# Lower Bounds on the Approximation of the Exemplar Conserved Interval Distance Problem of Genomes*

Zhixiang Chen[1], Richard H. Fowler[1], Bin Fu[2], and Binhai Zhu[3]

[1] Department of Computer Science, University of Texas-American, Edinburg, TX 78541-2999, USA
chen@panam.edu, fowler@cs.panam.edu

[2] Department of Computer Science, University of New Orleans, New Orleans, LA 70148 and Research Institute for Children, 200 Henry Clay Avenue, New Orleans, LA 70118, USA
fu@cs.uno.edu

[3] Department of Computer Science, Montana State University, Bozeman, MT 59717-3880, USA
bhz@cs.montana.edu

**Abstract.** In this paper we present several lower bounds on the approximation of the exemplar conserved interval distance problem of genomes. We first prove that the exemplar conserved interval distance problem cannot be approximated within a factor of $c \log n$ for some constant $c > 0$ in polynomial time, unless P=NP. We then prove that it is NP-complete to decide whether the exemplar conserved interval distance between any two sets of genomes is zero or not. This result implies that the exemplar conserved interval distance problem does not admit any approximation in polynomial time, unless P=NP. In fact, this result holds even when a gene appears in each of the given genomes at most three times. Finally, we strengthen the second result under a weaker definition of approximation (which we call weak approximation). We show that the exemplar conserved interval distance problem does not admit a weak approximation within a factor of $m$, where $m$ is the maximum length of the given genomes.

## 1 Introduction

A central problem in the genome comparison and rearrangement area is to compute the number (i.e., genetic distances) and the actual sequence of genetic operations needed to convert a source genome to a target genome. This problem originates from evolutionary molecular biology. In the past, typical genetic distances studied include edit [10], signed reversal [13,9,1] and breakpoint [17],

* This research is supported in part by FIPSE Congressional Award P116Z020159, NSF CNS-0521585, Louisiana Board of Regents under contract number LEQSF(2004-07)-RD-A-35 and MSU-Bozeman's Short-term Professional Development Leave Program.

D.Z. Chen and D.T. Lee (Eds.): COCOON 2006, LNCS 4112, pp. 245–254, 2006.

conserved interval [3,4], etc. (It was Sturtevant and Dobzhansky who came up with the idea of signed reversal and breakpoint distance, though implicitly, in 1936 [16].) Recently, conserved interval distance was also proposed to measure the similarity of multiple sequences of genes [3]. For an overview of the research performed in this area, readers are referred to [8,7] for a comprehensive survey.

Until a few years ago, in genome rearrangement research, people always assumed that each gene appears in a genome exactly once. Under this assumption, the genome rearrangement problem is essentially the problem of comparing and sorting signed/unsigned permutations [8,7]. However, this assumption is very restrictive and is only justified in several small virus genomes. For example, this assumption does not hold on eukaryotic genomes where paralogous genes exist [12,15]. Certainly, it is important in to compute genomic distances efficiently, e.g., by Hannenhalli and Pevzner's method [8], when no gene duplications arise; on the other hand, one might have to handle this gene duplication problem as well. A few years ago, Sankoff proposed a way to select, from the duplicated copies of genes, the common ancestor gene such that the distance between the reduced genomes (*exemplar genomes*) is minimized [15]. He also proposed a general branch-and-bound algorithm for the problem [15]. Recently, Nguyen, Tay and Zhang used a divide-and-conquer method to compute the exemplar breakpoint distance empirically [12]. As these problem seemed to be hard, theoretical research was followed almost immediately. It was shown that computing the signed reversals and breakpoint distances between exemplar genomes are both NP-complete [5]. Recently, Blin and Rizzi further proved that computing the conserved interval distance between exemplar genomes is NP-complete [4]; moreover, it is NP-complete to compute the minimum conserved interval matching (i.e., without deleting the duplicated copies of genes). There has been no formal theoretical results, before Nguyen [11] and our recent work [6], on the approximability of the exemplar genomic distance problems except the NP-completeness proofs [5,4]. Nguyen [11] proved that exemplar breakpoint distance cannot be approximated within constant ratio in polynomial time unless $P = NP$. Actually, the result was proved through a reduction from the set cover problem. This work was announced in [12].

In [6], we present the first set of inapproximability and approximation results for the Exemplar Breakpoint Distance problem, given two genomes each containing only one sequence of genes drawn from $n$ identical gene families. (Some of the results hold subsequently for the Exemplar Reversal Distance problem.) For the One-sided Exemplar Breakpoint Distance Problem, which is also known to be NP-complete, we obtain a factor-$2(1 + \log n)$, polynomial-time approximation. The approximation algorithm follows the greedy strategy for Set-Cover, but constructing the family of sets is non-trivial and is related to a new problem of *longest constrained common subsequences* which is related to but different from the recently studied *constrained longest common subsequences* [2].

