CMST 24(4) 259-271 (2018) DOI:10.12921/cmst.2018.0000055

Computational Aspects of DNA Sequencing by Hybridization – a Survey

Kamil Kwarciak[1], **Marcin Radom**[1,2], **Piotr Formanowicz**[1*,2]

[1] *Institute of Computing Science*
Poznan University of Technology

[2] *Institute of Bioorganic Chemistry*
Polish Academy of Sciences

**E-mail: piotr.formanowicz@cs.put.poznan.pl*

Received: 15 November 2018; revised: 14 December 2018; accepted: 17 December 2018; published online: 28 December 2018

Abstract: Sequencing by hybridization (SBH) is a method of reading DNA sequence from its smaller fragments. Such a method has been proposed in late 1980s and until the emergence of the new generation sequencing it has been widely used and improved. Since the initial, classical approach to SBH, many modifications and enhancements was proposed, aimed at improving the preciseness and the length of sequences which can be unambiguously read. Even now, for some DNA sequences sequencing by hybridization can still be used effectively and at a low cost. In this paper many different approaches to the SBH will be described, mainly from the points of view of algorithms and computational complexity.

Key words: DNA sequencing, hybridization errors, algorithms, computational complexity

I. INTRODUCTION

One of the most important biochemical procedure of molecular and computational biology is reading DNA sequences, both long and short ones. It is the first and necessary step of various biological projects, since many of them perform the analysis of information encoded in these sequences. Obviously, in order to analyze the information, the sequences have to be read first. It is the reason for the importance of DNA sequencing. This importance is confirmed by the rapid development of new technologies for DNA sequencing, which could be observed during the last fifteen years. Next generation sequencing methods are becoming more and more effective, which means that they allow the sequencing of whole genomes in a relatively short time [1]. However, the new sequencing techniques are very well suited for sequencing very long DNA molecules, while there is still a need for fast and low cost methods for sequencing or re-sequencing of relatively shorter DNA fragments. Such methods are needed, for example, in medical diagnostics. A good candidate for such a method is sequencing by hybridization (SBH). It has been proposed in the late 1980s when it was considered as a promising candidate for a universal DNA sequencing approach. When the next generation sequencing technologies have emerged, SBH can still be considered as an effective method for sequencing or re-sequencing shorter DNA molecules. For example, it is still possible to use it for reading whole bacterial genome as proposed in [2] or for virus identifications [3].

Since the development of the SBH method in the late 1980s many variants of this approach have been proposed and even recently new papers concerning, e.g., algorithms with novel SBH approaches are being published [4]. In this paper we provide a survey of SBH methods. For many of them interesting computational complexity problems have emerged, which will also be mentioned. In the following

paper we will start with the description of the classical sequencing by hybridization. Then, various proposed modifications to the initial methodology will be introduced, like, e.g., multistage sequencing, information about repetition or non-classical approaches to the DNA chip design.

II. CLASSICAL SEQUENCING BY HYBRIDIZATION

Sequencing by hybridization is a method composed of two stages. The first one is called a biochemical one, during which a hybridization experiment is performed. In such an experiment a DNA chip containing a full library of all oligonucleotides (i.e., short single-stranded nucleotide sequences) of a given length l is being used. In this stage cloned DNA sequences bind to the probes (cells in the DNA chip filled with copies of a known, specific type of oligonucleotide) according to the Watson-Crick complementary rule: adenine binds to thymine, while cytosine to guanine. In other words the DNA attaches itself to the probe, if its fragment of a given length is complementary to the oligonucleotide of the same length within the probe. As a result, a set called *spectrum* is obtained. In an ideal case it consists of all substrings of length l of the target DNA sequence.

Let $\Sigma = \{A, C, T, G\}$ be an alphabet of nucleotides. A DNA sequence is represented by a string over alphabet Σ_{DNA}. The length of a reconstructed sequence is denoted by n. An ideal spectrum $S^{is}(Q)$ of a sequence $Q = \langle q_1 q_2 \dots q_n \rangle$ is a set of all unique l-long substrings (l-mers) of Q.

Since the spectrum does not contain any information about the order of substrings in the analyzed a DNA molecule it is necessary to determine it. It is a goal of a second, computational stage of the SBH method. In this stage a permutation of the spectrum elements corresponding to their order in the target DNA sequence must be found. It can be done using some combinatorial algorithms. This task can be formulated as DNA sequencing problem without errors [5] formulated as follows.

Problem 1. DNA sequencing without errors - search version
INSTANCE: An ideal spectrum $S^{is}(Q)$ of elements of equal length l over the alphabet Σ_{DNA}, the length n of an original sequence Q, $|S^{is}| = n - l + 1$.
ANSWER: A sequence of length n containing all elements of $S^{is}(Q)$.

In an ideal case, where there are no errors in the spectrum, Problem 1 of finding DNA sequence Q can be solved in a polynomial time [6].

However, a spectrum obtained during the biochemical experiment may be affected by hybridization errors. Some substrings of Q may be missing (they will be called *negative errors*) and some additional l-mers which are not a part of Q may be included (*positive errors*). Such a spectrum is denoted by $S(Q)$. For each l-mer $s_i = \langle s_{i_1} s_{i_2} \dots s_{i_l} \rangle \in \Sigma^l$, $Q(s_i) = 1$ if s_i is a substring of Q and $Q(s_i) = 0$ in the other case.

Negative errors appear when the analyzed DNA sequence does not hybridize to the complementary oligonucleotide on the chip and as a result the spectrum does not contain information of the corresponding l-mer of the target sequence. Another type of negative errors is caused by repetitions of substrings of a given length l in the target sequence. Since the information obtained on the basis of reading the signal from the DNA chip probe is binary, i.e., it only indicates whether a given substring is present in the target sequence or not, the hybridization experiment usually does not provide information about the repeated fragments (a case when such information is available will be described later). As a result the spectrum contains information about only one occurrence of the repeated substring. It should be noted that errors resulting from repetitions do not follow from imperfectness of the hybridization experiment and they depend only on the nature of the analyzed DNA sequence. Hence, it is difficult or even impossible to avoid errors of this type in the first stage of the SBH method. On the other hand it may happen that the analyzed DNA molecule will hybridize to the oligonucleotide on the chip which is not perfectly complementary to it. In this the case spectrum contains information about a substring complementary to such an oligonucleotide, which in fact is not a part of the target sequence. Errors of this type are called positive errors.

