Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

An efficient algorithm for one-sided block ordering problem under block-interchange distance [☆]

Kun-Tze Chen [a], Chi-Long Li [a], Hsien-Tai Chiu [b,*], Chin Lung Lu [a,*]

[a] Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan
[b] Department of Chemistry, National Cheng Kung University, Tainan City 701, Taiwan

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 29 July 2015
Accepted 6 October 2015
Available online xxxx
Communicated by G.F. Italiano

Keywords:
Algorithm
Computational biology
One-sided block ordering problem
Block-interchange

ABSTRACT

In this work, we study the one-sided block ordering problem under block-interchange distance. Given two signed permutations π and σ of size n, where π represents a partially assembled genome consisting of several blocks (i.e., contigs) and σ represents a completely assembled genome, the one-sided block ordering problem under block-interchange distance is to order (i.e., assemble) the blocks of π such that the block-interchange distance between the assembly of π and σ is minimized. The one-sided block ordering problem is useful in genome resequencing, because its algorithms can be used to assemble the contigs of partially assembled resequencing genomes based on their completely assembled genomes. By using permutation groups in algebra, we design an efficient algorithm to solve the one-sided block ordering problem under block-interchange distance in $\mathcal{O}(n \log n)$ time. Moreover, we show that the assembly of π can be done in $\mathcal{O}(n)$ time and its block-interchange distance from σ can also be calculated in advance in $\mathcal{O}(n)$ time.

1. Introduction

During a past decade, the next-generation DNA sequencing techniques have greatly advanced [1–3], which allows an increasing number of draft genomes to be produced rapidly in an ever-decreasing cost. Usually, these draft genomes are partially sequenced and thus their released genomes are just collections of unassembled *contigs* (i.e., contiguous segments of chromosomes). However, these draft genomes in contig form cannot be used immediately by current algorithms for studying genome rearrangement, because these algorithms need completely assembled genomes as their input. Basically, in the studies of genome rearrangements [4], a gene is usually represented by a signed integer, with sign indicating on which of the two complementary DNA strands the gene resides, and a chromosome by a sequence of integers corresponding to those genes on the chromosome. During evolution, the gene order in a genome is subject to be changed by rearrangements, such as reversal, transposition, block-interchange, fusion/fission, and translocation (see Fig. 1 for graphical illustration). A *reversal* (also called *inversion*) rearranges a segment of consecutive integers on the chromosome by reversing the order and the signs of the integers [5–7]. A *transposition* moves a segment on a chromosome to another location or, equivalently, exchanges two adjacent and non-overlapping segments on the chromosome [8,9]. A *block-interchange* is a generalized transposition that exchanges two nonoverlapping but not necessarily adjacent segments on a chromosome [10–12]. A *fusion* joins two chromosomes into a bigger one and a fission breaks a chromosome into two smaller ones [13,14]. A *transloca-*

[☆] The preliminary version of this paper was presented at the 19th Annual International Computing and Combinatorics Conference (COCOON 2013).

[*] Corresponding authors. Tel.: +886 6 2757575x65322, +886 3 5731205.
E-mail addresses: holystu@gmail.com (K.-T. Chen), 496470140@ntnu.edu.tw (C.-L. Li), hchiu@mail.ncku.edu.tw (H.-T. Chiu), cllu@cs.nthu.edu.tw (C.L. Lu).

http://dx.doi.org/10.1016/j.tcs.2015.10.010
0304-3975/© 2015 Elsevier B.V. All rights reserved.