## 2 Preliminaries

In the genome comparison and rearrangement problem, we are given a set of genomes, each of which is a signed sequence of genes. The order of the genes corresponds to their positions on the linear chromosome and the signs correspond to which of the two DNA strands the genes are located. While most of the past research are under the assumption that each gene occurs in a genome once, this assumption is problematic in reality for eukaryotic genomes or the likes where duplications of genes exist [15]. Sankoff proposed a method to select an *exemplar genome*, by deleting redundant copies of a gene, such that in an exemplar genome any gene appears exactly once; moreover, the resulting exemplar genomes should have a property that certain genetic distance between them is minimized [15].

The following definitions are very much following those in [3,4]. Given $n$ *gene families* (alphabet) $\mathcal{F}$, a genome $G$ is a sequence of elements of $\mathcal{F}$ such that each element is with a sign (+ or −). In general, we allow the repetition of a gene family in any genome. Each occurrence of a gene family is called a *gene*, though we will not try to distinguish a gene and a gene family if the context is clear. Given a genome $G = g_1 g_2 ... g_m$ with no repetition of any gene, we say that gene $g_i$ *immediately precedes* $g_j$ if $j = i + 1$. Given genomes $G$ and $H$, if gene $a$ immediately precedes $b$ in $G$ but neither $a$ immediately precedes $b$ nor $-b$ immediately precedes $-a$ in $H$, then they constitute a *breakpoint* in $G$. The *breakpoint distance* is the number of breakpoints in $G$ (symmetrically, it is the number of breakpoints in $H$).

The number of a gene $g$ appearing in a genome $G$ is called the cardinality of $g$ in $G$, written as $card(g, G)$. A gene in $G$ is called *trivial* if $g$ has cardinality exactly 1; otherwise, it is called *non-trivial*. In this paper, we assume that all the genomes we discuss could contain both trivial and non-trivial genes. A genome $G$ is called *r-repetitive*, if all the genes from the same gene family appear at most $r$ times in $G$. A genome $G$ is called a *k-span* genome, if all the genes from the same gene family are within distance at most $k$ in $G$. For example, $G = -adc - bdaeb$ is 2-repetitive and it is a 5-span genome.

Given a genome $G = g_1 g_2 \cdots g_m$, an interval $[g_i, g_j]$ is simply the substring $g_i g_{i+1} \cdots g_j$ (which will also be denoted as $G[i, j]$). For example, given $G' = bdc - ag - e - fh, G'' = bdce - gafh$, between the two intervals $I_1 = dc - ag - e - f$ and $I_2 = dce - gaf$, there are 2 breakpoints $c - a$ and $-e - f$. A *signed reversal* on a genome $G$ simply reverses the order and signs of all the elements in an interval of $G$. In the previous example, if a signed reversal operation is conducted in $I_1$ on $G'$, then we obtain a new genome $G^* = bfe - ga - c - dh$. (All the reversals concerned in this paper are signed reversals. Henceforth, we simply use *reversal* to make the presentation simpler.) The *reversal distance* between genomes $G$ and $H$ is the minimum number of reversals to transfer $G$ into $H$.

Given a genome $G$ over $\mathcal{F}$, an *exemplar genome* of $G$ is a genome $G'$ obtained from $G$ by deleting duplicating genes such that each gene family in $G$ appears exactly once in $G'$. For example, let $G = bcaadagef$ there are two exemplar genomes: $bcadgef$ and $bcdagef$.

Given a set of genomes $\mathcal{G}$ and two gene families $a, b \in \mathcal{F}$, an interval $[a, b]$ is a *conserved interval* of $\mathcal{G}$ if (1) $a$ precedes $b$ or $-b$ precedes $-a$ in any genome in $\mathcal{G}$; and (2) the set of unsigned genes (i.e., ignoring signs) between $a$ and $b$ are the same for all genomes in $\mathcal{G}$. Let $\mathcal{G} = \{G_1, G_2\}$, where $G_1 = bc-ag-e-fdh, G_2 = b-ce-gaf-dh$, there are three conserved intervals between $G_1$ and $G_2$: $[e, a], [b, h]$ and $[-a, g]$.

Given two sets of genomes $\mathcal{G}$ and $\mathcal{H}$, the *conserved interval distance* between $\mathcal{G}$ and $\mathcal{H}$ is defined as

$$d(\mathcal{G}, \mathcal{H}) = N_{\mathcal{G}} + N_{\mathcal{H}} - 2N_{\mathcal{G}\cup\mathcal{H}},$$

where $N_{\mathcal{G}}$ (resp. $N_{\mathcal{H}}$ and $N_{\mathcal{G}\cup\mathcal{H}}$) is the number of conserved intervals in $\mathcal{G}$ (resp. $\mathcal{H}$ and $\mathcal{G} \cup \mathcal{H}$). Continuing the example in the previous paragraph, let $\mathcal{H} = \{H_1, H_2\}$, where $H_1 = b - cg - af - edh, H_2 = bagcdefh$, then there are two conserved intervals between $H_1$ and $H_2$: $[b, h]$ and $[a, c]$. There is only one conserved interval in $\mathcal{G} \cup \mathcal{H}$: $[b, h]$. Therefore, $d(\mathcal{G}, \mathcal{H}) = 3 + 2 - 2 \times 1 = 3$.

