Inapproximability of (1,2)-Exemplar Distance

Laurent Bulteau, Minghui Jiang

Inapproximability of $(1, 2)$-Exemplar Distance

Laurent Bulteau[1] and Minghui Jiang[2]

[1] 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@univ-nantes.fr
[2] Department of Computer Science, Utah State University
Logan, UT 84322-4205, USA
mjiang@cc.usu.edu

Abstract. Given two genomes possibly with duplicate genes, the exemplar distance problem is that of removing all but one copy of each gene in each genome, so as to minimize the distance between the two reduced genomes according to some measure. Let (s, t)-Exemplar Distance denote the exemplar distance problem on two genomes G_1 and G_2 where each gene occurs at most s times in G_1 and at most t times in G_2. We show that the simplest non-trivial variant of the exemplar distance problem, $(1, 2)$-Exemplar Distance, is already hard to approximate for a wide variety of distance measures, including popular genome rearrangement measures such as adjacency disruptions and signed reversals, and classic string edit distance measures such as Levenshtein and Hamming distances.

Keywords: comparative genomics, hardness of approximation, adjacency disruption, sorting by reversals, edit distance, Levenshtein distance, Hamming distance.

1 Introduction

In the study of genome rearrangement, a *gene* is usually represented by a signed integer: the absolute value of the integer (the unsigned integer) denotes the gene family to which the gene belongs; the sign of the integer denotes the orientation of the gene in its chromosome. Then a *chromosome* is a sequence of signed integers, and a *genome* is a collection of chromosomes. Given two genomes possibly with duplicate genes, the *exemplar distance* problem [14] is that of removing all but one copy of each gene in each genome, so as to minimize the distance between the two reduced genomes according to some measure. The reduced genomes are said to be *exemplar subsequences* of the original genomes. This approach amounts to considering that, in the evolution history, duplications have taken place after the speciation of the genomes (or more generally, that we are able to distinguish genes that have been duplicated before the speciation). Hence, in each genome, only one copy of each gene may be matched to an ortholog gene in the other genome.

For example, the following two monochromosomal genomes

$$G_1: \quad -4\ +1\ +2\ +3\ -5\ +1\ +2\ +3\ -6$$
$$G_2: \quad -1\ -4\ +1\ +2\ -5\ +3\ -2\ -6\ +3$$

can both be reduced to the same genome

$$G': \quad -4\ +1\ +2\ -5\ +3\ -6$$

by removing duplicates, thus they have exemplar distance zero for any reasonable distance measure. In general, unless we are to decide simply whether two genomes can be reduced to the same genome by removing duplicates, the exemplar distance problem is not a single problem but a group of related problems because the choice of the distance measure is not unique.

We denote by (s,t)-Exemplar Distance the exemplar distance problem on two genomes G_1 and G_2 where each gene occurs at most s times in G_1 and at most t times in G_2. It is known [5, 12] that for any reasonable distance measure, $(2,2)$-Exemplar Distance does not admit any approximation. This is because to decide simply whether two genomes with maximum occurrence 2 can be reduced to the same genome by removing duplicates is already NP-hard. In this paper, we focus on the simplest non-trivial variant of the exemplar distance problem: $(1,2)$-Exemplar Distance.

The problem $(1,t)$-Exemplar Distance has been studied for several distance measures commonly used in genome rearrangement. Angibaud et al. [2] showed that $(1,2)$-Exemplar Breakpoint Distance, $(1,2)$-Exemplar Common Interval Distance, and $(1,2)$-Exemplar Conserved Interval Distance are all APX-hard. Blin et al. [4] showed that $(1,9)$-Exemplar MAD Distance is NP-hard to approximate within $2-\epsilon$ for any $\epsilon > 0$, and that $(1,\infty)$-Exemplar SAD Distance is NP-hard to approximate within $c \log n$ for some constant $c > 0$, where n is the number of genes in G_1. See also [8, 6, 7] for related results.

