On the Approximability of Comparing Genomes with Duplicates

Sébastien Angibaud, Guillaume Fertin, Irena Rusu

On the Approximability of Comparing Genomes with Duplicates

Sébastien Angibaud, Guillaume Fertin, Irena Rusu

Laboratoire d'Informatique de Nantes-Atlantique (LINA), FRE CNRS 2729
Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3 - France

— ComBi —

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

RESEARCH REPORT

N° Numéro du rapport

June 2007

UNIVERSITÉ DE NANTES

ECOLE DES MINES DE NANTES

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Sébastien Angibaud, Guillaume Fertin, Irena Rusu

On the Approximability of Comparing Genomes with Duplicates

20 p.

Les rapports de recherche du Laboratoire d'Informatique de Nantes-Atlantique sont disponibles aux formats PostScript® et PDF® à l'URL :
http://www.sciences.univ-nantes.fr/lina/Vie/RR/rapports.html

Research reports from the Laboratoire d'Informatique de Nantes-Atlantique are available in PostScript® and PDF® formats at the URL:
http://www.sciences.univ-nantes.fr/lina/Vie/RR/rapports.html

On the Approximability of Comparing Genomes with Duplicates

Sébastien Angibaud, Guillaume Fertin, Irena Rusu

sebastien.angibaud@univ-nantes.fr,guillaume.fertin@univ-nantes.fr,irena.rusu@univ-nantes.fr

Abstract

A central problem in comparative genomics consists in computing a (dis-)similarity measure between two genomes, e.g. in order to construct a phylogenetic tree. A large number of such measures has been proposed in the recent past: *number of reversals*, *number of breakpoints*, *number of common* or *conserved intervals*, *SAD* etc. In their initial definitions, all these measures suppose that genomes contain no duplicates. However, we now know that genes can be duplicated within the same genome. One possible approach to overcome this difficulty is to establish a one-to-one correspondence (i.e. a matching) between genes of both genomes, where the correspondence is chosen in order to optimize the studied measure. Then, after a gene relabeling according to this matching and a deletion of the unmatched signed genes, two genomes without duplicates are obtained and the measure can be computed.
In this paper, we are interested in three measures (*number of breakpoints*, *number of common intervals* and *number of conserved intervals*) and three models of matching (*exemplar* model, *maximum matching* model and *non maximum matching* model). We prove that, for each model and each measure, computing a matching between two genomes that optimizes the measure is **APX-Hard**. We show that this result remains true even for two genomes G_1 and G_2 such that G_1 contains no duplicates and no gene of G_2 appears more than twice. Therefore, our results extend those of [5, 6, 7]. Finally, we propose a 4-approximation algorithm for a measure closely related to the *number of breakpoints*, the *number of adjacencies*, under the *maximum matching* model, in the case where genomes contain the same number of duplications of each gene.

Additional Key Words and Phrases: genome rearrangement, APX-Hardness, duplicates, breakpoints, adjacencies, common intervals, conserved intervals, approximation algorithm

1 Introduction and Preliminaries

In comparative genomics, computing a measure of (dis-)similarity between two genomes is a central problem; such a measure can be used for instance to construct phylogenetic trees. The measures defined so far fall into two categories: the first one contains distances, for which we count the number of operations needed to transform a genome into another (see for instance *edit distance* [12] or *number of reversals* [3]). The second one contains (dis-)similarity measures based on the genome structure, such as *number of breakpoints* [5], *conserved intervals distance* [4], *number of common intervals* [6], *SAD* and *MAD* [15] etc.

When genomes contain no duplicates, most measures can be computed in polynomial time. However, assuming that genomes contain no duplicates is too limited, as it has been shown that a great number of duplicates exists in some genomes. For example, in [11], authors estimate that fifteen percent of genes are duplicated in the human genome. A possible approach to overcome this difficulty is to specify a one-to-one correspondence (i.e. a matching) between genes of both genomes and to remove the remaining genes, thus obtaining two genomes with identical gene composition and no duplicates. This matching is chosen in order to optimize the studied measure. Three models achieving this correspondence have been proposed : *exemplar* model [14], *maximum matching* model [16] and *non maximum matching* model [2].

Let $\mathcal{F}$ be a set of *genes*, where each gene is represented by an integer. A genome G is a sequence of signed elements (*signed genes*) from $\mathcal{F}$. Let $occ(g, G)$ be the number of occurrences of a gene g in a genome G and let $occ(G) = \max\{occ(g, G) | g \text{ is present in } G\}$. Two genomes G_1 and G_2 are called *balanced* iff, for each gene g, we have $occ(g, G_1) = occ(g, G_2)$. Denote η_G the size of genome G. Let $G[p]$, $1 \leqslant p \leqslant \eta_G$, be the signed gene that occurs at position p on genome G. For any signed gene g, let $\overline{g}$ be the signed gene having the opposite sign. Given a genome G without duplicates and two signed genes a, b such that a is located before b, let $G[a, b]$ be the set of genes located between genes a and b in G. We also note $[a, b]_{G_1}$ the substring (i.e. the sequence of consecutive elements) of G_1 starting by a and finishing by b.
For example, consider the set $\mathcal{F} = \{1, 2, 3, 4, 5, 6\}$ and the genome $G_1 = +1+2+3+4+5-1-2+6-2$. Then, $occ(1, G_1) = 2$, $occ(G_1) = 3$, $G_1[5] = +5$ and $\overline{G_1[5]} = -5$. Now, consider the genome $G_2 = +3-2+6+4-1+5$ without duplicates. We have $G_2[+6, -1] = \{1, 4, 6\}$ and $[+6, -1]_{G_2} = (+6, +4, -1)$.

Breakpoints, adjacencies, common and conserved intervals. Let us now define the four measures we will study in this paper. Let G_1, G_2 be two genomes without duplicates and with the same gene composition.

Breakpoint and Adjacency. Let (a, b) be a pair of consecutive signed genes in G_1. We say that the pair (a, b) induces a *breakpoint* of (G_1, G_2) if neither (a, b) nor $(\overline{b}, \overline{a})$ is a pair of consecutive signed genes in G_2. Otherwise, we say that (a, b) induces an *adjacency* of (G_1, G_2). For example, when $G_1 = +1 + 2 + 3 + 4 + 5$ and $G_2 = +5 - 4 - 3 + 2 + 1$, the pair $(2, 3)$ in G_1 induces a breakpoint of (G_1, G_2) while $(3, 4)$ in G_1 induces an adjacency of (G_1, G_2). We note $B(G_1, G_2)$ the number of breakpoints that exist between G_1 and G_2.

Common interval. A *common interval* of (G_1, G_2) is a substring of G_1 such that G_2 contains a permutation of this substring (not taking signs into account). For example, consider $G_1 = +1 + 2 + 3 + 4 + 5$ and $G_2 = +2 - 4 + 3 + 5 + 1$. The substring $[+3, +5]_{G_1}$ is a common interval of (G_1, G_2). We notice that the notion of common interval does not consider the sign of genes.

Conserved interval. Consider two signed genes a and b of G_1 such that a precedes b, where the precedence relation is large in the sense that, possibly, $a = b$. The substring $[a, b]_{G_1}$ is called a *conserved interval* of (G_1, G_2) if it satisfies the two following properties: first, either a precedes b or $\overline{b}$ precedes $\overline{a}$ in G_2; second, the set of genes located between genes a and b in G_2 is equal to $G_1[a, b]$. For example, if $G_1 = +1 + 2 + 3 + 4 + 5$ and $G_2 = -5 - 4 + 3 - 2 + 1$, the substring $[+2, +5]_{G_1}$ is a conserved interval of (G_1, G_2).
Note that a conserved interval is actually a common interval, but with additional restrictions on its extremities. An interval of a genome G which is either of length one (i.e. a singleton) or the whole genome G is called a *trivial interval*.

