J Comb Optim (2008) 15: 201–221
DOI 10.1007/s10878-007-9077-1

On the inapproximability of the exemplar conserved interval distance problem of genomes

Zhixiang Chen · Richard H. Fowler · Bin Fu · Binhai Zhu

Published online: 30 May 2007

Abstract In this paper we present two main results about the inapproximability of the exemplar conserved interval distance problem of genomes. First, we prove that it is NP-complete to decide whether the exemplar conserved interval distance between any two 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 every gene appears in each of the given genomes at most three times. Second, we strengthen the first result under a weaker definition of approximation, called weak approximation. We show that the exemplar conserved interval distance problem does not admit any weak approximation within a super-linear factor of $\frac{2}{7}m^{1.5}$, where m is the maximal length of the given genomes. We also investigate polynomial time algorithms for solving the exemplar conserved interval distance problem when certain constrains are given. We prove that the zero exemplar conserved interval distance problem of two genomes is decidable in polynomial time when one genome is $O(\log n)$-spanned. We also prove that one can solve the constant-sized exemplar conserved interval distance problem in polynomial time, provided that one genome is trivial.

Keywords Genome rearrangement · Exemplar conserved interval distance · Approximation algorithm · Inapproximability · Weak approximation

Z. Chen (✉) · R.H. Fowler · B. Fu
Department of Computer Science, University of Texas-Pan American, Edinburg, TX 78541-2999, USA
e-mail: chen@cs.panam.edu

R.H. Fowler
e-mail: fowler@cs.panam.edu

B. Fu
e-mail: binfu@cs.panam.edu

B. Zhu
Department of Computer Science, Montana State University, Bozeman, MT 59717-3880, USA
e-mail: bhz@cs.montana.edu

1 Introduction

Genome rearrangement was pioneered by Sturtevant and Dobzhansky (1936) in 1936. 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 (Marron et al. 2004), signed reversal (Palmer and Herbon 1988; Makaroff and Palmer 1988; Bafna and Pevzner 1995) and breakpoint (Watterson et al. 1982) distances. Recently, conserved interval distance was also proposed to measure the similarity of multiple sequences of genes (Bergeron and Stoye 2003; Blin and Rizzi 2005). For an overview of the research performed in this area, readers are referred to (Li et al. 2006; Hannenhalli and Pevzner 1999; Gascuel 2004) 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 (Hannenhalli and Pevzner 1999; Gascuel 2004). 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 (Sankoff 1999; Nguyen et al. 2005; Nguyen 2005). Certainly, it is important in practice to compute genomic distances efficiently, e.g., by Hannenhalli and Pevzner's method (Hannenhalli and Pevzner 1999), 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 (Sankoff 1999). He also proposed a general branch-and-bound algorithm for the problem (Sankoff 1999). Recently, Nguyen, Tay and Zhang used a divide-and-conquer method to compute the exemplar breakpoint distance empirically (Nguyen et al. 2005). As these problems seemed to be hard, theoretical research followed almost immediately. It was shown that computing signed reversals and breakpoint distances between exemplar genomes are both NP-complete (Bryant 2000). Recently, Blin and Rizzi further proved that computing conserved interval distances between exemplar genomes is NP-complete (Blin and Rizzi 2005); 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 (2005) and our recent work (Chen et al. 2006a, 2006b), on the approximability of the exemplar genomic distance problems except the NP-completeness proofs (Bryant 2000; Blin and Rizzi 2005). Nguyen (2005) proved that the exemplar breakpoint distance problem cannot be approximated within a constant ratio in polynomial time unless $P = NP$. Actually, this result was proved through a reduction from the set cover problem. This work was announced in (Nguyen et al. 2005).

In (Chen et al. 2006b), 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 the set cover problem, 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* (Bereg and Zhu 2005). More recently, non-breaking similarity of genomes has been studied in (Chen et al. 2007). It was proved in this paper that the exemplar non-breaking similarity problem of genomes does not admit any $n^{1-\varepsilon}$ factor approximation, unless $P = NP$. Polynomial time algorithms were also obtained for several practically interesting cases of the problem.

In this paper, we study the inapproximability of the exemplar conserved interval distance problem of genomes. we first prove that deciding zero exemplar conserved interval distance between two genomes is NP-complete. 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 every gene appears in each of the given genomes at most three times. This result significantly improves the NP-completeness result obtained by Blin and Rizzi (2005). We then strengthen the first result under a weaker definition of approximation (which we call weak approximation). We show that the exemplar conserved interval distance problem does not admit any weak approximation within a super-linear factor of $\frac{2}{7}m^{1.5}$, where m is the maximal length of the given genomes. We also investigate polynomial time algorithms for solving the exemplar conserved interval distance problem when certain constraints are given. We prove that the zero exemplar conserved interval distance problem of two genomes is decidable in polynomial time when one genome is $O(\log n)$-spanned. Blin and Rizzi (2005) proved that the exemplar conserved interval distance problem of two genomes is NP-complete, even when one genome is trivial. We prove that one can find the exemplar conserved interval distance between two genomes in polynomial time, provided that one genome is trivial and the distance between the two is a constant that is not known beforehand.

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. Most of the past research are under the assumption that each gene occurs in a genome once. This assumption is not usually fulfilled for eukaryotic genomes or the likes where duplications of genes exist (Sankoff 1999). 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 (Sankoff 1999).

The following definitions are very much following those in (Bergeron and Stoye 2003; Blin and Rizzi 2005). Given a set of *gene families* (alphabet) $\mathcal{F}$, a genome G is a sequence of elements, called genes, of $\mathcal{F}$ such that each element is provided with

a sign (+ or −). In general, we allow duplicated genes to be present 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. For convenience, for any gene g in a genome, we let g stand for $+g$. Given a genome $G = g_1 g_2 \cdots g_m$ with no duplication 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* between G and H is the number of breakpoints in G (symmetrically, it is the number of breakpoints in H).

Let G be a genome built over a set of gene families $\mathcal{F}$. Given a gene family $f \in \mathcal{F}$, the number of occurrences of f in G is called the cardinality of f in G and is denoted by $\text{card}(f, G)$. A gene family of cardinality equal to (resp. greater than) one in G is called trivial (resp. non-trivial) in G. For commodity, a gene belonging to a trivial (resp. non-trivial) gene family will also be called trivial (resp. non-trivial). A genome G is called trivial if every gene in G is trivial.

Moreover, a genome G is said to be k-duplicated if $\max_{f \in \mathcal{F}} \text{card}(f, G) = k$. If all the genes of a given gene family $f \in \mathcal{F}$ are pair-wise distant of at most k positions in G then G is said to be a k-span genome, or k-spanned. For example, the following genome $G = -adc\text{-}bdaeb$ is 2-duplicated and it is a 5-span genome.

Given a genome $G = g_1 g_2 \cdots g_m$, for $1 \le i \le m$, let $G[i]$ denote the gene g_i at position i in G. For any two positions i and j with $1 \le i \le j \le m$, we will alternatively denote the substring $G[i]G[i+1]\cdots G[j] = g_i g_{i+1} \cdots g_j$ by the interval $G[i, j]$, or simply by $[G[i], G[j]]$ when no confusion arises. When G is trivial, for any two genes a and b in G such that a precedes b, we let $[a, b]$ denote the interval between a and b in G. For example, given $G' = bdc\text{-}ag\text{-}e\text{-}fh$, $G'' = bdce\text{-}gafh$, there are 2 breakpoints $c\text{-}a$ and $-e\text{-}f$ between G' and G'' within the two intervals $I_1 = dc\text{-}ag\text{-}e\text{-}f$ in G' and $I_2 = dce\text{-}gaf$ in G''. A *signed reversal* on a genome G simply reverses the order and signs of all the elements in an interval of G, i.e. between two positions in G. In the previous example, if a signed reversal operation is conducted in I_1 on G', then we obtain a new genome $G^* = bfe\text{-}ga\text{-}c\text{-}dh$.

Given a genome G built over a set of gene families $\mathcal{F}$, an *exemplar genome* G' of G is a genome obtained from G by deleting all but one occurrences of each gene family. In other words, G' is 1-duplicated. For example, let $G = bcaadagef$, the two exemplar genomes of G are: $bcadgef$ and $bcdagef$.

