]). Fixed parameter tractable algorithms have also been sought for CMSR, [29, 30, 33, 18, 17, 7] with in particular an FPT algorithm having a complexity of $O(2.36^k \text{poly}(n))$ [7].

The variant we introduce, δ-gap-MSR, takes into account the fact that it is unlikely for long sequences of markers to appear only from noise and errors. If a large number of elements is inserted between two consecutive elements of a prestrip (thus, if it has a large gap), then they are probably not errors, and the prestrip should not be considered a synteny block of the original genomes. We thus consider the restriction of the problem where the gap is bounded; the relevance

of applying this constraint on experimental data has been verified in [31]. The corresponding formal problems are defined as follows, where δ is any non-negative integer:

Problem: δ-gap-MSR
Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$ without duplicates, $\ell \in \mathbb{N}$.
Question: Is there a feasible set $\mathcal{S}$ of prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ such that every $\sigma \in \mathcal{S}$ has gap at most δ, and $\|\mathcal{S}\| \geq \ell$?

Problem: δ-gap-MSR-DU
Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$ (possibly with duplicates), $\ell \in \mathbb{N}$.
Question: Is there a feasible set $\mathcal{S}$ of prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ such that every $\sigma \in \mathcal{S}$ has gap at most δ, and $\|\mathcal{S}\| \geq \ell$?

With the gap constraint, only prestrips which are nearly contiguous are kept, while some noise in the input data is tolerated. There is no direct reduction from MSR to δ-gap-MSR or vice versa. However, the APX-hardness proof [21] can be extended to δ-gap-MSR with $\delta \geq 2$, and the FPT and approximation algorithms in [7] for CMSR also apply to δ-gap-CMSR. This paper focuses on the δ-gap-MSR and δ-gap-MSR-DU problems, especially for small values of δ.

3. Hardness Results

In this section we study the complexity of problems δ-gap-MSR and δ-gap-MSR-DU. We first observe in Section 3.1 that these problems become more difficult to solve when δ grows. Then in Section 3.2 we focus on 1-gap-MSR and prove that this problem is NP-hard. Finally, in Sections 3.3 and 3.4, we prove the APX-hardness of respectively δ-gap-MSR (for all $\delta \geq 2$) and δ-gap-MSR-DU (for all $\delta \geq 0$).

The APX-hardness results rely on the notion of L-reduction [26], defined below.

Given P an optimization problem, x an instance of P and y a feasible solution of x, we write $c_P(x, y)$ the cost of y. $opt_P(x)$ denotes the optimal value of $c_P(x, y)$ over all solutions y of x. Let P and Q be two optimization problems. An L-reduction from P to Q is a pair of polynomial time computable functions f and g such that:

- if x is an instance of P, then $f(x)$ is an instance of Q,
- if y is a solution of $f(x)$ for some x, then $g(y)$ is a solution of x,
- there exists a positive constant α such that

$$opt_Q(f(x)) \leq \alpha \; opt_P(x),$$

- there exists a positive constant β such that

$$|opt_P(x) - c_P(x, g(y))| \leq \beta \; |opt_Q(f(x)) - c_P(f(x), y)|.$$

Given such an L-reduction from P to Q, if P is NP-hard to approximate within $1 + \delta$, then Q is NP-hard to approximate within $1 + \delta/(\alpha\beta)$.

3.1. Hardness increases with the gap

In this section, we show that the problems δ-gap-MSR and δ-gap-MSR-DU become more and more difficult as δ increases. However, this result does not allow us to compare those problems to MSR and MSR-DU, for which the hardness results are quite independent (see [19] for the APX-hardness of those problems).

Theorem 1. *Let $0 \leq \delta < \delta'$. Then there exists an L-reduction from δ-gap-MSR to δ'-gap-MSR, and from δ-gap-MSR-DU to δ'-gap-MSR-DU.*

Note that the L-reduction [26] refers to the *optimization* versions of problems δ-gap-MSR and δ-gap-MSR-DU, which are easy to deduce from the decision versions presented in Section 2.

Proof. Let $(\mathcal{M}_1, \mathcal{M}_2)$ be an instance of δ-gap-MSR (resp. δ-gap-MSR-DU). For $i \in \{1, 2\}$ and any $k \geq 0$, we write $K_i^\delta(k)$ the sequence $\langle \mathcal{M}_i[\delta k + 1], \mathcal{M}_i[\delta k + 2], \ldots, \mathcal{M}_i[\delta k + \delta] \rangle$. We construct a pair of comparative maps $(\mathcal{M}_1{}', \mathcal{M}_2{}')$ in the following way:

$$\begin{aligned} \mathcal{M}_1{}' &= \langle K_1^\delta(0), \times^{\delta'-\delta}, K_1^\delta(1), \times^{\delta'-\delta}, \ldots \rangle \\ \mathcal{M}_2{}' &= \langle K_2^\delta(0), \times^{\delta'-\delta}, K_2^\delta(1), \times^{\delta'-\delta}, \ldots \rangle \end{aligned}$$

Consider that $(\mathcal{M}_1{}', \mathcal{M}_2{}')$ is an instance of δ'-gap-MSR (resp. δ'-gap-MSR-DU).

We show that there is a one-to-one correspondence between prestrips of $(\mathcal{M}_1, \mathcal{M}_2)$ with gap at most δ and prestrips of $(\mathcal{M}_1{}', \mathcal{M}_2{}')$ with gap at most δ'. Let σ be a prestrip of $(\mathcal{M}_1, \mathcal{M}_2)$ with gap at most δ, then it is also a prestrip of $(\mathcal{M}_1{}', \mathcal{M}_2{}')$. Moreover, let a and b be two consecutive markers of σ, such that a appears in $K_1^\delta(k)$ for some k, then b is in $K_1^\delta(k)$, $K_1^\delta(k-1)$ or $K_1^\delta(k+1)$. In the first case the gap between a and b is the same in $\mathcal{M}_1{}'$ as in $\mathcal{M}_1$, and otherwise the gap is increased by exactly $\delta' - \delta$. Hence, since the gap between a and b is at most δ in $\mathcal{M}_1$, it is at most δ' in $\mathcal{M}_1{}'$. We have the same property with $\mathcal{M}_2{}'$, thus σ has gap at most δ' in $(\mathcal{M}_1{}', \mathcal{M}_2{}')$.

Conversely, a prestrip σ' of gap δ' in $(\mathcal{M}_1{}', \mathcal{M}_2{}')$ corresponds to a prestrip in $(\mathcal{M}_1, \mathcal{M}_2)$, and if a, b are two consecutive elements in σ' with a gap strictly greater than δ in $(\mathcal{M}_1{}', \mathcal{M}_2{}')$, then they cannot appear in the same $K_i^\delta(k)$, and thus the gap between a and b is reduced by at least $\delta' - \delta$. Hence σ' has gap at most δ in $(\mathcal{M}_1, \mathcal{M}_2)$.

This one-to-one correspondence is enough to prove the fact that we have an L-reduction, since it preserves the prestrip lengths and the overlapping relation. □

3.2. NP-hardness of 1-gap-MSR

In this section, we prove the following theorem.

Theorem 2. *1-gap-MSR is NP-hard.*

The proof uses a reduction from a variant of MAXIMUM INDEPENDENT SET, 3-colored-MIS, which is defined below. A *3-edge-coloring* (also known as Tait Coloring) of a cubic graph $G = (V, E)$ is a partition of its edges in three classes $E = E^{\text{A}} \cup E^{\text{B}} \cup E^{\text{C}}$ such that if two edges $e_1, e_2 \in E$ are incident to a common vertex, they belong to different classes. Note that if a cubic graph with n vertices admits a 3-edge-coloring, then each class contains $n/2$ edges.

Problem: 3-colored-MIS
Input: A cubic graph $G = (V, E)$, provided with a 3-edge-coloring $(E^{\text{A}}, E^{\text{B}}, E^{\text{C}})$ of G, an integer k.
Question: Is $\alpha(G) \geq k$?

Lemma 3. *3-colored-MIS is NP-hard, even when restricted to cubic planar 2-connected graphs.*

Proof. To prove this lemma, we consider the class of cubic, planar, and 2-connected graphs: we present a reduction from the variant of the VERTEX COVER problem on this class of graphs (which is known to be NP-hard [5]) to 3-colored-MIS. This reduction uses a well-known equivalence between the 4-coloring of a planar graph and the 3-edge-coloring of a cubic graph [6].

Let $G = (V, E)$ be a cubic, planar, and 2-connected graph. The Four Color Theorem [3] ensures that its region graph admits a 4-coloring, and compute such a coloring, with colors taken in the set $\{0, 1, 2, 3\}$, using e.g. the quadratic time algorithm from Robertson et al. [27].

For every edge $e \in E$, we write $\phi(e)$ the pair of colors associated to its two adjacent faces (since the graph is 2-connected, e is adjacent to two different faces, which are colored with different values). We deduce a 3-edge-coloring of G with the following formulae:

- If $\phi(e) = \{0, 1\}$ or $\phi(e) = \{2, 3\}$, give to e the color A.
- If $\phi(e) = \{0, 2\}$ or $\phi(e) = \{1, 3\}$, give to e the color B.
- If $\phi(e) = \{0, 3\}$ or $\phi(e) = \{1, 2\}$, give to e the color C.

If two edges e_1 and e_2 are incident to the same vertex, then, since this vertex has degree 3, there are 3 faces f_0, f_1, f_2 (with 3 different colors) such that e_1 is adjacent to f_0 and f_1, and e_2 is adjacent to f_0 and f_2. So $\phi(e_1) \cap \phi(e_2)$ has size 1, and e_1 and e_2 have different colors. We have a 3-edge-coloring of G, which is an instance of 3-colored-MIS, and an optimal solution to 3-colored-MIS(G) of cardinality k gives an optimal solution to VERTEX COVER(G) of cardinality $|V| - k$: the reduction is complete and Lemma 3 is proved. □

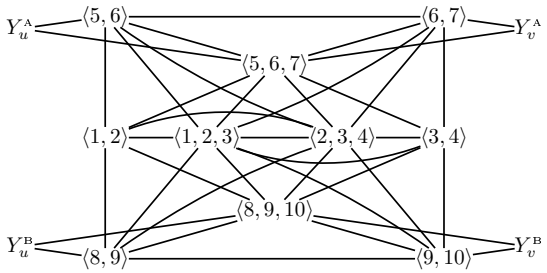
Starting from any instance of 3-colored-MIS, we construct two comparative maps as follows. First, we assign a list of 4 distinct positive integers (or 4 "markers") to each vertex $u \in V$: they are denoted y_u^{A1}, y_u^{A2}, y_u^{B1} and y_u^{B2}. We also assign a list of 10 distinct integers $x_{uv}^1, \ldots, x_{uv}^{10}$ to each edge $(u, v) \in E^{\mathrm{C}}$, in such a way that no integer appears in two different lists.

We construct the comparative maps with the following iterative procedure. Suppose we have arbitrarily ordered the vertices in V. In that case:

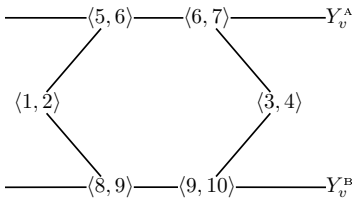
1. For all $(u, v) \in E^{\mathrm{A}}$ such that $u < v$, add $\langle y_u^{\mathrm{A1}}, y_v^{\mathrm{A1}}, y_u^{\mathrm{A2}}, y_v^{\mathrm{A2}}, \times, \times \rangle$ to $\mathcal{M}_1$.
2. For all $(u, v) \in E^{\mathrm{B}}$ such that $u < v$, add $\langle y_u^{\mathrm{B1}}, y_v^{\mathrm{B1}}, y_u^{\mathrm{B2}}, y_v^{\mathrm{B2}}, \times, \times \rangle$ to $\mathcal{M}_2$.
3. For all $(u, v) \in E^{\mathrm{C}}$ such that $u < v$, add $\Gamma_1(u, v)$ to $\mathcal{M}_1$, $\Gamma_2(u, v)$ to $\mathcal{M}_2$, where Γ_1 and Γ_2 are defined as:

$$\begin{aligned}\Gamma_1(u, v) = \langle & x_{uv}^1, x_{uv}^5, x_{uv}^2, x_{uv}^6, x_{uv}^3, x_{uv}^7, x_{uv}^4, \times, \times, \\ & y_u^{\mathrm{B1}}, x_{uv}^8, y_u^{\mathrm{B2}}, x_{uv}^9, y_v^{\mathrm{B1}}, x_{uv}^{10}, y_v^{\mathrm{B2}}, \times, \times \rangle \\ \Gamma_2(u, v) = \langle & x_{uv}^1, x_{uv}^8, x_{uv}^2, x_{uv}^9, x_{uv}^3, x_{uv}^{10}, x_{uv}^4, \times, \times, \\ & y_u^{\mathrm{A1}}, x_{uv}^5, y_u^{\mathrm{A2}}, x_{uv}^6, y_v^{\mathrm{A1}}, x_{uv}^7, y_v^{\mathrm{A2}}, \times, \times \rangle.\end{aligned}$$

Property 4. *Let $G = (V, E)$ be an n-vertex cubic graph with a 3-edge-coloring, and let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the two comparative maps obtained by the construction defined above. Then the optimal value of 1-gap-MSR over $(\mathcal{M}_1, \mathcal{M}_2)$ equals $4n + 2\alpha(G)$.*



(a) All length-2 and length-3 prestrips of γ_{uv} and their links to prestrips of $\mathcal{Y}$



(b) Induced subgraph with prestrips of γ_{uv}^{01}, γ_{uv}^{10} and γ_{uv}^{00}

Figure 2: Overlapping prestrips of γ_{uv} for an arc $(u,v) \in E^{\mathrm{C}}$, with notation $x_{uv}^i = i$ for all $1 \leq i \leq 10$

Proof. In the proof of this property, we use the following notations: for $u \in V$, $Y_u^{\mathrm{A}} = \langle y_u^{\mathrm{A}1}, y_u^{\mathrm{A}2}\rangle$ and $Y_u^{\mathrm{B}} = \langle y_u^{\mathrm{B}1}, y_u^{\mathrm{B}2}\rangle$. We say that σ_1 is a sub-prestrip of σ_2 if σ_1 and σ_2 are prestrips of $(\mathcal{M}_1, \mathcal{M}_2)$, and σ_1 is a subsequence of σ_2. We write $\mathcal{Y} = \{Y_u^{\mathrm{A}} \mid u \in V\} \cup \{Y_u^{\mathrm{B}} \mid u \in V\}$, and Ω the set of all prestrips of $(\mathcal{M}_1, \mathcal{M}_2)$ with gap at most 1. We also write $\ell_1(\mathcal{M}_1, \mathcal{M}_2)$ the optimal value of 1-gap-MSR($\mathcal{M}_1$,$\mathcal{M}_2$).