The problem of finding a sequence of a given length from the non-ideal spectrum (i.e., one that contains both positive and negative errors) can be formulated as follows [5].

Problem 2. DNA sequencing with negative and positive errors - search version
INSTANCE: Spectrum $S(Q)$ of elements of length l over the alphabet Σ_{DNA}, the length n of an original sequence Q.
ANSWER: A sequence of length $\leq n$ containing the maximum number of elements of $S(Q)$.

When positive or negative errors resulting from the imperfectness of the hybridization process are present in the spectrum, the computational problem which must be solved in the second phase of the method becomes strongly NP-hard. The computational complexity of the problem with only negative errors resulting from repetitions remains an open question. It is worth mentioning, that SBH methods sometimes result in the formulation of combinatorial problems with the computational complexity being difficult to establish, for example [7].

The errors are the main drawback of the SBH approach. The method is very sensitive to them, meaning that they significantly influence the obtained DNA sequence. Moreover, they make the computational problem which must be solved in the second stage a more difficult task. Hence, new versions of the SBH method have been developed in order to

make SBH more resistant to errors. In some of them the hybridization experiment is also modified in such a way that the probability of errors occurrence is decreased.

III. INTERACTIVE PROTOCOLS, MULTISTAGE SEQUENCING

The general idea of interactive protocols described in [8] (also being called sequencing by hybridization in rounds) is to perform multiple hybridization experiments using differently designed DNA chips. Each hybridization stage is being followed by the separate computational phase, the result of which gives the background for chip redesign for the next turn of hybridization. Each turn is performed on a much smaller scale in terms of the number of probes on a DNA chip. Using the spectrum from the first round one usually cannot reconstruct the whole DNA sequence unambiguously. However, after the first turn there are new data available, or in other words, there are new questions that can be answered by a specifically constructed DNA chip in the next round. Authors prove that the total capacity of chips used in all rounds to successfully reconstruct the DNA is much smaller than the capacity of the classical chip used in the same task. An interesting summary Table 1 is given in [8]. It shows the advantages of such an approach when it comes to the scale of the hybridization experiment, i.e., it compares the total number of generated probes on the DNA chips for both methods - proposed and classical one.

Authors also propose two algorithms designed for such an approach and present the results obtained in computational experiments. The first algorithm is called the *Doubling Algorithm* and is based on the theorem (proved in the paper) that $O(\lg n)$ rounds of $n^2/\lg n$ substring queries per round are enough to reconstruct any string of length n on an alphabet of size $\alpha \leq n$. A *query* is a simple question whether a specific substring exists or does not exist within a given string. In practice such a query takes the form of a specific probe on a sequencing chip. The second algorithm called *Adaptive Length Algorithm* uses another theorem which states that if S is a random string over an alphabet of size α, then with a probability of $1 - 1/n^{\epsilon}$ string S can be determined using $O(\alpha \cdot \log_{\alpha} n)$ rounds of n queries per round.

The algorithms have been tested on real sequences obtained from GenBank. For most sequences the Adaptive Algorithm required between 9 and 11 rounds with 50 to 150 thousands queries. The longest DNA fragment successfully sequenced came from Bacteriophage Lambda having length of 48502bp it required 11 rounds and about 386 thousands queries.

The sole problem of a string reconstruction in rounds on the basis of questions concerning its substrings was formulated and analyzed two years earlier in [9]. That paper, however, did not provide a direct algorithm for the application in SBH, gave rather extensive theoretical knowledge about such a reconstruction. Tight bounds on the complexity of reconstructing an unknown string from substrings queries have been given, for example, the authors provided a pattern for a maximum number of queries necessary for reconstruction, given the size of the string and the alphabet building it.

Another example of research on a field of sequencing in rounds is [10] by Kruglyak. The author further investigates the concept of the Doubling Algorithm ([8]) for the multistage SBH. The problem of estimating the number of the l-tuples necessary to sequence the DNA depending on its length and the length of the used oligonucleotides is described in detail. Further progress has been made and described in [11]. The main thesis of the paper is that with a "high probability" a string of length n can be successfully reconstructed using up to seven different hybridization chips, each containing $O(n)$ probes. Authors introduce an algorithm solving the DNA sequencing problem in rounds and show experimental results for different DNA sequences. For the first round a DNA chip having all strings (oligonucleotides within probes) of length $\lceil \log_4 n \rceil + c$ is required, where n is the length of the analyzed DNA, c is an arbitrary number from a range 0 to 4. The greater the c, the longer the length of oligonucleotides in a given round, but fewer rounds will be necessary in order to sequence the DNA. In subsequent rounds at most $4^c \cdot n$ probes are necessary.

Another paper is [12] where the author further extended the method of sequencing in rounds. In the proposed approach one can ask a question not only if a substring exists in an analyzed string (the DNA sequence), but also if it exists more than once, thus introducing a concept of substring repetitions.

IV. POSITIONAL SEQUENCING BY HYBRIDIZATION

Another variant of the SBH is called positional sequencing by hybridization (PSBH), described first in [13] by Broude et al., in 1994 and in [14] by Hannenhalli et al. in 1996. Authors proposed using additional data for the DNA spectrum from the hybridization experiment. In this modified approach, for every oligonucleotide in the spectrum there is also information about the probable location of the fragment within the original sequence. In [14] the so called *positional Eulerian path* problem is defined as follows:

Problem 3. Positional Eulerian path problem
INSTANCE: A directed multigraph $G(V, E)$ and an interval $I_e = \{l_e, h_e\}$, $l_e \leq h_e$ associated with every edge $e \in E$.
ANSWER: An Eulerian path P in G such that for all $e \in E$, $l_e \leq \pi(P, e) \leq h_e$.