Given a set of trivial 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 genes between a and b is the same—regardless to the signs—among all the genomes in $\mathcal{G}$. For example, let $\mathcal{G} = \{G_1, G_2\}$, where $G_1 = bc\text{-}ag\text{-}e\text{-}fdh$, $G_2 = b\text{-}ce\text{-}gaf\text{-}dh$, there are three conserved intervals between G_1 and G_2: $[e, a]$, $[b, h]$ and $[-a, g]$.

Given two sets of trivial 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 $\mathcal{G} \cup \mathcal{H}$).

Let $\mathcal{H} = \{H_1, H_2\}$, where $H_1 = b\text{-}cg\text{-}af\text{-}edh$, $H_2 = bagcdefh$, 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 singletons, we use, for convenience, $d(G, H)$ instead of $d(\mathcal{G}, \mathcal{H})$ to denote the conserved interval distance between $\mathcal{G} = \{G\}$ and $\mathcal{H} = \{H\}$. Note that when only one genome G is considered, every interval in G is a conserved interval. This implies that when both G and H are composed of 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 genomes G and H built over a set of n gene families, 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 G' of G and H' of H such that the conserved interval distance $d(G', H')$ is at most K?

Note that the above ECID problem can be easily extended to two sets of genomes. However, in this paper we will focus on this formulation of two single genomes.

In the next two sections, we present inapproximability results about the optimization version of the ECID problem, namely, to compute or approximate the minimum value K in the above formulation.

Given any two genomes G and H built over a set of gene families $\mathcal{F}$, we define the *exemplar conserved interval distance* between them as

$$ECID(G, H) = \min\{d(G', H') \mid G' \text{ and } H' \text{ are respective exemplar genomes for } G \text{ and } H\}.$$

The *exemplar breakpoint distance* between G and H is the minimal breakpoint distance between any two respective exemplar genomes for G and H.

Given a minimization problem Π, let OPT be the optimal solution of Π. We say that an approximation algorithm $\mathcal{A}$ provides a *performance guarantee* of α for Π if for every instance of Π, the solution value returned by $\mathcal{A}$ is at most $\alpha \times OPT$. (Usually we say that $\mathcal{A}$ is a factor-α approximation algorithm for Π.) 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 *weak approximation* (Chen et al. 2006b). We say that a weak approximation algorithm B provides a *performance guarantee* of α for Π if for every instance of Π, the solution value returned by B is at most $\alpha \times (OPT + 1)$.

3 The zero exemplar conserved interval distance (ZECID) problem

Recently, Chen et al. (2006b) proved that the zero exemplar breakpoint distance problem is NP-complete. Following the spirit of this work, in this section we shall con-

sider the *zero exemplar conserved interval distance problem*, denoted by the *ZECID* problem, in the following:

Instance: Two genomes G and H built over a set of n gene families, 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).

Question: Are there respective exemplar genomes G' of G and H' of H such that the conserved interval distance $d(G', H')$ is zero? In other words, is $ECID(G, H) = 0$?

We first give the following property about the zero conserved interval distance between two trivial genomes.

Proposition 1 *Let G and H be two trivial genomes built over the same set of genes (i.e., G is a signed permutation of H). Then, the conserved interval distance between G and H is zero iff 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 the signed reversal of H, then every two genes in G form a conserved interval in G and H. Thus, $N_{G\cup H} = \binom{n}{2}$, implying $d(G, H) = 0$.

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 both G and H. We prove by induction on n that either $G = H$ or G is the singed reversal of H.

The property is trivially true for $n = 1$. When $n = 2$, let $G = a_1a_2$. In this case, $[a_1, a_2]$ must be a conserved interval in H. Hence, H is either a_1a_2 or $\text{-}a_2\text{-}a_1$, i.e., either $G = H$ or G is the signed reversal of H.

Assume that the property is true for $n \geq 2$. Now we consider the case of $n + 1 \geq 3$. Let $G = a_1a_2 \cdots a_na_{n+1}$ and $H = b_1b_2 \cdots b_nb_{n+1}$. $N_{G\cup H} = \binom{n}{2}$ implies that $[a_i, a_j]$ is a conserved interval in both G and H for $1 \leq i < j \leq n+1$. In particular, $[a_1, a_{n+1}]$ is a conserved interval in H. Thus, either $b_1 = a_1$ and $b_{n+1} = a_{n+1}$, or $b_1 = \text{-}a_{n+1}$ and $b_{n+1} = \text{-}a_1$. Let $G' = a_2 \cdots a_n$ and $H' = b_2 \cdots b_n$. We have $N_{G'\cup H'} = \binom{n-2}{2}$, hence $d(G', H') = 0$. By the assumption, we have either $G' = H'$ or G' is the signed reversal of H'. If $b_1 = a_1$ and $b_{n+1} = a_{n+1}$, and $G' = H'$, then we have $G = H$. If $b_1 = \text{-}a_{n+1}$ and $b_{n+1} = \text{-}a_1$, and G' is the signed reversal of H', then G is the signed reversal of H. If $b_1 = a_1$ and $b_{n+1} = a_{n+1}$, but G' is the signed reversal of H', then $G = a_1a_2 \cdots a_na_{n+1}$ and $H = a_1\text{-}a_n \cdots \text{-}a_2a_{n+1}$. In this case, $[a_1, a_2]$ is not a conserved interval in H. Similarly, If $b_1 = \text{-}a_{n+1}$ and $b_{n+1} = \text{-}a_1$, but $G' = H'$, then $[a_1, a_2]$ is not a conserved interval in H. Therefore, only the first two cases can be true, which imply the property. □

We now show the NP-completeness of the ZECID problem.

Theorem 1 *The ZECID problem is NP-complete. More precisely, given any two 3-duplicated genomes G and H, 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 the ZECID problem is in NP. To prove its NP-hardness, we propose a reduction to it from the NP-complete 3SAT problem (Garey and Johnson 1979): Given a collection $F = \{f_1, f_2, \ldots, f_q\}$ of q clauses, where each clause

consists of a set of 3 literals (representing the disjunction of those literals) over a finite set of n Boolean variables $X = \{x_1, x_2, \ldots, x_n\}$, is F satisfiable? That is, is there an assignment of truth values to X which makes every clause true? This reduction is inspired of the one proposed in (Chen et al. 2006b) for the zero exemplar breakpoint distance problem.

Let (F, X) be any instance of the 3SAT problem such that $F = \{f_1, f_2, \ldots, f_q\}$ and $X = \{x_1, x_2, \ldots, x_n\}$. Let $\mathcal{F} = \{f_i, g_j \mid 1 \leq i \leq q, 1 \leq j \leq n-1\}$ be the set of gene families. We construct two non-trivial genomes G and H built over $\mathcal{F}$ as follows:

$$G = S_1 g_1 S_2 g_2 \cdots g_{n-1} S_n,$$
$$H = T_1 g_1 T_2 g_2 \cdots g_{n-1} T_n,$$

where $g_1, g_2, \ldots, g_{n-1}$ are trivial genes (that would have a peg purpose), $S_i = V_i \overline{V}_i$ and $T_i = \overline{V}_i V_i$ with V_i (resp. $\overline{V}_i$) being the sequence of elements of $\{f_j \mid x_i \in f_j\}$ (resp. $\{f_j \mid \overline{x}_i \in f_j\}$) ordered by j. In other words, V_i (resp. $\overline{V}_i$) represents in order the clauses of F containing x_i (resp. $\overline{x}_i$). The obtained genomes G and H are indeed 3-duplicated, since any clause of F has 3 literals and therefore G and H are both composed of exactly 3 occurrences of each f_i of $\mathcal{F}$ for $1 \leq i \leq q$.

We now show that F is satisfiable iff G and H have zero exemplar conserved interval distance.