The two distance measures we first consider, MAD and SAD, were introduced by Sankoff and Haque [15]. For two permutations $\pi' = \pi'_1 \dots \pi'_n$ and $\pi'' = \pi''_1 \dots \pi''_n$ of n distinct elements, define $\tau'(i)$ as the index j such that $\pi'_j = \pi''_i$, and $\tau''(i)$ as the index j such that $\pi''_j = \pi'_i$. Then the maximum adjacency disruption (MAD) and the summed adjacency disruption (SAD) between π' and π'' are

$$\mathrm{MAD}(\pi',\pi'') = \max_{1 \le i \le n-1} \big\{\, |\tau'(i) - \tau'(i+1)|,\, |\tau''(i) - \tau''(i+1)| \,\big\},$$
$$\mathrm{SAD}(\pi',\pi'') = \sum_{1 \le i \le n-1} \big(\, |\tau'(i) - \tau'(i+1)| + |\tau''(i) - \tau''(i+1)| \,\big).$$

Our first two theorems sharpen the previous results on the inapproximability on $(1,t)$-Exemplar Distance for both MAD and SAD measures:

Theorem 1. $(1,2)$-Exemplar MAD Distance *is NP-hard to approximate within* $2-\epsilon$ *for any* $\epsilon > 0$.

Theorem 2. $(1,2)$-Exemplar SAD Distance *is NP-hard to approximate within* $10\sqrt{5} - 21 - \epsilon = 1.3606\ldots - \epsilon$, *and is NP-hard to approximate within* $2 - \epsilon$ *if the unique games conjecture is true, for any* $\epsilon > 0$.

For an unsigned permutation $\pi = \pi_1 \ldots \pi_n$, an *unsigned reversal* (i, j) with $1 \leq i \leq j \leq n$ turns it into $\pi_1 \ldots \pi_{i-1}\ \pi_j \ldots \pi_i\ \pi_{j+1} \ldots \pi_n$, where the substring $\pi_i \ldots \pi_j$ is reversed. For a signed permutation $\sigma = \sigma_1 \ldots \sigma_n$, a *signed reversal* (i, j) with $1 \leq i \leq j \leq n$ turns it into $\sigma_1 \ldots \sigma_{i-1}\ -\sigma_j \ldots -\sigma_i\ \sigma_{j+1} \ldots \sigma_n$, where the substring $\sigma_i \ldots \sigma_j$ is reversed and negated. The *unsigned reversal distance* (resp. *signed reversal distance*) between two unsigned (resp. signed) permutations is the minimum number of unsigned (resp. signed) reversals required to transform one to the other. Computing the unsigned reversal distance is APX-hard [3], although the signed reversal distance can be computed in polynomial time [11].

Our next theorem answers an open question of Blin et al. [4] on the inapproximability of the exemplar reversal distance problem:

Theorem 3. $(1,2)$-Exemplar Signed Reversal Distance *is NP-hard to approximate within* $1237/1236 - \epsilon$ *for any* $\epsilon > 0$.

In the last theorem of this paper, we present the first inapproximability result on the exemplar distance problem using the classic string edit distance measure:

Theorem 4. $(1,2)$-Exemplar Edit Distance *is APX-hard to compute when the cost of a substitution is* 1 *and the cost of an insertion or a deletion is at least* 1.

Note that both Levenshtein distance and Hamming distance are special cases of the string edit distance: for Levenshtein distance, the cost of every operation (substitution, insertion, or deletion) is 1; for Hamming distance, the cost of a substitution is 1 and the cost of an insertion or a deletion is $+\infty$. Thus we have the following corollaries:

Corollary 1. $(1,2)$-Exemplar Levenshtein Distance *is APX-hard.*

Corollary 2. $(1,2)$-Exemplar Hamming Distance *is APX-hard.*

2 MAD Distance

In this section we prove Theorem 1. We prove that Exemplar MAD Distance is NP-hard by a reduction from the well-known NP-hard problem 3SAT [10]. Let (V, C) be a 3SAT instance, where $V = \{v_1, \ldots, v_n\}$ is a set of n boolean variables, $C = \{c_1, \ldots, c_m\}$ is a conjunctive boolean formula of m clauses, and each clause in C is a disjunction of exactly three literals of the variables in V. The problem 3SAT is that of deciding whether (V, C) is satisfiable, i.e., whether there is a truth assignment for the variables in V that satisfies all clauses in C.