We first enumerate the possible prestrips of $(\mathcal{M}_1, \mathcal{M}_2)$ appearing in Ω:

- For all $(u,v) \in E^{\mathrm{A}}$, both Y_u^{A} and Y_v^{A} belong to Ω. Moreover, Y_u^{A} and Y_v^{A} overlap in $\mathcal{M}_1$ (see step 1 of the construction).
- For all $(u,v) \in E^{\mathrm{B}}$, both Y_u^{B} and Y_v^{B} belong to Ω. Moreover, Y_u^{B} and Y_v^{B} overlap in $\mathcal{M}_2$ (see step 2 of the construction).
- For all $(u,v) \in E^{\mathrm{C}}$, $\langle x_{uv}^1, x_{uv}^2, x_{uv}^3, x_{uv}^4\rangle$, $\langle x_{uv}^5, x_{uv}^6, x_{uv}^7\rangle$ and $\langle x_{uv}^8, x_{uv}^9, x_{uv}^{10}\rangle$ belong to Ω. We write γ_{uv} the set containing those three prestrips and all their sub-prestrips (see Figure 2a).

Because of the gap condition, which prevents prestrips from overlapping the peg markers, there are no other prestrips in Ω.

For an edge $(u,v) \in E^{\mathrm{C}}$, we give names to different feasible subsets of γ_{uv} (see Figure 2b):

$$\gamma_{uv}^{01} = \left\{\langle x_{uv}^1, x_{uv}^2\rangle, \langle x_{uv}^6, x_{uv}^7\rangle, \langle x_{uv}^9, x_{uv}^{10}\rangle\right\},$$
$$\gamma_{uv}^{10} = \left\{\langle x_{uv}^3, x_{uv}^4\rangle, \langle x_{uv}^5, x_{uv}^6\rangle, \langle x_{uv}^8, x_{uv}^9\rangle\right\},$$
$$\gamma_{uv}^{00} = \left\{\langle x_{uv}^1, x_{uv}^2\rangle, \langle x_{uv}^3, x_{uv}^4\rangle\right\}.$$

The first inequality we need to prove is the following:

$$\ell_1(\mathcal{M}_1, \mathcal{M}_2) \geq 4n + 2\alpha(G).$$

Consider X a maximal independent set of G ($|X| = \alpha(G)$). Construct a set of prestrips $\mathcal{S}$ in the following way:

1 For all $(u, v) \in E^{\mathrm{A}}$, if $u \notin X$, then add Y_u^{A} to $\mathcal{S}$. Else, $v \notin X$: add Y_v^{A} to $\mathcal{S}$.

2 For all $(u, v) \in E^{\mathrm{B}}$, if $u \notin X$, then add Y_u^{B} to $\mathcal{S}$. Else, $v \notin X$: add Y_v^{B} to $\mathcal{S}$.

3 For all $(u, v) \in E^{\mathrm{C}}$, there are three possible cases:

- If $u \notin X$ and $v \notin X$, add γ_{uv}^{00} to $\mathcal{S}$.
- If $u \in X$ and $v \notin X$, add γ_{uv}^{10} to $\mathcal{S}$.
- If $u \notin X$ and $v \in X$, add γ_{uv}^{01} to $\mathcal{S}$.

Before considering the overlaps in $\mathcal{S}$, we compute its total size: $\|\mathcal{S}\|$ is increased by 2 for each edge in E^{A} and in E^{B} (steps 1 and 2), and it is increased by either 6 or 4 for each edge in E^{C}, depending on whether this edge is incident to a vertex of X. As each vertex is incident to exactly one edge in E^{C}, we have the following formula:

$$\|\mathcal{S}\| = 2|E^{\mathrm{A}}| + 2|E^{\mathrm{B}}| + 4|E^{\mathrm{C}}| + 2|X|.$$

Since each class E^{A}, E^{B} and E^{C} contains exactly $n/2$ edges, and $|X| = \alpha(G)$, we have

$$\|\mathcal{S}\| = 4n + 2\alpha(G).$$

We now prove that $\mathcal{S}$ is a feasible set of prestrips. First note that prestrips of $\mathcal{S} \cap \mathcal{Y}$ are pairwise non-overlapping, since $\mathcal{S}$ never contains both Y_u^{A} and Y_v^{A} (resp. Y_u^{B} and Y_v^{B}) for $(u, v) \in E^{\mathrm{A}}$ (resp. $(u, v) \in E^{\mathrm{B}}$).

If $\sigma_1, \sigma_2 \in \mathcal{S} - \mathcal{Y}$, then there exist (u, v) and (u', v') such that $\sigma_1 \in \gamma_{uv}^{i,j}$ and $\sigma_2 \in \gamma_{u'v'}^{i',j'}$ are prestrips of $\mathcal{S}$ (where (i, j) and (i', j') are in $\{(0, 0), (0, 1), (1, 0)\}$). They also are non-overlapping: if $(u, v) \neq (u', v')$ then they appear in different sequences Γ_1 and Γ_2, and thus they cannot overlap. Otherwise, that is if $(u, v) = (u', v')$, then they appear in the same set $\gamma_{uv}^{i,j}$ which is, by construction, a set of non-overlapping prestrips.

Now suppose $\sigma_1 = \mathcal{S} \cap \mathcal{Y}$ (e.g. $\sigma_1 = Y_w^{\mathrm{A}}$ for some vertex w, the case $\sigma_1 = Y_w^{\mathrm{B}}$ is similar) and $\sigma_2 \in \gamma_{uv}^{i,j}$ are overlapping prestrips of $\mathcal{S}$. Then they can only overlap in $\Gamma_2(u, v)$, and $w = u$ or $w = v$. In the case $w = u$, (resp. $w = v$), σ_2 necessarily contains the element x_{uv}^5 (resp. x_{uv}^7), and thus γ_{uv}^{10} (resp. γ_{uv}^{01}) has been selected. It implies that $u \in X$ (resp. $v \in X$): in both cases, $w \in X$, which is a contradiction since Y_w^{A} can only be selected if $w \notin X$.

We conclude that prestrips in $\mathcal{S}$ are non-overlapping: consequently, $\mathcal{S}$ is a feasible set and we have

$$\ell_1(\mathcal{M}_1, \mathcal{M}_2) \geq \|\mathcal{S}\| = 4n + 2\alpha(G).$$

To prove the other inequality of Property 4, that is

$$\alpha(G) \geq \frac{\ell_1(\mathcal{M}_1, \mathcal{M}_2) - 4n}{2},$$

we consider $\mathcal{S}$, a maximal feasible set of prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ with gap at most 1. Then $\|\mathcal{S}\| = \ell_1(\mathcal{M}_1, \mathcal{M}_2)$.

We first enumerate and name the feasible subsets of γ_{uv} with total size at least 5, for some $(u, v) \in E^{\mathrm{C}}$. They are:

Name	Subset	Overlaps with			
γ_{uv}^{10}	$\left\{\langle x_{uv}^3, x_{uv}^4\rangle, \langle x_{uv}^5, x_{uv}^6\rangle, \langle x_{uv}^8, x_{uv}^9\rangle\right\}$	Y_u^{A},	Y_u^{B}		
γ_{uv}^{1a}	$\left\{\langle x_{uv}^5, x_{uv}^6, x_{uv}^7\rangle, \langle x_{uv}^8, x_{uv}^9\rangle\right\}$	Y_u^{A},	Y_u^{B},	Y_v^{A}	
γ_{uv}^{1b}	$\left\{\langle x_{uv}^5, x_{uv}^6\rangle, \langle x_{uv}^8, x_{uv}^9, x_{uv}^{10}\rangle\right\}$	Y_u^{A},	Y_u^{B},		Y_v^{B}
γ_{uv}^{01}	$\left\{\langle x_{uv}^1, x_{uv}^2\rangle, \langle x_{uv}^6, x_{uv}^7\rangle, \langle x_{uv}^9, x_{uv}^{10}\rangle\right\}$			Y_v^{A},	Y_v^{B}
γ_{uv}^{a1}	$\left\{\langle x_{uv}^5, x_{uv}^6, x_{uv}^7\rangle, \langle x_{uv}^9, x_{uv}^{10}\rangle\right\}$	Y_u^{A},		Y_v^{A},	Y_v^{B}
γ_{uv}^{b1}	$\left\{\langle x_{uv}^6, x_{uv}^7\rangle, \langle x_{uv}^8, x_{uv}^9, x_{uv}^{10}\rangle\right\}$		Y_u^{B},	Y_v^{A},	Y_v^{B}
γ_{uv}^{11}	$\left\{\langle x_{uv}^5, x_{uv}^6, x_{uv}^7\rangle, \langle x_{uv}^8, x_{uv}^9, x_{uv}^{10}\rangle\right\}$	Y_u^{A},	Y_u^{B},	Y_v^{A},	Y_v^{B}

There is no feasible subset of γ_{uv} with total size 7 or more.

We construct a feasible set of prestrips $\mathcal{S}'$ in the following way. Add all prestrips of $\mathcal{S} \cap \mathcal{Y}$ to $\mathcal{S}'$. Then, for all $(u, v) \in E^{\mathrm{C}}$, three cases are possible:
If $\|\mathcal{S} \cap \gamma_{uv}\| \leq 4$, add γ_{uv}^{00} to $\mathcal{S}'$ (case 1).
Else, if $\mathcal{S} \cap \gamma_{uv}$ is γ_{uv}^{01}, γ_{uv}^{a1}, γ_{uv}^{b1} or γ_{uv}^{11}, add γ_{uv}^{01} to $\mathcal{S}'$ (case 2).
Else, $\mathcal{S} \cap \gamma_{uv}$ is either γ_{uv}^{10}, γ_{uv}^{1a} or γ_{uv}^{1b}. Add γ_{uv}^{10} to $\mathcal{S}'$ (case 3).

Note that it is impossible to have overlapping prestrips in $\mathcal{S}'$ after these steps: in case 1, because $\gamma_{uv}^{00} = \langle x_{uv}^1, x_{uv}^2, x_{uv}^3, x_{uv}^4\rangle$ does not overlap with any prestrip except in γ_{uv}. In case 2, the prestrips in γ_{uv}^{01} overlap with Y_v^{A} and Y_v^{B}, but it is also the case of the sets γ_{uv}^{a1}, γ_{uv}^{b1} and γ_{uv}^{11}: neither Y_v^{A} nor Y_v^{B} belongs to $\mathcal{S}$ (and they do not belong to $\mathcal{S}'$). And in case 3, the prestrips in γ_{uv}^{10} overlap with Y_u^{A} and Y_u^{B}, but it is also the case for the sets γ_{uv}^{1a} and γ_{uv}^{1b}: neither Y_u^{A} nor Y_u^{B} belongs to $\mathcal{S}$ (nor $\mathcal{S}'$).

At this point, we have a set $\mathcal{S}'$ which is feasible and that satisfies $\|\mathcal{S}'\| \geq \|\mathcal{S}\|$: indeed, each time we do not directly include a subset of $\mathcal{S}$, we include a set of prestrips with greater or equal total size. Hence, $\|\mathcal{S}'\| = \ell_1(\mathcal{M}_1, \mathcal{M}_2)$.

We now create a first set of vertices $X_1 \subseteq V$ from $\mathcal{S}'$ with the following construction procedure. Start with $X_1 = \emptyset$, and for all $(u, v) \in E^{\mathrm{C}}$:

- If $\mathcal{S}' \cap \gamma_{uv} = \gamma_{uv}^{00}$, do nothing.
- If $\mathcal{S}' \cap \gamma_{uv} = \gamma_{uv}^{10}$, add u to X_1.
- If $\mathcal{S}' \cap \gamma_{uv} = \gamma_{uv}^{01}$, add v to X_1.

Two interesting remarks can be made about X_1. The first one is about its cardinality: since $\|\gamma_{uv}^{00}\| = 4$ and $\|\gamma_{uv}^{01}\| = \|\gamma_{uv}^{10}\| = 6$, then

$$|X_1| = \sum_{(u,v)\in E^{\mathrm{C}}} \frac{\|\mathcal{S}' \cap \gamma_{uv}\| - 4}{2} \ .$$

The other remark is that, if $u \in X_1$, then $Y_u^{\mathrm{A}}, Y_u^{\mathrm{B}} \notin \mathcal{S}'$: indeed, let v be the vertex such that $(u, v) \in E^{\mathrm{C}}$ (the case $(v, u) \in E^{\mathrm{C}}$ is similar). Since $u \in X_1$, $\gamma_{uv}^{10} \subseteq \mathcal{S}'$: the prestrips in γ_{uv}^{10} overlap Y_u^{A} and Y_u^{B}, so none of them is in $\mathcal{S}'$.

