Maximal Strip Recovery Problem with Gaps: Hardness and Approximation Algorithms

Laurent Bulteau, Guillaume Fertin, Irena Rusu

To cite this version:

Laurent Bulteau, Guillaume Fertin, Irena Rusu. Maximal Strip Recovery Problem with Gaps: Hardness and Approximation Algorithms. ISAAC 2009 (20th International Symposium on Algorithms and Computation), 2009, Hawaii, United States. Springer-Verlag, 5878, pp.710-719, 2009, Lecture Notes in Computer Science (LNCS). <hal-00425145>

HAL Id: hal-00425145

https://hal.archives-ouvertes.fr/hal-00425145

Submitted on 20 Oct 2009

Maximal Strip Recovery Problem with Gaps: Hardness and Approximation Algorithms

Laurent Bulteau[1,2], Guillaume Fertin[2] and Irena Rusu[2]

[1] École Normale Supérieure, 45 rue d'Ulm, 75000 Paris, France
[2] Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR CNRS 6241
Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3 - France
Laurent.Bulteau@ens.fr, {Guillaume.Fertin,Irena.Rusu}@univ-nantes.fr

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-overlapping 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. Depending on the nature of the comparative maps (i.e., with or without duplicates), we show that δ-gap-MSR is NP-complete for any $\delta \geq 1$, and even APX-hard for any $\delta \geq 2$. We also provide two approximation algorithms, with ratio 1.8 for $\delta = 1$, and ratio 4 for $\delta \geq 2$.

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 hold 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. [9] 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 *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. [9] 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 are defined as *strips*, which are contiguous sequences of *at least two* markers that occur on each genome either in the same order, or in reverse order and with a reversed sign. Zheng et al [9] and Choi et al. [4] propose two heuristics to solve the MSR problem. Chen et al. [3] devise a 4-approximation algorithm for it, propose its extension, called MSR-d, to an arbitrary number $d \geq 2$ of genomes and show that MSR-3 is NP-complete. The NP-completeness of MSR (or equivalently MSR-2) is a result obtained by Wang et al. [8], who also propose FPT algorithms for MSR-d (with arbitrary d) and MSR-DU, the variant of MSR where duplicated markers are allowed in the maps and in different synteny blocks.

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. For the NP-complete or APX-hard cases, we provide two approximation algorithms, whose approximation ratios are given in Table 2.

Table 1. Complexity of variants of MSR.

Problem	Without duplicates	With duplicates (-DU variant)
0-gap-MSR	P (Section 4.2)	?
1-gap-MSR	NP-complete (Section 3.1)	NP-complete (Section 3.1)
δ-gap-MSR ($\delta \geq 2$)	APX-complete (Section 3.2)	APX-complete (Section 3.2)
MSR	NP-complete [8]	NP-complete [8]

Table 2. Best approximation ratios of variants of MSR.

Problem	Without duplicates	With duplicates (-DU variant)
0-gap-MSR	-	4 (Section 4.2)
1-gap-MSR	1.8 (Section 4.1)	4 (Section 4.2)
δ-gap-MSR ($\delta \geq 2$)	4 (Section 4.2)	4 (Section 4.2)
MSR	4 [3]	4 [3]

The organization of the paper is as follows. In Section 2, we introduce some notations, and we define formally MSR, MSR-DU, δ-gap-MSR and δ-gap-MSR-DU. We prove in Section 3 the hardness results (NP-completeness for $\delta = 1$ in Section 3.1, APX-completeness for $\delta \geq 2$ in Section 3.2). We then give approximation algorithms in Section 4: a 1.8-approximation for 1-gap-MSR in Section 4.1, and a general 4-approximation in Section 4.2. Due to space constraints, most of the proofs are omitted from this paper.