In the above problem, the $\pi(P, e)$ denotes the position of end edge e in path P. The theorem that Problem 3 is NP-complete, even if each vertex has in-degree and out-degree at most 2 and intervals associated with edges are of the same

Tab. 1. Characteristic length of unambiguously reconstructed DNA as a function of the size classical and interactive SBH as given in [8]

DNA	Classical SBH		Interactive SBH	
size	Oligo length	Chip size	Rounds	Chip size
80	7	16 384	7	560
180	8	65 536	8	1 140
260	9	262 144	8	2 080
560	10	1 048 576	8	4 480
1300	11	4 194 304	9	11 700
2450	12	16 777 216	9	22 050

size, has been formally proved in the paper. The authors proposed two algorithms solving the sequencing problem using the information about the substrings locations.

In the paper [15] the idea has been extended. The authors have given a linear time algorithm for a case when for every element of spectrum there are at most two possible location. They have also proved that the problem is NP-complete if for every element there are at most three possible locations. The *positional SBH* problem has been defined as follows.

Problem 4. Positional SBH problem
INSTANCE: A multiset S of strings having length p. For each $s \in S$ there is a set $P(s) \subseteq \{0, ..., |S| - 1\}$.
ANSWER: Yes, if S is the ideal spectrum of substrings having length p of some string X such that for each $s \in S$ its position along X is in $P(s)$, no in other case.

If the set of allowed positions for each string is of size at most k, then Problem 4 is called *k-positional SBH*. If, for each $s \in S$, $P(s)$ is a sub-interval of $\{0, ..., |S| - 1\}$, then the problem is called *interval PSBH* [15]. Further in the paper there are proofs that *2-positional SBH* problem is solvable in linear time, while 3-positional SBH problem is NP-complete. Also the *interval PSBH* problem is proved to be NP-complete. In such a version all the positions are intervals of equal length.

In paper [16] published in 2006 by Zhang et al., a novel positional SBH problem was introduced, this time handling both negative and positive errors. Authors extensively discussed the mathematical formulation of the problem and proposed a dynamic programming method for fixing the optimal solution and a branch and bound algorithm for the positional SBH reconstruction. Results of an extensive computational experiments for the proposed algorithm also given.

V. INFORMATION ABOUT REPETITIONS

Real DNA sequences are repetitious. A given substring may occur more than once in an analyzed DNA. If the length of the repetitive fragment is longer than the length of oligonucleotides on a DNA chip then some information about spectrum composition may be lost. In the classical SBH the data from the hybridization experiment consists of the binary information about l-mers included in a spectrum, i.e., a given oligonucleotide is or is not a part of a target DNA. Besides imperfect hybridization, this is the second source of negative errors.

The consequence of repetitions longer than the length of oligonucleotides on a DNA chip is often an ambiguity of obtained results. In general, the ideal solution reconstructed in the computational stage should have the same length as a target sequence, containing as many oligonucleotides from a spectrum as possible (i.e., $n - l + 1$ oligonucleotides in the idea case, cf. 1). However, if a target sequence contains multiple occurrences of a given l-mer then a reconstructed sequence may meet the above requirements, but it may still be significantly different from the target. Consequently, it is more difficult to reconstruct a repetitious sequence than a sequence without repetitions and one should expect that repetitions lead to a lower quality of obtained results.

Some of studies related to the classical SBH take into consideration multiple occurrences of l-mers in DNA sequences and examine their influence on the algorithm performance, especially on the spectrum elements utilization and the similarity of a reconstructed sequence to its real counterpart. Błażewicz et al. compared in [17] a hybrid genetic algorithm [18] and a tabu and scatter search algorithm [19]. Błażewicz et al. presented in [20] outcomes for the same tabu and scatter search algorithm [19], a hybrid genetic algorithm with isothermic libraries [21] and a revised hybrid genetic algorithm with standard libraries [20]. Zhang et al. described results of a branch and bound algorithm for the classical SBH approach in [22] and the positional SBH in [16]. Note that although all the above-mentioned works deal with the repetitions in sequences they use the standard model of binary information about spectrum composition.

The idea to extend the classical sequencing by hybridization with the information about repetitions was proposed by Formanowicz [23]. The current development of the DNA chip technology makes information about the intensity of chip signals available. This information can be, at least to

some extent, correlated with the number of repetitions of a given oligonucleotide in a target DNA sequence. Unfortunately, the intensity information is not very precise. The determination of the exact number of occurrences of a given l-mer becomes more difficult, the bigger the number of repetitions is and the stronger the signal is. It is easy to distinguish the signal coming from one occurrence and many occurrences. However, to differ the signal representing, for example, six and seven occurrences may be very hard or even impossible. Nevertheless, even partial information about repetitions should be useful.

Taking into account the information about l-mers multiplicity requires distinguishing between four types of spectra [23]. Let $S(Q)$ denote a classical spectrum of a sequence Q and $S^{is}(Q)$ denote an ideal spectrum of this sequence. The ideal spectrum $S^{is}(Q)$ contains all and only these l-mers which are a part of the sequence Q. However, the ideal spectrum is a set but not a multiset and each l-mer occurs in the ideal spectrum once, even if it occurs in the sequence Q multiple times. The complete information about the composition of the sequence Q is provided by the ideal multispectrum $S^{im}(Q)$.Such a multiset consists of all and only these l-mers which are subsequences of length l of the sequence Q. Note that the number of occurrences of a given l-mer in the ideal multispectrum is equal to the multiplicity of this l-mer in the sequence Q. The last type of spectra is a multispectrum of the sequence Q denoted by $S^m(Q)$. It may be affected by hybridization errors so some l-mers which are a part of the sequence Q may be missed. Moreover, the multispectrum may contain some additional oligonucleotides which are not subsequences of Q. Finally, let us define for every oligonucleotide $s_i \in S^m(Q)$ a parameter m_i which is equal to the number of occurrences of s_i in $S^m(Q)$.

