Available online at www.sciencedirect.com
SCIENCE DIRECT
Discrete Applied Mathematics 154 (2006) 718–729
DISCRETE APPLIED MATHEMATICS
www.elsevier.com/locate/dam

Computational complexity of isothermic DNA sequencing by hybridization☆

Jacek Blazewicz[a, b, *], Marta Kasprzak[a, b]

[a] *Institute of Computing Science, Poznan University of Technology, Poznan, Poland*
[b] *Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

Received 5 November 2002; received in revised form 30 October 2003; accepted 31 May 2005
Available online 23 November 2005

Abstract

In the paper, the computational complexity of several variants of the problem of isothermic DNA sequencing by hybridization, is analyzed. The isothermic sequencing is a recent method, in which isothermic oligonucleotide libraries are used during the hybridization with an unknown DNA fragment. The variants of the isothermic DNA sequencing problem with errors in the hybridization data, negative ones or positive ones, are proved to be strongly NP-hard. On the other hand, the polynomial time algorithm for the ideal case with no errors is proposed.

Keywords: DNA sequencing; Computational complexity; Eulerian path

1. Introduction

The *DNA sequencing by hybridization* (SBH) is a basic problem in one of the approaches leading to a reconstruction of DNA chains. It consists in determining a sequence of nucleotides of an unknown DNA fragment [3,24,31,13,14,2,5,20,27,28,15,19,30,34] (see also an excellent overview of this subject in a recent book [26]). Its input data come from a biochemical *hybridization experiment*, and they can be viewed as a set (called *spectrum*) of words (*oligonucleotides*) over the alphabet {A, C, G, T}, being short subsequences of the studied DNA fragment. The aim is to reconstruct the original DNA sequence of a known length on the basis of these overlapping words.

In the standard approach to the DNA sequencing by hybridization, the oligonucleotide library used in the hybridization experiment contains all possible oligonucleotides of a given constant length (cf. [1,6,18,29,33]). The spectrum being output of the experiment is a subset of the library, i.e. the set of words of equal length composing the original sequence. For the standard DNA sequencing, the computational complexity of several variants of the problem is already known. The variant with no errors in the spectrum is polynomially solvable [25], while the variants assuming presence of errors in the data (negative ones, positive ones, or both) are all strongly NP-hard [11]. The present work concerns the computational complexity of several variants of the isothermic DNA sequencing by hybridization.

☆ The research has been supported by KBN grant 3T11F00227.
* Corresponding author. Tel.: +48 61 8790 790; fax: +48 61 8771 525.
E-mail addresses: jblazewicz@cs.put.poznan.pl (J. Blazewicz), marta@cs.put.poznan.pl (M. Kasprzak).

0166-218X/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2005.05.028

In the isothermic version of the DNA sequencing, the hybridization experiment is performed with *isothermic oligonucleotide libraries*, which contain oligonucleotides of equal "temperatures" (in fact, melting temperatures of oligonucleotide duplexes), but different lengths. The isothermic approach is a novel method [7–9], in which oligonucleotides contained in an isothermic library should form duplexes with their complements (after a hybridization) in a more narrow range of experimental conditions (temperature, salt concentration etc.) than that characteristic for an oligonucleotide library with oligomers of the same length. Therefore, the hybridization experiments performed with isothermic libraries should result in a smaller number of experimental errors. The use of such libraries should substantially limit a number of these errors to be considered in the computational phase of the SBH approach. On the other hand, this approach may also avoid some repetitions, being a serious drawback of the standard approach (cf. Example 1).

To better describe the problem we recall the following definition and basic properties of the hybridization process.

Definition 1 (*Blazewicz et al. [9]*). An *isothermic oligonucleotide library L* of temperature $\mathscr{T}_L$ is a library of all oligonucleotides satisfying the relations

$$w_A x_A + w_C x_C + w_G x_G + w_T x_T = \mathscr{T}_L,$$
$$w_A = w_T,$$
$$w_C = w_G,$$
$$\text{and } 2w_A = w_C,$$

where w_A, w_C, w_G, and w_T are increments of nucleotides A, C, G, and T, respectively, and x_A, x_C, x_G, and x_T denote numbers of these nucleotides in the oligonucleotide. It is assumed that $w_A = w_T = 2$ degrees and $w_C = w_G = 4$ degrees [32].

Claim 1 (*Blazewicz et al. [9]*). *One isothermic library is not sufficient to cover all DNA sequences.*

As the example for the above claim we can give any sequence of the type $[CG]^+$ (e.g. CCGCGGG), which is not possible to be covered by oligonucleotides from a library of a temperature not divisible by 4. On the other side, a library of a temperature divisible by 4 does not cover any sequence of the type $[CG]^+[AT][CG]^+$ (e.g. CCCACGG).

Claim 2 (*Blazewicz et al. [9]*). *It is always possible to cover any DNA sequence by probes coming from two isothermic libraries of temperatures differing by* 2 *degrees. Moreover, this coverage is such that in the sequence two consecutive oligonucleotides* (*from the libraries*) *have starting points shifted by at most one position.*

The reconstruction is complicated by the presence of *errors* within the spectrum, coming from the hybridization experiment. We distinguish two types of the errors: *negative* ones, i.e. missing words in the spectrum (the words which are parts of the original sequence but are not present in the set), and *positive* ones, i.e. additional words in the spectrum (the words which are not parts of the original sequence but are present in the set). The repetitions of oligonucleotides in the original sequence are treated as negative errors, because the experiment cannot count the number of occurrences of oligonucleotides, it only checks their presence. Of course, we do not have information which words are missing or erroneous within the set.

Example 1 illustrates the difference between two SBH approaches: using the standard and isothermic libraries, respectively. We see, that the usage of isothermic libraries enables one to avoid some negative errors coming from repetitions.

Example 1. Let our sequence to be reconstructed be AAATGTAAA. This will result in one negative error (AAA), since repeated oligonucleotides cannot be measured twice in the hybridization experiment. Additionally, let us assume that a positive error (GTT) appeared. Thus, our spectrum will now have the form {AAA, AAT, ATG, GTA, GTT, TAA, TGT}. The reconstruction phase is shown in Fig. 1.