Note that X_1 is not necessarily independent (we only know that for every edge $(u, v) \in E^{\mathrm{C}}$, u and v cannot both be in X_1). If an edge $(u, v) \in E^{\mathrm{A}} \cup E^{\mathrm{B}}$ is such that $u, v \in X_1$, we call it a *bad* edge. We call n_b the number of bad edges, and for each bad edge we arbitrarity remove one of its end vertices from X_1. The result is an independent set X with cardinality

$$|X| = |X_1| - n_b \ .$$

By the previous remark about X_1, we know that if $(u, v) \in E^{\mathrm{A}}$ is a bad edge, then neither Y^{A}_u nor Y^{A}_v belongs to $\mathcal{S}'$. In any other case, at most one of Y^{A}_u and Y^{A}_v belongs to $\mathcal{S}'$, since they overlap in $\mathcal{M}_1$. And it can be seen that the same occurs with edges of E^{B}, thus the number of prestrips in $\mathcal{S}' \cap \mathcal{Y}$ is at most $|E^{\mathrm{A}}| + |E^{\mathrm{B}}| - n_b$.

We have:

$$\begin{aligned}
\|\mathcal{S}'\| &= \|\mathcal{S}' \cap \mathcal{Y}\| + \sum_{(u,v)\in E^{\mathrm{C}}} \|\mathcal{S}' \cap \gamma_{uv}\| \\
&= 2|\mathcal{S}' \cap \mathcal{Y}| + \sum_{(u,v)\in E^{\mathrm{C}}} \|\mathcal{S}' \cap \gamma_{uv}\| \\
&\leq 2(|E^{\mathrm{A}}| + |E^{\mathrm{B}}| - n_b) + \sum_{(u,v)\in E^{\mathrm{C}}} \|\mathcal{S}' \cap \gamma_{uv}\|
\end{aligned}$$

Hence,

$$\sum_{(u,v)\in E^{\mathrm{C}}} \|\mathcal{S}' \cap \gamma_{uv}\| \geq \|\mathcal{S}'\| - 2(n - n_b)$$

Finally,

$$\begin{aligned}
\alpha(G) &\geq |X| \\
&= \left(\sum_{(u,v)\in E^{\mathrm{C}}} \frac{\|\mathcal{S}' \cap \gamma_{uv}\| - 4}{2} \right) - n_b \\
&\geq \frac{\|\mathcal{S}'\| - 2(n - n_b) - 4|E^{\mathrm{C}}|}{2} - n_b \\
&= \frac{\|\mathcal{S}'\| - 4n}{2} \\
&= \frac{\ell_1(\mathcal{M}_1, \mathcal{M}_2) - 4n}{2}
\end{aligned}$$

This last inequality achieves the proof of Property 4. □

Proof (of Theorem 2). The above property directly implies that our construction (which can clearly be achieved in polynomial time) leads to a reduction from 3-colored-MIS to 1-gap-MSR, which proves Theorem 2. □

3.3. δ-gap-MSR is APX-hard for δ ≥ 2

The δ-gap-MSR problem is known to be APX-hard, by extending the APX-hardness proof for MSR [21], which uses a reduction from a variant of SAT. We present here an alternative proof, using a reduction based on a graph approach, which leads to a larger inapproximability lower bound: 1.0106382 instead of 1.000625.

Theorem 5. *δ-gap-MSR is APX-hard for any $\delta \geq 2$. More precisely, it is NP-hard to approximate within $95/94 \simeq 1.0106382$.*

To prove this theorem, we present an L-reduction to 2-gap-MSR from the variant of MAXIMUM INDEPENDENT SET restricted to cubic graphs, that we call 3-MIS here. Using Theorem 1, we extend the APX-hardness to δ-gap-MSR for any $\delta \geq 2$.

Problem: 3-MIS
Input: A cubic graph $G = (V, E)$, an integer k.
Question: Is $\alpha(G) \geq k$?

The 3-MIS problem is APX-hard [2], and NP-hard to approximate within 95/94 [12]. Given a cubic graph $G = (V, E)$, our reduction consists in constructing two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$, having properties P1, P2 and P3 described below, where Ω denotes the set of all prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ having gap at most δ:

P1. There exists a bijection Φ between V and Ω

P2. Every prestrip in Ω has length 2

P3. Two prestrips σ_1 and σ_2 of Ω are overlapping iff $\left(\Phi^{-1}(\sigma_1), \Phi^{-1}(\sigma_2)\right) \in E$

Let P_k denote the path graph with k vertices.

Lemma 6. *Given a cubic graph $G = (V, E)$, one can compute in polynomial time a partition of E into two classes E^B and E^W (for "Black" and "White" edges), such that (1) each connected component of (V, E^B) (called "black component") is isomorphic to a path P_k , and (2) each connected component of (V, E^W) (called "white component") is isomorphic to a path $P_{k'}$, with $k' \leq 4$.*

Proof. Given a cubic graph $G = (V, E)$, we can compute in polynomial time a bipartition of the edges $E = E^{\text{B}} \cup E^{\text{W}}$ such that both (V, E^{B}) and (V, E^{W}) are linear forests (i.e. acyclic graphs of maximum degree 2, see [1]). At this point every black and white component is isomorphic to a path.

Suppose there exist 5 vertices a, b, c, d, e such that edges (a, b), (b, c), (c, d), (d, e) are white. We deduce that b, c and d cannot belong to the same black component (they are three different degree-1 vertices of (V, E^{B}), and a path graph has only 2 vertices of degree 1). Then either b and c, or c and d, are in different black components. In the first case, we can switch the color of (b, c) from white to black, and we can switch (c, d) in the second case. The result is that (V, E^{B}) and (V, E^{W}) are still linear forests, and we have strictly reduced the size of a white component. We can apply this process until no white component is longer than P_4: Lemma 6 is proved. □

The first step of the reduction from 3-MIS to 2-gap-MSR is to compute a partition of E into two classes E^{B} and E^{W} according to Lemma 6. We then construct two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$, satisfying properties P1, P2 and P3. Moreover, incompatibilities in $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) will correspond to black (resp. white) edges. We begin by assigning a different pair of integers (x_a, x'_a) to every vertex $a \in V(G)$; we write $\Phi(a) = \langle x_a, x'_a \rangle$.

Then, for every black component B_i of order k, let $V(B_i) = \{a_h \mid 1 \leq h \leq k\}$ and let $(a_h, a_{h+1}) \in E^{\text{B}}$ for $1 \leq h < k$; we construct the following sequence (see Figure 3):

$$I_i = \langle x_{a_1}, \times, x_{a_2}, x'_{a_1}, \ldots, x_{a_h}, x'_{a_{h-1}}, x_{a_{h+1}}, x'_{a_h}, \ldots, x_{a_k}, x'_{a_{k-1}}, \times, x'_{a_k} \rangle$$

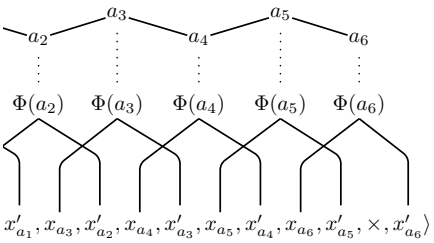


Figure 3: Transformation of a black component B_i (top) into the sequence I_i (bottom)

The full comparative map $\mathcal{M}_1$ is given by $\mathcal{M}_1 = \langle I_1, \times^3, I_2, \times^3, \ldots\rangle$.

For $\mathcal{M}_2$, we use a similar construction, but we need to take the reversed opposite of some subsequences to avoid creating undesired prestrips. For a white component W_j having 4 vertices, say a, b, c and d with $(a, b), (b, c), (c, d) \in E^{\mathrm{W}}$, we create the following sequence:

$$J_j = \langle x_a, \times, x_b, x'_a, -x'_c, x'_b, -x'_d, -x_c, \times, -x_d\rangle.$$

If W_j is of order three (resp. two), we remove the extra elements from J_j, i.e. we obtain $J_j = \langle x_a, \times, x_b, x'_a, -x'_c, x'_b, \times, -x_c\rangle$ (resp. $J_j = \langle x_a, \times, x_b, x'_a, \times, x'_b\rangle$). Finally, $\mathcal{M}_2$ is created in the same way as $\mathcal{M}_1$: $\mathcal{M}_2 = \langle J_1, \times^3, J_2, \times^3, \ldots\rangle$.

Lemma 7. *The set Ω of the prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ with gap at most 2 is exactly $\{\Phi(a) \mid a \in V\}$. Moreover, $\Phi(a)$ and $\Phi(b)$ overlap in $\mathcal{M}_1$ iff $(a, b) \in E^{B}$, and $\Phi(a)$ and $\Phi(b)$ overlap in $\mathcal{M}_2$ iff $(a, b) \in E^{W}$.*

Proof. Suppose, by contradiction, that a prestrip of Ω contains markers corresponding to two different vertices u and v: then there exists $\sigma \in \Omega$ such that $\sigma = \langle \sigma_1, \sigma_2\rangle$, with $\sigma_1 \in \{x_u, x'_u\}$ and $\sigma_2 \in \{x_v, x'_v\}$.

First note that σ_1 and σ_2 appear with the same orientation, since every element of $\mathcal{M}_1$ is positive. Because of the gap condition, both elements must appear in the same I_i in $\mathcal{M}_1$, and in the same J_j in $\mathcal{M}_2$. In J_j, the markers with a positive orientation come from the two prestrips associated to vertices (called a and b in the construction) linked by a white edge. Similarly, the negative markers come from two vertices c and d with $(c, d) \in E^{\mathrm{W}}$. So, whatever the orientation of σ in $\mathcal{M}_2$, there must be an edge $(u, v) \in E^{\mathrm{W}}$, and consequently this edge does not belong to E^{B}.

We look at the subsequences in I_i with gap at most 2 which do not contain any peg marker. Using the notations of the construction, they are of one of the following kinds:

1. $\langle x_{a_h}, x'_{a_{h-1}}\rangle$
2. $\langle x_{a_h}, x_{a_{h+1}}\rangle$
3. $\langle x_{a_h}, x'_{a_h}\rangle$
4. $\langle x'_{a_h}, x_{a_{h+2}}\rangle$
5. $\langle x'_{a_h}, x'_{a_{h+1}}\rangle$
6. $\langle x'_{a_h}, x_{a_{h+3}}\rangle$

If we write $u = a_h$, then v can be none of a_{h-1}, a_h or a_{h+1}, since $u \neq v$ and $(u, v) \notin E^{\mathrm{B}}$. Only possibilities 4. and 6. remain, that is a prestrip of the form $\sigma = \langle x'_u, x_v\rangle$. However this kind of prestrip does not appear in J_j, thus we have proved that each prestrip of $\mathcal{M}_1$ and $\mathcal{M}_2$ with gap

at most 2 is either of the kind $\langle x_u, x'_u \rangle$ or $\langle x'_u, x_u \rangle$. Moreover, for any $u \in V$, $\langle x'_u, x_u \rangle$ is not a subsequence of $\mathcal{M}_1$ nor $\mathcal{M}_2$, thus each prestrip of Ω can be written $\langle x_u, x'_u \rangle = \Phi(u)$ with $u \in V$.

Conversely, for every $a \in V$, $\langle x_a, x'_a \rangle$ is a subsequence of $\mathcal{M}_1$ with gap 2, and either $\langle x_a, x'_a \rangle$ or $\langle -x'_a, -x_a \rangle$ is a subsequence of $\mathcal{M}_2$ with gap 2. So $\Phi(a)$ is a prestrip of $\mathcal{M}_1$ and $\mathcal{M}_2$ with gap 2: it belongs to Ω.

Finally, the second part of Lemma 7 is deduced from the construction of the sequences $\mathcal{M}_1$ and $\mathcal{M}_2$. □

The consequence of Lemma 7 is that $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy the three properties P1, P2 and P3 defined above. The reduction we have described is an L-reduction from 3-MIS to 2-gap-MSR: indeed, Φ transforms an independent set of cardinality k into a feasible set of prestrips with gap 2 of total size $\ell = 2k$, and Φ^{-1} does the reverse operation. We conclude that 2-gap-MSR and, by Theorem 1, δ-gap-MSR for $\delta \geq 2$ is APX-hard. More precisely, these problems are, like 3-MIS, NP-hard to approximate within 95/94. Thus Theorem 5 is proved.

3.4. δ-gap-MSR-DU is APX-hard, for all δ

In this section, we focus on the variant of MSR allowing duplicates in the input sequences. The problem becomes much harder, even with a *0-gap* constraint. The 0-gap-MSR-DU shares similarities with a well-known string comparison problem: Minimum Common String Partition (MCSP), see e.g. [15]. Both problems deal with two sequences with duplicates, and aim at matching markers in order to reconstruct common strips. However, they differ both in the input sequences and in the optimization function. Indeed, each marker in an MCSP instance should have the same number of occurrences in both sequences, which is not necessary in MSR-DU. Moreover, in MCSP, one wants to create a minimum number of strips, using length-1 strips if necessary (and all elements are covered), while in MSR-DU the number of elements covered by the strips (each strip having length at least 2) has to be maximized.

Theorem 8. *δ-gap-MSR-DU is APX-hard for any $\delta \geq 0$. More precisely, it is NP-hard to approximate within $8649/8648 \simeq 1.000115$ for $\delta = 0, 1$, and within $95/94 \simeq 1.0106382$ for $\delta \geq 2$.*

We note that we need to consider only 0-gap-MSR-DU since APX-hardness of δ-gap-MSR-DU directly follows from APX-hardness of 0-gap-MSR-DU (see Theorem 1). Moreover, the inapproximability bound for $\delta \geq 2$ is directly deduced from Theorem 5.