If $\mathcal{G}$ and $\mathcal{H}$ are both a singleton, i.e., $\mathcal{G}$ contains only a genome $G$, and $\mathcal{H}$ contains only a genome $H$, then we simply use the notation $d(G, H) = N_G + N_H - 2N_{G\cup H}$ to stand for $d(\mathcal{G}, \mathcal{H})$. Note that when only one genome $G$ is considered, every interval in $G$ is a conserved interval. This implies that when $G$ (resp. $H$) has $n$ trivial genes, then $d(G, H) = 2\binom{n}{2} - 2N_{G\cup H}$.

The Exemplar Conserved Interval Distance Problem, denoted as the *ECID problem*, is defined as follows:

**Instance:** Two sets of genomes $\mathcal{G}$ and $\mathcal{H}$, each genome is of length $O(m)$ and covers $n$ identical gene families (i.e., it contains at least one gene from each of the $n$ gene families); an integer $K$.
**Question:** Are there respective exemplar genomes $\mathcal{G}^*$ of $\mathcal{G}$ and $\mathcal{H}^*$ of $\mathcal{H}$, such that the conserved interval distance distance $d(\mathcal{G}^*, \mathcal{H}^*)$ is at most $K$?

In the next three sections, we present lower bounds on the approximation of the optimization version of the ECID problem, namely, to compute or approximate the minimum value $K$ in the above formulation. Given a minimization problem $\Pi$, let the optimal solution of $\Pi$ be $OPT$. We say that an approximation algorithm $\mathcal{A}$ provides a *performance guarantee* of $\alpha$ for $\Pi$ if for every instance $I$ of $\Pi$, the solution value returned by $\mathcal{A}$ is at most $\alpha \times OPT$. (Usually we say that $\mathcal{A}$ is a factor-$\alpha$ approximation for $\Pi$.) Typically we are interested in polynomial time approximation algorithms.

In many biological problems, the optimal solution value $OPT$ could be zero. (For example, in some minimum recombination haplotype reconstruction problems the optimal solution could be zero.) In that case, if computing such a zero optimal solution value is NP-complete then the problem does not admit *any* approximation unless P=NP. However, in reality one would be happy to obtain a solution with value one or two. Due to this reason, we relax the above (traditional) definition of approximation to a *weak approximation*. Given a minimization problem $\Pi$, let the optimal solution of $\Pi$ be $OPT$. We say that a weak approximation algorithm $B$ provides a *performance guarantee* of $\alpha$ for

$\Pi$ if for every instance $I$ of $\Pi$, the solution value returned by $B$ is at most $\alpha \times (OPT + 1)$.

## 3 A $c \log n$ Lower Bound on Approximating ECID

**Theorem 1.** *It is NP-complete to approximate the Exemplar Conserved Interval Distance problem within a factor of $c \log n$ for some constant $c > 0$.*

*Proof.* We use a reduction from the Dominating Set problem to the ECID problem for two sets of genomes $\mathcal{G} = \{G_1, G_2\}$ and $\mathcal{H} = \{H_1, H_2\}$ that will be constructed from the given graph.

Let $T = (V, E)$ be any given graph with $V = \{v_1, v_2, \cdots, v_n\}$ and $E = \{e_1, e_2, \cdots, e_m\}$. We assume that vertices and edges in $T$ are sorted by their corresponding indices. We construct two sets of genomes $\mathcal{G} = \{G_1, G_2\}$ and $\mathcal{H} = \{H_1, H_2\}$ as follows. For each $v_i \in V$, we have four corresponding genes $v_i^1, v_i^2, v_i^3$, and $v_i^4$. We enforce a rule that $v_i^k$ is incident to $v_j^l$ if and only if $k = l$ and $v_i$ is incident to $v_j$ in $T$. Let $B_i^j$ be the sorted sequence of vertices incident to $v_i^j$ and $\mathcal{B}_i^j$ be the unsigned reversal of $B_i^j$. ("|" is not a gene and is used for readability purpose.)