Dealing with duplicates in genomes. When genomes contain duplicates, we cannot directly compute the measures defined previously. A solution consists in finding a one-to-one correspondence (i.e. a matching) between duplicated genes of G_1 and G_2, and use this correspondence to rename genes of G_1 and G_2 and to delete the unmatched signed genes in order to obtain two genomes G'_1 and G'_2 such that G'_2 is a permutation of G'_1; thus, the measure computation becomes possible. In this paper, we will focus on three models of matching : the *exemplar*, *maximum matching* and *non maximum matching* models.

- The *exemplar model* [14]: for each gene g, we keep in the matching only one occurrence of g in G_1 and in G_2, and we remove all the other occurrences. Hence, we obtain two genomes G_1^E and G_2^E without duplicates. The pair (G_1^E, G_2^E) is called an *exemplarization* of (G_1, G_2).
- The *maximum matching model* [16]: in this case, we keep in the matching the maximum number of genes in both genomes. More precisely, we look for a one-to-one correspondence between genes of G_1 and G_2 that, for each gene g, matches exactly $min(occ(g, G_1), occ(g, G_2))$ occurrences. After this operation, we delete each unmatched signed genes. The pair (G_1^E, G_2^E) obtained by this operation is called a *maximum matching* of (G_1, G_2).
- The *non maximum matching model* [2]: this model is an intermediate between the *exemplar* and the *maximum matching* models. In this new model, for each gene family g, we keep an arbitrary number k_g, $1 \leqslant k_g \leqslant min(occ(g, G_1), occ(g, G_2))$, of genes in G_1^E and in G_2^E. We call the pair (G_1^E, G_2^E) a *non maximum matching* of (G_1, G_2).

Problems studied in this paper. Consider two genomes G_1 and G_2 with duplicates.

Let *EBD* (resp. MBD, $NMBD$) be the problem which consists in finding an exemplarization (G_1^E, G_2^E) of (G_1, G_2) (resp. maximum matching, non maximum matching) that minimizes the number of breakpoints between G_1^E and G_2^E. EBD is proved to be **NP-Complete** even if $occ(G_1) = 1$ and $occ(G_2) = 2$ [5]. Some inapproximability results are given: it has been proved in [7] that, in the general case, EBD cannot be approximated within a factor $c \log n$, where $c > 0$ is a constant, and cannot be approximated within a factor 1.36 when $occ(G_1) = occ(G_2) = 2$. Likewise, the problem consisting in deciding if there exists an exemplarization (G_1^E, G_2^E) of (G_1, G_2) such that there is no breakpoint between G_1^E and G_2^E is **NP-Complete** even when $occ(G_1) = occ(G_2) = 3$. Moreover, for two balanced genomes G_1 and G_2 such that $k = occ(G_1) = occ(G_2)$, several approximation algorithms for MBD are given. Those approximation algorithms admit respectively a ratio of 1.1037 when $k = 2$ [9], 4 when $k = 3$ [9] and $4k$ in the general case [10].

Let $EComI$ (resp. $MComI$, $NMComI$) be the problem which consists in finding an exemplarization (G_1^E, G_2^E) of (G_1, G_2) (resp. maximum matching, non maximum matching) such that the number of common intervals of (G_1^E, G_2^E) is maximized. $EComI$ and $MComI$ are proved to be **NP-Complete** even if $occ(G_1) = 1$ and $occ(G_2) = 2$ in [6].

Let $EConsI$ (resp. $MConsI$, $NMConsI$) be the problem which consists in finding an exemplarization (G_1^E, G_2^E) of (G_1, G_2) (resp. maximum matching, non maximum matching) such that the number of conserved intervals of (G_1^E, G_2^E) is maximized. In [4], Blin and Rizzi have studied the problem of computing a *distance* built on the number of conserved intervals. This distance differs from the *number of conserved intervals* we study in this paper, mainly in the sense that (i) it can be applied to two *sets* of genomes (as opposed to two genomes in our case), and (ii) the distance between two identical genomes of length n is equal to 0 (as opposed to $\frac{n(n+1)}{2}$ in our case). Blin and Rizzi [4] proved that finding the minimum distance is **NP-Complete**, under both the *exemplar* and *maximum matching* models. A closer analysis of their proof shows that it can be easily adapted to prove that *EConsI* and *MConsI* are NP-complete, even in the case $occ(G_1) = 1$.

We can conclude from these results that the MBD, $NMBD$, $NMComI$ and $NMConsI$ problems are also **NP-Complete**, since when one genome contains no duplicates, *exemplar*, *maximum matching* and *non maximum matching* models are equivalent.

In this paper, we study the approximation complexity of three measure computations: *number of breakpoints*, *number of conserved intervals* and *number of common intervals*. In Section 2 and 3, we prove the **APX**-Harness of $EComI$, $EConsI$ and EBD even when applied on genomes G_1 and G_2 such that

$occ(G_1) = 1$ and $occ(G_2) = 2$, which induce the **APX**-Harness under the other models. These results extend those of papers [5, 6, 7]. In Section 4, we consider the *maximum matching* model and a fourth measure, the *number of adjacencies* for which we give a 4-approximation algorithm when genomes are balanced. Hence, we are able to provide an approximation algorithm with *constant* ratio, even when the number of occurrences of genes is unbounded.

2 $EComI$ and $EConsI$ are APX-Hard

In this section, we prove the following theorem:

Theorem 1 *$EComI$ and $EConsI$ are **APX-Hard** even when applied to genomes G_1, G_2 such that $occ(G_1) = 1$ and $occ(G_2) = 2$.*

We prove Theorem 1 by using an *L-reduction* [13] from the $Minimum\ Vertex\ Cover$ problem on cubic graphs, denoted by VC_3. Let $G = (V, E)$ be a cubic graph, i.e. for all $v \in V, degree(v) = 3$. A set of vertices $V' \subseteq V$ is called a *vertex cover* of G if for each edge $e \in E$, there exists a vertex $v \in V'$ such that e is incident to v. The problem VC_3 is defined as follows:

Problem: VC_3
Input: A cubic graph $G = (V, E)$, an integer k.
Question: Does there exist a vertex cover V' of G such that $|V'| \leqslant k$?

VC_3 was proved **APX-Complete** in [1].

2.1 Reduction

Let (G, k) be an instance of VC_3, where $G = (V, E)$ is a cubic graph with $V = \{v_1 \dots v_n\}$ and $E = \{e_1 \dots e_m\}$. Consider the transformation R which associates to the graph G two genomes G_1 and G_2 in the following way, where each gene has a positive sign.

$$G_1 = b_1, b_2, \dots, b_m, x, a_1, C_1, f_1, a_2, C_2, f_2, \dots, a_n, C_n, f_n, y, b_{m+n}, b_{m+n-1}, \dots, b_{m+1} \quad (1)$$
$$G_2 = y, a_1, D_1, f_1, b_{m+1}, a_2, D_2, f_2, b_{m+2}, \dots, b_{m+n-1}, a_n, D_n, f_n, b_{m+n}, x \quad (2)$$

with :

- for each i, $1 \leqslant i \leqslant n, a_i = 6i - 5$, $f_i = 6i$ and $C_i = (a_i + 1), (a_i + 2), (a_i + 3), (a_i + 4)$
- for each i, $1 \leqslant i \leqslant n + m, b_i = 6n + i$
- $x = 7n + m + 1$ and $y = 7n + m + 2$
- for each i, $1 \leqslant i \leqslant n, D_i = a_i + 3, b_{j_i}, a_i + 1, b_{k_i}, a_i + 4, b_{l_i}, a_i + 2$ where e_{j_i}, e_{k_i} and e_{l_i} are the edges which are incident to v_i in G, with $j_i < k_i < l_i$.

In the following, genes b_i, $1 \leqslant i \leqslant m$, are called *markers*. There is no duplicated gene in G_1 and the markers are the only duplicated genes in G_2; these genes occur twice in G_2. Hence, we have $occ(G_1) = 1$ and $occ(G_2) = 2$.