In [23], Formanowicz described two models of additional information about repetitions. According to the first one, called "one and many", there is available information if a given l-mer occurs in a target sequence exactly once or at least twice. In the second model, called "one, two and many", it is assumed that the results coming from the biochemical experiment allow for distinguishing between one, two and more than two occurrences of any oligonucleotide in an analyzed DNA sequence. The current DNA chip technology justifies these assumptions and makes these models of multiplicity information realistic. It is a common practice to take into consideration such information in a gene expression analysis [24].

The variants of the classical SBH problem were reformulated in [23] for both above models of information. The considered problems include:

- problem without any errors,
- problem with negative errors resulting from repetitions,
- problem with negative errors resulting from hybridization,
- problem with negative errors of arbitrary types,
- problem with positive errors,
- problem with positive errors and negative once resulting from repetitions,
- problem with errors of arbitrary types.

In this paper only the most general problems with errors of arbitrary types are presented. See [23] for definitions of the other ones. The combinatorial problem related to the classical SBH approach (no information about repetitions) has already been formulated as Problem 1 and 2. Let us assume that the additional multiplicity information according to the model "one and many" is available. Then the combinatorial problem may be stated as follows:

Problem 5. 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 n containing at most one occurrence of s_i if $m_i = 1$ and at least one occurrence of s_i if $m_i = 2$. Moreover, Q' can contain some l-mers which are not elements of $S(Q)$.

Assuming that one is able to obtain in the biochemical experiment information of the type "one, two and many" the sequencing problem may be defined as follows:

Problem 6. 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 n containing at most one occurrence of s_i if $m_i = 1$, one or two occurrences of s_i if $m_i = 2$ and at least two occurrences of of s_i if $m_i = 3$. Moreover, Q' can contain some l-mers which are not elements of $S(Q)$.

Problems 5 and 6 may be transformed to a variant of the traveling salesman problem (TSP). An instance of the classical TSP consists of a directed or undirected graph and a cost defined for each: an arc (in the case of a directed graph) or an edge (in the case of an undirected graph). The goal is to find the lowest cost Hamiltonian cycle in the given graph. In order to obtain a computational problem representing the sequencing by hybridization with information about repetitions, TSP has to be modified as follows. Firstly, the solution to be found is a path instead of a cycle. The cost of the path may not exceed a length n of a reconstructed sequence decreased by an oligonucleotide length l. Additionally, if the first l-mer is a part of input data then the first node in the path should be the one corresponding to the given oligonucleotide. Moreover, some vertices may not be visited at all and others may be visited more than once according to parameter m_i.

If oligonucleotides from spectrum are represented by nodes in a directed graph then the traveling salesman problem customized as described above corresponds to the SBH

problem with additional multiplicity information. Cost c_{ij} of an arc connecting nodes representing oligonucleotides i and j depends on their overlapping. The cost is equal to their length l decreased by the length of their common subsequence. For example, the cost of an arc from a node representing oligonucleotide $CGCTTA$ to a node representing $GCTTAT$ is equal to 1 because the sequences have common substring $GCTTA$ of length 5.

The classical TSP is strongly NP-hard. The computational problem related to the classical SBH is also strongly NP-hard [5]. The above variants of problems for sequencing by hybridization with multiplicity information are strongly NP-hard [23] too so there does not exist a polynomial time exact algorithm to solve them (assuming $P \neq NP$).

The additional information about repetitions may be used also in the case of sequencing by hybridization with isothermic libraries. The computational problems for this extension with the multiplicity information have been formulated in [25].

V. 1. Algorithms for SBH problems with information about repetitions

There exist results confirming usefulness of the additional information about repetitions and several algorithms have been implemented. Most of them are heuristics because of the time complexity of the computational problem.

V. 1. 1. Branch and bound algorithm

One exact algorithm has been proposed: a branch and bound method [26] which solves the problem optimally. The implementation takes into consideration negative errors of arbitrary types and the multiplicity information model of the type "one and many". The algorithm explores the space of all solutions and stores these which utilize all spectrum elements and represent a sequence not longer than the analyzed one. It is verified if a given exploration path may lead to a feasible solution. If not then it is terminated to reduce the computation time. Subsequently, the stored solutions are validated using the multiplicity information, i.e., each spectrum element should be used a given number of times. Inconsonant solutions are discarded. Finally, the set of stored solutions is returned. Using multiplicity information leads to a significantly reduced number of acceptable solutions.

V. 1. 2. Greedy algorithm

Another implemented algorithm is a greedy heuristic for SBH with errors of arbitrary types [27]. It is able to use as an input a multispectrum with multiplicity information of the type "one and many" or "one, two and many"

The heuristic starts at an initial oligonucleotide and it iteratively extends a current solution by adding l-mers. A reconstructed sequence cannot be longer than an analyzed one of length n so the process stops when appending another oligonucleotide violates the maximum length constraint. The criterion to choose the next oligonucleotide is the cost of overlapping of the last l-mer in the current solution and a new one plus the smallest overlapping cost of the new one and one of its possible successors. The next l-mer is chosen from a set of oligonucleotides not used yet. For each l-mer the number of occurrences in the current solution is monitored. If it reaches the maximum value (m_i, provided as the parameter) then a given oligonucleotide is not taken into consideration during selection any more.

The algorithm has been tested in a computational experiment and the results have shown that even the partial multiplicity information leads to better sequence reconstruction, i.e., it increases the alignment score. Moreover, using the more precise model of multiplicity information "one, two and many" enables to obtain slightly better results.