$$
\begin{aligned}
G_1 &= v_1^1 B_1^1 v_1^3 | v_1^2 B_1^2 v_1^4 | \cdots | v_{n-1}^1 B_{n-1}^1 v_{n-1}^3 | v_{n-1}^2 B_{n-1}^2 v_{n-1}^4 | v_n^1 B_n^1 v_n^3 | v_n^2 B_n^2 v_n^4 \\
G_2 &= -v_1^3 \mathcal{B}_1^1 - v_1^1 | - v_1^4 \mathcal{B}_1^2 - v_1^2 | \cdots | - v_{n-1}^3 \mathcal{B}_{n-1}^1 - v_{n-1}^1 | - v_{n-1}^4 \mathcal{B}_{n-1}^2 - v_{n-1}^2 | \\
&\quad -v_n^3 B_n^1 - v_n^1 | - v_n^4 \mathcal{B}_n^2 - v_n^2 \\
H_1 &= v_1^1 B_1^1 v_1^3 | \cdots | v_{n-1}^1 B_{n-1}^1 v_{n-1}^3 | v_n^1 B_n^1 v_n^3 | v_1^2 - v_1^4 | \cdots | v_{n-1}^2 - v_{n-1}^4 | v_n^2 - v_n^4 \\
H_2 &= -v_1^3 \mathcal{B}_1^1 - v_1^1 | \cdots | - v_{n-1}^3 \mathcal{B}_{n-1}^1 - v_{n-1}^1 | - v_n^3 \mathcal{B}_n^1 - v_n^1 | \\
&\quad -v_n^4 v_n^2 | - v_{n-1}^4 v_{n-1}^2 | \cdots | - v_1^4 v_1^2
\end{aligned}
$$

Fig. 1 shows a simple graph with six vertices $v_1, v_2, \ldots, v_6$. The corresponding genomes for this graph are given as follows.

$$
\begin{aligned}
G_1 &= v_1^1 v_3^1 v_1^3 | v_1^2 v_3^2 v_1^4 | v_2^1 v_3^1 v_2^3 | v_2^2 v_3^2 v_2^4 | v_3^1 v_1^1 v_2^1 v_5^1 v_3^3 | v_3^2 v_1^2 v_2^2 v_5^2 v_3^4 | \\
&\quad v_4^1 v_5^1 v_6^1 v_4^3 | v_4^2 v_5^2 v_6^2 v_4^4 | v_5^1 v_3^1 v_4^1 v_6^1 v_5^3 | v_5^2 v_3^2 v_4^2 v_6^2 v_5^4 | v_6^1 v_4^1 v_5^1 v_6^3 | v_6^2 v_4^2 v_5^2 v_6^4 | \\
G_2 &= -v_1^3 v_3^1 - v_1^1 | - v_1^4 v_3^2 - v_1^2 | - v_2^3 v_3^1 - v_2^1 | - v_2^4 v_3^2 - v_2^2 | \\
&\quad -v_3^3 v_5^1 v_2^1 v_1^1 - v_3^1 | - v_3^4 v_5^2 v_2^2 v_1^2 - v_3^2 | - v_4^3 v_6^1 v_5^1 - v_4^1 | - v_4^4 v_6^2 v_5^2 - v_4^2 | \\
&\quad -v_5^3 v_6^1 v_4^1 v_3^1 - v_5^1 | - v_5^4 v_6^2 v_4^2 v_3^2 - v_5^2 | - v_6^3 v_5^1 v_4^1 - v_6^1 | - v_6^4 v_5^2 v_4^2 - v_6^2 | \\
H_1 &= v_1^1 v_3^1 v_1^3 | v_2^1 v_3^1 v_2^3 | v_3^1 v_1^1 v_2^1 v_5^1 v_3^3 | v_4^1 v_5^1 v_6^1 v_4^3 | v_5^1 v_3^1 v_4^1 v_6^1 v_5^3 | v_6^1 v_4^1 v_5^1 v_6^3 | \\
&\quad v_1^2 - v_1^4 | v_2^2 - v_2^4 | v_3^2 - v_3^4 | v_4^2 - v_4^4 | v_5^2 - v_5^4 | v_6^2 - v_6^4 \\
H_2 &= -v_1^3 v_3^1 - v_1^1 | - v_2^3 v_3^1 - v_2^1 | - v_3^3 v_5^1 v_2^1 v_1^1 - v_3^1 | \\
&\quad -v_4^3 v_6^1 v_5^1 - v_4^1 | - v_5^3 v_6^1 v_4^1 v_3^1 - v_5^1 | - v_6^3 v_5^1 v_4^1 - v_6^1 | \\
&\quad -v_6^4 v_6^2 | - v_5^4 v_5^2 | - v_4^4 v_4^2 | - v_3^4 v_3^2 | - v_2^4 v_2^2 | - v_1^4 v_1^2
\end{aligned}
$$

**Claim A.** The given graph $T$ has a dominating set of size $K$ if and only if there are exemplar genomes $g_i$ for $G_i$ and $h_i$ for $H_i$ for $i = 1, 2$, such that we have, letting $\mathcal{G}^* = \{g_1, g_2\}$ and $\mathcal{H}^* = \{h_1, h_2\}$,

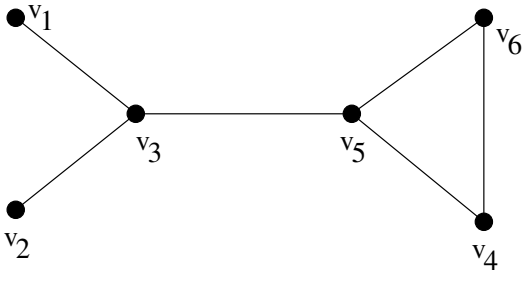


**Fig. 1.** Illustration of a simple graph for the reduction

(1) $N_{\mathcal{G}^*} = 2K$
(2) $N_{\mathcal{H}^*} = K$
(3) $N_{\mathcal{G}^* \cup \mathcal{H}^*} = K$

Note that (1), (2) and (3) together imply $d(\mathcal{G}^*, \mathcal{H}^*) = K$.