As in the previous section, we use an L-reduction from 3-MIS, the variant of MAXIMUM INDEPENDENT SET restricted to cubic graphs.

This L-reduction is done in two steps. First, we transform the input graph such that it admits a partition of its edges and a labelling of its vertices with good properties (see below for the corresponding definitions). Then, using these partitions and labellings, we can create an instance of 0-gap-MSR-DU which simulates the behaviour of 3-MIS. Finally, Lemma 15 gives the whole L-reduction from the APX-hard problem 3-MIS [2] to 0-gap-MSR-DU, which achieves the proof of Theorem 8.

The first transformation of the reduction defines an oriented graph, for which we use the following definitions. If $G = (V, A)$ is a loopless oriented graph, $a = (u, v) \in A$ corresponds to an arc from u to v, of which u is the *source*, and v the *target*. The *degree* of a vertex $u \in V$ is the number of arcs $a \in A$ of which u is the source or the target. A subset X of V is *independent* if for all $(u, v) \in A$, X does not contain both u and v.

Definition 1. *Let $G = (V, A)$ be a loopless directed graph. We say that $A = A_1 \cup A_2$ is a* good partition *of A if (i) $A_1 \cap A_2 = \emptyset$, (ii) for any $p \in \{1, 2\}$ and $a, b \in A_p$, a and b neither have the same source nor the same target, and (iii) (V, A_2) contains no cycle.*

Note that if $A = A_1 \cup A_2$ is a good partition of $G = (V, A)$, then every $u \in V$ has degree at most two in (V, A_1) and in (V, A_2) (using condition ii). Moreover, with condition iii, if $C \subseteq V$ is a connected component in the underlying undirected graph of (V, A_2), then we can write $C = \{u_0, u_1, \ldots, u_k\}$, such that the vertices $u_0, u_1, \ldots, u_k$ form a directed path: $(u_i, u_j) \in A_2 \Leftrightarrow j = i+1$.

Definition 2. *Let $G = (V, A)$ be a loopless directed graph, Σ a set of labels, and $\phi : V \to \Sigma \times \Sigma$, where we write $\phi(u) = (u^1, u^2)$ the image of a vertex u.*

Then ϕ is said to be a good labelling *of G if*

1. *$u^1 \neq u^2$ for all $u \in V$,*
2. *$(v^1, v^2) \neq \phi(u)$ and $(v^2, v^1) \neq \phi(u)$ for any $u, v \in V$ such that $u \neq v$,*
3. *$u^2 = v^1$ for $(u, v) \in A$.*

Lemma 9. *Let $G = (V, E)$ be an undirected graph with maximum degree 3. Then we can compute in polynomial time the following entities:*

- *a directed graph $G' = (V', A')$,*
- *a good partition $A' = A'_1 \cup A'_2$ of G',*
- *a good labelling ϕ of G',*

with the following properties:

- *$|V'| = |V| + 2|E|$, $|A'| = 3|E|$,*
- *the maximum degree of G' is 3*
- *$\alpha(G') \geq \alpha(G) + |E|$.*
- *If X' is an independent set of G', we can deduce an independent set X of G such that $|X'| \leq |X| + |E|$.*

Proof. We first use Vizing's theorem (see [25]) to obtain a 4-coloring of the edges of (V, E), that is a partition $E = E_1 \cup E_2 \cup E_3 \cup E_4$, such that two edges appearing in the same E_i are not incident.

To create ϕ, we need a numbering of the vertices $y : V \to \{0, \ldots, |V| - 1\}$ and a numbering of the edges $x : E \to \{0, \ldots, |E| - 1\}$. For each $u \in V$, we choose $\phi(u) = (2y(u), 2y(u) + 1)$.

For each $e = \{u, v\} \in E$, we create two vertices u_e and v_e, and three arcs a_e, b_e, c_e such that (see Figure 4a, and an example in Figure 4b):

- If $e \in E_1 \cup E_2$, then $a_e = (u, u_e)$, $b_e = (v, v_e)$, $c_e = (u_e, v_e)$. Moreover, $\phi(u_e) = (u^2, v^2)$ and $\phi(v_e) = (v^2, 2|V| + x(e))$.
- If $e \in E_3 \cup E_4$, then $a_e = (u_e, u)$, $b_e = (v_e, v)$, $c_e = (v_e, u_e)$. Moreover, $\phi(u_e) = (v^1, u^1)$, and $\phi(v_e) = (2|V| + x(e), v^1)$.

We add each arc a_e, b_e and c_e to either A'_1 or A'_2, according to the following rules:

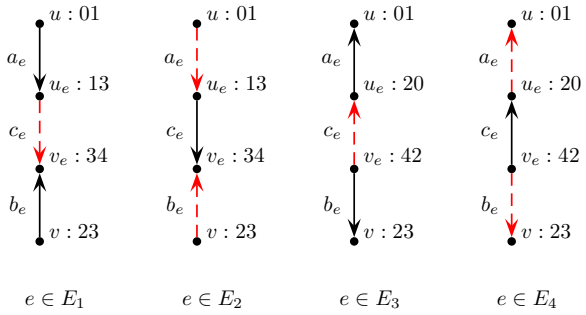
- If $e \in E_1 \cup E_3$, then $a_e, b_e \in A'_1$ and $c_e \in A'_2$.
- If $e \in E_2 \cup E_4$, then $a_e, b_e \in A'_2$ and $c_e \in A'_1$.

Thus we have created a graph $G' = (V', A')$ with the set of vertices $V' = V \cup \{u_e, v_e \mid e \in E\}$ and the set of arcs $A' = A'_1 \cup A'_2$. We now prove that this graph has the required properties:

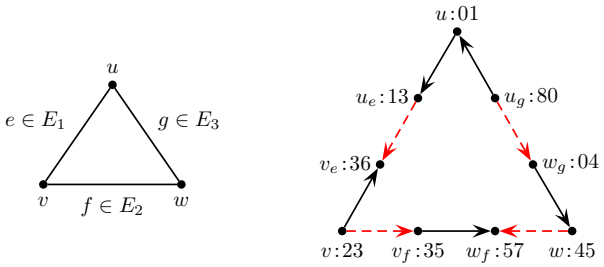
- The cardinality conditions on V' and A' are satisfied by construction.
- The degree of $u \in V$ in (V', A') is the same as in (V, E), thus it is at most 3. And the degree of $u_e \in V' - V$ in (V', A') is 2.
- The independence number of the graph is increased by at least 1 each time we split an edge e into 3 arcs a_e, b_e, c_e (we can add either u_e or v_e to any independent set).
- Let X' be an independent set of G'. We create X in the following way: start with $X = X'$. Then, consider each edge $e = (u, v)$ of G. Several cases are possible: if X' contains both u and v, then it contains neither u_e nor v_e, and we remove u from X. Otherwise, X' contains at most one element among $\{u_e, v_e\}$, and we remove this element. We obtain a set X which is a subset of V (it does not contain any vertex u_e or v_e) and is independent in G (it cannot contain both u and v for $(u, v) \in E$). Finally, we have removed at most one vertex per edge $e \in E$, so $|X| \geq |X'| - E$.
- $A' = A'_1 \cup A'_2$ is a good partition of G'. (i) $A'_1 \cap A'_2 = \emptyset$ by construction. (ii) Each vertex $u_e \in V' - V$ is adjacent to exactly one arc of A'_1 and one arc of A'_2. Each vertex $u \in V$ is adjacent in G to at most one edge in E_1 (resp. E_2), thus it is the source in G' of at most one arc in A'_1 (resp. A'_2). And it is also adjacent in G to at most one edge in E_3 (resp. E_4), so it is the target in G' of at most one arc in A'_1 (resp. A'_2). (iii) There are no cycles in (V', A'_2), since each connected component of this graph contains at most two arcs.
- ϕ is a good labelling of G': first remark that for $u, v \in V$, $\{u^1, u^2\} \cap \{v^1, v^2\} = \emptyset$. It implies that condition 1. is true for all $u \in V'$, and that condition 2. is true for all $u \in V'$ and $v \in V$. For $u_e \in V' - V$ and $u'_{e'} \in V' - V$, with $e = \{u, v\}$ and $e' = \{u', v'\} \neq e$, we have $u' \notin \{u, v\}$ or $v' \notin \{u, v\}$. It implies that one element of $\phi(u'_{e'})$ does not appear in $\phi(u_e)$. Finally, condition 3. is verified by construction.

□

Let $G = (V, A)$ be a directed graph, with $A = A_1 \cup A_2$ a good partition of A (such that (V, A_2) is a degree-2 acyclic graph) and $\phi : V \to \Sigma \times \Sigma$ a good labelling of G. We give an arbitrary order over each set A_1, A_2 and V, and we construct two gene maps $\mathcal{M}_1$, $\mathcal{M}_2$ with the following procedure (see for example the directed graph in Figure 5a and the resulting maps in Figure 5b):
$\mathcal{M}_1 = \langle\rangle$;
For each $u \in V$
  $\mathcal{M}_1 = \langle \mathcal{M}_1, \times, u^2, u^1 \rangle$;
For each $(u, v) \in A_1$
  $\mathcal{M}_1 = \langle \mathcal{M}_1, \times, u^1, u^2, v^2 \rangle$;
For each $u \in V$ s.t. u has no incoming arc in A_1
  $\mathcal{M}_1 = \langle \mathcal{M}_1, \times, u^1, u^2 \rangle$;
For each $u \in V$ s.t. u has no outgoing arc in A_1



(a) Construction of vertices u_e, v_e, arcs a_e, b_e, c_e, and labelling for an edge $e = \{u, v\}$, and for each case $e \in E_1, E_2, E_3$ or E_4. We assume $y(u) = 0$, $y(v) = 1$ and $x(e) = 0$.



(b) Example on the triangle graph, with $y(u) = 0$, $y(v) = 1$, $y(w) = 2$ and $x(e) = 0$, $x(f) = 1$, $x(g) = 2$.

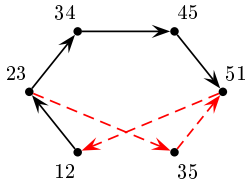
Figure 4: Construction of a good partition and a good labelling ϕ for an undirected graph of maximum degree 3. We write $u : u^1u^2$ for $\phi(u) = (u^1, u^2)$; arcs of A'_1 are solid, those of A'_2 are dashed.

$\mathcal{M}_1 = \langle \mathcal{M}_1, \times, u^1, u^2 \rangle$;
$\mathcal{M}_2 = \langle\rangle$;
For each connected component $\{u_0, u_1, \ldots, u_k\}$ in (V, A_2)
//such that $u_0, u_1, \ldots, u_k$ is a path in (V, A_2), with $k \geq 0$
$\mathcal{M}_2 = \langle \mathcal{M}_2, \times, u_0^1 \rangle$;
For $i = 0$ to k
$\mathcal{M}_2 = \langle \mathcal{M}_2, u_i^2, u_i^1, u_i^2 \rangle$;

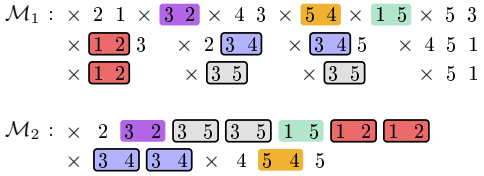
The resulting maps have the property that, for all $u \in V$:

- there is exactly one occurrence of $\langle u^2, u^1 \rangle$ in $\mathcal{M}_1$, and exactly two occurrences of $\langle u^1, u^2 \rangle$ in $\mathcal{M}_1$ (recall that, for $(u, v) \in A_1$, $u^2 = v_1$). Moreover, these three subsequences are non-overlapping;
- there is exactly one occurrence of $\langle u^1, u^2, u^1, u^2 \rangle$ in $\mathcal{M}_2$, and no other occurrence of $\langle u^1, u^2 \rangle$ or $\langle u^2, u^1 \rangle$.

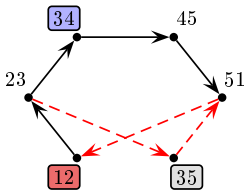
Moreover, the strip $\langle u^2, u^1 \rangle$ does not intersect any occurrence of $\langle v^1, v^2 \rangle$ or $\langle v^2, v^1 \rangle$ for $v \neq u$.



