BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 52, No. 1, 2004

# An algorithm for isothermic DNA sequencing

M. KASPRZAK*

Poznań University of Technology, Institute of Computing Science, 3 Piotrowo St., 60-965 Poznań, Poland

**Abstract.** In the paper, the problem of isothermic DNA sequencing by hybridization, without any errors in its input data, is presented and an exact polynomial-time algorithm solving the problem is described. The correctness of the algorithm is confirmed by an enumerative proof.

**Keywords:** computational molecular biology, DNA sequencing by hybridization.

## 1. Introduction

The *DNA sequencing by hybridization*, one of important problems from the computational molecular biology domain, consists in determining a sequence of nucleotides of an unknown DNA fragment [1–4]. 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. The spectrum may contain *positive errors*, i.e. oligonucleotides present in the spectrum but absent in the original sequence, and *negative errors*, i.e. oligonucleotides not present in the spectrum, but possible to distinguish in the original sequence. Since the spectrum is a set, repetitions of oligonucleotides in the sequence are also treated as negative errors.

In the standard approach to the DNA sequencing, the oligonucleotide library used in the hybridization experiment contains all possible oligonucleotides of a given constant length (cf. [5–9]). 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 [10], while the variants assuming presence of errors in the data (negative ones, positive ones, or both) are all strongly NP-hard [11].

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 [12], which allows to avoid a lot of experimental errors at the biochemical level of the DNA sequencing process. Then, the sequencing results are usually much more similar to the original sequences. In Section 2 the problem of isothermic DNA sequencing without errors in experimental data is formulated, while Section 3 contains an exact polynomial-time algorithm for the problem together with its proof of correctness.

* e-mail: Marta.Kasprzak@cs.put.poznan.pl

## 2. Formulation of the isothermic DNA sequencing problem without errors

Definition 1 [12]. An *isothermic oligonucleotide library* $L$ of temperature $\mathcal{T}_L$ is a library of all oligonucleotides satisfying the relations:

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

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 [13].

Claim 1. [12] 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. [12] 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.

When the variant of the problem with no error within experimental data is considered, the spectrum contains all oligonucleotides being members of the two libraries which can be distinguished in the DNA sequence, and it does not contain any other oligonucleotides. Moreover, the DNA

sequence cannot contain repetitions of oligonucleotides. Such spectrum is called the *ideal spectrum*. Example 1 shows the reconstruction of a DNA sequence on the base of a spectrum without errors containing oligonucleotides of temperatures differing by 2 degrees.

*Example* 1. Let our original sequence be CCTACGT. We assume the following temperatures of the oligonucleotide libraries used in a hypothetical hybridization experiment: 10 and 12 deg. Thus, the spectrum without any experimental errors created in the ideal experiment for our sequence would be {ACG, ACGT, CCT, CCTA, CGT, CTAC, TACG}. The spectrum contains only all oligonucleotides of temperatures 10 and 12 deg. appearing in the original sequence. The goal of the isothermic DNA sequencing without errors is to reconstruct the sequence using all the words from the spectrum, and it must be not possible to distinguish in the solution any oligonucleotide of temperature 10 or 12 deg. not belonging to the spectrum. Also no repetitions of the oligonucleotides are allowed. The only possible solution is shown below.

```
CCTACGT
CCT
CCTA
 CTAC
  TACG
   ACG
   ACGT
    CGT
```
□

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

PROBLEM 1. *Isothermic DNA sequencing without errors — search version.*

*Instance:* The ideal spectrum $S$ of oligonucleotides, each of them of temperature $\mathcal{T}$ or $\mathcal{T}+2$.

*Answer:* A sequence containing all elements of $S$ exactly once as subsequences and such, that all oligonucleotides of temperatures $\mathcal{T}$ or $\mathcal{T}+2$ appearing in this sequence are elements of $S$.

In the next section an exact, polynomial-time algorithm solving the above problem is presented.

## 3. The algorithm

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.) Without this knowledge we just run the algorithm $O(|S|^2)$ times with all possible pairs of oligonucleotides at the ends of the solution.

*The algorithm for isothermic DNA sequencing without errors*

(1) Create for every oligonucleotide $o_i$ from the spectrum vertex $v_i$ of graph $G$.

(2) Introduce arcs to graph $G$ according to the following rules ($\forall i, j$):
— 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 $v_j$;
— 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 $v_j$;
— 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.

(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. 1, and add to the graph some temporary arcs according to the rules shown in Fig. 2.