We first assume that F is satisfiable. In this case, let $x_1 = b_1, \ldots, x_n = b_n$ be the assignment of truth values that makes each clause of F true. For $1 \leq i \leq n$, if $x_i = b_i = 1$, then deleting all the genes in $\overline{V}_i$ from S_i and T_i to obtain $S'_i = V_i$ and $T'_i = V_i$, respectively. If $x_i = b_i = 0$, then deleting all the genes in V_i from S_i and T_i to obtain $S'_i = \overline{V}_i$ and $T'_i = \overline{V}_i$, respectively. Let

$$G' = S'_1 g_1 S'_2 g_2 \cdots g_{n-1} S'_n,$$
$$H' = T'_1 g_1 T'_2 g_2 \cdots g_{n-1} T'_n.$$

It is easy to see that G' is the same as H'. Since the assignment of truth values makes F true, it makes each clause $f_t \in F = \{f_1, \ldots, f_q\}$ true. That is, there is at least one Boolean variable x_i with either $x_i \in f_t$ or $\overline{x}_i \in f_t$ such that the truth assignment $x_i = b_i$ makes f_t true. If $x_i \in f_t$, then $b_i = 1$. This means that f_t is in V_i, hence it is in both S'_i and T'_i. Similarly, if $\overline{x}_i \in f_t$, then $b_i = 0$, implying that f_t is in $\overline{V}_i$, hence it in both S'_i and T'_i. It follows from the above analysis that each gene $f_t \in \mathcal{F}$ (which represents the clause f_t in F) must occur in both S' and H'. If f_t occurs more than once in G' or H' then one has to delete all but one of its occurrences in G and to delete all but one of its corresponding occurrences in H'. The two resulting genomes, still denoted by G' and H', are trivial and still the same. Finally, notice that both G' and H' contain all $q + n - 1$ genes in $\mathcal{F}$. By Property 1, we have $d(G', H') = 0$. Hence, $ECID(G, H) = 0$.

We then assume that G and H have zero exemplar conserved interval distance. In this case, let G'' (resp. H'') be an exemplar genome of G (resp. H) such that $ECID(G, H) = d(G'', H'') = 0$. Since G'' and H'' contain all the genes in $\mathcal{F} = \{f_1, \ldots, f_q, g_1, \ldots, g_{n-1}\}$ without duplications, we have by Property 1 that either

$G'' = H''$ or G'' is the signed reversal of H''. Because both G'' and H'' contain unsigned peg genes g_i, we must have $G'' = H''$. Let

$$G'' = S_1'' g_1 S_2'' g_2 \cdots g_{n-1} S_n'',$$
$$H'' = T_1'' g_1 T_2'' g_2 \cdots g_{n-1} T_n'',$$

where S_i'' (resp. T_i'') denotes the substring in G'' (resp. H'') corresponding to S_i (resp. T_i) in G (resp. H). $G'' = H''$ implies $S_i'' = T_i''$ for $1 \le i \le n$. This further implies that $S_i'' = T_i''$ is a subsequence of either V_i or $\overline{V}_i$ because $S_i = V_i \overline{V}_i$ and $T_i = \overline{V}_i V_i$. We now consider how to choose an assignment of truth values to Boolean variables x_i, $1 \le i \le n$, to make F true. If S_i'' is empty then we can assign either 1 or 0 to x_i arbitrarily. If S_i'' is not empty and is a subsequence of V_i then we assign $x_i = 1$. If S_i'' is not empty and is a subsequence of $\overline{V}_i$ then we assign $x_i = 0$. Because each gene $f_t \in \mathcal{F}$ (that represents the clause $f_t \in F = \{f_1, \ldots, f_q\}$) occurs in G'' (and H'') once, it must occur in a non-empty S_j''. If S_j'' is a subsequence of V_j, then $x_j \in f_t$, this means that one can set $x_j = 1$ to make f_t true. Similarly, if S_j'' is a subsequence of $\overline{V}_j$, then $\overline{x}_j \in f_t$, this means that one can set $x_j = 0$ to make f_t true. Thus, the assignment of truth values obtained above make every clause $f_t \in F$ truth, hence it makes F true.

The above reduction takes linear time in the length of F. Each clause $f_i \in F, i = 1, \ldots, q$, appears in G (resp. H) exactly 3 times. There are $n-1$ additional peg genes in G and in H. Therefore, the length of G (resp. H) is bounded by $3q + n - 1 \le |F|$, since $|F| = 3q + n$. □

Example 1 Given $F = \{f_1, f_2, f_3\}$, $X = \{x_1, x_2, x_3, x_4\}$, where $f_1 = \{x_1, \overline{x}_2, x_4\}$, $f_2 = \{\overline{x}_1, x_3, x_4\}$, $f_3 = \{x_2, x_3, \overline{x}_4\}$, and $f_4 = \{\overline{x}_1, \overline{x}_2, \overline{x}_3\}$, we have two 3-duplicated genomes in the following:

$$G = f_1 f_2 f_4 g_1 f_3 f_1 f_4 g_2 f_2 f_3 f_4 g_3 f_1 f_2 f_3,$$
$$H = f_2 f_4 f_1 g_1 f_1 f_4 f_3 g_2 f_4 f_2 f_3 g_3 f_3 f_1 f_2.$$

$d(G'', H'') = 0$, with $G'' = H'' = f_4 g_1 f_3 g_2 g_3 f_1 f_2$, corresponding to the assignment of truth values (that makes F true) with $x_1 = 0$, $x_3 = 0$ (or 1), and $x_2 = x_4 = 1$.

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 1 and the above corollary imply that the ECID problem does not admit any polynomial time approximation unless $P = NP$—if such a polynomial time approximation existed, then it would be able to decide, given any two genomes G and H, whether G and H have zero exemplar conserved interval distance in polynomial time, hence contradicting Theorem 1. We should point out that it remains open whether the zero exemplar conserved interval distance problem of two 2-duplicated genomes is NP-hard or not. Interestingly, the similar problem for the zero exemplar breakpoint distance problem of two 2-duplicated genomes is also open (Chen et al. 2006b).

4 Weak inapproximability

Given two genomes G and H built over a set of gene families $\mathcal{F}$, let $\text{opt}(G, H)$ be the optimal exemplar conserved interval distance between G and H, i.e.

$$\begin{aligned}\text{opt}(G, H) &= ECID(G, H)\\ &= \min\{d(G', H') \mid G' \textit{ (resp. } H') \textit{ is an exemplar genome for } G \textit{ (resp. } H)\}.\end{aligned}$$

We shall prove a super-linear $\frac{2}{7}m^{1.5}$ factor inapproximability bound on the optimal exemplar conserved interval distance between two genomes under the weak approximation model we proposed in Sect. 2. Notice that the super-linear factor $\frac{2}{7}m^{1.5}$ in the bound we obtain in Theorem 2 and Corollary 2 for the optimal exemplar conserved interval distance problem is substantially stronger than the sublinear $m^{1-\epsilon}$ factor, $\epsilon > 0$, in the bound obtained in (Chen et al. 2006b) for the optimal exemplar breakpoint distance problem.

Theorem 2 *Let $t(x) : N \to N$ be a function computable in polynomial time. If there is a polynomial time algorithm such that, given two genomes A and B built over a set of gene families $\mathcal{F}$ with length at most m, it can return two exemplar genomes A' and B' for A and B respectively such that $d(A', B') \leq t(m)\text{opt}(A, B) + \frac{2}{7}m^{1.5}$, then $P = NP$.*

Proof Let (F, X) be any instance of the 3SAT problem (Garey and Johnson 1979) such that $X = \{x_1, x_2, \ldots, x_n\}$ is a set of n Boolean variables and $F = \{f_1, f_2, \ldots, f_q\}$ is a set of q clauses over X, where each clause in F has three literals. We assume without loss of generality that $q \geq 2n$. (We can add additional clauses such as $\{x_i, \overline{x_i}, x_j\}$ to F to make sure $q \geq 2n$ so that the expanded formula is equivalent to F.) Let $G(F)$, $H(F)$ be the two genomes as constructed in the proof of Theorem 1 for (F, X) such that F is satisfiable iff there are two exemplar genomes $G'(F)$ and $H'(F)$ for $G(F)$ and $H(F)$ respectively such that $d(G'(F), H'(F)) = 0$. Let $|G(F)| = |H(F)| = u$, i.e., the number of occurrences of all the genes in $G(F)$ (or $H(F)$). Since each clause in F has three literals, we have

$$u + 1 \leq 3q + n \leq 3q + \frac{1}{2}q \leq \frac{7}{2}q.$$

Hence,

$$q \geq \frac{2}{7}(u + 1). \tag{1}$$

We now give several notations that are needed in the following analysis. Given a genome S, let $\Sigma(S)$ be the set of all the distinct genes occurred in S. If Σ' is a different set of genes such that $\Sigma(S) \cap \Sigma' = \emptyset$ and $|\Sigma'| = |\Sigma(S)|$, we define $S(\Sigma')$ to be a new genome obtained by replacing all the genes in S, in one to one fashion, by those in Σ'. For example, let $S = ab\text{-}acd\text{-}bc$, then $\Sigma(S) = \{a, b, c, d\}$. Let $\Sigma' = \{w, x, y, z\}$, then $S(\Sigma') = wx\text{-}wyz\text{-}xy$.