(a) Directed graph $G = (V, A)$ with a good partition of A and a good labelling of the vertices



(b) Maps created for the reduction



(c) Maximum Independent Set corresponding to a solution of 0-gap-MSR-DU($\mathcal{M}_1$,$\mathcal{M}_2$)

Figure 5: Reduction from MIS to 0-gap-MSR-DU

Definition 3. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the maps constructed by the above procedure from a graph $G = (V, A)$, and let O be a solution of 0-gap-MSR-DU($\mathcal{M}_1, \mathcal{M}_2$). We say that $u \in V$ is* selected *in O if both occurrences of $\langle u^1, u^2\rangle$ appear in O, we say it is unselected if only the strip $\langle u^2, u^1\rangle$ appears in O.*

Lemma 10. *All strips in a feasible solution have length 2. Moreover, each of the strips is of one of the following kinds: $\langle u^1, u^2\rangle$ or $\langle u^2, u^1\rangle$ for $u \in V$.*

Proof. Since the peg markers × cannot be selected in map $\mathcal{M}_1$, and a strip cannot overlap them (the gap constraint is $\delta = 0$), all strips are either length-2 strips of the kind $\langle u^1, u^2\rangle$ or $\langle u^2, u^1\rangle$, or length-3 strips of the kind $\langle u^1, u^2, v^2\rangle$, with $(u, v) \in A_1$. We show that strips of this last kind are in fact impossible. Indeed, we have $v^2 \neq u^1$, and three different non-peg markers are never consecutive in map $\mathcal{M}_2$, except for the sequences $\langle u_i^1, u_i^2, u_{i+1}^2\rangle$, where (u_i, u_{i+1}) is an arc of A_2. Hence if we have such a length-3 strip, $u^1 = u_i^1$, $u^2 = u_i^2 = v^1 = u_{i+1}^1$, and $v^2 = u_{i+1}^2$. So $u = u_i$ and $v = u_{i+1}$, which implies that the arc (u, v) appears both in A_1 and in A_2, a contradiction. □

Lemma 11. *Given a feasible solution $\mathcal{S}$ of 0-gap-MSR-DU($\mathcal{M}_1$,$\mathcal{M}_2$) of total size ℓ, we can create a feasible solution $\mathcal{S}'$ of total size at least ℓ where each vertex $u \in V$ is either selected or unselected.*

Proof. We start with $\mathcal{S}' = \mathcal{S}$. Remember that for all u, the strip $\langle u^2, u^1\rangle$ only intersects the occurrences of $\langle u^1, u^2\rangle$, which already implies that a vertex u cannot be both selected and unselected. If $\mathcal{S}'$ uses at most one strip amongst $\langle u^2, u^1\rangle$ and the occurrences of $\langle u^1, u^2\rangle$, then we can replace it by $\langle u^2, u^1\rangle$ without creating conflicts, and the vertex u becomes unselected. Otherwise, $\mathcal{S}'$ uses two independent strips amongst $\langle u^2, u^1\rangle$ and the occurrences of $\langle u^1, u^2\rangle$, hence it cannot use $\langle u^2, u^1\rangle$, and u is selected. □

Lemma 12. *If $(u, v) \in A$, then u and v cannot be both selected in a feasible solution.*

Proof. Two cases are possible: $(u, v) \in A_1$ and $(u, v) \in A_2$. In the first case, one occurrence of $\langle u^1, u^2 \rangle$ intersects an occurrence of $\langle v^1, v^2 \rangle$ in map $\mathcal{M}_1$ (in the sequence $\langle \times, u^1, u^2, v^2, \times \rangle$). So both occurrences of $\langle u^1, u^2 \rangle$ and both occurrences of $\langle v^1, v^2 \rangle$ cannot be all selected in the same solution. The situation is similar if $(u, v) \in A_2$, since an occurrence of $\langle u^1, u^2 \rangle$ intersects an occurrence of $\langle v^1, v^2 \rangle$ in the sequence $\langle u^1, u^2, u^1, u^2, v^2, v^1, v^2 \rangle$ of map $\mathcal{M}_2$. □

Lemma 13. *If $X \subset V$ is an independent set of $G = (V, A)$, then the set of strips selecting every vertex $u \in X$ and unselecting every $u \in V - X$ is a feasible solution of 0-gap-MSR-DU($\mathcal{M}_1, \mathcal{M}_2$).*

Proof. We first create the feasible solution which keeps all the vertices $u \in V$ unselected (this solution contains all strips $\langle u^2, u^1 \rangle$, which are pairwise non-overlapping). For each $u \in X$, we replace $\langle u^2, u^1 \rangle$ by the two occurrences of $\langle u^1, u^2 \rangle$: these two strips do not overlap any v^2, v^1 for $v \in V$, nor any $\langle v^1, v^2 \rangle$ for $v \in X$ (this is because there is no arc linking u and v). Thus we end with a feasible set of strips, and every $u \in X$ is selected, while the other vertices are unselected. □

Lemma 14. *There exists an independent set X of G of cardinality k if, and only if, there exists a feasible solution $\mathcal{S}$ of 0-gap-MSR-DU ($\mathcal{M}_1, \mathcal{M}_2$) of total size $\ell = 2(|V| + k)$. Moreover, given such an $\mathcal{S}$, the corresponding independent set X is computable in polynomial time.*

Proof. This is the corollary of the four previous lemmas: the "only if" part follows directly Lemma 13. For the "if" part, we start with $\mathcal{S}$ a feasible solution of 0-gap-MSR-DU($\mathcal{M}_1, \mathcal{M}_2$) of total size ℓ. Using Lemma 11, we obtain a feasible solution $\mathcal{S}'$ of total size at least ℓ, such that, if X_1 is the set of selected vertices in $\mathcal{S}'$ and X_2 the set of unselected vertices, then $X_1 \cup X_2$ is a partition of V. Then $\|\mathcal{S}'\| = 4|X_1| + 2|X_2|$, and X_1 is an independent set (by Lemma 12). Thus $\ell \leq 4|X_1| + 2|X_2| = 2(|V| + |X_1|)$. Then we can remove vertices from X_1 until we reach a set X such that $\ell = 2(|V| + |X|)$. □

Lemma 15. *There exists an L-reduction from 3-MIS to 0-gap-MSR-DU.*

Proof. We start with an instance $G = (V, E)$ of 3-MIS (G is a cubic graph). Using Lemma 9, we obtain a new directed graph $G' = (V', A')$, a good partition $A' = A'_1 \cup A'_2$ of G', and a good labelling ϕ of G'. Moreover, $|V'| = |V| + 2|E|$, and $\alpha(G') = \alpha(G) + |E|$. Since G is a cubic graph, we have $|E| = 3|V|/2$ and $|V| \leq 4\alpha(G)$, thus $|V'| = 4|V|$ and $\alpha(G') \leq 7\alpha(G)$. Next we create an instance $(\mathcal{M}_1, \mathcal{M}_2)$ of 0-gap-MSR-DU from $G' = (V', A')$ with the procedure described above. Let ℓ_0 be the optimal value of 0-gap-MSR-DU($\mathcal{M}_1, \mathcal{M}_2$). Applying Lemma 14 on the optimal solution we have,

$$\begin{aligned} \ell_0 &= 2(|V'| + \alpha(G')) \\ &\leq 2(4|V| + 7\alpha(G)) \\ &\leq 46\alpha(G). \end{aligned}$$

Hence we have the first inequality of the L-reduction. We now consider $\mathcal{S}$ a feasible solution of 0-gap-MSR-DU($\mathcal{M}_1, \mathcal{M}_2$). Using Lemma 14 we can construct an independent set X' of G' such that:

$$\|\mathcal{S}\| = 2(|V'| + |X'|),$$

and using Lemma 9, we can deduce from X' an independent set X of G of cardinality $|X| \geq |X'| - |E|$. Hence,

$$\begin{aligned} \ell_0 - \|\mathcal{S}\| &= 2(|V'| + \alpha(G') - |V'| - |X'|) \\ &= 2(\alpha(G) + |E| - |X'|) \\ &\geq 2(\alpha(G) + |E| - |X| - |E|) \\ &= 2(\alpha(G) - |X|). \end{aligned}$$

This proves the second inequality of the L-reduction from 3-MIS to 0-gap-MSR-DU. Moreover, since 3-MIS is not approximable within 95/94 [12], 0-gap-MSR-DU is not approximable within $1 + 1/(94 \times 46 \times 2) = 8649/8648$. □

4. Approximation Algorithms

4.1. Reduction to Maximum Weight Independent Set

In this section we consider the variants of Maximum Weight Independent Set on two classes of graphs: interval graphs and 2-interval graphs. An *interval graph* is a graph $G = (V, E)$, where every vertex in V is seen as an interval I of $\mathbb{R}$, and such that $(I, J) \in E$ iff (1) I and J are distinct intervals from V, and (2) $I \cap J \neq \emptyset$. A *2-interval graph* is a graph $G = (V, E)$, where every vertex in V is seen as a pair of disjoint intervals (I_1, I_2) of $\mathbb{R}$ (also called a *2-interval*), and such that $((I_1, I_2), (J_1, J_2)) \in E$ iff (1) (I_1, I_2) and (J_1, J_2) are distinct 2-intervals from V, and (2) $(I_1 \cup I_2) \cap (J_1 \cup J_2) \neq \emptyset$.

Problem: Interval-MWIS
Input: An interval graph $G = (V, E)$, a weight function $w : V \to \mathbb{R}^+$, $k \in \mathbb{R}^+$
Question: Is there an independent set X of G such that $\sum_{x \in X} w(x) \geq k$?

Problem: 2-Interval-MWIS
Input: A 2-interval graph $G = (V, E)$, a weight function $w : V \to \mathbb{R}^+$, $k \in \mathbb{R}^+$
Question: Is there an independent set X of G such that $\sum_{x \in X} w(x) \geq k$?

The problem Interval-MWIS is known to be polynomial [16]. On the other hand, 2-Interval-MWIS is APX-hard, and we know a 4-approximation for it [4].

The following construction follows the one used by Chen et al. [11] to design a 4-approximation algorithm for MSR and MSR-DU. We use this construction in order to extend the 4-approximation algorithm to the δ-gap variants of the problems (Theorem 17), and to design an exact polynomial-time algorithm for 0-gap-MSR (Theorem 18).

Given a pair of comparative maps $(\mathcal{M}_1, \mathcal{M}_2)$ and a gap δ, we construct a set of 2-intervals in the following way. First, compute the set Ω of all prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ having gap at most δ. Then, to each prestrip $\sigma \in \Omega$, assign the intervals I^1_σ and I^2_σ described below, the 2-interval $I_\sigma = (I^1_\sigma, I^2_\sigma)$, and the weight $w(I_\sigma) = |\sigma|$. We write respectively $\min(\mathrm{idx}(\sigma, \mathcal{M}))$ and $\max(\mathrm{idx}(\sigma, \mathcal{M}))$ the indices of the first and last element of σ in $\mathcal{M}$, and $l = |\mathcal{M}_1| + 1$.

$$\begin{aligned} I^1_\sigma &= [\min(\mathrm{idx}(\sigma, \mathcal{M}_1)), \max(\mathrm{idx}(\sigma, \mathcal{M}_1))], \\ I^2_\sigma &= [\min(\mathrm{idx}(\sigma, \mathcal{M}_2)) + l, \max(\mathrm{idx}(\sigma, \mathcal{M}_2)) + l], \end{aligned}$$

We denote $G_\delta(\mathcal{M}_1, \mathcal{M}_2)$ the weighted 2-interval graph with vertex set $\{I_\sigma : \sigma \in \Omega\}$ and weight w. It has the following property:

Property 16. *Let $\mathcal{S} \subseteq \Omega$ and $X = \{I_\sigma \mid \sigma \in \mathcal{S}\}$. The set X is an independent set of $G_\delta(\mathcal{M}_1, \mathcal{M}_2)$ with weight W iff $\mathcal{S}$ is feasible with total size W.*

Proof. With the definitions of I^1_σ and I^2_σ, intervals I^1_σ and I^1_τ intersect iff σ and τ overlap in $\mathcal{M}_1$, and I^2_σ and I^2_τ intersect iff σ and τ overlap in $\mathcal{M}_2$. Moreover, with the chosen value of l, I^1_σ never intersects with I^2_τ, for all prestrips σ and τ. Thus we have proved the following equivalence:

$$I_\sigma \text{ and } I_\tau \text{ intersect } \Leftrightarrow \sigma \text{ and } \tau \text{ overlap.}$$

Hence, a set of 2-intervals X is independent iff $\mathcal{S} = \{\sigma \mid I_\sigma \in X\}$ is feasible.

For the weight conservation, we have:

$$w(X) = \sum_{I_\sigma \in X} w(I_\sigma) = \sum_{\sigma \in \mathcal{S}} |\sigma| = \|\mathcal{S}\|$$

□

Theorem 17. *There exists a factor-4 approximation algorithm for δ-gap-MSR for all $\delta \geq 2$, and for δ-gap-MSR-DU for all $\delta \geq 0$.*

Proof. We use the construction described above in the following algorithm.

1. Given two comparative maps $(\mathcal{M}_1, \mathcal{M}_2)$, compute the weighted 2-interval graph $G_\delta(\mathcal{M}_1, \mathcal{M}_2)$.
2. Compute X, a 4-approximation to 2-Interval-MWIS($G_\delta(\mathcal{M}_1, \mathcal{M}_2)$).
3. Deduce a feasible set of prestrips $\mathcal{S} = \{\sigma \mid I_\sigma \in X\}$.

Property 16 yields that the total size of $\mathcal{S}$ is the weight of X, and that δ-gap-MSR-DU$(\mathcal{M}_1, \mathcal{M}_2)$ and 2-Interval-MWIS($G_\delta(\mathcal{M}_1, \mathcal{M}_2)$) have the same optimal values. Consequently, $\mathcal{S}$ is a 4-approximation of the optimal solution of δ-gap-MSR-DU$(\mathcal{M}_1, \mathcal{M}_2)$, and a 4-approximation of δ-gap-MSR$(\mathcal{M}_1, \mathcal{M}_2)$ when $\mathcal{M}_1$ and $\mathcal{M}_2$ do not contain duplicates. We have proved Theorem 17. □

Theorem 18. *There exists an exact polynomial-time algorithm for 0-gap-MSR.*

Proof. Let $(\mathcal{M}_1, \mathcal{M}_2)$ be a pair of comparative maps without duplicates. The graph $G_0(\mathcal{M}_1, \mathcal{M}_2)$ has the following property:

$$I_\sigma \text{ and } I_\tau \text{ intersect } \Leftrightarrow I^1_\sigma \cap I^1_\tau \neq \emptyset.$$

Indeed, if two prestrips overlap in $\mathcal{M}_2$, since they have gap zero in this map, they must have a common marker m appearing in $\mathcal{M}_2$. But since m can appear only once in $\mathcal{M}_1$, they also overlap in $\mathcal{M}_1$. Thus $I^2_\sigma \cap I^2_\tau \neq 0$ implies $I^1_\sigma \cap I^1_\tau \neq 0$, which suffices to prove the claim. Using this property, we can see that $G_0(\mathcal{M}_1, \mathcal{M}_2)$ is also an interval graph, with vertex set $\{I^1_\sigma \mid \sigma \in \Omega\}$. Hence, we can adapt the previous algorithm to obtain an optimal solution, and complete the proof of Theorem 18:

1. Given two comparative maps $(\mathcal{M}_1, \mathcal{M}_2)$, compute the weighted *interval* graph $G_0(\mathcal{M}_1, \mathcal{M}_2)$.
2. Compute X, an *optimal* solution to Interval-MWIS($G_0(\mathcal{M}_1, \mathcal{M}_2)$).
3. Deduce a *maximal* feasible set of prestrips $\mathcal{S} = \{\sigma \mid I_\sigma \in X\}$.