We now prove the above claim. The "only if part" is easy. We only show the proof for (2) as the proofs for (1) and (3) would be similar. If $T$ has a dominating set of size $K$, then for all those $v_j$ which is not in the dominating set we delete all $B_j^1$ in $H_1$ and $\mathcal{B}_j^1$ in $H_2$. For those remaining $B_i^1$ in $H_1$ and $\mathcal{B}_i^1$ in $H_2$, we delete the duplications (say $v_l^1$) consistently in $H_1$ and $H_2$. It is easy to see that the conserved intervals we have are $[v_i^1, v_i^3]$ in $H_1$ and $[-v_i^3, -v_i^1]$ in $H_2$, which correspond to the vertices in the dominating set.

The "if part" is slightly more tricky. Assume that (1), (2) and (3) are all true. In this case, we only need to focus on (2), i.e., $N_{\mathcal{H}^*} = K$. First, notice that the second half of $H_1$ and $H_2$ (i.e., those involved with $v_j^2 - v_j^4$ or their unsigned reversals) will not contribute anything to the number of conserved intervals in $\mathcal{H}^*$. Notice also that only these $[v_i^1, v_i^3]$ from the first half of $H_1$ can possibly form conserved intervals with the corresponding $[-v_i^3, -v_i^1]$ from the first half of $H_2$. If the number of conserved intervals in $\mathcal{H}^*$ is $K$, then those $K$ conserved intervals must come from $[v_i^1, v_i^3]$ in $h_1$ and $[-v_i^3, v_i^1]$ in $h_2$. Moreover, if there is any deletion in $B_i^1$ in $H_1$ and in $\mathcal{B}_i^1$ in $H_2$, then the deletion has to be consistent. If $v_j^1$ appears in $B_i^1$ and $\mathcal{B}_i^1$, then unless it appears in $B_l^1$ and $\mathcal{B}_l^1$ we must keep them to avoid extra conserved intervals in the form of $[v_j^1, v_j^3]$ in $H_1$ and $[-v_j^3, -v_j^1]$ in $H_2$. Therefore, from the $K$ conserved intervals in $\mathcal{H}^*$ we can construct the $K$ vertices which form the dominating set for $T$.

Let $opt(T)$ denote the size of the minimum dominating set of the graph $T$, and let $opt(\mathcal{G}, \mathcal{H})$ denote the minimum exemplar conserved interval distance between $\mathcal{G}$ and $\mathcal{H}$. It follows from Claim A that $opt(T) = opt(\mathcal{G}, \mathcal{H})$. The size of $T$ is $|V| + |E| = n + m$. It is easy to see that the size of $\mathcal{G}$ and $\mathcal{H}$ is at most $8(n + m)$. Raz and Safra [14] proved that the Dominating Set Problem cannot be approximated within a factor of $c_1 \log(n + m)$ from some constant $c_1 > 0$. Let $c = c_1/4$. If there is an algorithm that can approximate the exemplar conserved interval distance problem within a factor of $c_1 \log(|\mathcal{G}| + |\mathcal{H}|)$, where

$|\mathcal{G}|$ (resp. $|\mathcal{H}|$) denotes size of $\mathcal{G}$ (resp. $\mathcal{H}$), i.e., the number of genes in it. Then, this algorithm can be used to solve the Dominating Set Problem: the returned exemplar conserved interval distance for $opt(\mathcal{G}, \mathcal{H})$ is also for $opt(T)$. Let $app(T)$, which is $app(\mathcal{G}, \mathcal{H})$, denote the result returned by the algorithm. Then, we have

$$\begin{aligned} app(T) = app(\mathcal{G}, \mathcal{H}) &\leq c \log(8(n+m)) opt(\mathcal{G}, \mathcal{H}) = c \log(8(n+m)) opt(T) \\ &\leq 4c \log |T| opt(T) = c_1 \log |T| opt(T) \end{aligned}$$

Hence, $opt(T)$ can be approximated within a factor of $c_1 \log |T|$, a contradiction to the result obtained by Raz and Safra [14]. Therefore, the exemplar conserved interval distance problem cannot be approximated with a factor of $c \log(|\mathcal{G}|+|\mathcal{H}|)$ for a constant $c > 0$. □

## 4 The Zero Exemplar Conserved Interval Distance Problem

Recently, Chen, Fu and Zhu proved in [6] that the zero exemplar breakpoint distance problem is NP-complete. Following the spirit of [6], in this section we shall consider the *zero exemplar conserved interval distance problem*, i.e., the problem of deciding whether the exemplar conserved interval distance between any two given sets of genomes $\mathcal{G}$ and $\mathcal{H}$ is zero or not. We shall show that this problem, like the zero exemplar breakpoint distance problem, is also NP-complete.