Let M be a large number to be specified. We will construct two sequences (genomes) G_1 and G_2 over $L = 3m + (n+1) + (2n+1) + (m+1) + (2M+2) = 2M + 3n + 4m + 5$ distinct markers (genes):

- 3 *literal markers* r_j, s_j, t_j for the 3 literals of each clause c_j, $1 \leq j \leq m$;
- $n+1$ *variable markers* x_i, $0 \leq i \leq n$;
- $2n+1$ *separator markers* y_i, $0 \leq i \leq 2n$;
- $m+1$ *clause markers* z_j, $0 \leq j \leq m$;
- $2M+2$ *dummy markers* ϕ_k and ψ_k, $0 \leq k \leq M$.

For each clause c_j, let $O_j = r_j s_j t_j$ be the concatenation of the three literal markers of c_j. For each variable v_i, let $P_i = p_{i,1} \dots p_{i,k_i}$ be the concatenation of the k_i literal markers of the positive literals of v_i, and let $Q_i = q_{i,1}, \dots, q_{i,l_i}$ be the concatenation of the l_i literal markers of the negative literals of v_i. Without loss of generality, assume that $\min\{k_i, l_i\} \geq 1$. Note that the two concatenated sequences $O_1 \dots O_m$ and $P_1 Q_1 \dots P_n Q_n$ are both permutations of the $3m$ literal markers.

The two sequences G_1 and G_2 are represented schematically as follows. G_1 contains exactly one copy of each marker, and has length L; G_2 contains exactly two copies of each literal marker and exactly one copy of each non-literal marker, and has length $L + 3m$.

G_1 : $\dots z_3 z_1 \quad \phi_0 \quad \dots x_2 x_0 \quad \phi_M \dots \phi_1 \quad y_0 P_1 y_1 Q_1 y_2 \dots P_n y_{2n-1} Q_n y_{2n} \quad \psi_1 \dots \psi_M \quad z_0 z_2 \dots \quad \psi_0 \quad x_1 x_3 \dots$
G_2 : $x_n P_n Q_n \dots x_1 P_1 Q_1 x_0 \quad \phi_M \dots \phi_1 \phi_0 \quad y_0 y_1 y_2 \dots y_{2n-1} y_{2n} \quad \psi_0 \psi_1 \dots \psi_M \quad z_0 O_1 z_1 \dots O_m z_m$

Lemma 1. *If (V, C) is satisfiable, then G_2 has an exemplar subsequence G_2' that satisfies* $\mathrm{MAD}(G_1, G_2') \leq M + 3n + 4m + 5$.

Proof. Let f be a truth assignment for the variables in V that satisfies all clauses in C. For each variable v_i, compose a subsequence V_i of $P_i Q_i$ such that $V_i = Q_i$ if $f(v_i)$ is true and $V_i = P_i$ if $f(v_i)$ is false. For each clause c_j, compose a subsequence C_j of O_j containing only the literal markers of the literals that are true under the assignment f. Then $V_1 \dots V_n C_1 \dots C_m$ is a permutation of the $3m$ literal markers. It is straightforward to verify that the exemplar subsequence G_2' of G_2 in the following satisfies $\mathrm{MAD}(G_1, G_2') \leq L - M = M + 3n + 4m + 5$:

G_1 : $\dots z_3 z_1 \quad \phi_0 \quad \dots x_2 x_0 \quad \phi_M \dots \phi_1 \quad y_0 P_1 y_1 Q_1 y_2 \dots P_n y_{2n-1} Q_n y_{2n} \quad \psi_1 \dots \psi_M \quad z_0 z_2 \dots \quad \psi_0 \quad x_1 x_3 \dots$
G_2 : $x_n P_n Q_n \dots x_1 P_1 Q_1 x_0 \quad \phi_M \dots \phi_1 \phi_0 \quad y_0 y_1 y_2 \dots y_{2n-1} y_{2n} \quad \psi_0 \psi_1 \dots \psi_M \quad z_0 O_1 z_1 \dots O_m z_m$
G_2' : $x_n V_n \dots x_1 V_1 x_0 \quad \phi_M \dots \phi_1 \phi_0 \quad y_0 y_1 y_2 \dots y_{2n-1} y_{2n} \quad \psi_0 \psi_1 \dots \psi_M \quad z_0 C_1 z_1 \dots C_m z_m$

□

Lemma 2. *If (V, C) is not satisfiable, then every exemplar subsequence G_2' of G_2 satisfies* $\mathrm{MAD}(G_1, G_2') > 2M$.