□

We can see that this proof does not use all the hypothesis. We have in fact proven that the following problem, which is more general than 0-gap-MSR, is also polynomial:

Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$, such that $\mathcal{M}_1$ has no duplicates, $\ell \in \mathbb{N}$.
Question: Is there a feasible set $\mathcal{S}$ of prestrips of $(\mathcal{M}_1, \mathcal{M}_2)$ such that the gap of each $\sigma \in \mathcal{S}$ is at most δ in $\mathcal{M}_1$ and 0 in $\mathcal{M}_2$, and $\|\mathcal{S}\| \geq \ell$?

4.2. 1.8-approximation for 1-gap-MSR

In this section, we prove the following result.

Theorem 19. *There exists a factor-1.8 approximation algorithm for 1-gap-MSR.*

Proof. Our algorithm makes uses of an exact algorithm to solve MAXIMUM WEIGHT INDEPENDENT SET (MWIS) on claw-free graphs. A *claw* is the 4-vertex graph (V, E) with $V = \{a, b, c, d\}$ and $E = \{(a, b), (a, c), (a, d)\}$. A graph is said to be claw-free if none of its induced subgraphs is isomorphic to a claw. The variant of MWIS on claw-free graphs, Claw-Free-MWIS (for which we know a polynomial algorithm, [24]), is stated as follows:

Problem: Claw-Free-MWIS
Input: A claw-free graph $G = (V, E)$, a weight function $w : V \to \mathbb{R}^+$, $k \in \mathbb{R}^+$
Question: Is there an independent set X of G such that $\sum_{x \in X} w(x) \geq k$?

Our 1.8-approximation algorithm (given in Algorithm 1) works as follows. Given two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$, compute the set Ω of all prestrips with length 2 or 3 (and gap at most 1). Longer prestrips are ignored, since they can be split into smaller ones appearing in Ω. Select a subset $V^\lambda \subseteq \Omega$ (according to some parameter λ: see the selection process described below), and create E^λ, the set of all overlapping pairs of prestrips of V^λ. The pair (V^λ, E^λ) forms a graph which is claw-free (see Lemma 20). An independent set for this graph (computable in polynomial time) yields a feasible set of prestrips V^λ_{Ind}.

The selection of V^λ amongst Ω is done as follows: given a prestrip σ of $\mathcal{M}_1$ and $\mathcal{M}_2$, take the values of $\text{idx}(\sigma, \mathcal{M}_2) - \lambda$ modulo 9. This is done by the arithmetic function π_9, which takes the values of a list modulo 9: for example, if σ has indices $(30, 32, 33)$ in $\mathcal{M}_2$, and $\lambda = 5$, then $\text{idx}(\sigma, \mathcal{M}_2) - \lambda = (25, 27, 28)$, and $\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) = (7, 0, 1)$. If the result of $\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda)$ belongs to some list (the list T in Algorithm 1), add σ to V^λ. We only need to test the 9 different values of λ to obtain 9 different feasible sets of prestrips.

Finally, Lemma 28 proves that there exists some λ for which the total size of the corresponding V^λ_{Ind} is at least $5/9^{th}$ of a maximum feasible set of prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$. Thus, Algorithm 1 is a polynomial-time algorithm giving a 1.8-approximation to 1-gap-MSR, and Theorem 19 is proved. □

Lemma 20. *For each λ, the graph (V^λ, E^λ) created by Algorithm 1 is claw-free.*

Proof. Although it is not necessary for the algorithm, we assume here, without loss of generality, that $\mathcal{M}_1$ is the identity permutation:

$$\mathcal{M}_1 = \langle 1, 2, \ldots, |\mathcal{M}_1| \rangle \ .$$

We use this to simplify somehow the notations: $\sigma = \text{idx}(\sigma, \mathcal{M}_1)$.

Algorithm 1 A factor-1.8 approximation algorithm for 1-gap-MSR

Input: Two comparative maps $\mathcal{M}_1$, $\mathcal{M}_2$ without duplicates.
$T \leftarrow \{(0,1,2),(1,2,3),(2,3,4),(0,2),(1,2),(1,3),(2,3),(2,4),$
$(5,6),(5,7),(6,7),(6,8),(7,8)\}$;
$\Omega \leftarrow$ set of all prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ of length 2 or 3, with gap at most 1;
for $\lambda \leftarrow 1$ **to** 9 **do**
  $V^\lambda \leftarrow \{\sigma \mid \sigma \in \Omega,\ \pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) \in T\}$;
  $E^\lambda \leftarrow \{(\sigma_1, \sigma_2) \mid \sigma_1, \sigma_2$ overlapping prestrips of $V^\lambda\}$;
  $w(\sigma) \leftarrow |\sigma|$ (for all $\sigma \in V^\lambda$);
  $V^\lambda_{Ind} \leftarrow$MAXIMUM WEIGHT INDEPENDENT SET of (V^λ, E^λ) with weight w;
end for
return $\max\{\|V^\lambda_{Ind}\| \mid 1 \leq \lambda \leq 9\}$;

(V^λ, E^λ) is the graph created by the algorithm. In the following we consider only prestrips contained in V^λ, and they are written σ and τ. We can see that all the elements of T have values included either in $\{0,1,2,3,4\}$ or in $\{5,6,7,8\}$. This means, that for every σ, we have

$$\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) \subseteq [0,4]$$
$$\text{or } \pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) \subseteq [5,8].$$

Thanks to the gap condition, there exists some integer k such that the indices of σ in $\mathcal{M}_2$ are all in one of the following size-5 or size-4 intervals:

$$\text{idx}(\sigma, \mathcal{M}_2) \subseteq [0 + 9k + \lambda, 4 + 9k + \lambda]$$
$$\text{or idx}(\sigma, \mathcal{M}_2) \subseteq [5 + 9k + \lambda, 8 + 9k + \lambda].$$

We write $K^\lambda(\sigma)$ the size-5 or size-4 interval of $\mathcal{M}_2$ containing σ, such that:

$$K^\lambda(\sigma) = \langle \mathcal{M}_2[0 + 9k + \lambda], \ldots, \mathcal{M}_2[4 + 9k + \lambda] \rangle$$
$$\text{or } K^\lambda(\sigma) = \langle \mathcal{M}_2[5 + 9k + \lambda], \ldots, \mathcal{M}_2[8 + 9k + \lambda] \rangle.$$

The last notations we use are for edges of E^λ: if $(\sigma, \tau) \in E^\lambda$, we write $\sigma \text{ — } \tau$. If σ and τ have a common element, we say that they *intersect*, and we write $\sigma \overset{\cap}{\text{—}} \tau$. Otherwise, they must overlap in $\mathcal{M}_1$ or $\mathcal{M}_2$ (possibly both): we write respectively $\sigma \overset{\mathcal{M}_1}{\text{—}} \tau$ or $\sigma \overset{\mathcal{M}_2}{\text{—}} \tau$.

Before proving that (V^λ, E^λ) is claw-free, we first give a series of properties over this graph.

Property 21. *If for some σ and τ, the sequences $K^\lambda(\sigma)$ and $K^\lambda(\tau)$ share a common element (we write $K^\lambda(\sigma) \cap K^\lambda(\tau) \neq \emptyset$), then $K^\lambda(\sigma) = K^\lambda(\tau)$.*

Proof. This property is obvious by the definition of K^λ, since the intervals $[0 + 9k + \lambda, 4 + 9k + \lambda]$ and $[5 + 9k + \lambda, 8 + 9k + \lambda]$ form a partition of the indices over $\mathcal{M}_2$, and this comparative map does not contain duplicates. □

Property 22. *If σ and τ overlap in $\mathcal{M}_i$ for some $i \in \{1,2\}$, without intersecting each other, then they both have gap 1 in $\mathcal{M}_i$.*

Proof. All the prestrips considered have gap at most 1. If one of them (say σ) has gap 0, then it would contain two consecutive elements in $\mathcal{M}_i$ which τ overlaps: consequently, τ would have gap at least 2, a contradiction. □

Property 23. *If $\sigma \overset{\cap}{—} \tau$ or $\sigma \overset{\mathcal{M}_2}{=} \tau$, then $K^\lambda(\sigma) = K^\lambda(\tau)$.*

Proof. Both cases imply $K^\lambda(\sigma) \cap K^\lambda(\tau) \neq \emptyset$, hence Property 21 applies. □

Property 24. *For $\sigma \neq \tau$, if $K^\lambda(\sigma) = K^\lambda(\tau)(= K)$ and $|K| = 5$, then $\sigma \overset{\cap}{—} \tau$ or $\sigma \overset{\mathcal{M}_2}{=} \tau$.*

Proof. In this proof we write $K = \langle K[0], K[1], K[2], K[3], K[4]\rangle$ (hence we have $K[i] = \mathcal{M}_2[9k+\lambda+i]$ for some integer k). This property is deduced from the list T of Algorithm 1: every element in T which is included in $[0, 4]$ either contains 2, or is $(1, 3)$. If both σ and τ are different from $\langle K[1], K[3]\rangle$, then they both contain $K[2]$ and have a non-empty intersection: $\sigma \overset{\cap}{—} \tau$. Otherwise, we can assume wlog that σ is the prestrip $\langle K[1], K[3]\rangle$, and that τ contains the element $K[2]$. If σ and τ do not intersect, then $\tau = \langle K[0], K[2]\rangle$ or $\tau = \langle K[2], K[4]\rangle$, and thus they overlap in $\mathcal{M}_2$. Consequently, $\sigma \overset{\cap}{—} \tau$ or $\sigma \overset{\mathcal{M}_2}{=} \tau$. □

Property 25. *Let σ, τ_1 and τ_2 be pairwise distinct prestrips of V^λ. If $\sigma \overset{\mathcal{M}_2}{=} \tau_1$ and either $\sigma \overset{\mathcal{M}_2}{=} \tau_2$ or $\sigma \overset{\cap}{—} \tau_2$, then $\tau_1 \overset{\cap}{—} \tau_2$ or $\tau_1 \overset{\mathcal{M}_2}{=} \tau_2$.*

Proof. Let $K = K^\lambda(\sigma)$. Using Property 23 both on (σ, τ_1) and (σ, τ_2), we have $K = K^\lambda(\tau_1)$ and $K = K^\lambda(\tau_2)$. If σ, τ_1 and τ_2 do not intersect, they correspond to 3 disjoint subsets (each of cardinality 2 or 3) of K (which is of cardinality 4 or 5): a contradiction. Since σ and τ_2 do not intersect, either $\tau_1 \overset{\cap}{—} \tau_2$, (in which case the property is proved), or $\sigma \overset{\cap}{—} \tau_2$.

If K has size 5, since $K^\lambda(\tau_1) = K^\lambda(\tau_2)$, we can directly use Property 24.

If K has size 4, then σ, τ_1 and τ_2 have length 2. Since σ and τ_1 do not intersect, and $|K| = |\sigma| + |\tau_1|$, every element of K appears either in σ or τ_1. Now τ_2 is a subsequence of K, and there is at least one element of τ_2 which does not appear in σ, so τ_2 and τ_1 intersect.

□

Property 26. *Let $|\sigma| = 2$. If $\sigma \overset{\mathcal{M}_1}{=} \tau$ then there exists $x \in \{1, \ldots, |\mathcal{M}_1|-2\}$ such that $\sigma = \langle x, x+2\rangle$, and τ contains as sub-prestrip $\langle x-1, x+1\rangle$ or $\langle x+1, x+3\rangle$.*

Proof. By Property 22, σ has gap 1 in $\mathcal{M}_1$, so there exists x such that $\sigma = \langle x, x+2\rangle$. The only two prestrips overlapping $\langle x, x+2\rangle$ without intersecting it are $\langle x-1, x+1\rangle$ and $\langle x+1, x+3\rangle$: τ must contain one of those as sub-prestrip. □

Property 27. *If $\sigma \overset{\mathcal{M}_1}{=} \tau_1$ and $\sigma \overset{\mathcal{M}_1}{=} \tau_2$ then either $\tau_1 \overset{\cap}{—} \tau_2$, or there exists x such that $\sigma = \langle x, x+2, x+4\rangle$, one of $\{\tau_1, \tau_2\}$ contains $\langle x-1, x+1\rangle$, and the other contains $\langle x+3, x+5\rangle$.*

Proof. If $|\sigma| = 2$ we can use Property 26 twice: there exists x such that $\sigma = \langle x, x+2\rangle$, and both τ_1 and τ_2 contain $x+1$: $\tau_1 \overset{\cap}{—} \tau_2$.

Suppose now that τ_1 and τ_2 do not intersect, which implies $|\sigma| = 3$. Since σ and τ_1 are overlapping in $\mathcal{M}_1$, there exists an element $x \in \sigma$ appearing between the first and the last elements of τ_1, that is, $\min \tau_1 \leq x \leq \max \tau_1$. With the same arguments as previously, τ_1 contains $\langle x-1, x+1\rangle$. There also exists $x' \in \sigma$ such that τ_2 contains $\langle x'-1, x'+1\rangle$. We can assume wlog that $x' \geq x$. Since the three prestrips do not intersect, $x' \notin \{x, x+1, x+2\}$.