The cardinality of the standard library used in this example is $4^3 = 64$ oligonucleotides. The closest (from the cardinality point of view) is to perform the hybridization experiment with two isothermic libraries of melting temperatures equal to, respectively, 6 and 8 degrees. Their cardinalities are: $card(6) = 16$ and $card(8) = 44$. We see, that the total cardinality of these two libraries is smaller than that of the standard library used above. Now, in the new experiment the obtained spectrum will be as follows: {AAA, AAAT, AAT, ATG, GT, GTA, GTT, TAA, TAAA, TG, TGT}. As before,

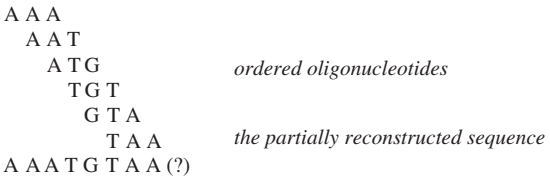


Fig. 1. The erroneous spectrum results in a partial reconstruction of the unknown sequence.

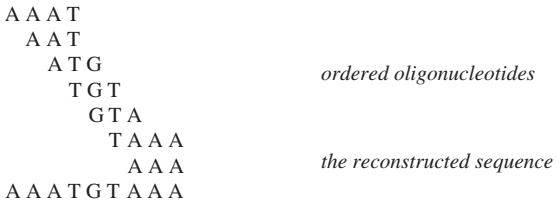


Fig. 2. Isothermic libraries allow for an errorless reconstruction of the sequence with repetitions given in Fig. 1.

we added one positive error — GTT, while AAA appeared only once, but this time this fact will not have an influence on the reconstructed sequence (see Fig. 2). Basing on Claim 2, we could even omit (in the reconstruction phase) some oligonucleotides (TAA, GT, TG), which were fully contained in longer ones, their beginning not being shifted to the right with respect to the latter.

Example 1 illustrates well the good feature of the analyzed isothermic libraries, i.e., their ability to avoid some short repetitions which standard libraries cannot. What is more, this new method improves also biochemical conditions in which the hybridization experiment takes place, thus, reducing a number of experimental errors. On the other hand, long repetitions (longer than the oligonucleotides in the library used) will still cause a problem for a proper reconstruction of the original sequence in practice. At this point let us also mention that the formula defining the isothermic library can be also more complicated, reflecting better a structure of a hybridized sequence (cf. [12]), but this context dependent measure will be much more complicated to handle mathematically.

As we mentioned, the present paper is devoted to the complexity analysis of the computational phase of the sequencing by hybridization with isothermic libraries. While the problems with different types of errors occurring in the spectrum are NP-hard in the strong sense, the ideal case involving no errors is polynomially solvable. This last result is quite surprising, provided the complicated nature of the outcome of the biochemical experiment proposed and the resulting overlaps.

The organization of the paper is as follows. Section 2 contains proofs of strong NP-hardness of isothermic DNA sequencing problem with negative errors or with positive ones. In Section 3 a polynomial time algorithm for isothermic DNA sequencing without errors in the hybridization data, is proposed. We conclude in Section 4.

2. Isothermic DNA sequencing with errors

The general problem of isothermic DNA sequencing by hybridization, assuming both types of error in the instance, negative and positive ones, has been proved to be strongly NP-hard in [9]. Note, that in this case both versions of the problem, i.e. the decision and the search one, respectively, are NP-hard. In this section, we prove that the isothermic sequencing problems with errors of one type, only negative ones or only positive ones, are also strongly NP-hard, but only in their search versions.

Considering decision versions of the subproblems assuming errors of only one type, Π_{nisd} (assuming only negative errors) and Π_{pisd} (assuming only positive errors, see the following subsections), one has the additional information not present in the decision formulation of the general isothermic DNA sequencing problem. It is the knowledge that for the instances defined for these subproblems at least one solution always exists. The spectrum with only negative errors is — by the definition — a subset of the *ideal multispectrum* for an original sequence. Such a multispectrum is a multiset containing all members of two oligonucleotide libraries of temperatures $\mathscr{T}$ and $\mathscr{T}+2$, which can be distinguished in an

original sequence (of the known length n). It is obvious, that all oligonucleotides from the ideal multispectrum always build some sequence of length n. Similarly, all oligonucleotides from the spectrum with only negative errors always can be included in some sequence of length n. In the case of only positive errors, the *ideal spectrum* for an original sequence is always a subset of the spectrum from the instance, the ideal spectrum being a set defined analogously to the ideal multispectrum. (Note, that the ideal spectrum does not exist for sequences with repetitions of oligonucleotides of temperature $\mathscr{T}$ or $\mathscr{T}+2$; in such case we can consider the ideal multispectrum.) Having the ideal spectrum, it is always possible to construct some sequence of length n with all its oligonucleotides of temperature $\mathscr{T}$ or $\mathscr{T}+2$ present in the ideal spectrum, and the oligonucleotides will be unique within the sequence. The original sequence is the example of a possible solution. In the analyzed problem we do not have to use all oligonucleotides from the spectrum during a construction of the sequence, hence a solution for the problem always exists.

Therefore, the two decision problems Π_{nisd} and Π_{pisd} are trivially solvable. However, the time complexity of their search counterparts Π_{niss} and Π_{piss} (see the following subsections), cannot be bounded by a polynomial function. To study the computational complexities of the two problems in search versions, two additional decision *quasi-sequencing* problems Π_{niqd} and Π_{piqd} must be introduced. In the quasi-sequencing problems an arbitrary set of oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T}+2$ is assumed to be the spectrum (no information about errors is known), thus the answer to the problems is not always "yes". The approach applied here has been previously used in [11], where strong NP-hardness of DNA sequencing problems with erroneous data has been proved for standard oligonucleotide libraries (with a constant length of oligonucleotides). In fact, a similar approach has been applied in [21], Hamiltonian circuit being one of the examples. As it has been stated there, even if we knew a graph contains a Hamiltonian circuit, we could not find it in polynomial time unless P = NP. For "if we had such an algorithm A, we could use it to tell in polynomial time whether an arbitrary graph G has a Hamiltonian circuit. Let p be the polynomial that bounds A's running time on graphs with Hamiltonian circuits. Apply A to G. If G has a Hamiltonian circuit, A will find one in time $p(|G|)$. If G does not have such a circuit, then after $p(|G|)$ steps A could not have found one, and we will know that none exists".