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. A marker may appear several times in a comparative map, possibly with different orientations: in this case, we say that the comparative map $\mathcal{M}$ has *duplicates* (the presence of duplicates is useful if we do not want to distinguish paralogs in the comparative map). 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 $\max\{i_{k+1} - i_k - 1 \ : \ 1 \leq k < h\}$, its *length* $|\sigma|$ is h. Two subsequences σ and τ are *non-overlapping* in $\mathcal{M}$ if one appears strictly before the other (i.e., if the last element of $idx(\sigma, \mathcal{M})$ is strictly smaller than the first element of $idx(\tau, \mathcal{M})$ or vice-versa). 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 $\mathcal{M}_1$ such that either σ or its reversed opposite is a subsequence of $\mathcal{M}_2$, and such that the markers in σ are pairwise different. 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 finally 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 [9]) and MSR-DU [3] are defined, in their decision formulation, as follows:

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

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

The idea behind both those problems is that, if we find a set of compatible 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 comparative maps can be partitioned into common strips, i.e. we have decomposed both genomes into synteny blocks with the same set of blocks in both genomes. Heuristics for the first problem have been given in [9,4]. They have been improved in [3] into a 4-approximation algorithm. Finally, those problems have been proved NP-complete in [8], where an FPT algorithm is also provided.

The variant we introduce, δ-gap-MSR, takes into account the fact that it is unlikely that long sequences of markers can appear only from noise and errors. If a large number of elements are inserted between two consecutive elements of a prestrip (thus, if it has a large gap), then they are not errors, and the prestrip cannot be considered a synteny block of the original genomes. Thus we define the following two problems:

Problem: δ-gap-MSR
Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$ without duplicates, $k \in \mathbb{N}$.
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 $||\mathcal{S}|| \geq k$?

Problem: δ-gap-MSR-DU
Input: Two comparative maps $\mathcal{M}_1$ and $\mathcal{M}_2$ (possibly with duplicates), $k \in \mathbb{N}$.

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 $||\mathcal{S}|| \geq k$?

With the gap constraint we introduce, we keep only prestrips which are nearly contiguous, while tolerating some noise in the input data. Note that those problems are defined for uni-chromosomal genomes. However, algorithms can easily be adapted to handle multi-chromosomal instances.

3 Hardness Results

3.1 NP-hardness of 1-gap-MSR

In this section, we prove the following theorem.

Theorem 1 *1-gap-MSR and 1-gap-MSR-DU are NP-hard.*

Note that we need to consider only 1-gap-MSR (without duplicates) since NP-hardness of 1-gap-MSR-DU directly follows from NP-hardness of 1-gap-MSR.

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.

Problem: 3-colored-MIS
Input: A cubic graph $G = (V, E)$, 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 2 *3-colored-MIS is NP-hard.*

Starting from any instance of 3-colored-MIS, we construct two comparative maps as follows. First, we assign a list of 4 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 integers $x_{uv}^1, \ldots, x_{uv}^{10}$ to each edge $(u, v) \in E^{\text{C}}$, in such a way that no integer appears in two different lists. We will also use peg markers: written with the symbol $\times$, they are integers appearing only once, either in $\mathcal{M}_1$ or in $\mathcal{M}_2$ (and thus cannot belong to any prestrip).

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 $(u, v) \in E^{\text{A}}$ such that $u < v$, add $\langle y_u^{\text{A1}}, y_v^{\text{A1}}, y_u^{\text{A2}}, y_v^{\text{A2}}, \times, \times\rangle$ to $\mathcal{M}_1$.
2. For all $(u, v) \in E^{\text{B}}$ such that $u < v$, add $\langle y_u^{\text{B1}}, y_v^{\text{B1}}, y_u^{\text{B2}}, y_v^{\text{B2}}, \times, \times\rangle$ to $\mathcal{M}_2$.
3. For all $(u, v) \in E^{\text{C}}$ 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:

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

Property 3 *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* $4n + 2\alpha(G)$.

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

3.2 δ-gap-MSR and δ-gap-MSR-DU are APX-hard

In this section, we prove the following theorem.

Theorem 4 *δ-gap-MSR and δ-gap-MSR-DU are APX-hard for any $\delta \geq 2$.*

As in the previous section, we note that we need to consider only δ-gap-MSR (without duplicates) since APX-hardness of δ-gap-MSR-DU directly follows from APX-hardness of δ-gap-MSR. For this, we use an L-reduction [7] from the variant of Maximum Independent Set restricted to cubic graphs, that we call 3-MIS here. Note that the L-reduction 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 here.