V. 1. 3. Tabu search algorithm

The performance of the greedy algorithm encouraged its authors to implement a more sophisticated heuristic. A tabu and scatter search algorithm has been proposed [28]. It solves the problem with any kind of hybridization errors and is able to take into account the multiplicity information models "one and many" and "one, two and many". The input data for the algorithm consist of the multispectrum (i.e., a multiset of oligonucleotides), a length of a target DNA sequence and the first oligonucleotide of the target (optionally). The global criterion function is the number of utilized oligonucleotides. The goal of the algorithm is to maximize it and compose a sequence not longer than the target. An initial solution is obtained using the greedy algorithm and it is represented by an ordered list of oligonucleotides. The multiplicity model determines both the minimum and the maximum number of occurrences of a given oligonucleotide and this constraint is satisfied at each stage of computation. Additionally, a DNA sequence corresponding to a current solution can never be longer than the target.

This algorithm has been compared with several existing ones. First, it has been compared with a previous implementation of tabu and scatter search for the classical SBH [19]. Spectra coming from sequences without repetitions have been used but they have been affected by random positive errors and negative errors to simulate the imperfect hybridization experiment. The length of oligonucleotides in spectra has been equal to 10. For the longest considered sequences of length 500bp, the new algorithm solved optimally (i.e., it used the maximum number of spectrum elements) 26 of 40 instances and the average solution similarity to an analyzed sequence (i.e., average global alignment score) was 95.11%. The corresponding values for the previous algorithm were 18 of 40 instances and 85.50%, respectively.

The next comparison utilized 59 real DNA sequences of length 509bp with natural repetitions only. The number of optimally solved instances for a hybrid genetic algorithm

[18], the old tabu and scatter search algorithm [19] and for the new tabu and scatter search [28] were respectively: 26, 52 and 59 (the last one solved all instances optimally).

The tabu and scatter search using partial multiplicity information has been also compared to a revised hybrid genetic algorithm [20]. A test set consisted of 40 human DNA sequences. They contained from 1 to 17 repetitions. In the case of no hybridization errors, the new tabu and scatter search generates significantly better results. It reconstructed perfectly (i.e. obtained solution the same as a target sequence) 23 instances and the average similarity was 93.62%. The revised hybrid genetic algorithm solved perfectly 18 instances and the average similarity was 90.99%. The algorithms were also compared using spectra of the same sequences affected by 5% random positive errors and up to 5% random negative errors simulating a biochemical experiment. In this case, the new tabu and scatter search solved perfectly 19 instances and the average similarity was 91.56%. The revised hybrid genetic algorithm solved perfectly one instance less but the average similarity was slightly higher (92.60%).

The impact of the multiplicity information on the results generated by the tabu and scatter search algorithm has also been checked. The computational experiment results confirm that using even partial multiplicity information leads to improved sequence reconstruction. Moreover, the more precise model "one, two and many" enables to solve perfectly more instances and the average similarity of obtained solutions is higher in comparison with the model "one and many".

V. 1. 4. Ant Colony Optimization algorithm

The last currently implemented algorithm for sequencing by hybridization with information about repetitions is an ant colony optimization algorithm (ACO) [29]. This heuristic is a probabilistic, iterative search for a path in a given graph. It is based on ACO presented in [30]. It solves the problem with any kind of hybridization errors and is able to take into consideration the multiplicity information models "one and many" and "one, two and many". The ant colony optimization algorithm outperforms the greedy algorithm. Moreover, computational experiment results confirmed that applying even partial multiplicity information leads to better sequence reconstruction.

VI. RESEQUENCING

Another area where sequencing by hybridization has been and is still being used is the resequencing, i.e., reading DNA sequence under the assumption that some given reference sequence is available. Such SBH approach has also been used for the analysis of homologous sequences. During the process of evolution the DNA of any species is changing, yet the changes are very slight. Sequences which evolved from the same ancestral sequence are very similar to the source and thus are also alike to each other and they are called homologous. For example, in the human genome the forecast average number of differences in DNA coming from two people is 1 per 100-300 base pairs. Such a variation occurring when a single nucleotide differs in DNA of two individuals is called Single Nucleotide Polymorphism (SNP). The homology phenomenon is not limited only to the same species. The genome of different but phylogenetically related species may also have homologous regions. The similarity may be up to 100% in highly conserved segments. The obtained chimpanzee genome is different by 1.23% compared to the human genome only if direct sequence comparison is utilized [31].

VI. 1. Pe'er resequencing algorithm

Availability of the already sequenced DNA and its high similarity to homologous counterparts encourage to develop new sequencing methods. It is possible to use a known DNA sequence to determine the homologous one. The main contribution of applying homologous information should be more unambiguous solutions. This idea was utilized by Pe'er et al. to develop a polynomial dynamic programming algorithm [32, 33].

In the classical SBH a signal from a DNA chip is converted into binary information about spectrum composition. Pe'er et al. applied a stochastic signal quantification. They define two probabilities for each spectrum element s_i, i.e. probability $P_1(s_i)$ that a given oligonucleotide is a part of an analyzed sequence and probability $P_0(s_i)$ that it is not a part of the sequence. The output of the biochemical experiment is *probabilistic spectrum* (PS) as a result. One should note that the classical spectrum is a set of l-mers. The probabilistic spectrum is a pair (P_0, P_1) of functions $P_i : \Sigma_{DNA}^l \mapsto [0, 1]$, where l is the length of oligonucleotides. Note that these functions are defined over a full oligonucleotide library.

Pe'er et al. modeled the resequencing problem using the de Bruijn graph $G(V, E)$, where vertices are labeled by all $(l-1)$-mers. Each oligonucleotide $s_i = \langle s_{i_1} s_{i_2} \dots s_{i_l} \rangle$ from the library is represented by arc $e_i = (v', v)$ which connects two vertices $v' = \langle s_{i_1} s_{i_2} \dots s_{i_{l-1}} \rangle$ and $v = \langle s_{i_2} s_{i_3} \dots s_{i_l} \rangle$. Weight $w(e_i)$ of arc e_i representing s_i is related to the probabilities $P_0(s_i)$ and $P_1(s_i)$ and it is equal to $\log_2 \frac{P_1(s_i)}{P_0(s_i)}$. A sequence of length n is represented by a path in G of length $n - l + 1$.