(5) Transform the line graph $G$ to its original graph $H$. Now the oligonucleotides correspond to arcs in the new graph.

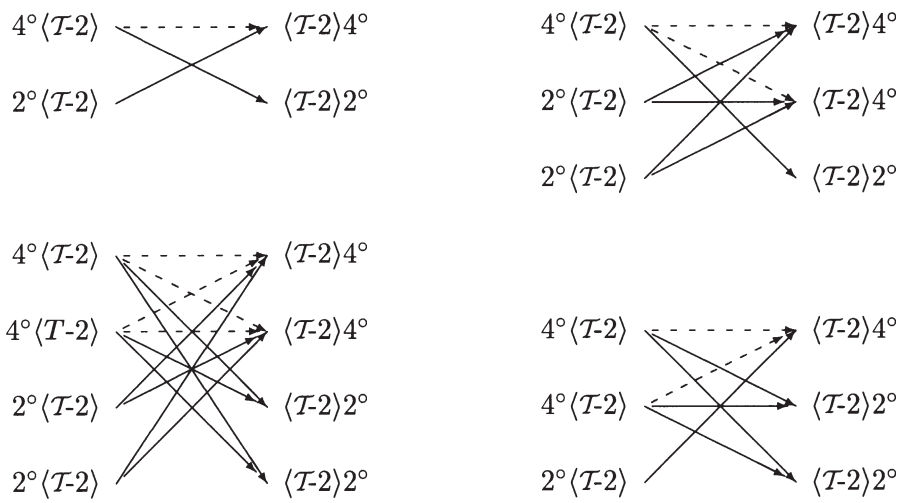



Fig. 1. The removal of excess arcs

(6) Use the modified algorithm searching for an Eulerian path in graph $H$ (i.e. accepting the exception from Fig. 2). The order of arcs in the path corresponds to the order of oligonucleotides in a DNA sequence being the solution of the problem.

The subgraphs from Fig. 1 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 these subgraphs (in the mentioned sense of arc completeness), the dash arcs must be removed. After that, the subgraphs become line graphs. $2^\circ$ and $4^\circ$ stand for a nucleotide of the increment 2 or 4 degrees, respectively; $\langle \mathcal{T}\text{-}2\rangle$ stands for a string of nucleotides of temperature $\mathcal{T}-2$.

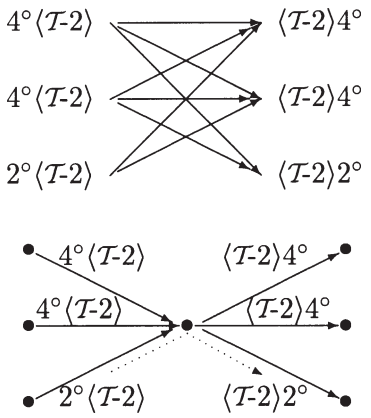



Fig. 2. The addition of temporary arcs

After the removal of excess arcs (cf. Fig. 1), the only obstacle on the way to make whole graph $G$ a line graph could be the subgraph from the top of the Fig. 2. If this subgraph is a part of graph $G$ from the algorithm (in the sense of arc completeness, cf. comments to Fig. 1), we add the temporary arc from $2^\circ\langle \mathcal{T}\text{-}2\rangle$ to $\langle \mathcal{T}\text{-}2\rangle 2^\circ$. After the transformation of the line graph to its original graph $H$ (step (5) of the algorithm), the top subgraph enlarged by the temporary arc becomes the bottom subgraph. Now, the transition from $2^\circ\langle \mathcal{T}\text{-}2\rangle$ to $\langle \mathcal{T}\text{-}2\rangle 2^\circ$ (the dotted one) must be forbidden during the searching for an Eulerian path in the new graph $H$.

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 [14] (cf. also [15]). The algorithm searching for an Eulerian path in a directed graph can be done in $O(n^2)$ time (cf. e.g. [16]). 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 bottom subgraph from Fig. 2. We must forbid the transition from $2^\circ\langle \mathcal{T}\text{-}2\rangle$ to $\langle \mathcal{T}\text{-}2\rangle 2^\circ$, what is always possible. If we reach the vertex by one of the arcs $4^\circ\langle \mathcal{T}\text{-}2\rangle$, and the arc $\langle \mathcal{T}\text{-}2\rangle 2^\circ$ is not yet traversed, we choose it as the next one in the path. If the vertex is reached by the arc $2^\circ\langle \mathcal{T}\text{-}2\rangle$, we choose this of the arcs $\langle \mathcal{T}\text{-}2\rangle 4^\circ$, which is not yet traversed.

