Theoretical Computer Science 550 (2014) 59–65

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the approximability of the exemplar adjacency number problem for genomes with gene repetitions

Zhixiang Chen[a], Bin Fu[a], Randy Goebel[b], Guohui Lin[b], Weitian Tong[b], Jinhui Xu[c], Boting Yang[d], Zhiyu Zhao[e], Binhai Zhu[f,*]

[a] *Department of Computer Science, University of Texas-American, Edinburg, TX 78739-2999, USA*
[b] *Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada*
[c] *Department of Computer Science, SUNY-Buffalo, Buffalo, NY 14260, USA*
[d] *Department of Computer Science, University of Regina, Regina, Saskatchewan S4S 0A2, Canada*
[e] *Department of Computer Science, University of New Orleans, New Orleans, LA 70148, USA*
[f] *Department of Computer Science, Montana State University, Bozeman, MT 59717-3880, USA*

ARTICLE INFO

Article history:
Received 21 October 2012
Received in revised form 8 July 2014
Accepted 9 July 2014
Available online 17 July 2014
Communicated by M. Crochemore

Keywords:
Genome comparison
Adjacency
Breakpoint
NP-hard
Approximation algorithm

ABSTRACT

In this paper, we apply a measure, *exemplar adjacency number*, which complements and extends the well-studied breakpoint distance between two permutations, to measure the similarity between two genomes (or in general, between any two sequences drawn from the same alphabet). For two genomes $\mathcal{G}$ and $\mathcal{H}$ drawn from the same set of n gene families and containing gene repetitions, we consider the corresponding Exemplar Adjacency Number problem (EAN), in which we delete duplicated genes from $\mathcal{G}$ and $\mathcal{H}$ such that the resultant exemplar genomes (permutations) G and H have the maximum adjacency number. We obtain the following results. First, we prove that the one-sided 2-repetitive EAN problem, i.e., when one of $\mathcal{G}$ and $\mathcal{H}$ is given exemplar and each gene occurs in the other genome at most twice, can be linearly reduced from the Maximum Independent Set problem. This implies that EAN does not admit any $O(n^{0.5-\epsilon})$-approximation algorithm, for any $\epsilon > 0$, unless P = NP. This hardness result also implies that EAN, parameterized by the optimal solution value, is W[1]-hard. Secondly, we show that the two-sided 2-repetitive EAN problem has an $O(n^{0.5})$-approximation algorithm, which is tight up to a constant factor.

1. Introduction

In genome comparison and rearrangement studies, the breakpoint distance is one of the most well-known distance measures [19]. The implicit idea of breakpoints was initiated as early as in 1936 by Sturtevant and Dobzhansky [18]. While gene duplication/loss, etc., is an inseparable part of evolution, due to the difficulty of handling duplicated genes, until only a few years ago, in computational genomics it had been largely assumed that every gene appears in a genome exactly once. Under this assumption, the genome rearrangement problem is essentially the problem of comparing and sorting unsigned

[*] Corresponding author.
E-mail addresses: chen@cs.panam.edu (Z. Chen), binfu@cs.panam.edu (B. Fu), rgoebel@ualberta.ca (R. Goebel), guohui@ualberta.ca (G. Lin), weitian@ualberta.ca (W. Tong), jinhui@cse.buffalo.edu (J. Xu), boting@cs.uregina.ca (B. Yang), sylvia@cs.uno.edu (Z. Zhao), bhz@cs.montana.edu (B. Zhu).

http://dx.doi.org/10.1016/j.tcs.2014.07.011
0304-3975/© 2014 Elsevier B.V. All rights reserved.

(or signed) permutations [12,10]. Computing the breakpoint distance between two permutations over the same alphabet can be done in linear time.

Genomes in the form of permutations are hard to obtain and so far, can only be obtained in several small virus genomes. In fact, these kinds of 'perfect' genomes do not occur on eukaryotic genomes where paralogous genes are common [16,17]. In practice, it is important to compute genomic distances between genomes in the form of permutations, such as is done by using the Hannenhalli–Pevzner method [12]. However, more often than never, one might have to handle the gene duplication problem. (Interested readers are referred to a recent survey on this topic [20].) In 1999, Sankoff proposed a way to select, from the duplicated copies of a gene, the common ancestral gene such that the distance between the reduced genomes (called *exemplar genomes*) is minimized. For this case, Sankoff produced a branch-and-bound algorithm [17]. In a subsequent work, Nguyen, Tay and Zhang proposed a divide-and-conquer method to compute the exemplar breakpoint distance empirically [16].