**Lemma 1.** *Let $G$ and $H$ be two genomes such that each has $n$ trivial genes and the set of genes in $G$ is the same as the set of genes in $H$. (In other word, $G$ is a signed permutation of $H$.) Then, the conserved interval distance between $G$ and $H$ is zero, i.e., $d(G, H) = 0$, if and only if either $G = H$ or $G$ is the signed reversal of $H$.*

*Proof.* It follows from the given condition that $d(G, H) = 2\binom{n}{2} - 2N_{G \cup H}$. If $G = H$ or $G$ is a signed reversal of $H$, then every two genes in $G$ form a conserved interval in $G$ and $H$. Thus, $N_{G \cup H} = \binom{n}{2}$. This implies $d(G, H) = 0$.

Now, suppose $d(G, H) = 0$. Then, we have $N_{G \cup H} = \binom{n}{2}$, i.e., every two genes in $G$ form a conserved interval in $G$ and $H$. We can prove by induction on $n$ that either $G = H$ or $G$ is the singed reversal of $H$. The details are omitted due to space limit. □

**Theorem 2.** *Given any two genomes $G$ and $H$ which are both 3-repetitive, it is NP-complete to decide whether the exemplar conserved interval distance between $G$ and $H$ is zero or not.*

*Proof.* It is easy to see that this ZECID problem is in NP. To prove its NP-hardness, we will construct a reduction from the 3SAT problem to the ZECID problem, following the reduction for proving the NP-hardness for the zero breakpoint distance problem in [6].

Let $F = f_1 \bigwedge f_2 \bigwedge \cdots \bigwedge f_q$ be a conjunctive normal form, where each $f_i$ is a 3-disjunctive clause like $(x_1 \bigvee x_4 \bigvee \neg x_7)$. We construct two genomes $G$ and $H$ such that $F$ is satisfiable iff $G$ and $H$ have zero exemplar conserved interval distance.

We consider $f_i, 1 \leq i \leq q$, as names of genes. Assume that $F$ has $n$ boolean variables $x_i, 1 \leq i \leq n$. Let $G = S_1 g_1 S_2 g_2 \cdots g_{n-1} S_n$ and $H = S_1^* g_1 S_2^* g_2 \cdots g_{n-1} S_n^*$, where $g_1, \cdots, g_{n-1}$ are peg genes that occur only once in $G$ or $H$. For $1 \leq i \leq n$, $S_i = f_{i_1} \cdots f_{i_u} f_{j_1} \cdots f_{j_v}$ and $S_i^* = f_{j_1} \cdots f_{j_v} f_{i_1} \cdots f_{i_u}$, where $f_{i_1}, \cdots, f_{i_u}$ are the clauses containing $x_i$, and $f_{j_1}, \cdots, f_{j_v}$ are the clauses containing $\neg x_i$. Since each clause has at most 3 literals, $S$ and $H$ are 3-repetitive.

Following the approach in [6], if F is satisfiable, then we have an exemplar genomes $G'$ and $H'$ such that $G' = H'$. Hence, by Lemma 1 we have $d(G', H') = 0$. If there are two exemplar genomes $G''$ and $H''$ such that $d(G'', H'') = 0$, then by Lemma 1 we have $G'' = H''$, because $G''$ and $H''$ contain all unsigned genes in the set $\{f_1, \cdots, f_q, g_1, \cdots, g_{n-1}\}$ and no genes are repetitive. If $S_i$ becomes empty in $G''$ then we can assign a value to $x_i$ arbitrarily. Otherwise, we assign $x_i = 1$ if it becomes a subsequence of $f_{i_1} \cdots f_{i_u}$ in $G''$, or we assign $x_i = 0$ if it becomes a subsequence of $f_{j_1} \cdots f_{j_v}$. It is easy to verify that such a truth assignment will make $F$ true. □

Example. $F = (x_1 \vee \neg x_2 \vee x_4) \bigwedge (\neg x_1 \vee x_3 \vee x_4) \bigwedge (x_2 \vee x_3 \vee \neg x_4) \bigwedge (\neg x_1 \vee \neg x_2 \vee \neg x_3)$, where $F_1 = (x_1 \vee \neg x_2 \vee x_4)$, $F_2 = (\neg x_1 \vee x_3 \vee x_4)$, $F_3 = (x_2 \vee x_3 \vee \neg x_4)$, and $F_4 = (\neg x_1 \vee \neg x_2 \vee \neg x_3)$.
$G = F_1 F_2 F_4 g_1 F_3 F_1 F_4 g_2 F_2 F_3 F_4 g_3 F_1 F_2 F_3$ and
$H = F_2 F_4 F_1 g_1 F_1 F_4 F_3 g_2 F_4 F_2 F_3 g_3 F_3 F_1 F_2$.
$d(G", H") = 0$, with $G" = H" = F_4 g_1 F_3 g_2 g_3 F_1 F_2$, corresponds to the truth assignment that $x_1 = \texttt{False(0)}$, $x_3 = \texttt{False(0)}$ or $\texttt{True(1)}$, and $x_2 = x_4 = \texttt{True(1)}$.