To illustrate the reduction, consider the cubic graph G of Figure 1. From G, we construct the following genomes G_1 and G_2:

$$\overbrace{25}^{b_1}\ \overbrace{26}^{b_2}\ \overbrace{27}^{b_3}\ \overbrace{28}^{b_4}\ \overbrace{29}^{b_5}\ \overbrace{30}^{b_6}\ \overbrace{35}^{x}\ 1\ \overbrace{2\,3\,4\,5}^{C_1}\ 6\ 7\ \overbrace{8\,9\,10\,11}^{C_2}\ 12\ 13\ \overbrace{14\,15\,16\,17}^{C_3}\ 18\ 19\ \overbrace{20\,21\,22\,23}^{C_4}\ 24\ \overbrace{36}^{y}\ \overbrace{34}^{b_{10}}\ \overbrace{33}^{b_9}\ \overbrace{32}^{b_8}\ \overbrace{31}^{b_7}$$

$$\underbrace{36}_{y}\ 1\ \underbrace{4\,25\,2\,26\,5\,27\,3}_{D_1}\ 6\ \underbrace{31}_{b_7}\ 7\ \underbrace{10\,25\,8\,28\,11\,29\,9}_{D_2}\ 12\ \underbrace{32}_{b_8}\ 13\ \underbrace{16\,26\,14\,17\,30\,15}_{D_3}\ 18\ \underbrace{33}_{b_9}\ 19\ \underbrace{22\,27\,20\,29\,23\,30\,21}_{D_4}\ 24\ \underbrace{34}_{b_{10}}\ \underbrace{35}_{x}$$

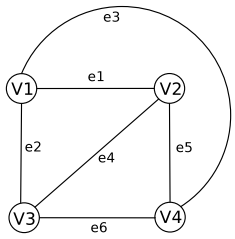

Figure 1: The cubic graph G.

2.2 Preliminary results

In order to prove Theorem 1, we first give four intermediate lemmas. Due to space constraints, the proofs are not given in the paper but can be found in appendix. In the following, a common interval for the $EComI$ problem or a conserved interval for $EConsI$ is called a $robust\ interval$.

Lemma 1 *For any exemplarization (G_1, G_2^E) of (G_1, G_2), the non trivial robust intervals of (G_1, G_2^E) are necessarily contained in some sequence $a_i C_i f_i$ of G_1 ($1 \leqslant i \leqslant n$).*

Proof. We start by proving the lemma for common intervals, and we will then extend it to conserved intervals. First, we prove that for any exemplarization (G_1, G_2^E) of (G_1, G_2), each common interval I such that $|I| \geqslant 2$ and which contains x (resp. y) also contains y (resp. x), which implies that I covers the whole genome. Suppose there exists a common interval I_x such that $|I_x| \geqslant 2$ and I_x contains x. Let PI_x be the permutation of I_x in G_2^E. The interval I_x must contain either b_m or a_1. Let us detail each of the two cases:

(a) If I_x contains b_m, then PI_x contains b_m too. Notice that there is some i, $1 \leqslant i \leqslant n$, such that b_m belongs to D_i in G_2^E. Then PI_x contains all genes between D_i and x in G_2^E. Thus PI_x contains b_{m+n}. Consequently, I_x contains b_{m+n} and it also contains y.

(b) If I_x contains a_1, then PI_x contains a_1 too. Then PI_x contains all genes between a_1 and x. Thus PI_x contains b_{m+n}. Hence, I_x contains b_{m+n} and then it also contains y.

Now, suppose that I_y is a common interval such that $|I_y| \geqslant 2$ and I_y contains y. Let PI_y be the permutation of I_y on G_2^E. The interval I_y must contain either b_{m+n} or f_n. Let us detail each of the two cases:

(a) If I_y contains b_{m+n}, then PI_y contains b_{m+n} too. Thus PI_y contains all genes between b_{m+n} and y. Hence PI_y contains all the sequences D_i, $1 \leqslant i \leqslant n$. In particular, PI_y contains all the markers and consequently I_y must contain x.

(b) If I_y contains f_n, then PI_y contains f_n too. Then PI_y contains all genes between f_n and y. In particular, PI_y contains b_{m+n-1} and then it contains I_y too. Hence, I_y also contains b_{m+n}, similarly to the previous case. Thus I_y contains x.

We conclude that each non singleton common interval containing either x or y necessarily contains both x and y. Therefore, and by construction of G_2, there is only one such interval, that is G_1 itself. Hence, any non trivial common interval is necessarily, in G_1, either strictly on the left of x, between x and y, or strictly on the right of y. Let us analyze these different cases:

- Let I be a non trivial common interval situated strictly on the left of x in G_1. Thus I is a sequence of at least two consecutive markers. Since in any exemplarization (G_1, G_2^E) of (G_1, G_2), every marker has neighboring genes which are not markers, this contradicts the fact that I is a common interval.
- Let I be a non trivial common interval situated strictly on the right of y in G_1. Then I is a substring of $b_{m+n}, \ldots, b_{m+1}$ containing at least two genes. In any exemplarization (G_1, G_2^E) of (G_1, G_2), for each pair (b_{m+i}, b_{m+i+1}), with $1 \leqslant i < n$, we have $a_{i+1} \in G_2^E[b_{m+i}, b_{m+i+1}]$. This contradicts the fact that I is strictly on the right of y in G_1.

- Let I be a non trivial common interval lying between x and y in G_1. For any exemplarization (G_1, G_2^E) of (G_1, G_2), a common interval cannot contain, in G_1, both f_i and a_{i+1} for some i, $1 \leqslant i \leqslant n-1$ (since b_{m+1} is situated between f_i and a_{i+1} in G_2^E and on the right of x in G_1). Hence, a non trivial common interval of (G_1, G_2^E) is included in some sequence $a_iC_if_i$ in G_1, $1 \leqslant i \leqslant n$. This proves the lemma for common intervals.
  By definition, any conserved interval is necessarily a common interval. So, a non trivial conserved interval of (G_1, G_2^E) is included in some sequence $a_iC_if_i$ in G_1, $1 \leqslant i \leqslant n$. The lemma is proved. □

Lemma 2 *Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and $i \in [1 \dots n]$. Let Δ_i be a substring of $[a_i + 3, a_i + 2]_{G_2^E}$ that does not contain any marker. If $|\Delta_i| \in \{2, 3\}$, then there is no robust interval I of (G_1, G_2^E) such that Δ_i is a permutation of I.*

Proof. First, we prove that there is no permutation I of Δ_i such that I is a common interval of (G_1, G_2^E). Next, we show that there is no permutation I of Δ_i such that I is a conserved interval. By Lemma 1, we know that a non trivial common interval of (G_1, G_2^E) is a substring of some sequence $a_iC_if_i$, $1 \leqslant i \leqslant n$. This substring contains only consecutive integers. Therefore, if there exists a permutation I of Δ_i such that I is a common interval of (G_1, G_2^E), then Δ_i must be a permutation of consecutive integers. If $|\Delta_i| = 2$, we have $\Delta_i = (p, q)$ where p and q are not consecutive integers and if $|\Delta_i| = 3$, then we have $\Delta_i = (a_i+3, a_i+1, a_i+4)$ or $\Delta_i = (a_i+1, a_i+4, a_i+2)$. In these three cases, Δ_i is not a permutation of consecutive integers. Hence, there is no permutation I of Δ_i such that I is a common interval of (G_1, G_2^E). Moreover, any conserved interval is also a common interval. Thus, there is no permutation I of Δ_i such that I is a conserved interval of (G_1, G_2^E). □

For more clarity, let us now introduce some notations. Given a graph $G = (V, E)$, let $VC = \{v_{i_1}, v_{i_2} \dots v_{i_k}\}$ be a vertex cover of G. Let $R(G) = (G_1, G_2)$ be the pair of genomes defined by the construction described in (1) and (2). Now, let F be the function which associates to VC, G_1 and G_2 an exemplarization $F(VC)$ of (G_1, G_2) as follows. In G_2, all the markers are removed from the sequences D_i for all $i \neq i_1, i_2 \dots i_k$. Next, for each marker which is still present twice, one of its occurrences is arbitrarily removed. Since in G_2 only markers are duplicated, we conclude that $F(VC)$ is an exemplarization of (G_1, G_2).