Obtained sequence Q is evaluated by the score which consists of two elements, the so called *experimental likelihood* computed for Q and homology information which represents the probability of a mutation on a given position in comparison with a homologous sequence. Details about such scoring function is given in [32]. Calculating only the optimal score requires memory space $O(|V|)$. However, in order to reconstruct the corresponding sequence trace back pointers for the full $n \times |V|$ matrix have to be stored. Pe'er et al. presented in [32, 33] how to reduce the required

space to $O(|V|)$ by increasing time complexity by the factor $O(\log n)$.

It should be noted that authors assumed utilization of standard DNA chips which can be mass produced. The cost of the biological experiment to obtain hybridization data are reduced in comparison to other SNP detection methods using special purpose chips [34, 35]. The use of universal DNA chips raise an issue. The number of probes on a chip of this type increases exponentially with the probe length. A DNA chip containing the full oligonucleotide library of length 9 consists of $4^9 \approx 2.6 \cdot 10^5$ probes. The development of the DNA chip technology enables to create a chip of this size [36, 37], but a cost-effective approach should utilize a significantly smaller one.

Pe'er at al. validated their approach by performing several biochemical experiments which utilized a library of all 5-mers [38]. However, if shorter oligonucleotides are used then experiment results are less specific and more noisy. They overcome these obstacles, at least partially, by applying the polymerase signaling assay (PSA) [39] instead of simple hybridization. PSA is a more sophisticated method and uses enzymatic discrimination, based on single-nucleotide primer extension, to identify oligonucleotides which are a part of a target sequence. The results of the experiment performed by Pe'er at al. presented in [38] shows that their method enables to successfully resequence in practice targets of length ca. 100 bp.

VI. 2. Shotgun SBH

Another approach to resequencing was proposed by Pihlak at al. [2]. Authors named their method *shotgun sequencing by hybridization* (shotgun SBH). It also utilizes the information about spectrum composition, but these data are obtained in a completely different way. The main difference lies in DNA chip composition. In the classical SBH approach it consist of l-mers representing elements of a oligonucleotide library and is put into a solution of fluorescent or radioactively labeled target DNA. In the case of shotgun SBH, a chip contains fragments of a target DNA sequence and is placed in contact with labeled probes being oligonucleotides of length 5 (*pentamers*). Shorter probes would have difficulties to hybridize properly, and longer ones would require a much larger probe set [2].

A chip used in a biochemical experiment of shotgun SBH is created as follows. In the beginning a DNA sequence is randomly fragmented and single-stranded subsequences of length 200bp are isolated. Every fragment is concatenated with a 50bp universal linker and formed as a closed circular molecule. The universal linker is used to bind the circle into a chip surface. Subsequently, *in-situ* rolling-circle amplification (RCA) is performed. As a result, the chip contains many tandem-repeated copies of a given feature which form loosely coiled balls.

In the classical approach, signals related to all probes are captured at once. Melting temperature of particular l-mers is different so adjusting experiment conditions is a challenge. Shotgun SBH tackle this issue. For the hybridization experiments on the DNA fragments, a set of 582 pentamer probes has been developed. As the authors explained, the DNA fragments were to be obtained from both strands of the genome, therefore, half of all 1024 possible pentamers sufficed to tile the reference genome at every position on either strand. A full set of probes consists of 512 pentamers plus additional 70 special-purposes probes. A hybridization experiment is performed independently for each probe s_i from the set so it is possible to optimize hybridization temperature according to a given l-mer.

In short, the shotgun SBH method proposed in [2] consist of four steps:

1. In situ rolling-circle amplification of circular single-stranded DNA fragments.
2. For each feature F (a small fragment of single-stranded DNA), spectrum $S(F)$ is generated by sequential hybridization of each probe independently.
3. Alignment of the spectra to reference sequence H.
4. Reconstruction of a target DNA using reference sequence H and the aligned spectra.

Pihlak et al. validated their approach in practice. They performed an experiment using a library of 5-mers to resequence 48.5kbp Bacteriophage λ genome and 4.6Mbp *Escherichia coli* genome. The results outperformed the classical methods - shotgun SBH decoded correctly in the first case ca. 96% of the analyzed genome and ca. 80% in the second case [2].

As a result, shotgun SBH provides a sequenced genome of a given individual. In particular, it identifies occurrences of single nucleotide polymorphisms (SNPs), i.e. variations in comparison to an already sequenced genome of another member of a species. These variations in human DNA influence how humans develop diseases and respond to drugs, vaccines, etc., so this information has very important practical applications in medicine. It may be used, for example, to:

- resequence bacterial or viral genome to identify drug resistance,
- identify SNP (Single Nucleotide Polymorphism) responsible for genetic diseases to enable medical diagnostics and prognostics,
- realize the concept of personalized medicine.

A resequencing platform may also be used to validate sequences which have already been determined. If a given DNA has been properly sequenced then resequencing it should produce the same result.

VII. NON-CLASSICAL SEQUENCING BY HYBRIDIZATION

One of the proposed ways to improve the efficiency of the original SBH methodology is the idea to modify the

DNA chip itself. Some different definitions of probes have been proposed. Classical DNA chips contain a number of cells where oligonucleotides are attached. These cells are probes and each of them contains a huge number of oligonucleotides of the same type, i.e. each probe is composed of many copies of same short single stranded DNA fragment. Every probe contains different oligonucleotides but all of them are of the same length. The probe is marked in the hybridization experiment when its oligonucleotides hybridize to the analyzed DNA. In many non-classical DNA sequencing chips one probe can be composed of many different types of oligonucleotides. The probe is then described by a specific *pattern* which defines what types of oligonucleotides compose it.

