Computers & Operations Research 47 (2014) 1–10

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Tabu search algorithm for DNA sequencing by hybridization with multiplicity information available

Kamil Kwarciak [a,*], Piotr Formanowicz [a,b]

[a] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[b] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland

ARTICLE INFO

Available online 22 January 2014

Keywords:
Sequencing by hybridization
Repetitions
Multiplicity information
Metaheuristic
Tabu search

ABSTRACT

The classical DNA sequencing by hybridization (SBH) uses a binary information about oligonucleotide presence in an analyzed DNA sequence. A given oligonucleotide is or is not a part of the sequence. However, the development of the DNA chip technology allows to take into consideration some information about repetitions in the target sequence. Currently, it is not possible to determine the exact data of such type but even partial multiplicity information should be very useful.

In this paper two simple but realistic multiplicity information models are taken into account. The first one assumes that it is known if a given oligonucleotide occurs in the analyzed sequence once or more than once. According to the second model it is possible to determine if a given oligonucleotide appears in the target sequence once, twice or at least three times. A tabu search algorithm has been implemented to verify these models. It solves the problem with any kind of hybridization errors. Computational experiment results confirm that the additional information leads to an improvement of the reconstruction process. They also show that the more precise model of information increases the quality of the obtained solutions.

Test data sets and the implemented tabu search algorithm are available on: http://bio.cs.put.poznan.pl/files/52234a7c9dfb89b808000001/download.

1. Introduction

The DNA sequencing is one of the most important problems of molecular and computational biology. Its goal is to determine a sequence of nucleotides an analyzed DNA sequence consists of. There exists many methods to obtain this information. The DNA sequencing by hybridization (SBH) is one of them [1,2]. It is comprised of two stages. The first one is a biochemical experiment. The result of this stage is a *spectrum*. It is a set of l-long subsequences of an examined DNA sequence. The elements of this set are also called l-*mers*. In the second stage a computational problem must be solved. The spectrum is used to reconstruct the sequence.

The biochemical experiment uses the ability of single stranded DNA molecules to join to complementary strands [3]. A single stranded DNA molecule is a sequence of nucleotides which are denoted by A, C, G and T. These letters represent four possible nucleotides: adenine, cytosine, guanine and thymine, respectively. According to the Watson–Crick complementarity rule adenine is complementary to thymine and cytosine is complementary to guanine. If two single stranded DNA molecules contain some substrings being complementary to each other then these DNA molecules may join by hydrogen bonds and create a double stranded molecule. This process is called hybridization.

The biological stage requires a DNA chip [4,5] and a solution containing a number of copies of a single stranded target DNA sequence. The DNA chip comprises a full l-long oligonucleotide library. It is a kind of matrix divided into cells. Each cell contains a number of copies of a given l-mer from the library. When such a chip is put into the DNA solution, some parts of the analyzed DNA hybridize to complementary oligonucleotides on the chip. If the copies of the examined DNA sequence are fluorescently or radioactively marked then one can obtain an image of the DNA chip. This image presents a spectrum composition for the target sequence.

In the ideal case a result of the biochemical experiment is a complete and correct information about l-long oligonucleotides present in an examined DNA sequence. However, some errors may occur during the experiment. There are two types of errors: positive ones and negative ones. A positive error is an issue when a spectrum contains an l-mer which is not a subsequence of the target DNA. It occurs when the analyzed sequence hybridizes to an oligonucleotide on the chip which is not perfectly complementary to it. A contrary case is possible too. The examined sequence may not hybridize to a perfectly complementary l-mer on the chip. In this case spectrum does not contain all l-long subsequences of the target as a result. Another cause of negative errors and missed information about target DNA composition is a repetitive sequence. In the classical SBH approach spectrum is a set, not a multiset, so the information about repetitions is lost.

* Corresponding author. Tel.: +48 61 8790790; fax: +48 61 8771525.
E-mail addresses: kamil.kwarciak@cs.put.poznan.pl (K. Kwarciak), piotr.formanowicz@cs.put.poznan.pl (P. Formanowicz).

0305-0548/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2014.01.007

In the classical sequencing by hybridization the result of the biochemical experiment is a binary information about l-mer presence in an examined DNA sequence (i.e. a given oligonucleotide is or is not included in an obtained spectrum). Consequently, repetitions in the target lead to negative errors. However, the current development of the DNA chip technology allows to take into consideration information about an intensity of the chip signals. This intensity can be, at least to some extent, correlated with the number of occurrences of a given oligonucleotide in the analyzed sequence. Unfortunately, the precision of this signal depends on l-mer multiplicity. The accuracy decreases when the number of repetitions of an oligonucleotide increases. It is possible to easily differentiate a shining of one and many occurrences but distinguishing the signal coming from, for example, seven and eight occurrences would be very difficult (or even not possible). Nevertheless, even partial information about repetitions can be very useful.

Formal definitions of SBH problems with the additional information about multiplicity have been formulated in [6,7]. To the best of authors' knowledge, only a greedy algorithm presented in [8] utilizes the additional information. The algorithm has been implemented to verify its usefulness and the results confirm that using it leads to an increased quality of the obtained solutions. However, a greedy approach is a simple algorithm. One could use its result as an input to a more sophisticated heuristic to obtain further solution improvement.