2.1. Negative errors

The isothermic DNA sequencing problem in the case of only negative errors (i.e. there is no false information in the spectrum, but some information is missing) is formulated below in search and decision versions.

Problem 1. Isothermic DNA sequencing by hybridization with only negative errors Π_{niss} — search version.

Instance: Set S (spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T}+2$, length n of an original sequence, where $S \subseteq S''$ and S'' is the ideal multispectrum for this sequence.

Answer: A sequence of length n containing all elements of S.

Problem 2. Isothermic DNA sequencing by hybridization with only negative errors Π_{nisd} — decision version.

Instance: Set S (spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T}+2$, length n of an original sequence, where $S \subseteq S''$ and S'' is the ideal multispectrum for this sequence.

Question: Is there a sequence of length n containing all elements of S?

We see that $D_{\Pi_{\text{nisd}}} = Y_{\Pi_{\text{nisd}}}$, where $D_{\Pi_{\text{nisd}}}$ is the set of all instances of Π_{nisd} and $Y_{\Pi_{\text{nisd}}}$ is the set of all instances of Π_{nisd} with the answer "yes". Thus, the complexity of the above problem is trivially polynomial. However, its search counterpart Π_{niss} is not easily solvable. To prove its strong NP-hardness, the following quasi-sequencing problem is introduced.

Problem 3. Negative isothermic quasi-sequencing Π_{niqd} — decision version.

Instance: Set S (spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T}+2$, length n of an original sequence.

Question: Is there a sequence of length n containing all elements of S?

The two problems Π_{nisd} and Π_{niqd} have the same sets of instances with the positive answer: $Y_{\Pi_{\text{nisd}}} = Y_{\Pi_{\text{niqd}}}$, because any instance of Π_{niqd} resulting in the required sequence belongs to $D_{\Pi_{\text{nisd}}}$. However, problem Π_{niqd} is much more

complicated because it contains also instances with the negative answer ($D_{\Pi_{\text{nisd}}} \subset D_{\Pi_{\text{niqd}}}$). In fact, this problem is strongly NP-complete.

Lemma 1. *Negative isothermic quasi-sequencing problem* Π_{niqd} *is strongly NP-complete.*

Proof. The proof is done by a polynomial transformation from the strongly NP-complete problem directed Hamiltonian path between two vertices [17] and uses some ideas presented in [16]. Given an instance of the problem, a 1-digraph $G = (V, A)$ (i.e. a digraph with set A of arcs not being a multiset) with two specified vertices s and t, the corresponding instance of Π_{niqd} is constructed as follows:

- To every vertex $v \in V$ assign two unique labels e_v and $\overline{e_v}$ of length $x = \lceil \log_2(2|V| + 1) \rceil$ over the alphabet {A, T}. Create the additional unique label $e_\#$.
- Let $R_v = \{w_0, \ldots, w_{\text{OUT}(v)-1}\}$ be a set of successors of v. Create for every vertex $v \in V \backslash \{t\}$ a subset of spectrum elements of the form $\{\overline{e_v} \cdot \text{C} \cdot e_{w_i} \cdot \text{C} \cdot \overline{e_v} \cdot \text{C} | w_i \in R_v\} \cup \{e_{w_i} \cdot \text{C} \cdot \overline{e_v} \cdot \text{C} \cdot e_{w_{i\oplus 1}} \cdot \text{C} | w_i \in R_v\}$ (where "$\oplus$" means addition mod OUT(v), "$\cdot$" is the symbol of concatenation and "C" is the nucleotide).
- Add to the spectrum the subset $\{e_v \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_v} \cdot \text{C} | v \in V \backslash \{s\}\}$ and the element $e_\# \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_s} \cdot \text{C}$.
- Denote by A' the set A without arcs leaving vertex t.

There exists a Hamiltonian path from s to t in graph G iff there exists a sequence of length $(x + 1)(2|A'| + 3|V|)$, including all elements of the spectrum.

Let us assume, that a Hamiltonian path from s to t in graph G exists. Then the solution for problem Π_{niqd} should be constructed in the following way. As the beginning of the sequence oligonucleotide $e_\# \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_s} \cdot \text{C}$ is taken. Next, the sequence of all oligonucleotides corresponding to arcs leaving s is added to the first oligonucleotide, with maximal possible overlap (i.e. $x+1$). These are the oligonucleotides $\{\overline{e_s} \cdot \text{C} \cdot e_{w_i} \cdot \text{C} \cdot \overline{e_s} \cdot \text{C} | w_i \in R_s\} \cup \{e_{w_i} \cdot \text{C} \cdot \overline{e_s} \cdot \text{C} \cdot e_{w_{i\oplus 1}} \cdot \text{C} | w_i \in R_s\}$, and they should be ordered in such way, that the first label appearing in the corresponding sequence is $\overline{e_s}$, the last label is e_v, where v is the second vertex in the Hamiltonian path, and that successive oligonucleotides overlap on $2(x + 1)$ nucleotides. To the created part of the solution, oligonucleotide $e_v \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_v} \cdot \text{C}$ is added, with maximal possible overlap (i.e. $x + 1$). The next parts of the solution are created analogously, and the solution ends with oligonucleotide $e_t \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_t} \cdot \text{C}$. In this way all oligonucleotides from the spectrum have been included into the resulting sequence, and it can be easily checked that its length is equal to $(x + 1)(2|A'| + 3|V|)$.