Given a cubic graph G and genomes G_1 and G_2 obtained by the transformation $R(G)$, let us define the function S which associates to an exemplarization (G_1, G_2^E) of (G_1, G_2) the vertex cover VC of G defined as follows: $VC = \{v_i | 1 \leqslant i \leqslant n \wedge \exists j \in \{1 \dots m\}, b_j \in G_2^E[a_i, f_i]\}$. In other words, we keep in VC the vertices v_i of G for which there exists some gene b_j such that b_j is in $G_2^E[a_i, f_i]$. We now prove that VC is a vertex cover. Consider an edge e_p of G. By construction of G_1 and G_2, there exists some i, $1 \leqslant i \leqslant n$, such that gene b_p is located between a_i and f_i in G_2^E. The presence of gene b_p between a_i and f_i implies that vertex v_i belongs to VC. We conclude that each edge is incident to at least one vertex of VC.

Let W be the function defined on $\{EConsI, EComI\}$ by $W(pb) = 1$ if $pb = EConsI$ and $W(pb) = 4$ if $pb = EComI$. Let $OPT_P(A)$ be the optimum result of an instance A for an optimization problem P, $P \in \{EcomI, EConsI, VC_3\}$.

We define the function T which associates to a problem $pb \in \{EConsI, EComI\}$ and a cubic graph G, the number of robust trivial intervals of an exemplarization of both genomes G_1 and G_2 obtained by $R(G)$ for the problem pb. Let n and m be respectively the number of vertices and the number of edges of G. We have $T(EConsI, G) = 7n + m + 2$ and $T(EComI, G) = 7n + m + 3$. Indeed, for $EComI$, there are $7n + m + 2$ singletons and we also need to consider the whole genome.

Lemma 3 *Let $pb \in \{EcomI, EConsI\}$. Let G be a cubic graph and $R(G) = (G_1, G_2)$. Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and let i, $1 \leqslant i \leqslant n$. Then only two cases can occur:*

1. *Either in G_2^E, all the markers from D_i were removed, and in this case, there are exactly $W(pb)$ non trivial robust intervals involving D_i.*
2. *Or in G_2^E, at least one marker was kept in D_i, and in this case, there is no non trivial robust interval involving D_i.*

Proof. We first prove the lemma for the $EComI$ problem and then we extend it to $EConsI$. Lemma 1 implies that each non trivial common interval I of (G_1, G_2^E) is contained in some substring of $a_iC_if_i$, $1 \leqslant i \leqslant n$. So, the permutation of I on G_2^E is contained in a substring of $a_iD_if_i$, $1 \leqslant i \leqslant n$. Consider i, $1 \leqslant i \leqslant n$, suppose that all the markers from D_i are removed on G_2^E. Thus, $a_iC_if_i$, C_i, a_iC_i and C_if_i are common intervals of (G_1, G_2^E). Let us now show that there is no other non trivial common interval involving D_i. Let Δ_i be a substring of $[a_i + 3, a_i + 2]_{G_2^E}$ such that $|\Delta_i| \in \{2, 3\}$. By Lemma 2, we know that Δ_i is not a common interval. The remaining intervals are $(a_i, a_i + 3)$, $(a_i, a_i + 3, a_i + 1)$, $(a_i, a_i + 3, a_i + 1, a_i + 4)$, $(a_i + 1, a_i + 4, a_i + 2, f_i)$, $(a_i + 4, a_i + 2, f_i)$ and $(a_i + 2, f_i)$. By construction, none of them can be a common interval, because none of them is a permutation of consecutive integers. Hence, there are only four non trivial common intervals involving D_i in G_2^E. Among these four common intervals, only $a_iC_if_i$ is a conserved interval too. In the end, if all the markers are removed from D_i, there are exactly four non trivial common intervals and one non trivial conserved interval involving D_i. So, given a problem $pb \in \{EcomI, EconsI\}$, there are exactly $W(pb)$ non trivial robust intervals involving D_i.

Now, suppose that at least one marker of D_i is kept in G_2^E. Lemma 1 shows that each non trivial common interval I of (G_1, G_2^E) is contained in some substring of $a_iC_if_i$, $1 \leqslant i \leqslant n$. Since no marker is present in a sequence $a_iC_if_i$, we deduce that there does not exist any trivial common interval containing a marker. So, a non trivial common interval involving D_i only, must contain a substring Δ_i of $[a_i + 3, a_i + 2]_{G_2^E}$ such that Δ_i contains no marker. Since no marker is an extremity of $[a_i + 3, a_i + 2]_{G_2^E}$, we have $|\Delta_i| \leqslant 3$. By Lemma 2, we know that Δ_i is not a common interval. The remaining intervals to be considered are the intervals $a_i\Delta_i$ and Δ_if_i. By construction of $a_iC_if_i$, these intervals are not common intervals (the absence of gene $a_i + 2$ for $a_i\Delta_i$ and of gene $a_i + 3$ for Δ_if_i implies that these intervals are not a permutation of consecutive integers). Hence, these intervals cannot be conserved intervals either. □

Lemma 4 *Let $pb \in \{EcomI, EConsI\}$. Let $G = (V, E)$ be a cubic graph with $V = \{v_1 \dots v_n\}$ and $E = \{e_1 \dots e_m\}$ and let G_1, G_2 be the two genomes obtained by $R(G)$.*

1. *Let VC be a vertex cover of G and denote $k = |VC|$. Then the exemplarization $F(VC)$ of (G_1, G_2) has at least $N = W(pb) \cdot n + T(pb, G) - W(pb) \cdot k$ robust intervals.*
2. *Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and let VC' be the vertex cover of G obtained by $S(G_1, G_2^E)$. Then $|VC'| = \frac{W(pb)\cdot n + T(pb,G) - N}{W(pb)}$, where N is the number of robust intervals of (G_1, G_2^E).*

Proof. 1. Let $pb \in \{EcomI, EConsI\}$. Let G be a cubic graph and let G_1 and G_2 be the two genomes obtained by $R(G)$. Suppose there is a vertex cover VC of G and denote $k = |VC|$. Let (G_1, G_2^E) be the exemplarization of (G_1, G_2) obtained by $F(VC)$. By construction, we have at least $(n - k)$ substrings D_i in G_2^E for which all the markers are removed. By Lemma 3, we know that each of these substrings implies the existence of $W(pb)$ non trivial robust intervals. So, we have at least $W(pb)(n - k)$ non trivial robust intervals. Moreover, it is easy to see that the number of trivial robust intervals of (G_1, G_2^E) is exactly $T(pb, G)$. Thus, we have at least $N = W(pb) \cdot n + T(pb, G) - W(pb) \cdot k$ robust intervals of (G_1, G_2^E).

2. Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and $n - j$ be the number of sequences D_i, $1 \leqslant i \leqslant n$, for which all markers have been deleted in G_2^E. Then, by Lemmas 1 and 3, the number of robust intervals of (G_1, G_2^E) is equal to $N = W(pb) \cdot n + T(pb, G) - W(pb) \cdot j$. Let VC' be the vertex cover obtained by $S(G_1, G_2^E)$. Each marker has one occurrence in G_2^E and these occurrences lie in j sequences D_i. So, by definition of S, we conclude that $|VC'| = j = \frac{W(pb)\cdot n + T(pb,G) - N}{W(pb)}$. □

2.3 Main result

Let us first define the notion of *L-reduction* [13]: let A and B be two optimization problems and c_A, c_B be respectively their cost functions. An *L-reduction* from problem A to problem B is a pair of polynomial functions R and S with the following properties:

(a) If x is an instance of A, then $R(x)$ is an instance of B ;

(b) If x is an instance of A and y is a solution of $R(x)$, then $S(y)$ is a solution of x;

(c) If x is an instance of A whose optimum is $OPT(x)$, then $R(x)$ is an instance of B such that $OPT(R(x)) \leqslant \alpha.OPT(x)$, where α is a positive constant ;

(d) If s is a solution of $R(x)$, then:
$|OPT(x) - c_A(S(s))| \leqslant \beta|OPT(R(x)) - c_B(s)|$ where β is a positive constant.

We prove Theorem 1 by showing that the pair (R, S) defined previously is an *L-reduction* from VC_3 to $EConsI$ and from VC_3 to $EComI$. First note that properties (a) and (b) are obviously satisfied by R and S.

Consider $pb \in \{EcomI, EConsI\}$. Let $G = (V, E)$ be a cubic graph with n vertices and m edges. We now prove properties (c) and (d). Consider the genomes G_1 and G_2 obtained by $R(G)$. First, we need to prove that there exists $\alpha \geqslant 0$ such that $OPT_{pb}(G_1, G_2) \leqslant \alpha.OPT_{VC_3}(G)$.

Since G is cubic, we have the following properties:

$$n \geqslant 4 \tag{3}$$

$$m = \frac{1}{2}\sum_{i=1}^{n} degree(v_i) = \frac{3n}{2} \tag{4}$$

$$OPT_{VC_3}(G) \geqslant \frac{m}{3} = \frac{n}{2} \tag{5}$$

To explain property (5), remark that, in a cubic graph G with n vertices and m edges, each vertex covers three edges. Thus, a set of k vertices covers at most $3k$ edges. Hence, any vertex cover of G must contain at least $\frac{m}{3}$ vertices.
By Lemma 3, we know that sequences of the form $a_iC_if_i$, $1 \leqslant i \leqslant n$ contain either zero or $W(pb)$ non trivial robust intervals. By Lemma 1, there are no other non trivial robust intervals. So, we have the following inequality: $OPT_{pb}(G_1, G_2) \leqslant \underbrace{T(pb, G)}_{trivial\ robust\ intervals} + W(pb) \cdot n$.

If $pb = EComI$, we have:

$$\begin{aligned} OPT_{EComI}(G_1, G_2) &\leqslant 7n + m + 3 + 4n \\ OPT_{EComI}(G_1, G_2) &\leqslant \frac{27n}{2} \ \ by\ (3)\ and\ (4) \end{aligned} \tag{6}$$

And if $pb = EConsI$, we have :

$$\begin{aligned} OPT_{EConsI}(G_1, G_2) &\leqslant 7n + m + 2 + n \\ OPT_{EConsI}(G_1, G_2) &\leqslant \frac{21n}{2} \ \ by\ (3)\ and\ (4) \end{aligned} \tag{7}$$

Altogether, by (5), (6) and (7), we prove property (c) with $\alpha = 27$.

Now, let us prove property (d). Let $VC = \{v_{i_1}, v_{i_2} \dots v_{i_P}\}$ be a minimum vertex cover of G. Denote $P = OPT_{VC_3}(G) = |VC|$ and let G_1 and G_2 be the genomes obtained by $R(G)$. Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and let k' be the number of robust intervals of (G_1, G_2^E). Finally, let VC' be the vertex cover of G such that $VC' = S(G_1, G_2^E)$. We need to find a positive constant β such that $|P - |VC'|| \leqslant \beta|OPT_{pb}(G_1, G_2) - k'|$.

For $pb \in \{EcomI, EConsI\}$, let N_{pb} be the number of robust intervals between the two genomes obtained by $F(VC)$. By the first property of Lemma 4, we have

$$OPT_{pb}(G_1, G_2) \geqslant N_{pb} \geqslant W(pb) \cdot n + T(pb, G) - W(pb) \cdot P$$

By the second property of Lemma 4, we have $|VC'| = \frac{W(pb)\cdot n + T(pb,G) - k'}{W(pb)}$.
Recall that $OPT_{pb}(G_1, G_2) \geqslant W(pb) \cdot n + T(pb, G) - W(pb) \cdot P$. So, it is sufficient to prove $\exists \beta \geqslant 0, |P - |VC'|| \leqslant \beta|W(pb) \cdot n + T(pb, G) - W(pb) \cdot P - k'|$. Since $P \leqslant |VC'|$, we have

$$|P - |VC'|| = |VC'| - P = \frac{W(pb) \cdot n + T(pb, G) - k'}{W(pb)} - P = \frac{1}{W(pb)}(W(pb)\cdot n + T(pb, G) - W(pb)\cdot P - k')$$

So $\beta = 1$ is sufficient in both cases, since $W(EComI) = 4$ and $W(EConsI) = 1$, which implies $\frac{1}{W(pb)} \leqslant 1$. Altogether, we then have $|OPT_{VC_3}(G) - |VC'|| \leqslant 1 \cdot |OPT_{pb}(G_1, G_2) - k'|$.

We proved that the reduction (R, S) is an *L-reduction*. This implies that for two genomes G_1 and G_2, both problems $EConsI$ and $EComI$ are **APX-Hard** even if $occ(G_1) = 1$ and $occ(G_2) = 2$. Theorem 1 is proved. □

We extend in Corollary 1 our results for the *maximum matching* and *non maximum matching* models.

Corollary 1 *$MComI$, $NMComI$, $MConsI$ and $NMConsI$ are **APX-Hard** even when applied to genomes G_1, G_2 such that $occ(G_1) = 1$ and $occ(G_2) = 2$.*

Proof. The *maximum matching* and *non maximum matching* models are identical to the *exemplar* model when one genome contains no duplicates. Hence, the **APX**-Hardness result for $EComI$ (resp. $EConsI$) also holds for $MComI$ and $NMComI$ (resp. $MConsI$ and $NMConsI$). □

3 EBD is APX-Hard

In this section, we prove the following theorem:

Theorem 2 *EBD is **APX-Hard** even when applied to genomes G_1, G_2 such that $occ(G_1) = 1$ and $occ(G_2) = 2$.*

To prove Theorem 2, we use an *L-Reduction* from the VC_3 problem to the EBD problem. Let $G = (V, E)$ be a cubic graph with $V = \{v_1 \dots v_n\}$ and $E = \{e_1 \dots e_m\}$. For each i, $1 \leqslant i \leqslant n$, let e_{f_i}, e_{g_i} and e_{h_i} be the three edges which are incident to v_i in G with $f_i < g_i < h_i$. Let R' be the polynomial transformation which associates to G the following genomes G_1 and G_2, where each gene has a positive sign:
$G_1 = a_0\ a_1\ b_1\ a_2\ b_2 \dots a_n\ b_n\ c_1\ d_1\ c_2\ d_2 \dots c_m\ d_m\ c_{m+1}$
$G_2 = a_0\ a_n\ d_{f_n}\ d_{g_n}\ d_{h_n}\ b_n \dots a_2\ d_{f_2}\ d_{g_2}\ d_{h_2}\ b_2\ a_1\ d_{f_1}\ d_{g_1}\ d_{h_1}\ b_1\ c_1\ c_2 \dots c_m\ c_{m+1}$
with :

- $a_0 = 0$, and for each i, $0 \leqslant i \leqslant n$, $a_i = i$ and $b_i = n + i$
- $c_{m+1} = 2n + m$, and for each i, $1 \leqslant i \leqslant m + 1$, $c_i = 2n + i$ and $d_i = 2n + m + 1 + i$

We remark that there is no duplication in G_1, so $occ(G_1) = 1$. In G_2, only the genes d_i, $1 \leqslant i \leqslant m$, are duplicated and occur twice. Thus $occ(G_2) = 2$.