For $M \geq 3$, let $\Sigma_1, \Sigma_2, \ldots, \Sigma_M$ be M pairwise disjoint sets of genes of size $|\Sigma(G(F))|$. Let $G_1 = G(F)(\Sigma_1), G_2 = G(F)(\Sigma_2), \ldots, G_M = G(F)(\Sigma_M)$ be the

sequences derived from $G(F)$. Let $H_1 = H(F)(\Sigma_1), H_2 = H(F)(\Sigma_2), \ldots, H_M = H(F)(\Sigma_M)$ be the sequences derived from $H(F)$. Let

$$\mathcal{F} = \bigcup_{i=1}^{M} \Sigma_i \cup \{s_j | 1 \leq j \leq M\}$$

be the set of gene families, where s_j is not in any Σ_i. Define two genomes A and B built over $\mathcal{F}$ as follows:

$$A = G_1 s_1 G_2 s_2 \cdots G_M s_M,$$
$$B = H_1 s_1 H_2 s_2 \cdots H_M s_M,$$

where each s_i is a trivial gene in A and B (that would have a peg purpose). Let $m = |A| = |B| = M(u+1)$. By the construction of $G(F)$ and $H(F)$ in the proof of Theorem 1, each of $G(F)$ and $H(F)$ has $q + n - 1$ distinct genes. Hence, each of $G_i s_i$ and $H_i s_i$ has $q + n$ distinct genes.

Assume that some polynomial time algorithm outputs two exemplar genomes A' and B' for A and B respectively with $d(A', B') \leq t(m)\text{opt}(A, B) + \frac{2}{7}m^{1.5}$, then one can decide whether F is satisfiable by checking whether $d(A', B') \leq \frac{2}{7}m^{1.5}$. The analysis is given as follows.

If F is satisfiable, then, as in the proof of Theorem 1, two identical exemplar genomes can be obtained respectively from G_i and H_i for $1 \leq i \leq M$. Hence, two identical exemplar genomes can be obtained respectively from A and B. Therefore, by Property 1, the conserved interval distance between those two identical genomes for A and B is zero. Hence, we have $\text{opt}(A, B) = 0$. This implies that $d(A', B') \leq t(m)\text{opt}(A, B) + \frac{2}{7}m^{1.5} = \frac{2}{7}m^{1.5}$.

We now consider that f is not satisfiable. We will show that $d(A', B') > \frac{2}{7}m^{1.5}$. Let

$$A' = A_1 A_2 \cdots A_M,$$
$$B' = B_1 B_2 \cdots B_M,$$

such that A_i (resp. B_i) is the exemplar genome corresponding to $G_i s_i$ (resp. $H_i s_i$), $1 \leq i \leq M$. Since F is not satisfiable, it follows from the proof of Theorem 1 that $d(A_i, B_i) \geq \text{opt}(G_i, H_i) \geq 1$ for $1 \leq i \leq M$; namely, there is at least one conserved interval in A_i but not in B_i. Recall that all the occurrences of genes in G_i and H_i (hence, A_i and B_i) are unsigned. This implies that there are two genes a and b in A_i, $1 \leq i \leq M$, such that one of the following is true:

(1) $[a, b]$ is a (conserved) interval in A_i but $[b, a]$ is in B_i (i.e. b precedes a in B_i).
(2) There exists some gene $c \in [a, b]$ in A_i but $c \notin [a, b]$ in B_i.
(3) There exists some genes c in A_i such that $c \notin [a, b]$ in A_i but $c \in [a, b]$ in B_i.

For case (1), for any gene d in A_j, $j \neq i$, if $j < i$, then $[d, a]$ does not contain b in A', but $[d, a]$ contains b in B', hence $[d, a]$ is a conserved interval in A' but not in both A' and B'; $[d, b]$ contains a in A', but $[d, b]$ does not contain a in B', hence $[d, b]$ is a conserved interval in A' but not in both A' and B'. Similarly, if $i < j$,

then $[a, d]$ and $[b, d]$ are two conserved intervals in A' but not in both A' and B'. Analogously, for any e in B_j, $j \neq i$, if $j < i$, then $[e, a]$ and $[e, b]$ are two conserved intervals in B' but not in both A' and B'; if $j > i$, then $[a, e]$ and $[b, e]$ are two conserved intervals in B' but not in both A' and B'. Recall that each of G_j and H_j has $q + n - 1$ distinct genes. This means that each of A_j and B_j has $q + n$ distinct genes. Thus, in this case, A_i and B_i contribute at least $4(q + n)(M - 1)$ conserved intervals in either A_i or B_i but not in both, $1 \leq i \leq M$. Taking off $\sum_{j=1}^{M-1} 8j$ many conserved intervals that are over counted, we have, when $q + n \geq 4$, at least

$$4(q+n)M(M-1) - \sum_{j=1}^{M-1} 8j = 4(q+n)M(M-1) - 4M(M-1)$$
$$\geq 3(q+n)M(M-1)$$

conserved intervals in either A' or B' but not in both. Hence, we have $d(A', B') \geq 3(q + n)M(M - 1)$, when $q + n \geq 4$.

For case (2), since c must also be in B_i, c occurs either before a or after b in B_i. If c occurs before a, then we have $[a, c]$ in A_i but $[c, a]$ in B_i. If c occurs after b, then we have $[c, b]$ in A_i but $[b, c]$ in B_i. In either case, following a similar analysis for case (1), we have $d(A', B') \geq 3(q + n)M(M - 1)$, when $q + n \geq 4$. Analogously, we also have $d(A', B') \geq 3(q + n)M(M - 1)$ for case (3), when $q + n \geq 4$.

In summary, in either of the three cases, when $q + n \geq 4$, we have

$$\begin{aligned}
d(A', B') &\geq 3(q+n)M(M-1) \\
&= 3\left(\frac{q}{2} + \frac{q}{2} + n\right)M(M-1) \\
&> 3\left(\frac{q}{2} + \frac{q}{2} - n\right)M(M-1) \\
&\geq \frac{3}{2}qM(M-1), \quad \text{for } q \geq 2n \\
&> \frac{3}{7}(u+1)M(M-1), \quad \text{by inequality (1)} \\
&> \frac{2}{7}(u+1)MM, \quad M \geq 3 \\
&= \frac{2}{7}mM.
\end{aligned}$$

Since $m = (u + 1)M$, setting $M = (u + 1)$, we have

$$d(A', B') > \frac{2}{7}m^{1.5}. \qquad \square$$

The following corollary follows directly from Theorem 2 by letting $t(m) = \frac{2}{7}m^{1.5}$.

Corollary 2 *If there is a polynomial time algorithm such that, given two genomes A and B of length at most m built over a set of gene families $\mathcal{F}$, it can re-*

turn exemplar genomes A' *and* B' *for* A *and* B *respectively satisfying* $d(A', B') \leq \frac{2}{7}m^{1.5}[\text{opt}(A, B) + 1]$, *then* $P = NP$.

This negative result shows that even under the weak approximation model proposed in Sect. 2 which is much weaker than the conventional approximation model, it is not possible to obtain a good approximation to the optimal exemplar conserved interval distance problem, unless $P = NP$.

5 When one genome is k-spanned

Theorem 1 implies it is impossible to decide whether the conserved interval distance between two genomes is zero or not in polynomial time, unless $P = NP$. Theorem 2 further implies that the problem of finding the exemplar conserved interval distance between two genomes does not admit any weak approximation, unless $P = NP$. Blin and Rizzi (2005) proved that the exemplar conserved interval distance problem between two genomes is NP-complete even when all non-trivial segments of the genomes are composed of only one duplicated gene. In fact, their proof implies that, even when one genome is trivial and both genomes consist of unsigned genes, the problem is still NP-complete. However, those negative results do not rule out the possibility of solving the exemplar conserved interval distance problem between two genomes in polynomial time when certain interesting constraint is imposed on the two genomes. Chen, Fu and Zhu (2006b) gave a $2(1 + \log n)$-approximation to the exemplar breakpoint distance problem for two genomes when one is $O(\log n)$-spanned. This motivates us to investigate whether one can solve in polynomial time the exemplar conserved interval distance problem between two genomes when one genome is $O(\log n)$-spanned.