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

It is proved in [1] that 3-MIS is APX-hard. 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 5 *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 $k' \leq 4$.*

The first step of the reduction is to compute a partition of E into two classes E^{B} and E^{W} according to Lemma 5. 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 $V(B_i) = \{a_h : 1 \leq h \leq k\}$ and let $(a_h, a_{h+1}) \in E^{\mathrm{B}}$ for $1 \leq h < k$; we construct the following sequence:

$$I_i = \left\langle x_{a_1}, \times, \times^{\delta-2}, x_{a_2}, x'_{a_1}, \times^{\delta-2}, \ldots, x_{a_h}, x'_{a_{h-1}}, \times^{\delta-2}, \ldots, \times, x'_{a_k} \right\rangle$$

where $\times^l$ represents l consecutive peg markers. The full comparative map $\mathcal{M}_1$ is given by $\mathcal{M}_1 = \left\langle I_1, \times^{\delta+1}, I_2, \times^{\delta+1}, \ldots \right\rangle$.

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:

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

If W_j is of order three (resp. two), we remove the extra elements from J_j, i.e. $J_j = \langle x_a, x_b, x'_a, -x'_c, x'_b, -x_c \rangle$ (resp. $J_j = \langle x_a, x_b, x'_a, x'_b \rangle$). Finally, $\mathcal{M}_2$ is created in the same way as $\mathcal{M}_1$: $\mathcal{M}_2 = \left\langle J_1, \times^{\delta+1}, J_2, \times^{\delta+1}, \ldots \right\rangle$.

Lemma 6 *The set Ω of the prestrips of $\mathcal{M}_1$ and $\mathcal{M}_2$ with gap less than or equal to δ is exactly $\{\Phi(a) : a \in V\}$. 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 $(a, b) \in E^W$.*

The consequence of this lemma 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 δ-gap-MSR: indeed, Φ transforms an independent set of size l into a feasible set of prestrips with gap δ of total size $2l$, and Φ^{-1} does the reverse operation. So δ-gap-MSR, like 3-MIS, is APX-hard for $\delta \geq 2$.

4 Approximation Algorithms

4.1 1.8-approximation for 1-gap-MSR

In this section, we present an approximation algorithm for 1-gap-MSR. Our result is the following.

Theorem 7 *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 (which is known to be in P, [6]), is stated as follows:

Problem: Claw-Free-MWIS
Input: A claw-free graph $G = (V, E)$, a weight function $w : V \rightarrow \mathbb{R}^+$, $k \in \mathbb{R}^+$
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 8). 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 $\mathrm{idx}(\sigma, \mathcal{M}_2) - \lambda$ 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 $\mathrm{idx}(\sigma, \mathcal{M}_2) - \lambda = (25, 27, 28)$, and $\pi_9(\mathrm{idx}(\sigma, \mathcal{M}_2) - \lambda) = (7, 9, 1)$. If the result of $\pi_9(\mathrm{idx}(\sigma, \mathcal{M}_2) - \lambda)$ 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 9 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 7 is proved. □

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

```
Input: Two comparative maps $\mathcal{M}_1$, $\mathcal{M}_2$ without duplicates.
$T \leftarrow \{(1,2,3),(2,3,4),(3,4,5),(1,3),(2,3),(2,4),(3,4),(3,5),$
        $(6,7),(6,8),(7,8),(7,9),(8,9)\}$;
$\Omega \leftarrow$ 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
    $V^\lambda \leftarrow \{\sigma \ : \ \sigma \in \Omega, \ \pi_9(\mathrm{idx}(\sigma, \mathcal{M}_2) - \lambda) \in T\}$;
    $E^\lambda \leftarrow \{(\sigma_1, \sigma_2) \ : \ \sigma_1, \sigma_2$ overlapping prestrips of $V^\lambda\}$;
    $w(\sigma) \leftarrow |\sigma|$ (for all $\sigma \in V^\lambda$);
    $V^\lambda_{Ind} \leftarrow$MAXIMUM WEIGHT INDEPENDENT SET of $(V^\lambda, E^\lambda)$ with weight $w$;
end for
return $\max\{||V^\lambda_{Ind}|| \ : \ 1 \le \lambda \le 9\}$;
```