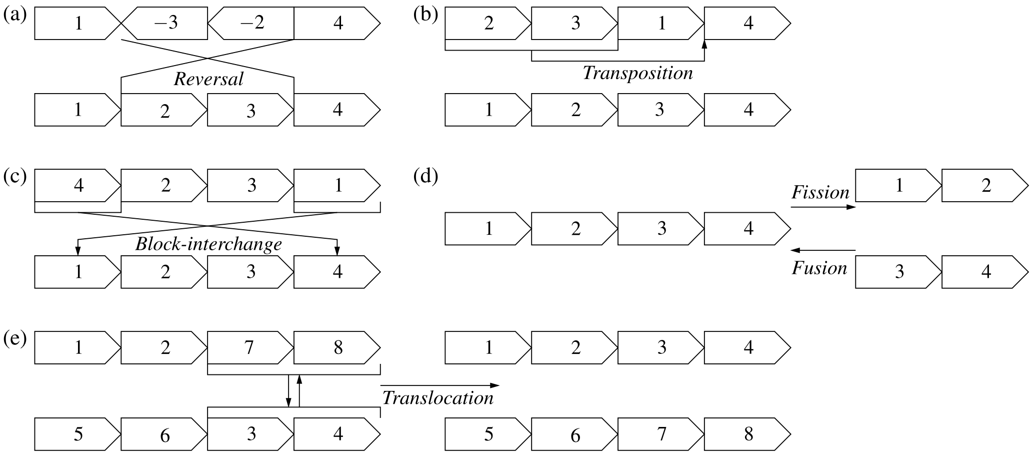

Fig. 1. Illustrated examples of chromosomal rearrangements: (a) reversal, (b) transposition, (c) block-interchange, (d) fission and fusion, and (e) translocation, where an integer represents a gene and its sign indicates the strandedness of the corresponding gene.

tion exchanges an end segment of a chromosome, which contains an end (i.e., telomere) of this chromosome, with an end segment of another chromosome [13,15–17]. Given two completely assembled genomes on the same set of genes and a set of possible rearrangements, the *genome rearrangement problem* aims to find a shortest series of rearrangements (or a series of rearrangements with minimum weight when rearrangements are weighted according to the probabilities of their occurrences) required to transform one genome into the other. The length (or weight) of an optimal series of rearrangements is then called *genome rearrangement distance*. Furthermore, the genome rearrangement distance is also called *block-interchange distance* (respectively, *reversal distance*) if only block-interchanges (respectively, reversals) are used to do the transformation. The genome rearrangement distances are useful in the studies of phylogeny reconstruction [11,18], because they can serve as an indicator of an evolutionary distance between organisms.

In [19], Gaul and Blanchette introduced the block ordering problem to address the issue caused by draft genomes as mentioned above. They used blocks to denote the contigs of a draft genome, that is, each *block* represents an ordered list of genes on the corresponding contig. Given two partially assembled genomes, with each being represented as an unordered set of blocks, the *block ordering problem* is to order and orient (i.e., assemble) the blocks of the two genomes such that the genome rearrangement distance between the two completely assembled genomes is minimized. In their work [19], Gaul and Blanchette studied this problem under reversal distance and proposed a linear-time algorithm to solve the block ordering problem when the problem is further simplified to maximize the number of cycles in the breakpoint graph of the completely assembled genomes. The rationale behind is based on a result obtained by Bourque and Pevzner [20], showing that the reversal distance between two completely assembled genomes can be approximated by maximizing the number of cycles in their corresponding breakpoint graph. In fact, in addition to the number of cycles, the number of hurdles, as well as the presence of a fortress or not, is also important and needed for determining the actual reversal distance [5]. Therefore, it is still a challenge to efficiently solve the block ordering problem by optimizing a true rearrangement distance.

In this paper, we study a variant of the block ordering problem, called as *one-sided block ordering problem*, for which one of the two input genomes is still partially assembled but the other is completely assembled, with optimizing the block-interchange distance (refer to the Preliminaries section for its formal definition). This one-sided block ordering problem has a useful application in genome resequencing [21], because the reference genome for a resequencing organism can be input as the completely assembled genome in the one-sided block problem so that the contigs of partially assembled genome for the resequencing organism can be assembled together. Once the complete genomes of resequencing organisms are obtained, they can be further used in the studies of genome rearrangement and phylogeny reconstruction. In [22,23], we have utilized permutation groups in algebra, instead of the breakpoint graphs used by Gaul and Blanchette [19], to design an $\mathcal{O}(\delta n)$ time algorithm for solving the one-sided block ordering problem under the genome rearrangement distance measured by weighted reversals and block-interchanges, whose weights are 1 and 2, respectively, where n is the number of genes and δ is the number of used reversals and block-interchanges. Note that $\delta \leq n$. In this study, we consider the one-sided block ordering problem under the genome rearrangement distance measured only by block-interchanges and design an efficient algorithm of $\mathcal{O}(n \log n)$ time to solve this problem. In addition, we show that the assembly of the partially assembled genome can be done in $\mathcal{O}(n)$ time and its block-interchange distance to the completely assembled genome can be calculated in advance in $\mathcal{O}(n)$ time.

Note that a problem related to the block ordering problem is that of computing the genome rearrangement distance between partially ordered genomes, which was introduced by Zheng et al. [24] and further studied by many researchers [25–29]. Here, we call this problem as minimum rearrangement linearization problem. In this problem, due to missing genes or missing order information of some neighboring genes, the input genomes to be compared are characterized only