Now, we assume that a sequence of length $(x + 1)(2|A'| + 3|V|)$ exists and that it includes all elements of the spectrum. If oligonucleotide $e_\# \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_s} \cdot \text{C}$ is assumed to be the first in the sequence, and it is assumed that all remaining oligonucleotides overlap on maximal number of nucleotides from their left side (i.e. on $2(x + 1)$ nucleotides for $2|A'| - |V| + 1$ oligonucleotides and on $x + 1$ nucleotides for remaining $2(|V| - 1)$ oligonucleotides — it follows from their structure), then the resulting sequence will have length $3(x + 1) + (2|A'| - |V| + 1)(x + 1) + 2(|V| - 1)2(x + 1) = (x + 1)(2|A'| + 3|V|)$. If oligonucleotide $e_\# \cdot \text{C} \cdot e_\# \cdot \text{C} \cdot \overline{e_s} \cdot \text{C}$ was not placed at the beginning, the length of the sequence would be greater. Joining the oligonucleotides on the assumed maximal possible overlap leads to the following partial orders. All oligonucleotides corresponding to arcs (i.e. the ones created according to the second item of the transformation) leaving the same vertex must be successive in the sequence. The subsequences created in this way must be joined by oligonucleotides corresponding to vertices (i.e. the ones created according to the third item). In order to attach one oligonucleotide corresponding to a vertex, to the left side of a subsequence and one another to the right side of the subsequence, the subset of arcs being source of the subsequence must include the arc from the "left" vertex to the "right" one. The whole sequence must end with the oligonucleotide corresponding to vertex t since nothing can leave this oligonucleotide with non-zero overlap. Therefore, the order of oligonucleotides corresponding to vertices within the sequence is equivalent to the order of the vertices within the Hamiltonian path. □

The transformation of an instance of directed Hamiltonian path between two vertices problem to an instance of negative isothermic quasi-sequencing problem Π_{niqd}, described in the proof of Lemma 1, is illustrated by the following example.

Example 2. A 1-digraph $G = (V, A)$ being an instance of the Hamiltonian path problem from s to t is given in Fig. 3.

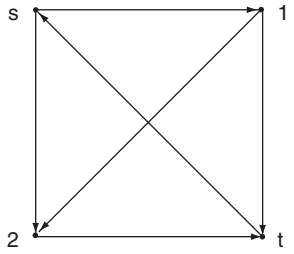


Fig. 3. The example graph $G = (V, A)$ with two specified vertices s and t.

Now we apply the transformation from Lemma 1 to this graph. First, we assign labels over the alphabet {A, T} to all its vertices, plus one extra. Because there are four vertices, the length of the labels is equal to $\lceil \log_2(2 \cdot 4 + 1) \rceil = 4$. The example labels are: $e_s = \text{AAAA}$, $\overline{e_s} = \text{AAAT}$, $e_1 = \text{AATA}$, $\overline{e_1} = \text{AATT}$, $e_2 = \text{ATAA}$, $\overline{e_2} = \text{ATAT}$, $e_t = \text{ATTA}$, $\overline{e_t} = \text{ATTT}$, $e_\# = \text{TAAA}$. The second step is to build the spectrum of the following form: {{AAATCAATACAAATC, AATACAAATCATAAC, AAATCATAACAAATC, ATAACAAATCAATAC, AATTCATAACAATTC,ATAACAATTCATTAC, AATTCATTACAATTC, ATTACAATTCATAAC, ATATCATTACATATC, ATTACATATCATTAC}. Finally, we add to the spectrum the following subset: {AATACTAAACAATTC, ATAACTAAACATATC, ATTACTAAACATTTC, TAAACTAAACAAATC}. As one can see, all oligonucleotides have the same corresponding temperature. There is only one sequence of length $(4 + 1)(2 \cdot 5 + 3 \cdot 4) = 110$ nucleotides, which includes all elements of the spectrum. It is TAAACTAAACAAATCAATACAAATCATAACAAATCAATACTAAACAATTCATAACAATTCATTACAATTCATAACTAAACATATCATTACATATCATTACTAAACATTTC. This sequence is equivalent to the Hamiltonian path $s \to 1 \to 2 \to t$ in graph G. The order of the vertices in the Hamiltonian path can be read over from the sequence by checking what labels follow $e_\#$. In this example we see the following sequence of such labels: $\overline{e_s}$, $\overline{e_1}$, $\overline{e_2}$, and $\overline{e_t}$, corresponding to the above path.

Theorem 1. *Isothermic DNA sequencing problem assuming only negative errors* Π_{niss} (*search version*) *is strongly NP-hard.*

Proof. Proving strong NP-completeness of problem Π_{niqd} (Lemma 1) directly leads to proving strong NP-hardness of the corresponding isothermic sequencing problem with only negative errors Π_{niss}. For, if we had an algorithm solving Π_{niss} in polynomial time, we could use it to solve problem Π_{niqd} in polynomial time in the following way. We could apply the algorithm to the instance of Π_{niqd}, and after a number of steps bounded by the polynomial function known for the algorithm we could have the answer for Π_{niqd}. Either the algorithm would find the solution and the answer would be "yes", or the algorithm would not find one and the answer would be "no". As pointed out earlier, a similar reasoning has been used in [21], where the problem of looking for a Hamiltonian circuit in a graph has been considered. □

2.2. Positive errors

In this subsection a restricted isothermic DNA sequencing problem, assuming only positive errors in the spectrum, will be considered. As in the previous case we define its search and decision versions.

Problem 4. Isothermic DNA sequencing by hybridization with only positive errors Π_{piss} — search version.

Instance: Set S (spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T} + 2$, length n of an original sequence, where $S' \subseteq S$ and S' is the ideal spectrum for this sequence.

Answer: A sequence of length n such that all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T} + 2$ appearing in this sequence are elements of S, and each of them occurs exactly once in this sequence.

Problem 5. Isothermic DNA sequencing by hybridization with only positive errors Π_{pisd} — decision version.

Instance: Set S (spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T} + 2$, length n of an original sequence, where $S' \subseteq S$ and S' is the ideal spectrum for this sequence.