Since x is not the last element in σ (because $x' > x$) and x, x' cannot be consecutive in σ (otherwise the gap would be at least 2), there exists x'' such that $x < x'' < x'$ and $\sigma = \langle x, x'', x' \rangle$. Now with the gap condition,

$$x'' \in \{x+1, x+2\} \cap \{x'-2, x'-1\},$$

and with the non intersecting condition,

$$x'' \notin \{x-1, x+1\},\ x'' \notin \{x'-1, x'+1\}.$$

Only one possibility remains:

$$x'' = x+2 = x'-2 \text{ and } \sigma = \langle x, x+2, x+4 \rangle.$$

This proves Property 27. □

We are now ready to prove Lemma 20: assume there exist four prestrips $\sigma, \tau_1, \tau_2, \tau_3$ forming a claw in (V^λ,E^λ), that is

$$\begin{array}{ccc} \sigma \text{ — } \tau_1, & \sigma \text{ — } \tau_2, & \sigma \text{ — } \tau_3, \\ (\tau_1, \tau_2) \notin E^\lambda, & (\tau_2, \tau_3) \notin E^\lambda, & (\tau_3, \tau_1) \notin E^\lambda. \end{array}$$

Let $n_{\mathcal{M}_1}$ be the number of prestrips in $\{\tau_1, \tau_2, \tau_3\}$ overlapping σ in $\mathcal{M}_1$ without intersecting it.

If $n_{\mathcal{M}_1} = 0$, then for all $j \in \{1, 2, 3\}$, either $\sigma \overset{\cap}{—} \tau_j$ or $\sigma \overset{\mathcal{M}_2}{—} \tau_j$. Hence we can use Property 23 to show that $K^\lambda(\sigma) = K^\lambda(\tau_1) = K^\lambda(\tau_2) = K^\lambda(\tau_3)$. In that case, τ_1, τ_2 and τ_3 are 3 prestrips of length 2 or 3 included in a set of size 4 or 5, so two of them must intersect, a contradicton.

The other trivial case is $n_{\mathcal{M}_1} = 3$: we use Property 27 and show that $\sigma = \langle x, x+2, x+4 \rangle$ and each one of τ_1, τ_2, τ_3 contains either $\langle x-1, x+1 \rangle$ or $\langle x+3, x+5 \rangle$. Again two of them must intersect, a contradiction.

Now we consider $n_{\mathcal{M}_1} = 2$: wlog, we can assume that $\sigma \overset{\mathcal{M}_1}{—} \tau_1$, $\sigma \overset{\mathcal{M}_1}{—} \tau_2$, and $\sigma \overset{\cap}{—} \tau_3$ or $\sigma \overset{\mathcal{M}_2}{—} \tau_3$. By Property 27, σ can be written $\sigma = \langle x, x+2, x+4 \rangle$, τ_1 contains $\langle x-1, x+1 \rangle$, and τ_2 contains $\langle x+3, x+5 \rangle$. Since σ has length 3, by definition of the vector T, see Algorithm 1, σ has gap 0 in $\mathcal{M}_2$. Hence by Property 22, the case $\sigma \overset{\mathcal{M}_2}{—} \tau_3$ is impossible, which implies $\sigma \overset{\cap}{—} \tau_3$. Moreover τ_3 does not overlap with τ_1 nor τ_2, so it contains neither x nor $x+4$. Necessarily, the common element between σ and τ_3 is $x+2$. Since $|\tau_3| \geq 2$ and since it has gap 0 or 1, it must contain an element amongst $\{x, x+1, x+3, x+4\}$ and thus overlap with τ_1 or τ_2: a contradiction.

Now, consider the last possible case, that is $n_{\mathcal{M}_1} = 1$ (wlog, assume that $\sigma \overset{\mathcal{M}_1}{—} \tau_1$). Property 25 applied to σ, τ_2 and τ_3, eliminates the cases where $\sigma \overset{\mathcal{M}_2}{—} \tau_2$ or $\sigma \overset{\mathcal{M}_2}{—} \tau_3$, since there can be no edge between τ_2 and τ_3. Hence we have $\sigma \overset{\cap}{—} \tau_2$ and $\sigma \overset{\cap}{—} \tau_3$. With Property 23, there exists a length-4 or lenght-5 sequence K such that $K = K^\lambda(\sigma) = K^\lambda(\tau_2) = K^\lambda(\tau_3)$.

If $|K| = 5$, Property 24 applies with τ_2 and τ_3, and we conclude that $\tau_2 \overset{\cap}{—} \tau_3$ or $\tau_2 \overset{\mathcal{M}_2}{—} \tau_3$, a contradiction.

Otherwise, $|K| = 4$. Since Algorithm 1 considers only length-2 prestrips in size-4 intervals $[5+9k+\lambda, 8+9k+\lambda]$, we have $|\sigma| = 2$. With Property 26, there exists an x such that $\sigma = \langle x, x+2 \rangle$, and τ_1 contains either $\langle x-1, x+1 \rangle$ or $\langle x+1, x+3 \rangle$ as sub-prestrip. We only consider the case where τ_1 contains $\langle x-1, x+1 \rangle$, the other being similar. Let $j \in \{2, 3\}$, since τ_j intersects σ without overlapping with τ_1, τ_j contains $x+2$. Hence τ_2 and τ_3 have a common element, $x+2$, a contradiction.

Altogether, we have shown that the graph (V^λ, E^λ) cannot contain any claw, and Lemma 20 is proved. □

Lemma 28. *Let O be an optimal solution of 1-gap-MSR($\mathcal{M}_1, \mathcal{M}_2$). Then Algorithm 1 provides a solution of total size at least $5\|O\|/9$.*

Proof. This proof relies on the construction of nine feasible sets of prestrips, $O^1, \ldots, O^9$, such that each prestrip in O^λ appears both in O (possibly as a sub-prestrip) and in V^λ. We also require that

$$\sum_{\lambda=1}^{9} \|O^\lambda\| \geq 5\|O\|.$$

First note that we can assume that each prestrip in O has length 2 or 3: a prestrip cannot be shorter, and we can split longer ones. The approach is as follows: we start with nine empty sets $O^1, \ldots, O^9$. Then, for each prestrip $\sigma \in O$, we enumerate the values of λ for which V^λ contains σ (or a sub-prestrip of σ), we add σ to the corresponding sets O^λ, and we measure the increase of the sum $\sum_{\lambda=1}^{9} \|O^\lambda\|$. Examples are given in Figure 6.

For a prestrip σ of length 2, we can add σ to O^λ only if $\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) \in T$. If $\text{idx}(\sigma, \mathcal{M}_2) = (x, x+1)$, then $\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda)$ takes the values (0,1), (1,2), …, (7,8), (8,0). Five of those nine pairs appear in the vector T of Algorithm 1: (1,2), (2,3), (5,6), (6,7) and (7,8). So we add σ to O^λ for 5 different values of λ: the total size $\sum_{\lambda=1}^{9} \|O^\lambda\|$ is increased by $10 = 5|\sigma|$.

The same goes for a prestrip of length 2 with indices $(x, x+2)$: it appears in V^λ for 5 different values of λ, with indices $(0,2), (1,3), (2,4), (5,7), (6,8)$. When added to the corresponding O^λ, the total size is again increased by $10 = 5|\sigma|$.

Now we consider a prestrip σ of length 3 with indices $(x, x+1, x+2)$. There are three values of λ for which $\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) \in T$ (because $(0,1,2)$, $(1,2,3)$, $(2,3,4)$ are in T): we add σ to O^λ in those three cases. We now consider the two sub-prestrips of σ: σ_1 with indices $(x, x+1)$, and σ_2 with indices $(x+1, x+2)$. Amongst the 6 remaining values of λ for which $\pi_9(\text{idx}(\sigma, \mathcal{M}_2) - \lambda) \notin T$, there are 3 for which σ_1 is selected (corresponding to pairs (5,6), (6,7) and (7,8) in T), and one more for which only σ_2 is selected (corresponding to the pair (5,6) in T). The total size of $\sum_{\lambda=1}^{9} \|O^\lambda\|$ is increased by $3\times3+4\times2 = 17$, which is greater than $5|\sigma| = 15$.

We use similar arguments for other prestrips of length 3. If $\text{idx}(\sigma, \mathcal{M}_2) = (x, x+2, x+3)$, we use pairs (0,2), (1,3), (2,4), (5,7) and (6,8) of T for σ_1, and (0,2), (5,6), (6,7) for σ_2. The quantity $\sum_{\lambda=1}^{9} \|O^\lambda\|$ is increased by 16.

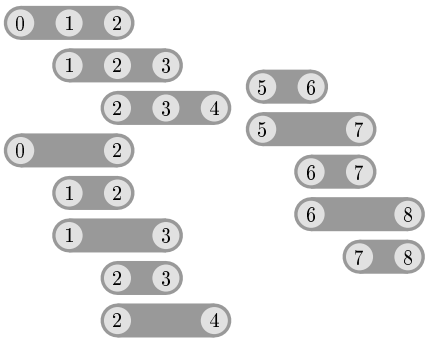
If $\text{idx}(\sigma, \mathcal{M}_2) = (x, x+1, x+3)$, we use pairs (1,2), (2,3), (5,6), (6,7) and (7,8) of T for σ_1, and (1,3), (2,4), (6,8) for σ_2. Again $\sum_{\lambda=1}^{9} \|O^\lambda\|$ is increased by 16.

Finally, if $\text{idx}(\sigma, \mathcal{M}_2) = (x, x+2, x+4)$, we use pairs (0,2), (1,3), (2,4), (5,7) and (6,8) of T for σ_1, and (0,2),(1,3), (5,7), (6,8) for σ_2. In that case, $\sum_{\lambda=1}^{9} \|O^\lambda\|$ is increased by 18.

For each strip σ of O, we have succeeded in adding σ, or sub-prestrips of σ, in several O^λ such that the total size is increased by at least $5|\sigma|$: we have 9 feasible sets (since O is feasible) satisfying the condition:

$$\sum_{\lambda=1}^{9} \|O^\lambda\| \geq 5\|O\|.$$

For each $\lambda \in \{1, \ldots, 9\}$, the prestrips of O^λ, being taken from a feasible set O, are pairwise



(a) Vector T defined in Algorithm 1

λ	$\pi_9(idx(\sigma, \mathcal{M}_2) - \lambda)$	Matching element in T
9	$(0,1)$	$\emptyset$
8	$(1,2)$	$(1,2)$
7	$(2,3)$	$(2,3)$
6	$(3,4)$	$\emptyset$
5	$(4,5)$	$\emptyset$
4	$(5,6)$	$(5,6)$
3	$(6,7)$	$(6,7)$
2	$(7,8)$	$(7,8)$
1	$(8,0)$	$\emptyset$

(b) Enumeration for idx$(\sigma, \mathcal{M}_2) = (x, x+1)$. We assume wlog that $\pi_9(x) = 0$.

λ	$\pi_9(idx(\sigma, \mathcal{M}_2) - \lambda)$	Matching element in T
9	$(0,2,3)$	$(0,2)$
8	$(1,3,4)$	$(1,3)$
7	$(2,4,5)$	$(2,4)$
6	$(3,5,6)$	$(5,6)$
5	$(4,6,7)$	$(6,7)$
4	$(5,7,8)$	$(5,7)$
3	$(6,8,0)$	$(6,8)$
2	$(7,0,1)$	$\emptyset$
1	$(8,1,2)$	$(1,2)$

(c) Enumeration for idx$(\sigma, \mathcal{M}_2) = (x, x+2, x+3)$. We assume wlog that $\pi_9(x) = 0$.

Figure 6: Enumeration of the prestrips of V^λ matching a prestrip $\sigma \in O$, for $\lambda \in \{1, \ldots, 9\}$. Note that V^λ contains all the possible prestrips whose indices taken modulo 9 are in T.

non-overlapping and form an independent set of (V^λ, E^λ). Thus we have $\|V^\lambda_{Ind}\| \geq \|O^\lambda\|$, and

$$\begin{aligned}\max\{\|V^\lambda_{Ind}\| \mid 1 \leq \lambda \leq 9\} &\geq \frac{1}{9}\sum_{\lambda=1}^{9}\|V^\lambda_{Ind}\| \\ &\geq \frac{1}{9}\sum_{\lambda=1}^{9}\|O^\lambda\| \\ &\geq \frac{5}{9}\|O\|.\end{aligned}$$

Hence the solution returned by the algorithm is at least $5\|O\|/9$. □

4.3. 2.25-approximation for 0-gap-MSR-DU

In this section, we prove the following result.