Given any two genome G and H built over a set of gene families $\mathcal{F}$, let $|G| = N$ and $|H| = M$. Assume that G is k-spanned. We investigate how to decide whether the conserved interval distance between G and H is zero or not.

We first consider that both G and H consist of unsigned genes only. Assume that $\mathcal{F}$ has m distinct genes. By Property 1, the exemplar conserved interval distance between G and H is zero iff there is an exemplar genome A (resp. B) for G (resp. H) such that $A = B$. Note that A and B, which consist of m distinct genes, are two subsequences of G and H respectively. Thus, the exemplar conserved interval distance between G and H is zero iff the length of a longest common subsequence, which consists of distinct genes, between G and H is m. This gives us the following idea to decide whether the exemplar conserved interval distance between G and H is zero or not: Compute the length of a longest common subsequence, which consists of distinct genes, between G and H. If the length is m, then the exemplar conserved interval distance between G and H is zero, otherwise it is not. It is well-known (see, for example, Cormen et al. 2002) that finding a longest common subsequence between two strings with respective lengths N and M can be done in $O(MN)$ time by means of dynamic programming. In our case of deciding zero exemplar conserved interval distance between two genomes G and H, a longest common subsequence is "*constrained*" to have distinct genes. Hence, we cannot apply the well-known $O(MN)$ algorithm for

finding a longest common subsequence to our "*longest constrained common subsequence*" problem.

Formally, given a segment (or a substring) A of G and a segment B of H, we call a subsequence S a "*constrained common subsequence*" between A and B if (1) S is a common subsequence between A and B and (2) S has no duplicated genes. Similarly, we call S a "*constrained subsequence*" of G (or H) if (1) S is a subsequence of G (or H) and (2) S has no duplicated genes.

To simplify presentation, we introduce a notation $\preceq$: For genome G, $S \preceq G$ denotes that S is a constrained subsequence of G.

Without loss of generality, we let $N = pk$. We divide G into p disjoint segments G_i of equal length k, $1 \le i \le p$, with $G_i = G[(i-1)k, ik]$. For any i, j, s, t, $1 \le i \le j \le p$ and $1 \le s \le t \le M$, let $A[i, j] = G_i G_{i+1} \cdots G_j$ and $H[s, t] = H[s]H[s+1] \cdots H[t]$. We consider how to find a longest constrained common subsequence between $A[i, j] = G_i G_{i+1} \cdots G_j$ and $H[s, t]$.

For $i = j$, $A[i, i] = G_i$. For each $B \preceq G_i$, let $lccs(i, i, B, B, s, t)$ be a longest constrained common subsequence between B and $H[s, t]$.

For $j = i + 1$, $A[i, j] = G_i G_{i+1}$. For each $B_1 \preceq G_i$ and each $B_2 \preceq G_{i+1}$ such that $B_1 B_2$ has no duplicated genes (i.e., $B_1 B_2$ is a constrained subsequence of $G_i G_j$), let $lccs(i, i+1, B_1, B_2, s, t)$ be a longest constrained common subsequence between $B_1 B_2$ and $H[s, t]$.

For $j = i + 2$, $A[i, j] = G_i G_{i+1} G_{i+2}$. For each $B_1 \preceq G_i$ and each $B_2 \preceq G_{i+2}$, let $lccs(i, i+2, B_1, B_2, s, t)$ be a longest constrained common subsequence between $B_1 G_{i+1} B_2$ and $H[s, t]$.

For $j > i + 2$, $A[i, j] = G_i G_{i+1} \cdots G_l G_{l+1} \cdots G_{j-1} G_j$ with $i + 1 \le l \le j - 2$, for each $B_1 \preceq G_i$ and each $B_2 \preceq G_j$, let $lccs(i, i+1, B_1, B_2, s, t)$ be a longest constrained common subsequence between $B_1 G_{i+1} \cdots G_l G_{l+1} \cdots G_{j-1} B_2$ and $H[s, t]$. For any l and x with $i + 1 \le l \le j - 2$ and $s \le x \le t$, for any $B_3 \preceq G_l$ and any $B_4 \preceq G_{l+1}$, let $lccs(i, l, B_1, B_3, s, x) = U_1 V_1$ and $lccs(l+1, j, B_4, B_2, x+1, t) = V_2 U_2$, where V_1 (resp. V_2) is a constrained subsequence for B_3 (resp. B_4), define

$$merge(lccs(i, l, B_1, B_3, s, x), lccs(l+1, j, B_4, B_2, x+1, t)) = \begin{cases} U_1 V_1 V_2 U_2, & \text{if } V_1 V_2 \text{ is a constrained subsequence of } B_3 B_4, \\ nil, & \text{otherwise.} \end{cases}$$

Lemma 1 *For any i, j, s, t with $1 \le i \le j \le p$ and $1 \le s \le t \le N$, $A[i, j] = G_i \cdots G_l G_{l+1} \cdots G_j$, $i + 1 \le l \le j - 2$, for each $B_1 \preceq G_i$ and $B_2 \preceq G_j$,*

$$lccs(i, j, B_1, B_2, s, t) = \max_{i<l<j-1; s \le x \le t} \{merge(lccs(i, l, B_1, B_3, s, x), lccs(l+1, j, B_4, B_2, x+1, t) \mid B_3 \preceq G_l, B_4 \preceq G_{l+1}\}. \quad (2)$$

Proof Note that $|G_l| = |G_{l+1}| = k$ and G is k-spanned. No genes occurred in $G_i G_{i+1} \cdots G_{l-1}$ will occur in $G_{l+1} \cdots G_{j-1} G_j$, and no genes occurred $G_{l+2} \cdots G_{j-1} G_j$ will occur in $G_i G_{i+1} \cdots G_l$. The only place for both $G_i G_{i+1} \cdots G_{l-1} G_l$

and $G_{l+1}G_{l+2}\cdots G_{j-1}G_j$ to have common genes is G_lG_{l+1}. This implies that $merge(lccs(i,l,B_1,B_3,s,x), lccs(l+1,j,B_4,B_2,x+1,t)$ yields a constrained common subsequence between $B_1G_{i+1}\cdots G_{l-1}G_lG_{l+1}\cdots G_{j-1}B_2$ and $H[s,t]$. Hence, the length of the constrained common subsequence obtained in the right hand side of expression (2) is no more than the length of the longest constrained common subsequence between $B_1G_{i+1}\cdots G_{l-1}G_l\ G_{l+1}G_{l+2}\cdots G_{j-1}B_2$ and $H[s,t]$.

Let $lccs(i,j,B_1,B_2,s,t) = UVWZ$ be a longest constrained common subsequence between $B_1G_{i+1}\cdots G_{l-1}G_lG_{l+1}\cdots G_{j-1}B_2$ and $H[s,t]$ such that, for some x with $s \le x \le t$, UV is a constrained common subsequence between $B_1G_{i+1}\cdots G_{l-1}G_l$ and $H[s,x]$ with $Z \preceq G_l$, and WZ is a constrained common subsequence between $G_{l+1}G_{l+2}\cdots G_{j-1}B_2$ and $H[x+1,t]$ with $W \preceq G_{l+1}$. Let $B_3 = V \preceq G_l$ and $B_4 = W \preceq G_{l+1}$. Let $lccs(i,l,B_1,B_3,s,x) = U'V'$ such that $U'V'$ is a longest constrained common subsequence between $B_1G_{i+1}\cdots G_{l-1}B_3$ and $H[s,x]$ with $V' \preceq B_3$, and $lccs(l,j,B_3,B_2,x+1,t) = W'Z'$ such that $W'Z'$ is a longest constrained common subsequence between $B_4G_{l+2}\cdots G_{j-1}B_2$ and $H[x+1,t]$ with $W \preceq G_{l+1}$. Then, $|UV| \le |U'V'|$ and $|WZ| \le |W'Z'|$. Since VW has no duplicated genes, $V' \preceq B_3$ and $W' \preceq B_4$ implies $V'W' \preceq B_3B_4$. Thus, merging $lccs(i,l,B_1,B_3,s,x)$ and $lccs(l+1,j,B_3,B_2,x+1,t)$ yields $merge(U'V',W'Z') = W'V'W'Z'$, which has length $|W'V'W'Z'| \ge |UVWZ|$.