Let G be a cubic graph and VC be a vertex cover of G. Let G_1 and G_2 be the genomes obtained by $R'(G)$. We define F' to be the polynomial transformation which associates to VC, G_1 and G_2 the exemplarization (G_1, G_2^E) of (G_1, G_2) as follows. For each i such that $v_i \notin VC$, we remove from G_2 the genes d_{f_i}, d_{g_i} and d_{h_i}. Then, for each $1 \leqslant j \leqslant m$ such that d_j still has two occurrences in G_2, we arbitrarily remove one of these occurrences in order to obtain the genome G_2^E. Hence, (G_1, G_2^E) is an exemplarization of (G_1, G_2).

Given a cubic graph G, we construct G_1 and G_2 by the transformation $R'(G)$. Given an exemplarization (G_1, G_2^E) of (G_1, G_2), let S' be the polynomial transformation which associates to (G_1, G_2^E) the set $VC = \{v_i | 1 \leqslant i \leqslant n, a_i$ and b_i are not consecutive in $G_2^E\}$. We claim that VC is a vertex cover of G. Indeed, let e_p, $1 \leqslant p \leqslant m$, be an edge of G. Genome G_2^E contains one occurrence of gene d_p since G_2^E is an exemplarization of G_2. By construction, there exists i, $1 \leqslant i \leqslant n$, such that d_p is in $G_2^E[a_i, b_i]$ and such that e_p is incident to v_i. The presence of d_p in $G_2^E[a_i, b_i]$ implies that vertex v_i belongs to VC. We can conclude that each edge of G is incident to at least one vertex of VC.

Lemmas 5 and 6 below are used to prove that (R', S') is an *L-Reduction* from the VC_3 problem to the EBD problem. Let $G = (V, E)$ be a cubic graph with $V = \{v_1, v_2 \dots v_n\}$ and $E = \{e_1, e_2 \dots e_m\}$ and let us construct (G_1, G_2) by the transformation $R'(G)$.

Lemma 5 *Let VC be a vertex cover of G and (G_1, G_2^E) the exemplarization given by $F'(VC)$. Then $|VC| = k \Rightarrow B(G_1, G_2^E) \leqslant n + 2m + k + 1$, where $B(G_1, G_2^E)$ is the number of breakpoints between G_1 and G_2^E.*

Proof. Suppose $|VC| = k$. Let us list the breakpoints between genomes G_1 and G_2^E obtained by $F'(R'(G), VC)$. The pairs (b_i, a_{i+1}), $1 \leqslant i \leqslant n-1$, and (b_n, c_1) induce one breakpoint each. For all $1 \leqslant i \leqslant m$, each pair of the form (c_i, d_i) (resp. (d_i, c_{i+1})) induces one breakpoint. For all $1 \leqslant i \leqslant n$ such that $v_i \in VC$, (a_i, b_i) induces at most one breakpoint. Finally, the pair (a_0, a_1) induces one breakpoint. Thus there are at most $n + 2m + k + 1$ breakpoints of (G_1, G_2^E). □

Lemma 6 *Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and VC' be the vertex cover of G obtained by $S'(G_1, G_2^E)$. We have $B(G_1, G_2^E) = k' \Rightarrow |VC'| = k' - n - 2m - 1$.*

Proof. Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and VC' be the vertex cover obtained by $S'(G_1, G_2^E)$. Suppose $B(G_1, G_2^E) = k'$. For any exemplarization (G_1, G_2^E) of (G_1, G_2), the following breakpoints always occur: the pair (a_0, a_1); for each i, $1 \leqslant i \leqslant m$, each pair (c_i, d_i) and (d_i, c_{i+1}); for each i, $1 \leqslant i \leqslant n-1$, the pair (b_i, a_{i+1}); the pair (b_n, c_1). Thus, we have at least $n + 2m + 1$ breakpoints. The other possible breakpoints are induced by pairs of the form of (a_i, b_i). Since we have $B(G_1, G_2^E) = k'$, there are exactly $k' - n - 2m - 1$ such breakpoints. By construction of VC', the cardinality of VC' is equal to the number of breakpoints induced by pairs of the form (a_i, b_i). So, we have: $|VC'| = k' - n - 2m - 1$. □

Lemma 7 *The inequality $OPT_{EBD}(G_1, G_2) \leqslant 12 \cdot OPT_{VC_3}(G)$ holds.*

Proof. For a cubic graph G with n vertices and m edges, we have $2m = 3n$ (see (4)) and $OPT_{VC_3}(G) \geqslant \frac{n}{2}$ (see (5)). By construction of the genomes G_1 and G_2, any exemplarization of (G_1, G_2) contains $2n + 2m + 1$ genes in each genome. Thus, we have $OPT_{EBD}(G_1, G_2) \leqslant 2n + 2m + 1 \leqslant 6n$. Hence, we conclude that $OPT_{EBD}(G_1, G_2) \leqslant 12 \cdot OPT_{VC_3}(G)$. □

Lemma 8 *Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and let VC' be the vertex cover of G obtained by $S'(G_1, G_2^E)$. Then, we have $|OPT_{VC_3}(G) - |VC'|| \leqslant |OPT_{EBD}(G_1, G_2) - B(G_1, G_2^E)|$*

Proof. Let (G_1, G_2^E) be an exemplarization of (G_1, G_2) and VC' be the vertex cover of G obtained by $S'(G_1, G_2^E)$. Let VC be a vertex cover of G such that $|VC| = OPT_{VC_3}(G)$. We know that $OPT_{VC_3}(G) \leqslant |VC'|$ and $OPT_{EBD}(G_1, G_2) \leqslant B(G_1, G_2^E)$. So, it is sufficient to prove $|VC'| - OPT_{VC_3}(G) \leqslant B(G_1, G_2^E) - OPT_{EBD}(G_1, G_2)$.

By Lemma 5, we have $B(F'(VC)) \leqslant n + 2m + 1 + OPT_{VC_3}$, which implies $OPT_{EBD}(G_1, G_2) \leqslant B(F'(VC)) \leqslant n + 2m + 1 + OPT_{VC_3}$, that is

$$B(G_1, G_2^E) - OPT_{EBD}(G_1, G_2) \geqslant B(G_1, G_2^E) - n - 2m - 1 - OPT_{VC_3}(G) \tag{8}$$

By Lemma 6, we have: $|VC'| = B(G_1, G_2^E) - n - 2m - 1$ which implies

$$|VC'| - OPT_{VC_3}(G) = B(G_1, G_2^E) - n - 2m - 1 - OPT_{VC_3}(G) \tag{9}$$

Finally, by (8) and (9), we get $|VC'| - OPT_{VC_3} \leqslant B(G_1, G_2^E) - OPT_{EBD}(G_1, G_2)$. □

Lemmas 7 and 8 prove that the pair (R', S') is an *L-reduction* from VC_3 to EBD. Hence, EBD is **APX-Hard** even if $occ(G_1) = 1$ and $occ(G_2) = 2$, and Theorem 2 is proved. We extend in Corollary 2 our results for the *maximum matching* and *non maximum matching* models.

Corollary 2 *The $MEBD$ and $NMEBD$ problems are **APX-Hard** even when applied to genomes G_1, G_2 such that $occ(G_1) = 1$ and $occ(G_2) = 2$.*

Proof. The *maximum matching* and *non maximum matching* models are identical to the *exemplar* model when one genome contains no duplicates. Hence, the **APX**-Hardness result for EBD also holds for MBD and $NMBD$. □

4 Approximating the number of adjacencies

For two balanced genomes G_1 and G_2, several approximation algorithms for computing the number of breakpoints between G_1 and G_2 are given for the *maximum matching* model [9, 10]. We propose in this section a 4-approximation algorithm to compute a maximum matching of two balanced genomes that maximizes the number of adjacencies (as opposed to minimizing the number of breakpoints). Remark that, as opposed to the results in [9, 10], our approximation ratio is independent of the maximum number of duplicates. We first define the problem $AdjD$ we are interested in as follows:

Problem: $AdjD$
Input: Two balanced genomes G_1 and G_2.
Question: Find a maximum matching (G'_1, G'_2) of (G_1, G_2) which maximizes the number of adjacencies between G'_1 and G'_2.