Theorem 29. *There exists a factor-2.25 approximation algorithm for 0-gap-MSR-DU.*

Proof. Algorithm 2 follows the same lines as Algorithm 1 does for 1-gap-MSR, that is it computes an exact maximum weight independent set of a subgraph (V^λ, E^λ) of the graph representing the possible strips and the overlapping relation. Due to the possibility of having duplicates in the input genomes, the graph considered can be significantly more complex, and thus Algorithm 2 uses a more selective contidition to create the set V^λ: the condition now bears on both $\text{idx}(\sigma, \mathcal{M}_1)$ and $\text{idx}(\sigma, \mathcal{M}_2)$. Lemma 30 proves that the subgraph (V^λ, E^λ) is indeed claw-free, which enables us to use a polynomial time algorithm to find a maximum weight independent set of it, which corresponds to a feasible set of strips. The approximation ratio of $2.25 = 9/4$ is given by Lemma 31. □

Algorithm 2 A factor-2.25 approximation algorithm for 0-gap-MSR-DU

Input: Two comparative maps $\mathcal{M}_1$, $\mathcal{M}_2$ (possibly with duplicates).
$T \leftarrow \{(0,1,2),(0,1),(1,2)\}$;
$\Omega \leftarrow$ set of all strips of $\mathcal{M}_1$ and $\mathcal{M}_2$ of length 2 or 3;
for $\lambda_1 \leftarrow 0$ **to** 2 **do**
  for $\lambda_2 \leftarrow 0$ **to** 2 **do**
    $\lambda \leftarrow 3\lambda_1 + \lambda_2$;
    $V^\lambda \leftarrow \{\sigma \mid \sigma \in \Omega,\ \pi_3(\text{idx}(\sigma, \mathcal{M}_1) - \lambda_1) \in T \text{ and } \pi_3(\text{idx}(\sigma, \mathcal{M}_2) - \lambda_2) \in T\}$;
    $E^\lambda \leftarrow \{(\sigma_1, \sigma_2) \mid \sigma_1, \sigma_2 \text{ intersecting strips of } V^\lambda\}$;
    $w(\sigma) \leftarrow |\sigma|$ (for all $\sigma \in V^\lambda$);
    $V^\lambda_{Ind} \leftarrow$MAXIMUM WEIGHT INDEPENDENT SET of (V^λ, E^λ) with weight w;
  end for
end for
return $\max\{\|V^\lambda_{Ind}\| \mid 0 \leq \lambda \leq 8\}$;

Lemma 30. *For each λ, the graph (V^λ, E^λ) created by Algorithm 2 is claw-free.*

Proof. Let $\lambda = 3\lambda_1 + \lambda_2$. For each $\sigma \in V^\lambda$, from the definiton of T, there exist two integers $k_1 = k_1(\sigma)$ and $k_2 = k_2(\sigma)$ such that:

$$\begin{aligned} idx(\sigma, \mathcal{M}_1) &\subseteq [3k_1 + \lambda_1, 3k_1 + \lambda_1 + 2] \\ idx(\sigma, \mathcal{M}_2) &\subseteq [3k_2 + \lambda_2, 3k_2 + \lambda_2 + 2] \end{aligned}$$

Moreover, σ contains the elements $\mathcal{M}_1[3k_1 + \lambda_1 + 1]$ and $\mathcal{M}_2[3k_2 + \lambda_2 + 1]$.

If σ and τ are two intersecting strips of V^λ, then they can intersect in $\mathcal{M}_1$ and $\mathcal{M}_2$, which leads respectively to $k_1(\tau) = k_1(\sigma)$ and $k_2(\tau) = k_2(\sigma)$. Hence if σ has at least three neighbours in (V^λ, E^λ), then two of them, written τ_1 and τ_2, are such that $k_1(\tau_1) = k_1(\tau_2)$ or $k_2(\tau_1) = k_2(\tau_2)$. So τ_1 and τ_2 share a common element, namely $\mathcal{M}_1[3k_1(\tau_1) + \lambda_1 + 1]$ or $\mathcal{M}_2[3k_2(\tau_1) + \lambda_2 + 1]$ respectively, and there is an edge between them in (V^λ, E^λ). □

Lemma 31. *If O is an optimal solution of 0-gap-MSR-DU($\mathcal{M}_1, \mathcal{M}_2$), Algorithm 2 provides a solution of total size at least $4\|O\|/9$.*

Proof. We can assume, wlog, that all strips in O have length 2 or 3. We now create nine sets of strips $O^0, \ldots, O^8$ such that each strip in O^λ appears both in V^λ and in O (possibly as a substrip), and such that

$$\sum_{\lambda=0}^{8} \|O^\lambda\| \geq 4\|O\|$$

Let σ be a strip of O, and r_1, r_2 two integers in $\{0, 1, 2\}$ such that σ starts at position r_1 modulo 3 in $\mathcal{M}_1$ and r_2 modulo 3 in $\mathcal{M}_2$.

First suppose σ has length 2. Then for $\lambda_1 \in \{r_1 - 1, r_1\}$ modulo 3 and for $\lambda_2 \in \{r_2 - 1, r_2\}$ modulo 3, we have $\pi_3(\text{idx}(\sigma, \mathcal{M}_1) - \lambda_1) \in T$ and $\pi_3(\text{idx}(\sigma, \mathcal{M}_2) - \lambda_2) \in T$, thus we have $\sigma \in V^\lambda$ for four different values of λ. We add σ to the corresponding sets O^λ, which increases the total size $\sum_{\lambda=0}^{8} \|O^\lambda\|$ by $8 = 4|\sigma|$.

Now suppose that σ has length 3. For $\lambda_1 = r_1$ and $\lambda_2 = r_2$, $\pi_3(\text{idx}(\sigma, \mathcal{M}_1) - \lambda_1) = \pi_3(\text{idx}(\sigma, \mathcal{M}_2) - \lambda_1) = (0, 1, 2)$, which is in T, thus we have $\sigma \in V^\lambda$. Moreover for $\lambda_1 \in \{r_1 - 1, r_1\}$ and $\lambda_2 \in \{r_2 - 1, r_2\}$, the beginning of σ, $\langle \sigma[1], \sigma[2] \rangle$, forms a length-2 strip appearing in V^λ. And for $\lambda_1 \in \{r_1, r_1 + 1\}$ and $\lambda_2 \in \{r_2, r_2 + 1\}$, the end of σ, $\langle \sigma[2], \sigma[3] \rangle$, forms a length-2 strip appearing in V^λ. Then for one value of λ (namely $3r_1 + r_2$), we add the length-3 strip σ to O^λ and for six other values of λ, we add one of the length-2 strips $\langle \sigma[1], \sigma[2] \rangle$ or $\langle \sigma[2], \sigma[3] \rangle$ to O^λ. Thus the total size $\sum_{\lambda=0}^{8} \|O^\lambda\|$ is increased by $3 + 6 \times 2 = 15 > 4|\sigma|$.

Thus we indeed have

$$\sum_{\lambda=0}^{8} \|O^\lambda\| \geq \sum_{\sigma \in O} 4|\sigma| = 4\|O\|$$

Hence there exists some λ such that $\|O^\lambda\| \geq \frac{4}{9}\|O\|$, and O^λ forms an independent set of (V^λ, E^λ) since a set of strips or substrips of O is necessarily feasible. Thus the size of the corresponding V^λ_{Ind} is at least $\|O^\lambda\|$ and Algorithm 2 gives a solution of size at least $\frac{4}{9}\|O\|$: it is indeed a $\frac{9}{4} = 2.25$-approximation. □

5. Conclusion

In this paper, we have introduced and studied δ-gap-MSR and δ-gap-MSR-DU, two variants of the MAXIMAL STRIP RECOVERY problem. These problems take into account biologically

sustained restrictions in the search for synteny blocks, namely the fact that two consecutive markers of a synteny block cannot appear at arbitrarily large distance from one another in a comparative map. We have proved that δ-gap-MSR and δ-gap-MSR-DU are APX-complete problems, with two exceptions: 0-gap-MSR is polynomial, and 1-gap-MSR may "only" be NP-complete. We also have given exact or approximation algorithms for all the variants: exact for 0-gap-MSR, 1.8-approximation for 1-gap-MSR, 2.25-approximation for 0-gap-MSR-DU, 4-approximation for other variants.

References

[1] J. Akiyama and V. Chvátal. A short proof of the linear arboricity for cubic graphs. *The bulletin of liberal arts & sciences, Nippon Medical School*, 2:1–3, 1982.

[2] P. Alimonti and V. Kann. Hardness of approximating problems on cubic graphs. In G. C. Bongiovanni, D. P. Bovet, and G. Di Battista, editors, *CIAC*, volume 1203 of *LNCS*, pages 288–298. Springer, 1997.

[3] K. Appel and W. Haken. Every planar map is four colorable. *Bulletin of the American Mathematics Society*, pages 82–711, 1976.

[4] R. Bar-Yehuda, M.M. Halldórsson, J. Naor, H. Shachnai, and I. Shapira. Scheduling split intervals. *SIAM J. Comput.*, 36(1):1–15, 2006.

[5] T.C. Biedl, G. Kant, and M. Kaufmann. On triangulating planar graphs under the four-connectivity constraint. In *Proc. 4th Scandinavian Workshop on Algorithm Theory (SWAT'94)*, volume 824 of *LNCS*, pages 83–94. Springer, 1994.

[6] N. L. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph theory, 1736-1936*. Oxford, Clarendon Press, 1976.

[7] L. Bulteau, G. Fertin, M. Jiang, and I. Rusu. Tractability and approximability of maximal strip recovery. In R. Giancarlo and G. Manzini, editors, *CPM*, volume 6661 of *Lecture Notes in Computer Science*, pages 336–349. Springer, 2011.

[8] L. Bulteau, G. Fertin, and I. Rusu. Maximal strip recovery problem with gaps: Hardness and approximation algorithms. In Dong et al. [14], pages 710–719.

[9] J. Chen and S. B. Cooper, editors. *Theory and Applications of Models of Computation, 6th Annual Conference, TAMC 2009, Changsha, China, May 18-22, 2009. Proceedings*, volume 5532 of *LNCS*. Springer, 2009.

[10] Z. Chen, B. Fu, M. Jiang, and B. Zhu. On recovering syntenic blocks from comparative maps. In B. Yang, D. Du, and C. A. Wang, editors, *COCOA*, volume 5165 of *LNCS*, pages 319–327. Springer, 2008.

[11] Z. Chen, B. Fu, M. Jiang, and B. Zhu. On recovering syntenic blocks from comparative maps. *Journal of Combinatorial Optimization*, 18(3):307–318, 2009.

[12] M Chlebík and J Chlebíková. Complexity of approximating bounded variants of optimization problems. *Theoretical Computer Science*, 354(3):320 – 338, 2006.

[13] V. Choi, C. Zheng, Q. Zhu, and D. Sankoff. Algorithms for the extraction of synteny blocks from comparative maps. In R. Giancarlo and S. Hannenhalli, editors, *WABI*, volume 4645 of *LNCS*, pages 277–288. Springer, 2007.

[14] Y. Dong, D.-Z. Du, and O. H. Ibarra, editors. *Algorithms and Computation, 20th International Symposium, ISAAC 2009, Honolulu, Hawaii, USA, December 16-18, 2009. Proceedings*, volume 5878 of *Lecture Notes in Computer Science*. Springer, 2009.

[15] A. Goldstein, P. Kolman, and J. Zheng. Minimum common string partition problem: Hardness and approximations. *Electr. J. Comb.*, 12, 2005.

[16] M. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York, 1980.

[17] H. Jiang, Z. Li, G. Lin, L. Wang, and B. Zhu. Exact and approximation algorithms for the complementary maximal strip recovery problem. *Journal of Combinatorial Optimization*, pages 1–14. to appear, 10.1007/s10878-010-9366-y.

[18] H. Jiang and B. Zhu. Weak kernels. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:5, 2010.

[19] M. Jiang. Inapproximability of maximal strip recovery. In Dong et al. [14], pages 616–625.

[20] M. Jiang. Inapproximability of maximal strip recovery: Ii. In D-T. Lee, D. Chen, and S. Ying, editors, *FAW*, volume 6213 of *Lecture Notes in Computer Science*, pages 53–64. Springer, 2010.

[21] M. Jiang. Inapproximability of maximal strip recovery. *Theoretical Computer Science*, 412(29):3759–3774, 2011.

[22] Z. Li, R. Goebel, L. Wang, and G. Lin. An improved approximation algorithm for the complementary maximal strip recovery problem. In M. J. Atallah, X.-Y. Li, and B. Zhu, editors, *FAW-AAIM*, volume 6681 of *Lecture Notes in Computer Science*, pages 46–57. Springer, 2011.

[23] G. Lin, R. Goebel, Z. Li, and L. Wang. An improved approximation algorithm for the complementary maximal strip recovery problem. *Journal of Computer and System Sciences*, 78(3):720 – 730, 2012.

[24] G. J. Minty. On maximal independent sets of vertices in claw-free graphs. *J. Comb. Theory, Ser. B*, 28(3):284–304, 1980.

[25] J. Misra and D. Gries. A constructive proof of Vizing's theorem. *Inf. Process. Lett.*, 41(3):131–133, 1992.

[26] C. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *J. Comput. Syst. Sci.*, 43(3):425–440, 1991.

[27] Neil Robertson, Daniel P. Sanders, Paul D. Seymour, and Robin Thomas. Efficiently four-coloring planar graphs. In Gary L. Miller, editor, *STOC*, pages 571–575. ACM, 1996.

[28] D. Sankoff, C. Zheng, A. Muñoz, Z. Yang, Z. Adam, R. Warren, V. Choi, and Q. Zhu. Issues in the reconstruction of gene order evolution. *J. Comput. Sci. Technol.*, 25(1):10–25, 2009.

[29] L. Wang and B. Zhu. On the tractability of maximal strip recovery. In Chen and Cooper [9], pages 400–409.

[30] L. Wang and B. Zhu. On the tractability of maximal strip recovery. *Journal of Computational Biology*, 17(7):907–914, 2010.

[31] C. Zheng, Q. Zhu, and D. Sankoff. Removing noise and ambiguities from comparative maps in rearrangement analysis. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 4(4):515–522, 2007.

[32] B. Zhu. Approximability and fixed-parameter tractability for the exemplar genomic distance problems. In Chen and Cooper [9], pages 71–80.

[33] B. Zhu. Efficient exact and approximate algorithms for the complement of maximal strip recovery. In B. Chen, editor, *AAIM*, volume 6124 of *Lecture Notes in Computer Science*, pages 325–333. Springer, 2010.