Proof. We prove the contrapositive. Suppose G_2 has an exemplar subsequence G_2' that satisfies $\mathrm{MAD}(G_1, G_2') \leq 2M$. We will find a truth assignment f for the variables in V that satisfies all clauses in C.

First, we claim that for each variable v_i, the literal markers of the positive literals of v_i must appear in G_2' either all before ϕ_M or all after ψ_M. Suppose

the contrary. Then there would be two literal markers of v_i, one before ϕ_M and one after ψ_M in G'_2, that are adjacent in the substring P_i in G_1, incurring a MAD distance larger than $2M$. Similarly, we claim that the literal markers of the negative literals of each variable v_i must appear in G'_2 either all before ϕ_M or all after ψ_M.

Next, we claim that for each variable v_i, the literal markers of either all positive literals of v_i or all negative literals of v_i must appear in G'_2 before ϕ_M, between x_i and x_{i-1}. Suppose the contrary that all literal markers of both the positive and the negative literals of v_i appear in G'_2 after ψ_M. Then the two variable markers x_i and x_{i-1}, one before ϕ_M and one after ψ_M in G_1, would become adjacent in G'_2, incurring a MAD distance larger than $2M$.

Finally, we claim that for each clause c_j, at least one of the three literal markers r_j, s_j, t_j must appear in G'_2 after ψ_M, between z_{j-1} and z_j. Suppose the contrary. Then the two clause markers z_{j-1} and z_j, one before ϕ_M and one after ψ_M in G_1, would become adjacent in G'_2, again incurring a MAD distance larger than $2M$.

Now compose a truth assignment f for the variables in V such that $f(v_i)$ is true if the literal markers for the negative literals of v_i appear before ϕ_M, and is false otherwise. Then f satisfies all clauses in C. □

For any constant ϵ, $0 < \epsilon < 2$, we can get a gap of $2M/(M + 3n + 4m + 5) = 2 - \epsilon$ by setting $M = (\frac{2}{\epsilon} - 1)(3n + 4m + 5)$. Thus the NP-hardness of 3SAT and the two preceding lemmas together imply that Exemplar MAD Distance is NP-hard to approximate within $2 - \epsilon$ for any $\epsilon > 0$.

3 SAD Distance

In this section we prove Theorem 2. We show that Exemplar SAD Distance is NP-hard to approximate by a reduction from another well-known NP-hard problem Minimum Vertex Cover [10]. Let (V, E) be a graph, where $V = \{v_1, \ldots, v_n\}$ is a set of n vertices, and $E = \{e_1, \ldots, e_m\}$ is a set of m edges. The problem Minimum Vertex Cover is that of finding a subset $C \subseteq V$ of the minimum cardinality such that each edge in E is incident to at least one vertex in C. Dinur and Safra [9] showed that Minimum Vertex Cover is NP-hard to approximate within any constant less than $10\sqrt{5} - 21 = 1.3606\ldots$. Khot and Regev [13] showed that Minimum Vertex Cover is NP-hard to approximate within any constant less than 2 if the unique games conjecture is true.

Let $M = 2(n + m)^2$. We will construct two sequences (genomes) G_1 and G_2 over $L = n + m + M + 1$ distinct markers (genes):

- *n vertex markers* v_i, $1 \le i \le n$;
- *m edge markers* e_j, $1 \le j \le m$;
- *M dummy markers* ϕ_k, $0 \le k \le M$.

For each vertex v_i, let $E_i = e_{i,1} \ldots e_{i,k_i}$ be the concatenation of the edge markers of all edges incident to v_i, where k_i is the degree of v_i. The two sequences

G_1 and G_2 are represented schematically as follows. G_1 contains exactly one copy of each marker, and has length L; G_2 contains exactly two copies of each edge marker and exactly one copy of each non-edge marker, and has length $L + m$.

$$\begin{aligned} G_1 &: \quad e_1 \dots e_m \quad \phi_0 \phi_1 \dots \phi_M \quad v_1 \dots v_n \\ G_2 &: \quad \phi_0 \phi_1 \dots \phi_M \quad E_1 v_1 \dots E_n v_n \end{aligned}$$

Lemma 3. *G has a vertex cover of size at most k if and only if G_2 has an exemplar subsequence G_2' that satisfies* $\mathrm{SAD}(G_1, G_2') \leq (2k+4)M$.