Question: Is there a sequence of length n such, that all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T}+2$ appearing in this sequence are elements of S, and each of them occurs exactly once in this sequence?

Once again $D_{\Pi_{\text{pisd}}} = Y_{\Pi_{\text{pisd}}}$, and the above problem is easily solvable. In order to prove strong NP-hardness of its search counterpart Π_{piss}, the new quasi-sequencing problem Π_{piqd} must be defined.

Problem 6. Positive isothermic quasi-sequencing Π_{piqd} — decision version.

Instance: Set S (spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T}+2$, length n of an original sequence.

Question: Is there a sequence of length n such, that all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T}+2$ appearing in this sequence are elements of S, and each of them occurs exactly once in this sequence?

Similarly, $Y_{\Pi_{\text{pisd}}} = Y_{\Pi_{\text{piqd}}}$ and $D_{\Pi_{\text{pisd}}} \subset D_{\Pi_{\text{piqd}}}$. Below we prove that problem Π_{piqd} is strongly NP-complete.

Lemma 2. *Positive isothermic quasi-sequencing problem* Π_{piqd} *is strongly NP-complete*.

Proof. The proof is done by a polynomial transformation from the mentioned above problem directed Hamiltonian path between two vertices. Given an instance of the problem, a 1-digraph $G = (V, A)$ with two specified vertices s and t, the corresponding instance of Π_{piqd} is constructed as follows:

- Add two vertices s' and t', and two arcs (s', s) and (t, t') to graph G, name the new graph $G' = (V', A')$.
- To every vertex $v \in V'$ assign a unique label e_v of length $x = \lceil \log_2 |V'| \rceil$ over the alphabet {A, T}.
- Create for every vertex $v \in V'$ a spectrum element of the form $e_v \cdot \mathrm{C} \cdot e_v \cdot \mathrm{G}$ (where "$\cdot$" is the symbol of concatenation and "C" or "G" is the nucleotide). Let the temperature corresponding to these elements be denoted by $\mathscr{T}$.
- To every arc $(u, v) \in A'$, introduce to the spectrum all oligonucleotides of temperature $\mathscr{T}$ or $\mathscr{T}+2$ which are included in the sequence $e_u \cdot \mathrm{C} \cdot e_u \cdot \mathrm{G} \cdot e_v \cdot \mathrm{C} \cdot e_v \cdot \mathrm{G}$ and which are not yet included in the spectrum (the spectrum must remain a set).

There exists a Hamiltonian path from s to t in graph G iff there exists a sequence of length $2|V'|(x+1)$ such that all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T}+2$ appearing in this sequence are elements of the spectrum, and each of them occurs exactly once in this sequence.

First, we assume that the Hamiltonian path from s to t exists in graph G. Then, the sequence of length $2|V'|(x+1)$ can be easily created. We start the construction by taking the oligonucleotide corresponding to vertex s', next we concatenate to the right side of the existing part of the sequence all oligonucleotides corresponding to vertices from V, in order as in the Hamiltonian path, and finally we concatenate the oligonucleotide corresponding to vertex t'. Obviously, because all the arcs joining successive vertices in the order exists in graph G', all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T}+2$ appearing in this sequence are elements of the spectrum. Moreover, because all used oligonucleotides contain at least one unique label of a vertex, which allows to place them in the sequence at unique positions, we have the certainty that they occur exactly once in this sequence.

Now, we assume that the sequence of length $2|V'|(x+1)$ exists, and that all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T}+2$ appearing in this sequence are elements of the spectrum, and additionally each of them occurs exactly once in this sequence. The structure of spectrum elements allows to build longer fragments of the sequence only if they correspond to arcs from graph G'. In other cases the condition that all oligonucleotides possible to distinguish in the sequence are present in the spectrum, is not satisfied. Thus, the existing sequence corresponds to a path in graph G'. Because its length is $2|V'|(x+1)$, the corresponding path includes $|V'|$ vertices. The vertices cannot be duplicated in the path, because their corresponding oligonucleotides cannot be duplicated within the sequence. Therefore, the path is a Hamiltonian path. Moreover, both vertices s' and t' can be joined with others only in one possible way: s' begins the path, and its immediate successor is s, t' ends the path, and its immediate predecessor is t. After removing vertices s' and t' from the path, we have the Hamiltonian path from s to t in graph G. □

The transformation described in the proof of Lemma 2 is illustrated by Example 3.

Example 3. Let us use once again the graph from Fig. 3, presenting an instance of directed Hamiltonian path problem from s to t. The transformation from Lemma 2 applied to this graph produces the following data. Six unique vertex labels of length $\lceil \log_2 6 \rceil = 3$ are created: $e_{s'} = \mathrm{AAA}$, $e_s = \mathrm{AAT}$, $e_1 = \mathrm{ATA}$, $e_2 = \mathrm{ATT}$, $e_t = \mathrm{TAA}$, $e_{t'} = \mathrm{TAT}$. We build one oligonucleotide for each vertex from V': {AAACAAAG, AATCAATG, ATACATAG, ATTCATTG, TAACTAAG, TATCTATG}. Next, we add to the spectrum a series of oligonucleotides corresponding to arcs from A'. The subset of oligonucleotides of temperature 20 or 22 degrees created for example arc (1, 2) has the following form: {ATACATAG, ATACATAGA, TACATAGA, TACATAGAT, ACATAGAT, ACATAGATT, CATAGATT, ATAGATTC, ATAGATTCA, TAGATTCA, TAGATTCAT, AGATTCAT, AGATTCATT, GATTCATT, ATTCATTG}. Because the spectrum is a set, no oligonucleotide is present there more than once. The unique sequence of length 48 nucleotides being the answer for Π_{piqd} is AAACAAAGAATCAATGATACATAGATTCATTGTAACTAAGTATCTATG, and it corresponds to the Hamiltonian path $s \to 1 \to 2 \to t$ in graph G (the order of oligonucleotides corresponding to vertices determines the order of the vertices in G', so in G).