In [8], a 4-approximation algorithm for the weighted 2-$interval\ Pattern$ problem ($W2IP$) is given. In the following, we first define $W2IP$, and then we present how we can relate any instance of $AdjD$ to an instance of $W2IP$.

The weighted 2-interval Pattern problem. A 2-*interval* is the union of two disjoint intervals defined over a single sequence. For a 2-interval $D = (I, J)$, we suppose that the interval I does not overlap J and that I precedes J. We will denote this relation by $I < J$. We say that two 2-intervals $D_1 = (I_1, J_1)$ and $D_2 = (I_2, J_2)$ are *disjoint* if D_1 and D_2 have no common point (i.e. $(I_1 \cup J_1) \cap (I_2 \cup J_2) = \emptyset$). Three possible relations exist between two disjoint 2-intervals: (1) $D_1 \prec D_2$, if $I_1 < J_1 < I_2 < J_2$; (2) $D_1 \sqsubset D_2$, if $I_2 < I_1 < J_1 < J_2$; (3) $D_1 \between D_2$, if $I_1 < I_2 < J_1 < J_2$.
We say that a pair of 2-intervals D_1 and D_2 is R-$comparable$ for some $R \in \{\prec, \sqsubset, \between\}$, if either $(D_1, D_2) \in R$ or $(D_2, D_1) \in R$. A set of 2-intervals $\mathcal{D}$ is $\mathcal{R}$-comparable for some $\mathcal{R} \subseteq \{\prec, \sqsubset, \between\}$, $\mathcal{R} \neq \emptyset$, if any pair of distinct 2-intervals in $\mathcal{D}$ is R-comparable for some $R \in \mathcal{R}$. The non-empty set $\mathcal{R}$ is called a $\mathcal{R}$-*model*. We can define $W2IP$ as follows:

Problem: Weighted 2-interval Pattern ($W2IP$)
Input: A set $\mathcal{D}$ of 2-intervals, a $\mathcal{R}$-model $\mathcal{R} \subseteq \{\prec, \sqsubset, \between\}$ with $\mathcal{R} \neq \emptyset$, a weighted function $w : \mathcal{D} \mapsto \mathbb{R}$.
Question: Find a maximum weight $\mathcal{R}$-comparable subset of $\mathcal{D}$.

Transformation. We now describe how to transform any instance of $AdjD$ into an instance of $W2IP$. Let G_1 and G_2 be two balanced genomes. Two intervals I_1 of G_1 and I_2 of G_2 are said to be *identical* if they correspond to the same string (up to a complete reversal, where a reversal also changes all the signs). We denote by $Make2I$ the construction of the 2-intervals set obtained from the concatenation of G_1 and G_2. $Make2I$ is defined as follows: for any pair (I_1, I_2) of identical intervals of G_1, G_2, we construct a 2-interval $D = (I_1, I_2)$ of weight $|I_1| - 1$. We note $\mathcal{D} = Make2I(G_1, G_2)$ the set of all 2-$intervals$ obtained in this way. Figure 2 gives an example of such a construction. We now define how

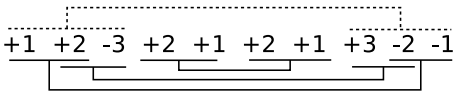

Figure 2: 2-intervals induced by genomes $G_1 = 1\ 2\ -3\ 2\ 1$ and $G_2 = 2\ 1\ 3\ -2\ -1$. For readability, singleton intervals are not drawn. The dotted 2-interval is of weight 2, while all the others are of weight 1.

to tranform any solution of $W2IP$ into a solution of $AdjD$. Let G_1 and G_2 be two balanced genomes and let $\mathcal{D} = Make2I(G_1, G_2)$. Let S be a solution of $W2IP$ over the $\{\prec, \sqsubset, \between\}$-model for $\mathcal{D}$. We denote by $W2IP\_to\_AdjD$ the transformation of S into a maximum matching (G'_1, G'_2) of $G_1, G_2)$ defined as follows. First, for each 2-interval $D = (I_1, I_2)$ of S, we match the genes of I_1 and I_2 in the natural way;

then, in order to achieve a maximum matching (since each gene is not necessarily covered by a 2-interval of S), we apply the following greedy algorithm: iteratively, we match arbitrarily two unmatched genes present in both G_1 and G_2, until no such gene exist. After a relabeling of signed genes, we obtain a maximum matching (G'_1, G'_2) of (G_1, G_2).

Lemma 9 *Let G_1 and G_2 be two balanced genomes and let $\mathcal{D} = Make2I(G_1, G_2)$. Let S be a solution of $W2IP$ over the $\{\prec, \sqsubset, \between\}$-model. Let W_S be the weight of S. Then the maximum matching (G'_1, G'_2) of (G_1, G_2) obtained by $W2IP\_to\_AdjD(S)$ induces at least W_S adjacencies.*

Proof. Let W_S be the weight of S. We construct the maximum matching (G'_1, G'_2) of (G_1, G_2) as using the transformation $W2IP\_to\_AdjD$. First, we have matched, for each 2-interval $D = (I_1, I_2)$ of S, the genes of I_1 and I_2 in the natural way. This operation implies, for each 2-interval $D = (I_1, I_2)$ of S, $|I_1| - 1$ adjacencies since I_1 and I_2 are identical. By construction of $\mathcal{D}$, this operation induces W_S adjacencies altogether. The second operation is the greedy algorithm for which no adjacency is suppressed (note that other adjacencies might be created). Hence, (G'_1, G'_2) induces at least W_S adjacencies. □

Lemma 10 *Let G_1 and G_2 be two balanced genomes and let (G'_1, G'_2) be a maximum matching of (G_1, G_2). Let $\mathcal{D} = Make2I(G_1, G_2)$.*
Let W be the number of adjacencies induced by (G'_1, G'_2) between G_1 and G_2. Then there exists a solution S of $W2IP$ over the $\{\prec, \sqsubset, \between\}$-model for $\mathcal{D}$ with weight equal to W.

Proof. [Lemma 10] Let (G'_1, G'_2) be a maximum matching of (G_1, G_2) and let n be the size of G'_1. Suppose that there exist W adjacencies between G'_1 and G'_2. There exists a unique partition $P_{(G'_1, G'_2)} = \{s_1, s_2 \dots s_p\}$ of genome G'_1 into p substrings such that for each i, $1 \leqslant i < p$, s_i and s_{i+1} are separated by one breakpoint and such that no breakpoint appears in s_i, $1 \leqslant i \leqslant p$. This partition implies that there exists $p - 1$ breakpoints between G'_1 and G'_2, and consequently, $n - p$ adjacencies. To each substring s_i of $P_{(G'_1, G'_2)}$ in G'_1, corresponds a unique substring t_i in G'_2, for which s_i and t_i are identical. Moreover, each substring s_i of size l_i, $1 \leqslant i \leqslant p$, contains $l_i - 1$ adjacencies. We construct the 2-interval set S as the union of $S_i = (\hat{s_i}, \hat{t_i})$, $1 \leqslant i \leqslant p$, where $\hat{s_i}$ (resp. $\hat{t_i}$) is the interval obtained from s_i (resp. t_i). The partition P implies that the 2-intervals created are disjoint and thus $\{\prec, \sqsubset, \between\}$-comparable and the weight of S is equal to $\sum_{i=1}^{p} (l_i - 1) = \sum_{i=1}^{p} l_i - \sum_{i=1}^{p} 1 = n - p = W$. □

We now describe the algorithm $ApproxAdjD$ and then prove that it is a 4-approximation of the problem $AdjD$ by Theorem 3.