**Corollary 1.** *Given any two sets of genomes $\mathcal{G}$ and $\mathcal{H}$, it is NP-complete to decide whether the exemplar conserved interval distance between $\mathcal{G}$ and $\mathcal{H}$ is zero or not.*

Theorem 2 and the above corollary imply that the ECID problem does not admit any approximation unless P=NP—if such a polynomial-time approximation existed then it would be able to decide whether $G$ and $H$ have zero exemplar conserved interval distance in polynomial time hence contradicting Theorem 2.

## 5 Weak Inapproximability Bound

Let $opt(\mathcal{G}, \mathcal{H})$ be the optimal exemplar conserved interval distance between $\mathcal{G}$ and $\mathcal{H}$. We also use $d(X, Y)$ to denote the minimum conserved interval distance between two genomes $X$ and $Y$, where $X$ and $Y$ do not have to be exemplar. We also adopt a similar approach as in [6] but with some more involved analysis. We obtain the following inapproximability bounds under a much weaker model of approximation. Notice that the $m$ factor in the bounds here are stronger than the $m^{1-\epsilon}$ factor in the bounds in [6] for exemplar break point distance problem.

**Theorem 3.** *Let $g(x) : N \to N$ be a function computable in polynomial time. If there is a polynomial time algorithm such that given two genomes $\mathcal{G}$ and $\mathcal{H}$ of length at most $m$ it can return exemplar genomes $G$ and $H$ satisfying $d(G, H) \leq g(m) opt(\mathcal{G}, \mathcal{H}) + m$, then P=NP.*

*Proof.* Let $f$ be a given CNF formula. Let $G(f), H(f)$ be the genomes as constructed in Theorem 2 such that $f$ is satisfiable if and only if $d(G(f), H(f)) = 0$. Let $|G(f)| = |H(f)| = u$, i.e., the number of all the genes occurred in $G(f)$ (or $H(f)$). Let $\Sigma(S)$ be the alphabet of a sequence $S$. If $\Sigma_i$ is a different set of letters with $|\Sigma_i| = |\Sigma(S)|$, we define $S(\Sigma_i)$ to be a new sequence obtained by replacing all letters in $S$, in one to one fashion, by those in $\Sigma_i$.

For $M \geq 1$, Let $\Sigma_1, \Sigma_2, \cdots, \Sigma_M$ be $M$ disjoint sets of letters of size $|\Sigma(G(f))|$. Let $G_1 = G(f)(\Sigma_1), G_2 = G(f)(\Sigma_2), \cdots, G_M = G(f)(\Sigma_M)$ be the sequences derived from $G(f)$. Let $H_1 = H(f)(\Sigma_1), H_2 = H(f)(\Sigma_2), \cdots, H_M = H(f)(\Sigma_M)$ be the sequences derived from $H(f)$.

Define $\mathcal{G} = G_1 s_1 G_2 s_2 \cdots G_M s_M$ and $\mathcal{H} = H_1 s_1 H_2 s_2 \cdots H_M s_M$, where $s_i$ is a peg gene appearing only once in $\mathcal{G}$ and $\mathcal{H}$, respectively. Let $m = |\mathcal{G}| = |\mathcal{H}|$. In fact, $m$ is the number of all the genes in $\mathcal{G}$ (or $\mathcal{H}$).