Proof. We first prove the direct implication. Let C be a vertex cover of size at most k in G. Extract a subsequence E_i' of E_i for each vertex v_i in C such that the concatenated sequence $E_1' \dots E_n'$ contains each edge marker e_j exactly once. From G_2, remove E_i for each vertex v_i not in C, and replace E_i by E_i' for each vertex v_i in C. Then we obtain an exemplar subsequence G_2' of G_2.

The two sequences G_1 and G_2' have the same length $L = n + m + M + 1$ and together have $2n + 2m + 2M$ adjacencies. The contributions of these adjacencies to $\mathrm{SAD}(G_1, G_2')$ are as follows:

1. The shared adjacencies $\phi_i \phi_{i+1}$ in G_1 and G_2', $0 \leq i \leq M - 1$, contribute a total value of exactly $2M$.
2. The adjacency $e_m \phi_0$ in G_1 contributes a value of at least M and at most $M + n + m$.
3. Each adjacency between an edge marker and a non-edge marker in G_2' contributes a value of at least M and at most $M + n + m$.
4. Each remaining adjacency contributes a value of at least 1 and at most $n+m$.

The number of adjacencies between an edge marker and a non-edge marker in G_2' is exactly twice the size of the vertex cover C. Thus we have

$$\begin{aligned} \mathrm{SAD}(G_1, G_2') &\leq 2M + (2k+1)(M+n+m) \\ &\quad + (2n + 2m + 2M - 2M - 2k - 1)(n+m) \\ &= (2k+3)M + 2(n+m)^2 = (2k+4)M. \end{aligned}$$

We next prove the reverse implication. Let G_2' be an exemplar subsequence of G_2 such that $\mathrm{SAD}(G_1, G_2') \leq (2k+4)M$. Refer back to the list of contributions to $\mathrm{SAD}(G_1, G_2')$. Let l be the number of adjacencies between an edge marker and a non-edge marker in G_2'. Then we have the following inequality:

$$\mathrm{SAD}(G_1, G_2') \geq 2M + (l+1)M = (l+3)M.$$

Since $\mathrm{SAD}(G_1, G_2') \leq (2k+4)M$, we have $l+3 \leq 2k+4$ and hence $l \leq 2k+1$. Note that l must be an even number: for each adjacency between an edge marker in E_i and a non-edge marker to its left, there must be another adjacency between an edge marker in E_i and a non-edge marker (indeed a vertex marker) to its right, and vice versa. It follows that there are at most k vertex markers v_i that are adjacent to an edge marker to its left. The corresponding at most k vertices v_i form a vertex cover of G. □

The inapproximability of Minimum Vertex Cover and the preceding lemma together imply that Exemplar SAD Distance is NP-hard to approximate within $10\sqrt{5}-21-\epsilon$, and is NP-hard to approximate within $2-\epsilon$ if the unique games conjecture is true, for any $\epsilon > 0$.

4 Signed Reversal Distance

In this section we prove Theorem 3. We show that $(1,2)$-Exemplar Signed Reversal Distance is APX-hard by a reduction from the problem Min-SBR [3], which asks for the minimum number of unsigned reversals to sort a given unsigned permutation into the identity permutation.

Let $\pi = \pi_1 \dots \pi_n$ be an unsigned permutation of $1 \dots n$. We construct two sequences $G_1 = 1 \dots n$ and $G_2 = \pi_1 - \pi_1 \dots \pi_n - \pi_n$.

Lemma 4. *π can be sorted into the identity permutation $1 \dots n$ by at most k unsigned reversals if and only if G_2 has an exemplar subsequence G_2' with signed reversal distance at most k from G_1.*

We leave the proof of Lemma 4 to the reader as an easy exercise. Since Min-SBR is NP-hard to approximate within $1237/1236-\epsilon$ for any $\epsilon > 0$ [3], $(1,2)$-Exemplar Signed Reversal Distance is NP-hard to approximate within $1237/1236-\epsilon$ for any $\epsilon > 0$ too.