Dealing with repetitions has also been identified as an important issue in the literature related to the classical sequencing by hybridization. Some results have been presented in [9,10]. However, the algorithms tested in these articles (a tabu search [11], a hybrid genetic algorithm [12] and a revised hybrid genetic algorithm [10]) do not use any information about repetitions.

In this paper DNA sequencing by hybridization with any kind of errors is taken into account. Positive errors as well as negative errors are accepted and the cause of an error (hybridization error or repetitious sequence) is not relevant. Two models of the multiplicity information are considered. They are the simplest ones but realistic and possible to apply in practice. The first one, called “one and many”, assumes it is known if a given oligonucleotide occurs in an analyzed sequence once or more than once. The second model, called “one, two and many”, uses the knowledge if a given oligonucleotide appears in a target sequence once, twice or at least three times. The current DNA chip technology justifies these assumptions. The foundation of these models is the signal intensity of DNA chip images. It is a common practice to take into account such information in gene expression analysis [13].

A tabu search algorithm presented in this paper uses as a part of the input data information about the first oligonucleotide of an examined sequence. The assumption that it is known is well justified since the PCR commonly used to produce a large number of copies of a target DNA requires the knowledge of the first l-mer. Moreover, if the information which oligonucleotide is the first one would be not available then the algorithm can be executed without it polynomial number of times, each time with a different l-mer from spectrum as the first one.

The next section contains formal problem definitions. The tabu search algorithm is described in Section 3. The results of computational experiments are discussed in Section 4. The last section contains a summary and conclusions.

2. Problem definition

Formal definitions of problems of DNA sequencing by hybridization with information about repetitions have been formulated in [6,7]. In this section only a brief overview of them is presented.

It is required to introduce additional types of spectra in order to precisely state the problems. Let $S(Q)$ denote a spectrum of sequence Q, $S^{(m)}(Q)$ be a multispectrum of sequence Q and m_i be the number of occurrences of a given sequence s_i (l-mer) in $S^{(m)}(Q)$. The spectrum is a set so every s_i may occur at most once in $S(Q)$. The multispectrum is a multiset. Consequently, a given l-mer may be present in $S^{(m)}(Q)$ more than once. Let $S^{(is)}(Q)$ denote an ideal spectrum of the sequence. The ideal spectrum contains all and only those types of l-mers but not all of these l-mers, which compose the target sequence Q. All of these l-mers compose an ideal multispectrum of sequence Q, which is denoted by $S^{(im)}(Q)$. Note that the number of occurrences m_i of any l-mer in the ideal multispectrum is equal to the number of its repetitions in sequence Q. The multispectrum may be affected by errors, so the value m_i for a given l-mer may differ from its repetition count in sequence Q. It may be smaller in the case of negative errors and greater in the case of positive errors.

Let us assume that the additional multiplicity information is not available. Then the combinatorial problem of sequencing by hybridization with errors may be stated as follows:

Problem 1. Lack of the multiplicity information

Instance: Set $S(Q)$, length n of sequence Q.

Answer: Sequence Q' of length at most n with the maximum utilization U of set $S(Q)$ defined as

$$U = \sum_{s_i \in S(Q)} u_i(Q')$$

where $u_i(Q') = 1$ if Q' contains s_i and $u_i(Q') = 0$ otherwise.

If the information about repetitions is available then the utilization of set $S(Q)$ for a sequence Q' is redefined as follows:

$$\tilde{U} = \sum_{s_i \in S(Q)} \tilde{u}_i(Q')$$

where $\tilde{u}_i(Q')$ is equal to the number of occurrences of s_i in Q'.

Assuming one is able to obtain in the biochemical experiment information according to the model “one and many” the sequencing problem may be defined as follows:

Problem 2. Multiplicity information of the type “one and many”

Instance: Set $S(Q)$, length n of sequence Q, parameter $m_i \in \{1, 2\}$ for every $s_i \in S(Q)$.

Answer: Sequence Q' of length at most n with the maximum utilization $\tilde{U}$ of set $S(Q)$ and $\min_i \le \tilde{u}_i(Q') \le \max_i$ for each $s_i \in S(Q)$. If $m_i = 1$ then $\min_i = 0$ and $\max_i = 1$. If $m_i = 2$ then $\min_i = 1$ and $\max_i = \infty$.

If there exists partial multiplicity information of the type “one, two and many” then the combinatorial problem may be stated as follows:

Problem 3. Multiplicity information of the type “one, two and many”

Instance: Set $S(Q)$, length n of sequence Q, parameter $m_i \in \{1, 2, 3\}$ for every $s_i \in S(Q)$.

Answer: Sequence Q' of length at most n with the maximum utilization $\tilde{U}$ of set $S(Q)$ and $\min_i \le \tilde{u}_i(Q') \le \max_i$ for each $s_i \in S(Q)$. If $m_i = 1$ then $\min_i = 0$ and $\max_i = 1$. If $m_i = 2$ then $\min_i = 1$ and $\max_i = 2$. If $m_i = 3$ then $\min_i = 2$ and $\max_i = \infty$.

All the problems defined above may be transformed into a variant of the traveling salesman problem (TSP). An instance of the classical TSP consists of a directed or undirected input graph and a weight (cost) assigned to each arc or edge. The goal of the classical TSP is to find the shortest (minimal cost) Hamiltonian cycle (i.e. a cycle containing all vertices exactly once).

One can obtain a problem corresponding to the sequencing by hybridization by modifying TSP as follows. Firstly, the goal is to