Assume that some polynomial time algorithm $\mathcal{A}$ outputs respectively two exemplar genomes $G$ and $H$ of $\mathcal{G}$ and $H$, and $d(G, H) \leq g(m)d(\mathcal{G}, \mathcal{H}) + m$, we can then decide if $f$ is satisfiable by checking whether $d(G, H) \leq m$. If $f$ is satisfiable, as in the proof of Theorem 2, two identical exemplar genomes can be obtained from $\mathcal{G}$ and $\mathcal{H}$. Hence, we have $d(\mathcal{G}, \mathcal{H}) = 0$ by Lemma 1. This implies that $d(G, H) \leq m$. If $f$ is not satisfiable, then from Theorem 2, $d(G_i, H_i) \geq 1$; namely, there is at least one conserved interval in $G_i$ but not in $H_i$. This implies one of the following is true: (1) $a \cdots b$ in $G_i$ but $b \cdots a$ in $H_i$; (2) $c \in [a, b]$ in $G_i$ but $c \notin [a, b]$ in $H_i$; and (3) $c \notin [a, b]$ in $G_i$ but $c \in [a, b]$ in $H_i$. For case (1), for any $d$ in $G_j s_j$, $j \neq i$, either $[a, d]$ or $[d, a]$is a conserved interval in $G_j$ but not in $H_j$. Similarly, for any $e$ in $H_j s_j$, $j \neq i$, either $[a, e]$ or $[e, a]$is a conserved interval in $H_j$ but not in $G_j$. Thus, in this case, we have at least $(u + 1)(M - 1)$ conserved interval in either $\mathcal{G}$ or $\mathcal{H}$ but not in both. Hence, we have $d(\mathcal{G}, \mathcal{H}) \geq 2(u + 1)(M - 1)$. It follows from some similar analysis that $d(\mathcal{G}, \mathcal{H}) \geq 2(u+1)(M-1)$ is true for the other two cases. Therefore, in either of the three cases, when $M \geq 2$, we have $d(\mathcal{G}, \mathcal{H}) \geq 2(u+1)(M-1) > (u+1)M = m$. Since $G, H$ are exemplar genomes of $\mathcal{G}$ and $\mathcal{H}$, we have $d(G, H) > m$. □

**Corollary 2.** *If there is a polynomial time algorithm such that given $\mathcal{G}$ and $\mathcal{H}$ of length at most $m$ it can return exemplar genomes $G$ and $H$ satisfying $d(G, H) \leq m[opt(\mathcal{G}, \mathcal{H}) + 1]$, then P=NP.*

This negative result shows that even under a much weaker model, unless P=NP, it is not possible to obtain a good approximation to the optimal exemplar conserved interval distance problem.

## 6 Concluding Remarks

We prove several lower bounds on the approximation of the Exemplar Conserved Interval Distance problem. Although it seems that the general problem does not admit any approximation, good approximation may exist for special cases of genomes, and good heuristics may perform well empirically or on average. It

would be interesting to study some meaningful special cases. For example, in real-world datasets repetitions of genes are typically pegged and not very far away [12]. Are these cases easier to solve/approximate?

## References


1. V. Bafna and P. Pevzner, Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome, *Mol. Bio. Evol.*, 12:239-246, 1995.
2. S. Bereg and B. Zhu. RNA multiple structural alignment with longest common subsequences. *Proc. 11th Intl. Ann. Comput. and Combinatorics (COCOON'05),* LNCS 3595, pp. 32-41, 2005.
3. A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. *Proc. 9th Intl. Ann. Comput. and Combinatorics (COCOON'03),* LNCS 2697, pp. 68-79, 2003.
4. G. Blin and R. Rizzi. Conserved interval distance computation between non-trivial genomes. *Proc. 11th Intl. Ann. Comput. and Combinatorics (COCOON'05),* LNCS 3595, pp. 22-31, 2005.
5. D. Bryant. The complexity of calculating exemplar distances. *In D. Sankoff and J. Nadeau, editors*, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, pp. 207-212. Kluwer Acad. Pub., 2000.
6. Z. Chen, B. Fu and B. Zhu, The approximability of the exemplar breakpoint distance problem, Proceedings of the Second Intl. Conf. Algorithmic Aspects in Information and Management (AAIM'06), LNCS 4041, pp. 291-302. Springer, 2006.
7. O. Gascuel, editor. *Mathematics of Evolution and Phylogeny.*Oxford University Press, 2004.
8. S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM,* **46**(1):1-27, 1999.
9. C. Makaroff and J. Palmer. Mitochondrial DNA rearrangements and transcriptional alternatives in the male sterile cytoplasm of Ogura radish. *Mol. Cell. Biol.*, **8**:1474-1480, 1988.
10. M. Marron, K. Swenson and B. Moret. Genomic distances under deletions and insertions. *Theoretical Computer Science,* **325**(3):347-360, 2004.
11. C.T. Nguyen, Algorithms for calculating exemplar distances, Honors Thesis, School of Computing, National University of Singapore, 2005.
12. C.T. Nguyen, Y.C. Tay and L. Zhang. Divide-and-conquer approach for the exemplar breakpoint distance. *Bioinformatics*, **21**(10):2171-2176, 2005.
13. J. Palmer and L. Herbon. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol. Evolut.*, **27**:87-97, 1988.
14. R. Raz and S. Safra. A sub-constant error-probability low-degree test, and subconstant error-probability PCP characterization of NP. In *Proc. 29th ACM Symp. on Theory Comput. (STOC'97)*, pages 475-484, 1997.
15. D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, **16**(11):909-917, 1999.
16. A. Sturtevant and T. Dobzhansky. Inversions in the third chromosome of wild races of *drosophila pseudoobscura*, and their use in the study of the history of the species. *Proc. Nat. Acad. Sci. USA,* 22:448-450, 1936.
17. G. Watterson, W. Ewens, T. Hall and A. Morgan. The chromosome inversion problem. *J. Theoretical Biology*, **99**:1-7, 1982.