Combining the analysis above, the right hand side of expression (2) yields a longest constrained common subsequence between $B_1G_{i+1}\cdots G_{l-1}G_lG_{l+1}G_{l+2}\cdots G_{j-1}B_2$ and $H[s,t]$. □

Lemma 2 *The exemplar conserved interval distance between $G = G_1\cdots G_p$ and H is zero if and only if the longest constrained common subsequence given by $\max_{B_1\preceq G_1, B_2\preceq G_p} lccs(1,p,B_1,B_2,1,M)$ has length n, where n is the total number of distinct genes occurred in G and H.*

Proof If the longest constrained common subsequence given by

$$\max_{B_1\preceq G_1, B_2\preceq G_p}\{lccs(1,p,B_1,B_2,1,M)$$

has length n, then both G and H have a common exemplar genome, so by Property 1, the exemplar conserved interval distance between G and H is zero.

If the exemplar conserved interval distance between G and H is zero, then by Property 1 there is a longest constrained common subsequence S of length n between G and H. If $G = G_1$, then $S \preceq G$, hence $lccs(1,1,S,S,1,M)$, which may not the same as S, is a longest constrained common subsequence of length n between G and H. If $G = G_1G_2$, let $S = S_1S_2$ such that $S_i \preceq G_i$, $i = 1,2$, then $lccs(1,2,S_1,S_2,1,M)$ is a longest constrained common subsequence of length n between G and H. Similarly, If $G = G_1G_2G_3$, let $S = S_1S_2S_3$ such that $S_i \preceq G_i$, $i = 1,2,3$, then $lccs(1,3,S_1,S_3,1,M)$ is a longest constrained common subsequence of length n between G and H. In general, when $G = G_1G_2\cdots G_{p-1}G_p$ with $p \ge 4$, let $S = S_1S_2S_3$ such that $S_1 \preceq G_1$, $S_2 \preceq G_p$ and $S_3 \preceq G_2\cdots G_{p-1}$. Then, by Lemma 1, $lccs(1,p,S_1,S_2,1,M)$ is a longest common subsequence of length n between G and H. In either of the four cases, the length of $\max_{B_1\preceq G_1, B_2\preceq G_2}\{lccs(1,p,B_1,B_2,1,M)$ is n. □

Let LCS denote the $O(MN)$ time algorithm for finding a longest common subsequence between two strings of respective lengths N and M (Cormen et al. 2002). We design the following algorithm LCCS (standing for Longest Constrained Common Subsequence) for computing a longest constrained common subsequence between two genomes G and H.

Algorithm $LCCS(G, H)$
1. for $(d = 0; d < q - 1; d = d + 1)$
2. {
3. for $(i = 1; i \leq q; i = i + 1)$
4. {
5. $j = i + d$;
6. if $(j > q)$
7. break;
8. for $(s = 1; s \leq M; s = s + 1)$ and for $(t = s; t \leq M; t = t + 1)$
9. {
10. if $(i == j)$
11. for any $B \preceq G_i$
12. call algorithm LCS to find $lccs(i, j, B, B, s, t)$;
13. if $(j == i + 1$ or $j == i + 2)$
14. for any $B_1 \preceq G_i$ and $B_2 \preceq G_j$ ($B_1B_2 \preceq G_iG_{i+1}$ if $j == i + 1$)
15. call algorithm LCS to find $lccs(i, j, B_1, B_2, s, t)$;
16. if $(j > i + 2)$
17. for any $B_1 \preceq G_i, B_2 \preceq G_j$,
18. $lccs(i, j, B_1, B_e2, s, t) =$
$\max_{i<l<j-1; s\leq x\leq t}\{merge(lccs(i, l, B_1, B_3, s, x),$
19. $lccs(l + 1, j, B_4, B_2, x + 1, t) | B_3 \preceq G_l, B_4 \preceq G_{l+1}\}$
20. }
21. }
22. }
23. return $\max\{lccs(1, p, B_1, B_2, 1, M) | B_1 \preceq G_1, B_2 \preceq G_p\}$
24. the end of algorithm LCCS

We are now ready to prove the following result:

Theorem 3 *Given any two genome G and H built over a set of n gene families $\mathcal{F}$, let $|G| = N$ and $|H| = M$. If G is k-spanned, then one can decide whether the exemplar conserved interval distance between G and H is zero or not in $O(2^{4k}N^3M^3/k^2)$ time.*

Proof We first consider that both G and H have unsigned genes. Assume without loss of generality $N = kp$. Let $G = G_1 \cdots G_p$ with $|G_i| = k$, $1 \leq i \leq p$. By Lemma 2, if the longest constrained common subsequence returned by algorithm LCCS has length n, then the exemplar conserved interval distance between G and H is zero, otherwise it is not. For each G_i, $1 \leq i \leq q$, there are 2^k possible ways to select a constrained subsequence for it. For each iteration of lines 10 to 15 dealing with the cases of $j = i, i + 1$ or $i + 2$, the algorithm performs an exhaustive search of all

constrained subsequences of $G_i \cdots G_j$ to find a longest constrained subsequence between $G_i \cdots G_j$ and $H[s,t]$. Thus, the time complexity at each iteration for those three cases is at most $O(2^{3k}kM)$. For each iteration of lines 16 to 19 dealing with the general case of $j > i+2$, the time complexity is $O(2^{4k}kpM)$. Hence, the total time complexity of algorithm LCCS is $O(2^{4k}kp^3M^3) = O(2^{4k}N^3M^3/k^2)$.

When G and H have singed genes, we run algorithm LCCS two times. The first time is to find a longest constrained common subsequence between G and H. The second time is to find a longest constrained common subsequence between G and the singed reversal of H. Here, we require algorithm LCCS to match an unsigned gene with an unsigned, and a signed gene with a signed gene. by Property 1, G and H have zero conserved interval distance iff there are exemplar genomes A and B respectively for G and H such that either $A = B$ or A is the signed reversal of B. In the former case, G and H has a longest constrained subsequence of length n. In the latter case, G and the signed reversal of H has a longest constrained subsequence of length n. Hence, the conserved interval distance between G and H is zero if we find a longest constrained common subsequence with length n, otherwise the distance is not zero. The time complexity is the same as $O(2^{4k}N^3M^3/k^2)$ time, when G and H have signed genes. □

The following corollary, followed directly from Theorem 3, implies that the zero exemplar conserved interval distance problem is decidable in polynomial time, when one genome is k-spanned.

Corollary 3 *Given any two genome G and H built over a set of n gene families $\mathcal{F}$, let $|G| = N$ and $|H| = M$. If G is $O(\log N)$-spanned, then one can decide whether the exemplar conserved interval distance between G and H is zero or not in $n^{O(1)}N^3M^3/\log^2 n$ time.*

6 When one genome is trivial

We want to know in this section whether we can improve the result obtained in the previous section when one genome is trivial. Blin and Rizzi (2005) proved that the exemplar conserved interval distance problem between two genomes is NP-complete, even when one genome is trivial and both genomes consist of unsigned genes. Hence, by their result, one shall not expect a polynomial time algorithm to find the exemplar conserved interval distance between two genomes with one being trivial, unless $P = NP$. Nevertheless, we shall prove in this section that one can find in polynomial time a constant-sized exemplar conserved interval distance between two genomes with one being trivial.

For convenience, we continue using the notation $\preceq$: $S \preceq G$ denotes that S is a constrained subsequence of G. We also introduce a new notation $\widetilde{\ }$: For any segment T of G, $\widetilde{T}$ denotes the signed reversal of T.

We first give a simple relation between exemplar breakpoint distance and the exemplar conserved interval distance between two genomes.