In the paper [40] Pevzner and Lipshutz have introduced three different non-classical DNA sequencing chips called Binary, Gapped and Alternating chip. The main reason for such a proposal has been to present chips that are able to unambiguously sequence longer DNA fragments that the classical SBH approach can. Any DNA sequencing chip can be described by a so called *branching probability*, which defines the probability of ambiguously extending a random n-sequence upon reconstruction with a spectrum coming from a specific biochip. The greater the branching probability is, the shorter DNA sequence can be unambiguously reconstructed. Three proposed chips in [40] have smaller branching probability than classical chips having the same number of probes. Each of those chips uses a so called *unspecific nucleotide* which can bind to more than one natural nucleotide described by a set $\Sigma_{DNA} = \{A, C, G, T\}$. An alternating chip uses an unspecified nucleotide denoted as X that can bind with any natural nucleotide from Σ_{DNA}. Chip *capacity* tells us how many probes the chip has. The total capacity of the alternating chip is $||C_{alt}(k)|| = 2 \cdot 4^k$. The chip is composed of all probes of two types, which can be described by the following general patterns:

$$N_1XN_2X...XN_k \text{ and } N_1XN_2X...XN_{k-1}N_k \qquad (1)$$

The number of X symbols is equal to $k-1$ for the first type of probes and $k-2$ for the second type. In both types of probes the number of known nucleotides denoted above as N ($N \in \Sigma_{DNA}$) is equal to k.

The next proposed chip has been called *Gapped chip*. It utilizes the same non-specific x nucleotide, but within a different pattern:

$$N_1N_2...N_k \text{ and } N_1N_2N_{k-1}...\underbrace{XX...X}_{k-1}N_k \qquad (2)$$

The total capacity $||C_{gap}(k)|| = 2 \cdot 4^k$ is the same as previously for the alternating chip. The third proposed chip is called *Binary chip*. The pattern describing its two halves is as follows:

$$\underbrace{\{W,S\},\{W,S\},...,\{W,S\}}_{k},N \quad \text{and} \quad \underbrace{\{R,Y\},\{R,Y\},...,\{R,Y\}}_{k},N \qquad (3)$$

where W, S, R and Y are two-letters sets such that $W = \{A, T\}$, $S = \{C, G\}$, $R = \{A, G\}$, $Y = \{C, T\}$, $N \in \Sigma_{DNA}$. Such two-element sets define their complementary nucleotides from Σ_{DNA}, e.g. W elements are complementary to T and A, respectively, while R elements to T and C. Binary chip capacity is defined as $||C_{bin}(k)|| = 2 \cdot 2^k \cdot 4$.

Such chips can have fewer probes compared to their classical counterparts in order to unambiguously sequence a given DNA fragment, but every probe can hybridize to more than one fragment of the DNA. The probe itself is described by a pattern which defines the set of oligonucleotides composing it. This can be achieved in two ways. First, one can put more than one type of oligonucleotide in a single probe, in order for the probe to be complementary to a given set of DNA fragments. There is, however, a problem with the strength of the hybridization signal, i.e., the more types of different oligonucleotides are in a single probe, the more difficult it is to detect such a hybridization event. In the classical SBH each probe consists of multiple copies of exactly one type of oligonucleotide. In order to simulate non-specific nucleotides behavior, many different oligonucleotides can be used within a single probe. For example, if the probe is denoted WA in the Binary chip and $W = \{A, T\}$, within such a probe two short oligonucleotides can be placed: AA and TA. However, due to technical difficulties, putting so many different oligonucleotides within a single probe weakens the hybridization signal during the hybridization experiment. Two algorithms have been proposed for Gapped chip [41] and recently for the Alternating chip data [42].

Another approach uses degenerate or universal bases as the building blocks of the oligonucleotides. The universal bases can (theoretically) bind to every natural nucleotide in the DNA. The degenerate bases are complementary to more than one type of nucleotide but not to all of them. Such artificial nucleotides have already been developed, but using them in the SBH methodology is uncommon [43, 44]. Such an approach has been proposed and analyzed in a different scientific papers, for example in [45] where different chip idea has been proposed and its properties analyzed. The algorithm for the DNA reconstruction basing on such an idea has been proposed in the same paper. In [46] there is an extensive analysis of the universal and so called *semidegenerate* bases. Authors present the results of various experiments and the analysis of the practical behavior of universal bases. The semidegenerate nucleotide idea is proposed in the same paper, basing on

the energy of the binding process in the hybridization phase. At the end of the paper algorithmic considerations are proposed. As for the Binary chip mentioned earlier, its idea has been also discussed in [47] by Sachadyn and Kur, where authors proposed yet another approach called sequencing by hybridization with oligonucleotides matrix.

VII. 1. Isothermic oligonucleotide libraries

In the hybridization experiment DNA fragments bind with the probes on the surface of a microarray. When such an experiment is considered, the process of DNA denaturation must be taken into account. Such a process takes place in the sufficiently high temperature and, in general, the longer the double helix, the more energy is necessary to achieve the denaturation. The melting temperature is the basis for another approach to construct DNA chips. Isothermic oligonucleotide libraries consist of oligonucleotides having different length but the same melting temperature. In order to construct such a library, a numerical value (a temperature) is assigned to each nucleotide. It corresponds to the increment which a particular nucleotide brings into the stability of oligonucleotide duplexes. In [48] a value equal to 2 is assigned to A and T nucleotides, a value equal to 4 to the C and G ones. This is a very simple thermodynamic model; however, previously none have been taken into account. A single isothermic library consists of oligonucleotides with the same melting temperature, which is a sum of degrees assigned to each nucleotide in the oligonucleotide. Obviously, such a library contains oligonucleotides of various length. In the classical SBH, the libraries of oligonucleotides of a given length are used, so their melting temperatures are different. Therefore, it is difficult to set the conditions of the biochemical experiment such that all the oligonucleotides create stable duplexes. This issue influences the number of hybridization errors: in general, isothermic libraries allow to decrease the number of hybridization errors. Because the DNA binding is more stable, the hybridization signal from a probe is stronger, which allows to create a spectrum with fewer hybridization errors than in the classical SBH experiment. The present paper has proved an important claim that it is always possible to cover the DNA sequence with the probes from two isothermic libraries having melting temperature which differ by two degrees. This allows to cover the sequence in such a way that two consecutive oligonucleotides from the libraries have starting points shifted by one position at most. In [48] the isothermic oligonucleotide libraries are studied and their various properties are being analyzed. The sequencing problem with and without errors for such libraries are formulated and followed by the complexity and computational results. For example, a proof of NP-completeness for a decision version of an isothermic SBH with positive and negative errors problem is given. Search versions of various isothermic SBH problems depending on the type of errors are also formulated and discussed. For the problem without errors a polynomial time algorithm has been proposed in [49].