Algorithm 1 $ApproxAdjD$

Require: Two balanced genomes G_1 and G_2.
Ensure: A maximum matching (G'_1, G'_2) of (G_1, G_2).

- Construct the set of weighted 2-intervals $\mathcal{D} = Make2I(G_1, G_2)$
- Invoke the 4-approximation algorithm of Crochemore et al. [8] to obtain a solution S of $W2IP$ over the $\{\prec, \sqsubset, \between\}$-model for $\mathcal{D}$
- Construct the maximal matching $(G'_1, G'_2) = W2IP\_to\_AdjD(S)$

Theorem 3 *Algorithm $ApproxAdjD$ is a 4-approximation algorithm for $AdjD$.*

Proof. Let G_1 and G_2 be two balanced genomes and let $\mathcal{D} = Make2I(G_1, G_2)$. We first prove that the optimum of $AdjD$ for (G_1, G_2) is equal to the optimum of $W2IP$. Let OPT_{AdjD} be the optimum of $AdjD$ for (G_1, G_2). By Lemma 10, we know that there exists a solution S for $W2IP$ with weight $W_S = OPT_{AdjD}$. Now, suppose that there exists a solution S' for $W2IP$ with weight $W_{S'} > W_S$. Then, by Lemma 9, there exists a solution for $AdjD$ with weight $W \geqslant W_{S'}$. However, $W_{S'} > W_S$ by hypothesis, a contradiction to the fact that $W_S = OPT_{AdjD}$. Therefore, the two problems have the same optimum and, as a result, any approximation ratio for $W2IP$ implies the same approximation ratio for $AdjD$. In [8], a 4-approximation algorithm is proposed for $W2IP$; this directly implies that $ApproxAdjD$ is a 4-approximation algorithm for $AdjD$. □

5 Conclusions and future work

In this paper, we have first given new approximation complexity results for several optimization problems in genomic rearrangement. We focused on breakpoints, conserved and common intervals measures and we took into account the presence of duplicates. We restricted our proofs to cases where one genome contains no duplicates and the other contains no more than two occurrences of each gene. With this assumption, we proved that the problems consisting in computing an exemplarization (resp. a maximum matching, a non-maximum matching) optimizing one of these measures is **APX-Hard**, thus extending the results of [5, 6, 7]. For that, we used an *L-reduction* from vertex cover on cubic graphs. In a second part of this paper, we gave a 4-approximation algorithm for computing the number of adjacencies of two balanced genomes under the *maximum matching* model. We note that our approximation ratio we obtain is constant, even when the number of occurrences in genomes is unbounded.
The problems studied in this paper are **APX-Hard**, but some approximation algorithms exist when genomes are balanced [9, 10]. However, it remains open whether approximation algorithms exist when genomes are not balanced. It has been shown in [7] that deciding if two genomes G_1 and G_2 have zero breakpoint under the *exemplar* model is **NP-Complete** even when $occ(G_1) = occ(G_2) = 3$ (problem $ZEBD$). This result implies that the EBD problem cannot be approximated in that case. Another open question is the complexity of $ZEBD$ when no gene appears more that twice in the genome.

References

[1] P. Alimonti and V. Kann. Some APX-completeness results for cubic graphs. *Theoretical Computer Science*, 237(1-2):123–134, 2000.

[2] S. Angibaud, G. Fertin, I. Rusu, and S. Vialette. A general framework for computing rearrangement distances between genomes with duplicates. *Journal of Computational Biology*, 14(4):379–393, 2007.

[3] V. Bafna and P. Pevzner. Sorting by reversals: genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, pages 239–246, 1995.

[4] G. Blin and R. Rizzi. Conserved interval distance computation between non-trivial genomes. In *Proc. COCOON 2005*, volume 3595 of *LNCS*, pages 22–31. Springer, 2005.

[5] D. Bryant. The complexity of calculating exemplar distances. In *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignement, and the Evolution of Gene Families*, pages 207–212. Kluwer Academic Publisher, 2000.

[6] C. Chauve, G. Fertin, R. Rizzi, and S. Vialette. Genomes containing duplicates are hard to compare. In *Proc. IWBRA 2006*, volume 3992 of *LNCS*, pages 783–790. Springer, 2006.

[7] Z. Chen, B. Fu, and B. Zhu. The approximability of the exemplar breakpoint distance problem. In *Proc. AAIM 2006*, volume 4041 of *LNCS*, pages 291–302. Springer, 2006.

[8] M. Crochemore, D. Hermelin, G. M. Landau, and S. Vialette. Approximating the 2-interval pattern problem. In *Proc. ESA 2005*, volume 3669 of *LNCS*, pages 426–437. Springer, 2005.

[9] A. Goldstein, P. Kolman, and Z. Zheng. Minimum common string partition problem: Hardness and approximations. In *Proc. ISAAC 2004*, volume 3341 of *LNCS*, pages 473–484. Springer, 2004.

[10] P. Kolman and T. Waleń. Reversal distance for strings with duplicates: Linear time approximation using hitting set. In *Proc. WAOA 2006*, volume 4368 of *LNCS*, pages 279–289. Springer, 2006.

[11] W. Li, Z. Gu, H. Wang, and A. Nekrutenko. Evolutionary analysis of the human genome. *Nature*, (409):847–849, 2001.

[12] M. Marron, K. M. Swenson, and B. M. E. Moret. Genomic distances under deletions and insertions. *Theoretical Computer Science*, 325(3):347–360, 2004.

[13] C. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43(3):425–440, 1991.

[14] D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, 1999.

[15] D. Sankoff and L. Haque. Power boosts for cluster tests. In *Proc. RECOMB-CG 2005*, volume 3678 of *LNCS*, pages 121–130. Springer, 2005.

[16] J. Tang and B. M. E. Moret. Phylogenetic reconstruction from gene-rearrangement data with unequal gene content. In *Proc. WADS 2003*, volume 2748 of *LNCS*, pages 37–46. Springer, 2003.

On the Approximability of Comparing Genomes with Duplicates

Sébastien Angibaud, Guillaume Fertin, Irena Rusu

Abstract

A central problem in comparative genomics consists in computing a (dis-)similarity measure between two genomes, e.g. in order to construct a phylogenetic tree. A large number of such measures has been proposed in the recent past: *number of reversals*, *number of breakpoints*, *number of common* or *conserved intervals*, *SAD* etc. In their initial definitions, all these measures suppose that genomes contain no duplicates. However, we now know that genes can be duplicated within the same genome. One possible approach to overcome this difficulty is to establish a one-to-one correspondence (i.e. a matching) between genes of both genomes, where the correspondence is chosen in order to optimize the studied measure. Then, after a gene relabeling according to this matching and a deletion of the unmatched signed genes, two genomes without duplicates are obtained and the measure can be computed.
In this paper, we are interested in three measures (*number of breakpoints*, *number of common intervals* and *number of conserved intervals*) and three models of matching (*exemplar* model, *maximum matching* model and *non maximum matching* model). We prove that, for each model and each measure, computing a matching between two genomes that optimizes the measure is **APX-Hard**. We show that this result remains true even for two genomes G_1 and G_2 such that G_1 contains no duplicates and no gene of G_2 appears more than twice. Therefore, our results extend those of [5, 6, 7]. Finally, we propose a 4-approximation algorithm for a measure closely related to the *number of breakpoints*, the *number of adjacencies*, under the *maximum matching* model, in the case where genomes contain the same number of duplications of each gene.

Additional Key Words and Phrases: genome rearrangement, APX-Hardness, duplicates, breakpoints, adjacencies, common intervals, conserved intervals, approximation algorithm

LINA, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3