Theorem 2. *Isothermic DNA sequencing problem assuming only positive errors* Π_{piss} (*search version*) *is strongly NP-hard.*

Proof. As in the case of negative errors (cf. Theorem 1), proving strong NP-completeness of problem Π_{piqd} (Lemma 2) directly leads to proving strong NP-hardness of the corresponding isothermic sequencing problem with only positive errors Π_{piss}. For, if we had an algorithm solving Π_{piss} in polynomial time, we could use it to solve problem Π_{piqd} in polynomial time, similarly as before. □

3. Isothermic DNA sequencing without errors

The problem of isothermic DNA sequencing without any errors in the hybridization data is formulated below as the search one.

Problem 7. Isothermic DNA sequencing by hybridization without errors Π_{iss} — search version.

Instance: Set S' (the ideal spectrum) of oligonucleotides, each of them of temperature $\mathscr{T}$ or $\mathscr{T} + 2$.

Answer: A sequence containing all elements of S' exactly once as subsequences and such that all oligonucleotides of temperatures $\mathscr{T}$ or $\mathscr{T} + 2$ appearing in this sequence are elements of S'.

As in the cases of negative or positive errors, the decision version of the above problem always has the answer "yes" (see the beginning of Section 2 for explanation). However, its search counterpart is not so trivially solvable. Below we propose an exact, polynomial time algorithm solving problem Π_{iss}.

The algorithm constructs a directed graph based on the spectrum, and after some transformations it searches for a path corresponding to a DNA sequence. Without loss of generality we can assume, that we know the first and the last oligonucleotide in the solution. (This knowledge can be provided by additional biochemical experiments. On the other hand, if we did not know the first and the last oligonucleotides we could repeat the presented algorithm $|S'|^2$ times, assuming each time a different pair of oligonucleotides playing these roles, and choosing the feasible solution.)

The algorithm for isothermic DNA sequencing by hybridization without errors

(1) Create for every oligonucleotide o_i from the spectrum a vertex v_i of graph G.
(2) Introduce arcs to graph G according to the following rules ($\forall i, j$):

   (a) if o_i contains o_j moved to its left end, then add the arc from v_j to v_i and prohibit all other arcs entering v_i or leaving v_j;
   (b) if o_i contains o_j moved to its right end, then add the arc from v_i to v_j and prohibit all other arcs leaving v_i or entering v_j;
   (c) if o_i and o_j have equal length and they overlap with shift by one letter (o_i is first), then add the arc from v_i to v_j on condition the overlap does not produce negative errors.

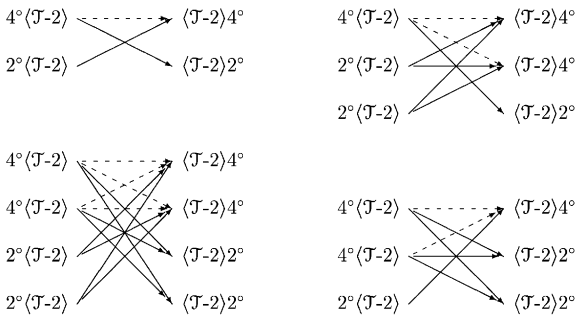


Fig. 4. The removal of excess arcs. The above subgraphs can be recognized in the whole graph only if there are no other arcs leaving the vertices on the left or entering the vertices on the right. The subgraphs contain some arcs (the dash ones), which for sure will not be used in the solution. Traversing these arcs, one makes impossible to collect all oligonucleotides from the spectrum. If graph G from the algorithm contains some of the above subgraphs (in the mentioned sense of arc completeness), the dash arcs must be removed. After that, the above subgraphs become line graphs. $2°$ and $4°$ stand for a nucleotide of the increment 2 or 4 degrees, respectively; $\langle \mathcal{T}-2\rangle$ stands for a string of nucleotides of temperature $\mathcal{T}-2$.

Fig. 5. The addition of temporary arcs. After the removal of excess arcs (cf. Fig. 4), the only obstacle on the way to make whole graph G a line graph could be the subgraph from the left side. If this subgraph is a part of graph G from the algorithm (in the sense of arc completeness, cf. caption to Fig. 4), we add the temporary arc from $2°\langle \mathcal{T}-2\rangle$ to $\langle \mathcal{T}-2\rangle 2°$. After the transformation of the line graph to its original graph H (step (5) of the algorithm), the left subgraph enlarged by the temporary arc becomes the right subgraph. Now, the transition from $2°\langle \mathcal{T}-2\rangle$ to $\langle \mathcal{T}-2\rangle 2°$ (the dotted one) must be forbidden during the searching for an Eulerian path in the new graph H.

(3) Remove from graph G all arcs entering the vertex corresponding to the first oligonucleotide in the solution, and all arcs leaving the vertex corresponding to the last oligonucleotide in the solution.
(4) In order to make graph G a line graph, remove from the graph some excess arcs according to the rules shown in Fig. 4 , and add to the graph some temporary arcs according to the rules shown in Fig. 5.
(5) Transform the line graph G to its original graph H. Now the oligonucleotides correspond to arcs in the new graph.
(6) Use the modified algorithm searching for an Eulerian path in graph H (i.e. accepting the exception from Fig. 5). The order of arcs in the path corresponds to the order of oligonucleotides in a DNA sequence being the solution of the problem.

Steps (5) and (6) require an additional comment. The transformation of a line graph to its original graph can be done in polynomial time, e.g. by the propagation algorithm proposed in [10] (cf. also [4]). The algorithm searching for an Eulerian path in a directed graph can be done in $O(n^2)$ time [23]. The modification mentioned does not affect its polynomial time complexity, and it is a simple rule of choosing the successor of a vertex (instead of choosing first available one in the standard approach). The rule concerns only the vertices like the one in the middle of the right subgraph from Fig. 5. We must forbid the transition from $2°\langle \mathcal{T}-2\rangle$ to $\langle \mathcal{T}-2\rangle 2°$, what is always possible. If we reach the vertex by one of the arcs $4°\langle \mathcal{T}-2\rangle$, and the arc $\langle \mathcal{T}-2\rangle 2°$ is not yet traversed, we choose it as the next one in the path. If the vertex is reached by the arc $2°\langle \mathcal{T}-2\rangle$, we choose this of the arcs $\langle \mathcal{T}-2\rangle 4°$, which is not yet traversed.