5 Edit Distance

In this section we prove Theorem 4. For any edit distance where the cost of a substitution is 1 and the cost of an insertion or a deletion is at least 1 (possibly $+\infty$), we show that the problem $(1,2)$-Exemplar Edit Distance is APX-hard by a reduction from the problem Minimum Vertex Cover in Cubic Graphs.

Let $G = (V, E)$ be a cubic graph of n vertices and m edges, where $3n = 2m$. We will construct two sequences (genomes) G_1 and G_2 over an alphabet of

$$3m + 4n + 2(m + 7n) + 2(m - 1) + (n - 1)$$

distinct markers (genes). For each edge $e = \{u, v\} \in E$, we have three edge markers e, e_u, and e_v. For each vertex $v \in V$, we have a vertex marker v and 3 dummy markers v_1', v_2', v_3'. In addition, we have $2(m+7n)+2(m-1)+(n-1)$ markers for separators.

The two sequences G_1 and G_2 are composed from $m+n+1$ gadgets: an edge gadget for each edge, a vertex gadget for each vertex, and a tail gadget. The $m+n+1$ gadgets are separated by $m+n$ separators of total length $2(m+7n)+2(m-1)+(n-1)$:

- two long separators, each of length $m+7n$: one between the last edge gadget and the first vertex gadget, one between the last vertex gadget and the tail gadget;

– $m+n-2$ short separators: a length-2 separator between any two consecutive edge gadgets, and a length-1 separator between any two consecutive vertex gadgets.

For each edge $e = \{u, v\}$, the edge gadget for e is

$$G_1\langle e\rangle = e$$
$$G_2\langle e\rangle = e_u e_v$$

For each vertex v incident to edges e, f, g, the vertex gadget for v is

$$G_1\langle v\rangle = v\, v'_1 v'_2 v'_3$$
$$G_2\langle v\rangle = e_v f_v g_v\, v\, e\, f\, g$$

Let V' be the $3n$ markers v'_1, v'_2, v'_3 for $v \in V$. Let E' be the $2m = 3n$ markers e_u and e_v for $e = \{u, v\} \in E$. The tail gadget is

$$G_1\langle tail\rangle = E'$$
$$G_2\langle tail\rangle = V'$$

This completes the construction.

Lemma 5. *G has a vertex cover of size at most k if and only if G_2 has an exemplar subsequence G'_2 with edit distance at most $m + 6n + k$ from G_1.*

Proof. We first prove the direct implication. Let X be a vertex cover of G with $|X| \leq k$. Create G'_2 as follows. For each edge $e = \{u, v\}$, at least one vertex, say u, is in X. Remove e_u and retain e_v in the edge gadget $G_2\langle e\rangle$, and correspondingly retain e_u in the vertex gadget $G_2\langle u\rangle$ and remove e_v in the vertex gadget $G_2\langle v\rangle$, then remove e in $G_2\langle u\rangle$ and retain e in $G_2\langle v\rangle$. We claim that the edit distance from G_1 to G'_2 is at most $m + 6n + k$.

It suffices to show that the Hamming distance of G_1 and G'_2 is at most $m+6n+k$ since, for the edit distance that we consider, the cost of a substitution is 1. Observe that in both G_1 and G'_2, each edge gadget has length 1, and each vertex gadget has length 4. Thus all gadgets are aligned and all separators are matched. The Hamming distance for each edge gadget is 1, so the total Hamming distance over all edge gadgets is m. The Hamming distance for each vertex gadget is at most 4. Moreover, for each vertex $v \notin X$ (v incident to edges e, f, g), since the markers e_v, f_v, g_v are removed (and the markers e, f, g are retained) in the vertex gadget, the marker v is matched, which reduces the Hamming distance by 1. Thus the total Hamming distance over all vertex gadgets is at most $4n - (n - |X|) = 3n + |X|$. Finally, since the Hamming distance for the tail gadget is $3n$, the overall Hamming distance of G_1 and G'_2 is at most $m + 6n + |X| \leq m + 6n + k$.