Lemma 3 *Given any two genomes G and H built over a set of gene families $\mathcal{F}$, the exemplar conserved interval distance between G and H is greater than or equal to the exemplar breakpoint distance between G and H.*

Proof Let A and B are two exemplar genomes respectively for G and H such that $d(A, B)$ is the exemplar conserved interval distance between G and H. Since every breakpoint between A and B is not, by definition, a conserved interval in both A and B, it contributes one to the value of $d(A, B)$. Hence, $d(A, B)$ is at least the number of breakpoints between A and B, and the latter is at least the exemplar breakpoint distance between G and H. □

Lemma 4 *Given any two genomes G and H built over a set of gene families $\mathcal{F}$ with G being trivial, assume that the exemplar breakpoint distance between G and H is $c \geq 0$. Then, G can be divided into $c+1$ segments $G = G_1 G_2 \cdots G_{c+1}$ such that the following two properties are true*:

(1) *H has a longest constrained subsequence $G'_{i_1} G'_{i_2} \cdots G'_{i_{c+1}}$, where $i_1, i_2, \ldots, i_{c+1}$ is a permutation of $1, 2, \ldots, c+1$, and $G'_{i_l} = G_{i_l}$ or $G'_{i_l} = \widetilde{G}_{i_l}$, $1 \leq l \leq c+1$.*
(2) *The last gene in G_j and the first gene in G_{j+1} form a breakpoint in G, $1 \leq j \leq c$.*

Proof Since the exemplar breakpoint distance between G and H is c, by definition, there are two exemplar genomes A and B respectively for G and H such that the number of breakpoints between A and B is c. Because G is trivial, we have $A = G$. Let $G = g_1 g_1 \cdots g_N$ and $g_{i_l} g_{i_l+1}$, $1 \leq l \leq c$, be c breakpoints. Then, $G_1 = g1 \cdots g_{i_1}$, $G_l = g_{i_{l-1}+1} \cdots g_{i_l}$, $2 \leq l \leq c$, $G_{c+1} = G_{i_c+1} \cdots g_N$ are $c+1$ segments of G satisfying (2). For each G_{i_l}, $1 \leq l \leq c+1$, any two consecutive genes in G_{i_l} do not form a breakpoint between G and B (otherwise the breakpoint distance between G and B is more than c), this means that either G_{i_l} or its signed reversal is a substring of B. Hence, $B = G'_{i_1} G'_{i_2} \cdots G'_{i_{c+1}}$, where $i_1, i_2, \ldots, i_{c+1}$ is a permutation of $1, 2, \ldots, c+1$, and $G'_{i_l} = G_{i_l}$ or $\widetilde{G}_{i_l}$, $1 \leq l \leq c+1$. Therefore, (1) is true. □

Theorem 4 *Given any two genomes G and H built over a set of gene families $\mathcal{F}$ with G being trivial. Let $|G| = N$ and $|H| = M$. If the exemplar conserved interval distance between G and H is a constant $c \geq 0$, which is unknown. Then, one can find this distance c in $O(N^{c+2} M^{c+2}(MN + N^3))$ time.*

Proof Both Lemmas 3 and 4 give us the following idea to find the exemplar conserved interval distance between G and H, provided that this distance is a constant $c \geq 0$: Lemma 3 implies that, if two genomes A and B, which are exemplar respectively for G and H, yield the exemplar conserved interval distance between G and H, then the number of breakpoints between A and B is no more than c. For any $j = 1, 2, \ldots$, we want to find all the possible exemplar genomes A and B respectively for G and H such that there are $j-1$ many breakpoints between them. Since G is trivial, any exemplar genome A for G must be G itself, i.e., $A = G$. For any two exemplar genomes G and B with $j-1$ breakpoints, by Lemma 4, G and B satisfy the two properties in the lemma. So, we can try all possible ways to divide G into

j many segments. Say, $G = G_1 G_2 \cdots G_j$ is one of such j-segment divisions satisfying property (2) in Lemma 4. Then, $B = G'_{i_1} G'_{i_2} \cdots G'_{i_{c+1}}$ is a longest constrained common subsequence satisfying property (1) in Lemma 4. We can find such a B, if it exists, through exhaustive search as follows. Try all possible ways to divide H into j segments. Say, one of such segment is $H = H_1 H_2 \cdots H_j$. For G_i and H_l, $1 \le i, l \le j$, decide whether G_i or its signed reversal is a subsequence of H_l or not. This can be done using the $O(MN)$ time algorithm for finding a longest common subsequence between two strings (Cormen et al. 2002). We obtain all possible $B = G'_{i_1} G'_{i_2} \cdots G'_{i_j}$ such that $i_1, i_2, \ldots, i_{c+1}$ is a permutation of $1, 2, \ldots, j$, $G'_{i_l} \preceq H_l$, and $G'_{i_l} = G_{i_l}$ or $\widetilde{G}_{i_l}$, $1 \le l \le j$. Once, we obtain a pair of exemplar genomes G and B with $j - 1$ many breakpoints, we find the conserved interval distance between them. This can be easily done in $O(N^3)$ time. We keep the smallest conserved interval distance we have found so far for $1, 2, \ldots, j$. If this distance is larger than j, then try the above process for $j + 1$. If this distance is less or equal to j, then stop and return it as the exemplar conserved interval distance between G and H. We shall prove that this termination condition is correct in Claim 5.

The following is the algorithm ConstECID (standing for Constant Exemplar Conserved Interval Distance) in detail.

Algorithm ConstECID(G, H)
/*precondition: G is trivial.*/
1. $D = \infty$;
2. for ($j = 1$; *true*; $j = j + 1$) /* the j-for-loop*/
3. {
4. for each j segments of G, $G = G_1 G_2 \cdots G_j$
5. for each j segments of H, $H = H_1 H_2 \cdots H_j$
6. {
7. for ($i = 1$; $i \le j$; $i = i + 1$) and for ($l = 1$; $l \le j$; $l = l + 1$)
8. {
9. decide whether $G_i \preceq H_l$ or $\widetilde{G}_i \preceq H_l$;
10. }
11. for each $B = G'_{i_1} G'_{i_2} \cdots G'_{i_j}$ such that $i_1, i_2, \ldots, i_j$ is a permutation of $1, 2, \ldots, j$,
12. and $G'_{i_l} \preceq H_l$, $G'_{i_l} = G_{i_l}$ or $\widetilde{G}_{i_l}$
13. {
14. find $d(G, B)$;
15. let $D = \min\{D, d(G, B)\}$;
16. }
17. }
18. if ($D \le j$)
19. return D and stop;
20. }
21. the end of algorithm ConstECID

Claim 5 *Let ECID be the exemplar conserved interval distance between G and H. Let D_j denote the value D found by algorithm ConstECID at the end of the j-th*

iteration of the j-for-loop. If for some $j^* \geq 1$, $D_{j^*} \leq j^*$ is true at line 18 *in the algorithm, then $ECID = D_{j^*}$.*

Proof of Claim 5 Recall that any exemplar genome for G is G itself for G is trivial. Let B' be an exemplar genome for H such that $ECID = d(G, B')$ and the number of breakpoints between G and B' is the smallest among all exemplar genomes B for H such that $ECID = d(G, B)$. Let j' be the number of the breakpoints between G and B'. By Lemma 3, $j' \leq ECID$. During the $(j'+1)$-th iteration of the j-for-loop, algorithm ConstECID must find B' via exhaustive search for all possible exemplar genomes B for H such that the number of breakpoints between G and B is j'. So, the algorithm finds $D_{j'+1} = d(G, B') = ECID$ at the $(j'+1)$-th iteration of the j-for-loop. Since the algorithm will never increase D_j when j gets larger, it keeps $ECID = D_{j'+1}$ for all $j \geq j'+1$. Let $j^* = \max\{j'+1, ECID = D_{j'+1}\}$. At the j^*-th iteration of the j-for-loop, the algorithm still finds $D_{j^*} = D_{j'+1} = ECID$. For any j with $1 \leq j < j^*$, we must have $D_j > j$. This can be shown in two cases. In the first case, we consider $j' < ECID$. In this case, $j^* = ECID$. Thus, $D_j \leq j$ implies $ECID \leq D_j \leq j < j^* = ECID$, a contradiction. In the second case, we consider $j' = ECID$. Here, $j^* = ECID + 1 = j' + 1$. Thus, $D_j \leq j$ implies $ECID \leq D_j \leq j < j^* = ECID + 1 = j' + 1$, hence $1 \leq j = D_j = ECID = j'$. Let B'' be the exemplar genome for H found by the algorithm at the j-th iteration of the j-for-loop such that $D_j = d(G, B)$. By Lemma 3, the breakpoint distance between G and B'' is $j - 1 \leq d(G, B) = ECID = j'$, contradicting to the fact that j' is the smallest breakpoint distance between G and any exemplar genome B for H such that $d(G, B) = ECID$. Therefore, we must have $D_j > j$. This implies that j^* is the smallest j such that $D_j \leq j$. Since the j-for-loop repeats for $j = 1, 2, \ldots, N$, the algorithm will find j^* at the j^*-th iteration of this loop. Again, for this j^*, we have $ECID = D_{j^*}$.