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.

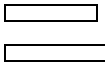
Proposition 3. All connections generating negative errors are forbidden.

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 [14]. In Proof 1 it will be shown, that graph $G$ is a line digraph. Propositions 2 and 3 have been considered in Proof 2.

P r o o f . 1 (Proposition 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 [14]. 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 Figures 1 and 2. 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. □

P r o o f . 2 (Propositions 2 and 3). On the base 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 $G$.)

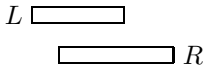
1. Left oligonucleotide $L$ is shorter than the right one $R$.
   (a) Shift = 0.




Here $L$ always have 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 deg. and $R$ must stand out two nucleotides (each of 2 deg.) 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.

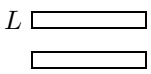
ii. $L$ and $R$ have the same temperature $\mathcal{T}+2$. Then again the first nucleotide of $L$ must have 4 deg. and $R$ must stand out two nucleotides to the right. It would generate a negative error of temperature $\mathcal{T}$, from the second position of $L$ to the last but one position of $R$, 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. $L$ have temperature $\mathcal{T}$, $R$ — temperature $\mathcal{T}+2$. This connection also must be forbidden. If the first nucleotide of $L$ would have 2 deg., $R$ would stand out two nucleotides (each of 2 deg.) to the right. This must be disabled, because it would create a negative error of temperature $\mathcal{T}$ (cf. item 1(b)ii). If the first nucleotide of $L$ would have 4 deg., $R$ 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. $L$ have temperature $\mathcal{T}+2$, $R$ — temperature $\mathcal{T}$. This case does not exist.

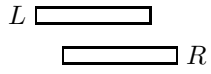
2. Both oligonucleotides $L$ and $R$ have the same length.

(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.

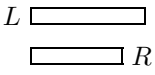
(b) Shift = 1.




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 $\mathcal{T}$ and the whole contig does not have temperature $\mathcal{T}+2$. 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.

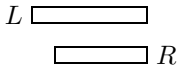
3. Left oligonucleotide $L$ is longer than the right one $R$.

(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 have temperature $\mathcal{T}+2$, $R$ — temperature $\mathcal{T}$, 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

The introduction of isothermic oligonucleotide libraries to the hybridization experiment with a great probability will result in more correct sequencing data than the standard approach. As it has been proved in the paper, from the computational point of view the isothermic DNA sequencing without errors can be solved in polynomial time, similarly like the DNA sequencing without errors basing on constant-length libraries. Therefore, the new proposition of the hybridization experiment has a big chance to become a widely used approach to DNA sequencing.

This paper was supported by the Polish State Committee for Scientific Research.

### References

[1] W. Bains and G. C. Smith, "A novel method for nucleic acid sequence determination", *J. Theor. Biol.* 135, 303–307 (1988).

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

[3] E. M. Southern, *United Kingdom Patent Application* GB8810400 (1988).

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

[5] A. Apostolico and R. Giancarlo, "Sequence alignment in molecular biology", in: *Mathematical Support for Molecular Biology*, ed.: M. Farach, F. Roberts, M. Waterman, *Am. Math. Soc. DIMACS* (1997).

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

[7] A. Guenoche, "Can we recover a sequence, just knowing all its subsequences of given length?", *CABIOS* 8, 569–574 (1992).

[8] J. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology*, Boston: PWS Publishing Company, (1997).

[9] M. S. Waterman, *Introduction to Computational Biology. Maps, Sequences and Genomes*, London: Chapman & Hall, (1995).

[10] P. A. Pevzner, "l-tuple DNA sequencing: computer analysis", *J. Biomol. Struct. Dyn.* 7, 63–73 (1989).

[11] J. Błażewicz and M. Kasprzak, "Complexity of DNA sequencing by hybridization", *Theor. Comp. Sci.* 290, 1459–1473 (2003).

[12] J. Błażewicz, P. Formanowicz, M. Kasprzak and W. T. Markiewicz, "Sequencing by hybridization with isothermic oligonucleotide libraries", *Disc. Applied Math.*, to be published.

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

[14] J. Błażewicz, A. Hertz, D. Kobler and D. de Werra, "On some properties of DNA graphs", *Disc. Applied Math.* 98, 1–19 (1999).

[15] J. Bang-Jensen and G. Gutin, *Digraphs: Theory, Algorithms and Applications*, Berlin: Springer-Verlag, 2001.

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