We next prove the reverse implication. Let G'_2 be an exemplar subsequence of G_2 with edit distance at most $m + 6n + k$ from G_1. Compute an alignment of G_1 and G'_2 corresponding to the edit distance, then obtain the following three sets $X_E(G'_2)$, $X_V(G'_2)$, and $X(G'_2)$:

- The set $X_E(G'_2) \subseteq E$ contains every edge $e = \{u, v\}$ such that either $G'_2\langle e\rangle$ contains both e_u and e_v, or $G'_1\langle e\rangle$ has an adjacent separator marker which is unmatched.
- The set $X_V(G'_2) \subseteq V$ contains every vertex v (v incident to edges e, f, g) such that either $G'_2\langle v\rangle$ contains one of $\{e_v, f_v, g_v\}$, or $G'_1\langle v\rangle$ has an adjacent separator marker (to its left) which is unmatched.
- The set $X(G'_2) \subseteq V$ is the union of $X_V(G'_2)$ and a set composed by arbitrarily choosing one vertex from each edge in $X_E(G'_2)$ (thus $|X(G'_2)| \leq |X_V(G'_2)| + |X_E(G'_2)|$).

We first show that the edit distance from G_1 to G'_2 is at least $m+6n+|X(G'_2)|$. If a long separator (with $m + 7n$ markers) is completely unmatched, then the edit distance is at least $m + 7n \geq m + 6n + |X(G'_2)|$. Hence we can assume that there is at least one matched marker in each long separator. Consequently, the markers e, e_u, e_v for all $e \in E$ and v'_1, v'_2, v'_3 for all $v \in V$ are unmatched.

Consider an edge $e = \{u, v\} \in E$. If $e \notin X_E(G'_2)$, then the edit distance for $G_1\langle e\rangle$ is at least 1 since the marker e is unmatched. If $e \in X_E(G'_2)$, then consider the substring of $G_1\langle e\rangle$ containing the marker e and the at most two separator markers adjacent to it (for the first edge gadget, there is only one separator marker adjacent to e, to its right). The edit distance for this substring is at least 2: the marker e is unmatched, and moreover either an adjacent separator marker is unmatched or an insertion is required. The total edit distance over all edge gadgets is at least $m + |X_E(G'_2)|$.

Consider a vertex $v \in V$ incident to three edges e, f, g. If $v \notin X_V(G'_2)$, then the edit distance for $G_1\langle v\rangle$ is at least 3 since the markers v'_1, v'_2, v'_3 are unmatched. If $v \in X_V(G'_2)$, then consider the substring of G_1 containing $G_1\langle v\rangle$ and the separator to its left. The edit distance for this substring is at least 4: the markers v'_1, v'_2, v'_3 are unmatched, and moreover at least one insertion is required unless either the marker v or the separator marker to its left is unmatched. The total edit distance over the vertex gadgets is at least $3n + |X_V(G'_2)|$.

Finally, the edit distance over the tail gadget is at least the length of $G_1\langle tail\rangle$, which is $3n$. Hence the overall edit distance is at least

$$m + |X_E(G'_2)| + 3n + |X_V(G'_2)| + 3n \geq m + 6n + |X(G'_2)|.$$

Since the edit distance from G_1 to G'_2 is at most $m + 6n + k$, it follows that $|X(G'_2)| \leq k$.

To complete the proof, we show that $X(G'_2)$ is a vertex cover of G. Consider any edge $e = \{u, v\}$. If $e \in X_E(G'_2)$, then, by our choice of $X(G'_2)$, either $u \in X(G'_2)$ or $v \in X(G'_2)$. Otherwise, if $e \notin X_E(G'_2)$, then in the edge gadget $G_2\langle e\rangle = e_u e_v$, at least one marker is removed to obtain $G'_2\langle e\rangle$. Assume that e_u is removed: then the second copy, in $G_2\langle u\rangle$, is retained, and $u \in X_V(G'_2) \subseteq X(G'_2)$. Likewise if e_v is removed, then $v \in X(G'_2)$. In summary, $X(G'_2)$ contains a vertex from every edge in E, hence it is a vertex cover of G. □