For the sake of completeness, we repeat here the proof of the correctness of the algorithm from [22]. The algorithm is correct if the following propositions are true.

Proposition 1. *The solution contains every oligonucleotide from the spectrum exactly once.*

Proposition 2. *All admissible connections between oligonucleotides are present in the graph.*

Proposition 3. *All connections generating negative errors are forbidden.*

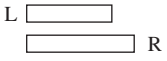
The correctness of Proposition 1 follows from the definition of graph G and from the equivalence of the problems of searching for the Hamiltonian path in a line digraph and searching for the Eulerian path in its original graph [10]. In the first proof of the following ones it will be shown, that graph G is a line digraph. Propositions 2 and 3 have been considered in the second proof.

Proof (*Propositions 1*). A digraph is a line graph if and only if for any pair of its vertices, their sets of successors are either the same or disjoint, and moreover its original graph is a 1-graph [10]. Simple paths created by two first rules from step (2) of the algorithm always satisfy the above condition. However, arcs added by third rule of step (2) can produce some incompatibility, removed as shown in Figs. 4 and 5. All other bipartite subgraphs either satisfy the above condition on sharing sets of successors or are not possible to build by the third rule when we assume no error within the data of the problem. A combination of the subgraphs into a greater structure does not affect the satisfiability of this condition. And because the spectrum is a set and no oligonucleotide is duplicated, the original graph of G must be a 1-graph and therefore graph G is a line digraph. □

Proof (*Propositions 2 and 3*). On the basis of Claim 2, the analysis of connections between oligonucleotides from the spectrum is restricted to the shifts of oligonucleotides by at most one position. Any other connections would produce negative errors. All possible cases of joining a pair of oligonucleotides are listed below. (Oligonucleotides correspond to vertices in digraph G.)

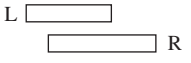
(1) Left oligonucleotide L is shorter than the right one R.

(a) Shift $= 0$.



Here L always has temperature $\mathcal{T}$, R — temperature $\mathcal{T}+2$, and R is always one nucleotide longer than L. L is contained in R and in order to avoid negative errors in the solution, L must immediately precede R within the solution. Thus, we introduce the arc from L to R and forbid the possibility of leaving L or entering R in any other way.

(b) Shift $= 1$.



(i) L and R have the same temperature $\mathcal{T}$. Then, the first nucleotide of L must have 4 degree and R must stand out two nucleotides (each of 2 degree) to the right. Such connection must be forbidden, because it would create an oligonucleotide of temperature $\mathcal{T}+2$ being a negative error (it would start at the first position of L and it would end at the last but one position of R). Either the created oligonucleotide would not be present in the spectrum, or it would be present, but this case is solved in item 1a and here it would cause a repetition of the oligonucleotide within the solution.

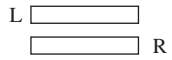
(ii) L and R have the same temperature $\mathcal{T}+2$. Then again the first nucleotide of L must have 4 degree and R must stand out two nucleotides to the right. It would generate a negative error of temperature $\mathcal{T}$, from the second position of L to the last but one position of R, and this connection must be also forbidden. Either the created oligonucleotide would not be in the spectrum, or it would be and this case should be solved as in item 1a.

(iii) L has temperature $\mathcal{T}$, R — temperature $\mathcal{T}+2$. This connection also must be forbidden. If the first nucleotide of L would have 2 degree, R would stand out two nucleotides (each of 2 degree) to the right. This must be disabled, because it would create a negative error of temperature $\mathcal{T}$ (cf. item 1b(ii)). If the first nucleotide of L would have 4 degree, R would stand out two or three nucleotides to the right, according to one of the following scheme: 4°2°, 2°4° or 2°2°2°. All these schemes would generate negative errors of temperatures, respectively $\mathcal{T}$, $\mathcal{T}+2$ and both.

(iv) L has temperature $\mathcal{T}+2$, R — temperature $\mathcal{T}$. This case does not exist.

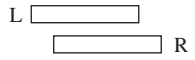
(2) Both oligonucleotides L and R have the same length.

(a) Shift $= 0$.



This case does not exist for a pair of oligonucleotides, the spectrum does not contain two identical oligonucleotides. On the other side, we admit the loop for an oligonucleotide composed of one kind of nucleotides (e.g. CCCCCC), see item 2b.

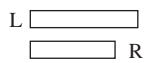
(b) Shift $= 1$.



Every such pair of oligonucleotides should be joined by the arc from L to R on the condition it will not create negative errors, i.e. the common part of L and R does not have temperature $\mathscr{T}$ and the whole contig does not have temperature $\mathscr{T}+2$. We admit also the loop for an oligonucleotide composed of one kind of nucleotides (i.e. standing simultaneously for L and R) — it is not significant for searching for a Hamiltonian path in graph G, but it contributes to making graph G a line graph.

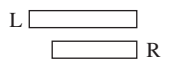
(3) Left oligonucleotide L is longer than the right one R.

(a) Shift $= 0$.



We do not consider this connection, having two such oligonucleotides we always pass from the shorter to the longer one, see item 1a.

(b) Shift $= 1$.



Here L always has temperature $\mathscr{T}+2$, R — temperature $\mathscr{T}$, and R is always one nucleotide shorter than L. R is contained in L and in order to avoid negative errors in the solution, we introduce the arc from L to R and forbid the possibility of leaving L or entering R in any other way.

All above rules constitute step (2) of the algorithm. □

4. Conclusions

Several variants of the isothermic DNA sequencing by hybridization problem have the same computational complexity as their traditional versions (with oligonucleotides of equal lengths). However, the hybridization experiment with isothermic libraries could produce fewer errors in the sequencing data (caused both by biochemical experiment and by repetitions) than the traditional one [9], what would lead to solutions more similar to original sequences. Thus, the isothermic DNA sequencing has an advantage over its traditional counterpart.