Lemma 8 *For each λ, the graph (V^λ, E^λ) created by Algorithm 1 is claw-free.*

Lemma 9 *If $\mathcal{O}$ is a feasible set of prestrips of $\mathcal{M}_1$, $\mathcal{M}_2$ with gap 1, Algorithm 1 provides a solution of total size at least $5||\mathcal{O}||/9$.*

4.2 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 [5]. On the other hand, 2-Interval-MWIS is APX-hard, and we know a 4-approximation for it [2].

Theorem 10 *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. In this proof, we describe a reduction from δ-gap-MSR to 2-Interval-MWIS. Given a pair of comparative maps and a maximal 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 following 2-interval (where l is $|\mathcal{M}_1| + 1$):

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

with the weight:

$$w(I_\sigma) = |\sigma| \,.$$

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 11 *The set $\{I_\sigma \ : \ \sigma \in \mathcal{S}\}$ is an independent set of $G_\delta(\mathcal{M}_1, \mathcal{M}_2)$ with weight W iff $\mathcal{S}$ is a feasible subset of Ω with total size W.*

The 4-approximation algorithm for δ-gap-MSR and δ-gap-MSR-DU is the following (adapted from the 4-approximation algorithm for MSR and MSR-DU [3]):

1. Compute the weighted 2-interval graph $G_\delta(\mathcal{M}_1, \mathcal{M}_2)$ as described above.
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 $\mathcal{S} = \{\sigma \ : \ I_\sigma \in X\}$.

Property 11 tells us 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: so $\mathcal{S}$ is indeed a 4-approximation of the optimal solution of δ-gap-MSR-DU$(\mathcal{M}_1, \mathcal{M}_2)$. We have proved Theorem 10. □

Theorem 12 *There exists an exact polynomial-time algorithm for 0-gap-MSR.*

Proof. We consider the case where $\mathcal{M}_1$ has no duplicates and the maximum gap is 0 (we only consider strips instead of prestrips): this is the case for instances of 0-gap-MSR.

We use the same reduction as for Theorem 10, with the difference that now, $G_0(\mathcal{M}_1, \mathcal{M}_2)$ is in fact an interval graph. It can be seen by considering intervals

$$I'_\sigma = [\min(\mathrm{idx}(\sigma, \mathcal{M}_1)), \max(\mathrm{idx}(\sigma, \mathcal{M}_1))].$$

We no longer need to consider the interval coming from $\mathcal{M}_2$ for the following reason. If two strips overlap in $\mathcal{M}_2$, since they have gap zero, 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_σ and I_τ intersect iff I'_σ and I'_τ intersect: $G_0(\mathcal{M}_1, \mathcal{M}_2)$ can thus be seen as an interval graph. Hence, we can adapt the previous algorithm to obtain an optimal solution, and complete the proof of Theorem 12:

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

□

References

1. 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.
2. R. Bar-Yehuda, M.M. Halldórsson, J. Naor, H. Shachnai, and I. Shapira. Scheduling split intervals. *SIAM J. Comput.*, 36(1):1–15, 2006.
3. 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.
4. 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.
5. M. Golumbic. *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, New York, 1980.
6. G. J. Minty. On maximal independent sets of vertices in claw-free graphs. *J. Comb. Theory, Ser. B*, 28(3):284–304, 1980.
7. C. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *J. Comput. Syst. Sci.*, 43(3):425–440, 1991.
8. L. Wang and B. Zhu. On the tractability of maximal strip recovery. In J. Chen and S. B. Cooper, editors, *TAMC*, volume 5532 of *LNCS*, pages 400–409. Springer, 2009.
9. 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.