The problem Minimum Vertex Cover in Cubic Graphs is APX-hard; see e.g. [1]. For a cubic graph G of n vertices and m edges, where $3n = 2m$,

the minimum size k^* of a vertex cover is $\Theta(m+n)$. By Lemma 5, the exemplar edit distance of the two sequences G_1 and G_2 in the reduced instance is also $\Theta(m+n)$. Thus by the standard technique of L-reduction, it follows that $(1,2)$-Exemplar Edit Distance, when the cost of a substitution is 1 and the cost of an insertion or a deletion is at least 1, is APX-hard too. Then the APX-hardness of $(1,2)$-Exemplar Levenshtein Distance and the APX-hardness of $(1,2)$-Exemplar Hamming Distance follow as special cases. Moreover, since the lengths of the two sequences G_1 and G_2 in the reduced instance are both $\Theta(m+n)$ as well, it follows that the complementary maximization problem $(1,2)$-Exemplar Hamming Similarity is also APX-hard, if we define the *Hamming similarity* of two sequences of the same length ℓ as ℓ minus their Hamming distance.

6 Concluding Remarks

We find it most intriguing that although the problem $(1,2)$-Exemplar Distance has been shown to be APX-hard for a wide variety of distance measures, including breakpoints, conserved intervals, common intervals, MAD, SAD, signed reversals, Levenshtein distance, Hamming distance..., no constant approximation is known for any one of these measures, while on the other hand, it seems difficult to improve the constant lower bound in any one of these APX-hardness results into a lower bound that grows with the input size similar to the logarithmic lower bound for Minimum Set Cover.

References

1. P. Alimonti and V. Kann. Some APX-completeness results for cubic graphs. *Theoretical Computer Science*, 237:123–134, 2000.
2. S. Angibaud, G. Fertin, I. Rusu, A. Thévenin, and S. Vialette. On the approximability of comparing genomes with duplicates. *Journal of Graph Algorithms and Applications*, 13:19–53, 2009.
3. P. Berman and K. Karpinski. On some tighter inapproximability results. In *Proceedings of the 26th International Colloquium on Automata, Languages and Programming*, LNCS 1644, pages 200-209, 1999.
4. G. Blin, C. Chauve, G. Fertin, R. Rizzi, and S. Vialette. Comparing genomes with duplications: a computational complexity point of view. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:523–534, 2007.
5. G. Blin, G. Fertin, F. Sikora, and S. Vialette. The Exemplar Breakpoint Distance for non-trivial genomes cannot be approximated. In *Proceedings of the 3rd Workshop on Algorithms and Computation (WALCOM'09)*, pages 357–368, 2009.
6. P. Bonizzoni, G. Della Vedova, R. Dondi, G. Fertin, R. Rizzi, and S. Vialette. Exemplar longest common subsequence. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:535–543, 2007.
7. Z. Chen, R. H. Fowler, B. Fu, and B. Zhu. On the inapproximability of the exemplar conserved interval distance problem of genomes. *Journal of Combinatorial Optimization*, 15:201–221, 2008. (A preliminary version appeared in *Proceedings*

of the 12th Annual International Conference on Computing and Combinatorics (COCOON'06), pages 245–254, 2006.)

8. Z. Chen, B. Fu, and B. Zhu. The approximability of the exemplar breakpoint distance problem. In *Proceedings of the 2nd International Conference on Algorithmic Aspects in Information and Management (AAIM'06)*, pages 291–302, 2006.
9. I. Dinur and S. Safra. On the hardness of approximating minimum vertex cover. *Annals of Mathematics*, 162:439–485, 2005.
10. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.
11. S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46:1–27, 1999.
12. M. Jiang. The zero exemplar distance problem. *Journal of Computational Biology*, 18:1077–1086, 2011.
13. S. Khot and O. Regev. Vertex cover might be hard to approximate to within $2-\epsilon$. *Journal of Computer and System Sciences*, 74:335–349, 2008.
14. D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15:909–917, 1999.
15. D. Sankoff and L. Haque. Power boosts for cluster tests. In *Proceedings of the RECOMB International Workshop on Comparative Genomics (RCG'05)*, pages 121–130, 2005.