Acknowledgements

The authors are grateful to anonymous referees for their constructive comments leading to a substantial improvement in the presentation of the results.

References

[1] A. Apostolico, R. Giancarlo, Sequence alignment in molecular biology, in: M. Farach, F. Roberts, M. Waterman (Eds.), Mathematical Support for Molecular Biology, American Mathematical Society DIMACS series, 1997.

[2] R. Arratia, B. Bollobas, D. Coppersmith, G.B. Sorkin, Euler circuits and DNA sequencing by hybridization, Discrete Appl. Math. 104 (2000) 63–96.

[3] W. Bains, G.C. Smith, A novel method for nucleic acid sequence determination, J. Theoret. Biol. 135 (1988) 303–307.

[4] J. Bang-Jensen, G. Gutin, Digraphs: Theory, Algorithms and Applications, Springer, London, 2001.

[5] A. Ben-Dor, I. Pe'er, R. Shamir, R. Sharan, On the complexity of positional sequencing by hybridization, J. Comput. Biol. 8 (2001) 361–371.

[6] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, J. Weglarz, DNA sequencing with positive and negative errors, J. Comput. Biol. 6 (1999) 113–123.

[7] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, Method of sequencing of nucleic acids, Polish Patent Application P335786, 1999.

[8] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, Isothermic oligonucleotide libraries, in: S. Miyano, R. Shamir, T. Takagi (Eds.), Currents in Computational Molecular Biology, poster proceedings of RECOMB, 2000, pp. 97–98.

[9] J. Blazewicz, P. Formanowicz, M. Kasprzak, W.T. Markiewicz, Sequencing by hybridization with isothermic oligonucleotide libraries, Discrete Appl. Math. 145 (2004) 40–51.

[10] J. Blazewicz, A. Hertz, D. Kobler, D. de Werra, On some properties of DNA graphs, Discrete Appl. Math. 98 (1999) 1–19.

[11] J. Blazewicz, M. Kasprzak, Complexity of DNA sequencing by hybridization, Theoret. Comput. Sci. 290 (2003) 1459–1473.

[12] K.J. Breslauer, R. Frank, H. Blöcker, L.A. Marky, Predicting DNA duplex stability from the base sequence, Proc. Nat. Acad. Sci. USA 83 (1986) 3746–3750.

[13] R. Drmanac, I. Labat, I. Brukner, R. Crkvenjakov, Sequencing of megabase plus DNA by hybridization: theory of the method, Genomics 4 (1989) 114–128.

[14] G.B. Fogel, K. Chellapilla, D.B. Fogel, Reconstruction of DNA sequence information from a simulated DNA chip using evolutionary programming, in: V.W. Porto, N. Saravanan, D. Waagen, A.E. Eiben (Eds.), Lecture Notes in Computer Science, vol. 1447, Springer, Berlin, 1998, pp. 429–436.

[15] A.M. Frieze, B.V. Halldorsson, Optimal sequencing by hybridization in rounds, J. Comput. Biol. 9 (2002) 355–369.

[16] J. Gallant, D. Maier, J.A. Storer, On finding minimal length superstrings, J. Comput. System Sci. 20 (1980) 50–58.

[17] M.R. Garey, D.S. Johnson, Computers and Intractability. A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979.

[18] A. Guenoche, Can we recover a sequence, just knowing all its subsequences of given length?, Comput. Appl. Biosci. 8 (1992) 569–574.

[19] E. Halperin, S. Halperin, T. Hartman, R. Shamir, Handling long targets and errors in sequencing by hybridization, in: Proc. of Sixth Annu. Internat. Conf. on Research in Computational Molecular Biology RECOMB, 2002, pp. 176–185.

[20] E. Hubbell, Multiplex sequencing by hybridization, J. Comput. Biol. 8 (2001) 141–149.

[21] D.S. Johnson, The NP-completeness column: an ongoing guide, J. Algorithms 6 (1985) 291–305.

[22] M. Kasprzak, An algorithm for isothermic DNA sequencing, Bull. Polish Acad. of Sci. Tech. Sci. 52 (2004) 31–35.

[23] E. Lawler, Combinatorial Optimization: Networks and Matroids, Holt, Rinehart and Winston, New York, 1976.

[24] Yu.P. Lysov, V.L. Florentiev, A.A. Khorlin, K.R. Khrapko, V.V. Shik, A.D. Mirzabekov, Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides, A new method, Dokl. Akad. Nauk SSSR 303 (1988) 1508–1511.

[25] P.A. Pevzner, l-tuple DNA sequencing: computer analysis, J. Biomol. Structure Dynam. 7 (1989) 63–73.

[26] P.A. Pevzner, Computational Molecular Biology: An Algorithmic Approach, MIT Press, Cambridge, 2000.

[27] V.T. Phan, S. Skiena, Dealing with errors in interactive sequencing by hybridization, Bioinformatics 17 (2001) 862–870.

[28] F.P. Preparata, E. Upfal, System and methods for sequencing by hybridization, United States Patent Application US 2001/0004728 A1, 2001.

[29] J. Setubal, J. Meidanis, Introduction to Computational Molecular Biology, PWS Publishing, Boston, 1997.

[30] R. Shamir, D. Tsur, Large scale sequencing by hybridization, J. Comput. Biol. 9 (2002) 413–428.

[31] E.M. Southern, United Kingdom Patent Application GB8810400, 1988.

[32] R.B. Wallace, M.J. Johnson, T. Hirose, T. Miyake, E.H. Kawashima, K. Itakura, The use of synthetic oligonucleotides as hybridization probes. Hybridization of oligonucleotides of mixed sequence to rabbit beta-globin DNA, Nucleic Acids Res. 9 (1981) 879–894.

[33] M.S. Waterman, Introduction to Computational Biology. Maps, Sequences, and Genomes, Chapman & Hall, London, 1995.

[34] J. Zhang, L. Wu, X. Zhang, Reconstruction of DNA sequencing by hybridization, Bioinformatics 19 (2003) 14–21.