Suppose the exemplar conserved interval distance between G and H is c. By Claim 5, algorithm ConstECID will find c at most the $(c+1)$-th iteration of the j-for-loop. Deciding whether $G_i \preceq H_l$ or $\widetilde{G}_i \preceq H_l$, i.e., whether G_i or its signed reversal is a subsequence of H_l can be done in $O(MN)$ time by the well-known algorithm for finding a longest common subsequence between two strings (Cormen et al. 2002). Given two exemplar genomes of length N, one can find its conserved interval distance in $O(N^3)$ time. The total time of algorithm ConstECID is $O(N^{c+2}M^{c+2}(MN + N^3))$. □

7 Concluding remarks

We prove two major lower bounds on the approximation of the exemplar conserved interval distance problem of genomes. The first result implies that the conserved interval distance problem of genomes does no admit any polynomial time approximation, unless $P = NP$. The second result further implies that this problem does not admit any weak approximation with a super-linear factor $\frac{2}{7}m^{1.5}$, unless $P = NP$. However, 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, can we obtain a good approximation when

one genome is 2-duplicated and the other is a 3-span genome? Such a special case partially conforms with the real-life dataset that duplications of genes are typically pegged and occur at not very far away positions (Nguyen et al. 2005).

To start with our effort on solving interesting special cases of the exemplar conserved interval distance problem, we present two positive results in this paper. The first is a polynomial time algorithm for deciding whether the exemplar conserved interval distance between two genomes is zero or not when one genome is $O(\log n)$-spanned. The second is a polynomial time algorithm for finding an unknown constant-sized exemplar conserved interval distance between two genomes. Blin and Rizzi (2005) proved that the exemplar conserved interval distance problem between two genomes is NP-complete, even when one genome is trivial and both genomes consist of unsigned genes. By our Theorem 1, the zero exemplar conserved interval distance problem between two genome is NP-complete, even when both are 3-duplicated. By our Theorem 2, it is impossible to give a good weak approximation to the exemplar conserved interval distance problem for two general genomes. Those two results together with Blin and Rizzi's result show that the two special cases solved by our two polynomial time algorithms are not trivial.

Given a *k-span* genome G and a general genome H, each is a sequence containing $O(m)$ signed or unsigned genes drawn from a set of n gene families, a $2(1+\log n)$-approximation algorithm was devised in (Chen et al. 2006b) to compute the exemplar breakpoint distance between G and H when $k = O(\log n)$. Can we improve Theorem 3 to approximate the non-zero exemplar conserved interval distance for such a k-span genome G and a general genome H when $k = O(\log n)$? Again, we should point out that computing the exemplar conserved interval distance of two genomes is more involved than computing their exemplar breakpoint distance. We should also point out that by Blin and Rizzi (2005) it is impossible to compute, in general, the exact exemplar conserved interval distance for two genomes when one is $O(\log n)$-spanned, unless $P = NP$, because a trivial genome is 0-spanned. We do not know whether Theorem 4 can be improved to find any non-constant (say, $\log n$) sized exemplar conserved interval distance between two genomes when one is trivial.

We prove that the zero exemplar conserved interval distance problem between two 3-duplicated genomes is NP-complete. Is this problem still NP-complete when the two genomes are 2-duplicated? The answer to this problem remains open. Interestingly, it also open whether the zero exemplar breakpoint distance problem of two 2-duplicated genomes is NP-complete or not (Chen et al. 2006b).

Acknowledgements We thank two anonymous reviewers for their valuable comments. Their comments help us revise the paper and improve its presentation quality. Theorem 1 and a linear factor lower bound on weak approximation were reported in the preliminary version of this paper appeared in (Chen et al. 2006b). In this paper, We have improved the linear factor lower bound on weak approximation to a super-linear factor $m^{1.5}$. Research in this paper 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. Part of Bin Fu's work was done when he was on the faculty in the Department of Computer Science, University of New Orleans, New Orleans, LA 70148, and was also affiliated with Research Institute for Children, 200 Henry Clay Avenue, New Orleans, LA 70118, USA.

References

Bafna V, Pevzner P (1995) Sorting by reversals: genome rearrangements in plant organelles and evolutionary history of X chromosome. Mol Biol Evol 12:239–246

Bereg S, Zhu B (2005) RNA multiple structural alignment with longest common subsequences. In: Proceedings of the 11th international annual confrerence on computers and combinatorics (COCOON'05). Lecture notes in computer science, vol 3595. Springer, Berlin, pp 32–41

Bergeron A, Stoye J (2003) On the similarity of sets of permutations and its applications to genome comparison. In: Proceedings of the 9th international annual confrerence on computers and combinatorics (COCOON'03). Lecture notes in computer science, vol 2697. Springer, Berlin, pp 68–79

Blin G, Rizzi R (2005) Conserved interval distance computation between non-trivial genomes. In: Proceedings of the 11th international annual confrerence on computers and combinatorics (COCOON'05). Lecture notes in computer science, vol 3595. Springer, Berlin, pp 22–31

Bryant D (2000) The complexity of calculating exemplar distances. In: Sankoff D, Nadeau J (eds) Comparative genomics: empirical and analytical approaches to gene order dynamics, map alignment, and the evolution of gene families. Kluwer Academic, Dordrecht, pp 207–212

Chen Z, Fowler RH, Fu B, Zhu B (2006a) Lowers bounds on the approximation of the conserved interval distance problem of genomes. In: Proceedings of the 12th international annual confrerence on computers and combinatorics (COCOON'06). Lecture notes in computer science, vol 4112. Springer, Berlin, pp 245–254

Chen Z, Fu B, Zhu B (2006b) The approximability of the exemplar breakpoint distance problem. In: Proceedings of the second international conference on algorithmic aspects in information and management (AAIM'06). Lecture notes in computer science, vol 4041. Springer, Berlin, pp 291–302

Chen Z, Fu B, Xu J, Yang B, Zhao Z, Zhu B (2007) Non-breaking similarity of genomes with gene repetitions, submitted for publication. In: Proceedings of the 18th international symposium on combinatorial pattern matching (CPM'07). Lecture notes in computer science, vol 4580. Springer, Berlin, pp 119–130

Cormen T, Leiserson C, Rivest R, Stein C (2002) Introduction to algorithms, 2nd edn. McGraw–Hill, Cambridge

Garey M, Johnson D (1979) Computers and intractability: a guide to the theory of NP-completeness. Freeman, San Francisco

Gascuel O (ed) (2004) Mathematics of evolution and phylogeny. Oxford University Press, Oxford

Hannenhalli S, Pevzner P (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. J ACM 46(1):1–27

Li Z, Wang L, Zhang K (2006) Algorithmic approaches for genome rearrangement: a review. IEEE Trans Sys Man Cybern Part C: Appl Rev 36(5):636–648

Makaroff C, Palmer J (1988) Mitochondrial DNA rearrangements and transcriptional alternatives in the male sterile cytoplasm of Ogura radish. Mol Cell Biol 8:1474–1480

Marron M, Swenson K, Moret B (2004) Genomic distances under deletions and insertions. Theor Comput Sci 325(3):347–360

Nguyen CT (2005) Algorithms for calculating exemplar distances. Honors thesis, School of Computing, National University of Singapore

Nguyen CT, Tay YC, Zhang L (2005) Divide-and-conquer approach for the exemplar breakpoint distance. Bioinformatics 21(10):2171–2176

Palmer J, Herbon L (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 27:87–97

Sankoff D (1999) Genome rearrangement with gene families. Bioinformatics 16(11):909–917

Sturtevant A, Dobzhansky T (1936) Inversions in the third chromosome of wild races of *drosophila pseudoobscura* and their use in the study of the history of the species. Proc Natl Acad Sci USA 22:448–450

Watterson G, Ewens W, Hall T, Morgan A (1982) The chromosome inversion problem. J Theor Biol 99:1–7