From the algorithmic complexity research point of view, it has been shown that computing the *exemplar signed reversal distance* and computing the *exemplar breakpoint distance* between two general genomes (i.e., with gene duplications) are both NP-hard [3]. A few years ago, Blin and Rizzi further proved that computing the *exemplar conserved interval distance* between two such genomes is NP-hard [2]; furthermore, it is NP-hard to compute the *minimum conserved interval matching*, that is, without deleting the duplicated copies of genes. On the approximability, for any exemplar genomic distance measure $d(\cdot,\cdot)$ satisfying coincidence axiom (i.e., $d(G,H)=0$ if and only if $G=H$ or the reversal of H), it was shown that the problem does not admit any approximation algorithms, even when each gene appears at most three times in each input genome unless P $=$ NP [8,6]. A few years later, this bound was tightened, as deciding when $d(\mathcal{G},\mathcal{H})=0$ is NP-complete even if each gene appears in the input genomes $\mathcal{G}$ and $\mathcal{H}$ at most twice [1,13]. It follows that for the exemplar breakpoint distance and the exemplar conserved interval distance problems, there are no polynomial time approximation algorithms. Furthermore, even under a weaker definition of polynomial time approximation algorithms, the exemplar breakpoint distance problem is shown not to admit any weak $O(n^{1-\epsilon})$-approximation algorithm, for any $0<\epsilon<1$, where n is the maximum length of the two input genomes [8]. The exemplar conserved interval distance problem is also shown not to admit any weak $O(n^{1.5})$-approximation algorithm [6,7].

Complementary to the genomic distances, computing certain genomic similarities between two genomes has also been studied in [4]. In general, genomic similarity measures do not satisfy coincidence axiom. Among others, Chauve *et al.* proved that computing the maximum *exemplar common interval similarity* between two general genomes is NP-hard, while leaving open the problem approximability [4].

Here we study the *exemplar adjacency number* between two (general) genomes, which complements the breakpoint distance measure. Formally, given an alphabet Σ of n genes and two genomes $\mathcal{G}$ and $\mathcal{H}$ drawn from Σ, the *Exemplar Adjacency Number* problem (*EAN* for short) is to delete duplicated genes from $\mathcal{G}$ and $\mathcal{H}$ such that the number of adjacencies between the two resultant exemplar genomes (i.e., permutations), G and H, is maximized. The EAN problem is NP-hard, and here we study the approximability. When one of the input genomes is already exemplar, the problem is called one-sided EAN; the general case is also called two-sided EAN. We first present a linear reduction from the *Maximum Independent Set* (MIS) problem to the one-sided 2-repetitive EAN problem. This reduction implies that the one-sided EAN problem is W[1]-hard, and that it does not admit an $O(n^{0.5-\epsilon})$-approximation algorithm, for any $\epsilon>0$, unless P $=$ NP. The W[1]-hardness (see [9] for details) and the recent lower bound results [5] imply that, if k is the optimal solution value to the one-sided EAN problem, then barring an unlikely collapse in the parameterized complexity hierarchy, the problem is not solvable in time $f(k)n^{o(k)}$, for any function f. Our second positive result is an $O(n^{0.5})$-approximation for the two-sided 2-repetitive EAN problem. Ignoring constants, the negative hardness result and the positive algorithmic result match perfectly for this case.

The rest of the paper is organized as follows. In Section 2, we summarize some of the necessary background definitions. Section 3 presents the linear reduction from the MIS problem to the one-sided EAN problem, and we draw the conclusion on inapproximability. The positive algorithmic result is presented in Section 4, with both the design and the analysis of the $O(n^{0.5})$-approximation algorithm. Section 5 concludes the paper with some discussions.

2. Preliminaries

In the (pairwise) genome comparison and rearrangement problems, we are given two genomes, each of which is a sequence of signed (or unsigned) genes. Note that in general a genome can be a set of such sequences (i.e., with multiple chromosomes); yet in this paper we focus on such one-sequence genomes, often called *singletons*. The order of the genes in one genome corresponds to their physical positions on the genome, and the sign of a gene indicates which one of the two DNA strands the gene is located. In the literature, most of the research assumes that each gene occurs exactly once in a genome; such an assumption is problematic in reality for eukaryotic genomes and the like where duplications of genes exist [17]. For such a general genome, Sankoff proposed to select an *exemplar genome*, by deleting duplicated copies of each gene, in which every gene appears exactly once. The deletion is to minimize certain genomic distance between the resultant exemplar genomes [17].

The following definitions are very much the same as those in [3,8]. In this paper, we consider only unsigned genomes, though our results can be applied to signed genomes. We assume a gene alphabet Σ that consists of n distinct genes. A genome $\mathcal{G}$ is a sequence of elements of Σ, under the constraint that each element occurs at least once in $\mathcal{G}$. We allow repetitions of every gene in any genome. Specifically, if each gene occurs exactly once in a genome, then the genome is