In [50] a new sequencing algorithm for the isothermic oligonucleotide libraries has been introduced, basing on the tabu search approach. The proposed algorithm deals with both types of hybridization errors, the negative and positive ones. Various experiments have been performed using DNA sequences obtained from GenBank, with percentage of errors in a spectrum set to 5%, 10% and 20%. Libraries with melting temperatures equal to $26/28$ and $36/38$ degrees, have been tested. The smaller ones, with temperatures equal to 26 and 28 degrees have similar cardinality as one classical library with oligonucleotides length equal to 10, which allows for comparison of the results for different libraries. Such an isothermic library for smaller sequences (200bp) allows reconstruction having almost 100% similarity to the original one if no more than 10% of errors of both types (positive and negatives) are present in the spectrum. If the number of errors is equal to 20% the similarity remained quite high - 93.25%. For longer sequences (600bp) similarity is between 81.88% and 76.47% depending on the number of errors. When larger isothermic libraries have been used, even for the long sequences and high rate of errors the similarity has been greater than 93%. The decrease of the quality has been less than 1.5% between two cases of hybridization errors rate: 5% and 20%. It should be stressed again that the purpose behind creating the isothermic libraries lies in the reduction of the hybridization errors. It means that a real experiment should have fewer hybridization errors than the theoretical levels used for computation purpose.

An interesting approach has been introduced in [51]. It joins isothermic libraries with a method described earlier: the multistage hybridization. Each method is resilient to a specific type of hybridization errors and more vulnerable to others. Hybridization in rounds tends to eliminate influence of substrings repetitions (i.e., of negative errors from repetitions) on the quality of obtained solutions, but it is quite susceptible to "normal" hybridization errors, especially negative errors resulting from imperfectness of probe detection technology. On the other hand, as it has been previously described in detail, isothermic libraries reduce such hybridization errors by taking into account the melting temperature in the hybridization experiment. The combined approach aims to be more efficient in handling both types of errors in a spectrum. The algorithms for the DNA sequence reconstruction and for the isothermic libraries creation for a multistage approach are proposed and extensively described, forming the so called *multistage isothermic SBH* approach.

VIII. SUMMARY

Sequencing by hybridization can be currently considered as a rather old sequencing method, yet it is still used for some specific tasks like, e.g., medical diagnosis, where at a low cost it can provide valuable data for diagnostic pur-

poses. However, even larger scale sequencing can be done using modified SBH approaches, as proved by recent and older publications. From the computer science perspective there are a lot of interesting combinatorial problems which appear in the computational phase of different variants of SBH. Finally, new algorithms for various SBH methods are still being proposed and the results often show that in some specific areas sequencing by hybridization can still be considered as a useful method of reading DNA sequences.

References

[1] J.T. Bosh and W. Grody, *Keeping up with the next generation: massively parallel sequencing in clinical diagnostics,* The Journal of Molecular Diagnostics **10**(6), 484–492 (2008).

[2] A. Pihlak, G. Baurén, E. Hersoug, P. Lönnerberg, A. Metsis, S. Linnarsson, *Rapid genome sequencing with short universal tiling probes.,* Nature biotechnology **26**, 676–684 (2008).

[3] J. Hardick, R. Woelfel, W. Gardner, S. Ibrahim, *Sequencing ebola and marburg viruses genomes using microarrays,* Journal of Medical Virology **88**(8), 1303–1308 (2016).

[4] V. Swaminathan, G. Rajaram, V. Abhishek, B.S. Reddy, K. Kannan, *A novel hypergraph-based genetic algorithm (hgga) built on unimodular and anti-homomorphism properties for DNA sequencing by hybridization, Interdisciplinary Sciences: Computational Life Sciences*, 2017.

[5] J. Błażewicz and M. Kasprzak, *Complexity of* DNA *sequencing by hybridization,* Theoretical Computer Science **290**(3), 1459–1473 (2003).

[6] P. Pevzner, *l-tuple DNA sequencing: a computer analysis,* Journal Of Biomolecular Structure & Dynamics **7**, 63–73 (1989).

[7] J. Błażewicz, P. Formanowicz, M. Kasprzak, P. Schuurman, G. Woeginger, *A polynomial time equivalence between DNA sequencing and the exact perfect matching problem,* Discrete Optimization **7**, 154–162 (2007).

[8] D. Margaritis and S. Skiena, *Reconstructing strings from substrings in rounds,* Proceedings 36th Symposium on Foundation of Computer Science **6**(2), 237–252 (1995).

[9] S. Skiena and G. Sundaram, *Reconstructing strings from substrings,* Journal of Computational Biology **2**, 333–353 (1995).

[10] S. Kruglyak, *Multistage sequencing by hybridization,* Journal of Compytational Biology **71**, 165–171 (1998).

[11] A. Frieze and B. Halldorsson, *Optimal sequencing by hybridization in rounds,* Journal of Computational Biology **9**, 355–369 (2002).

[12] D. Tsur, *Sequencing by hybridization in few rounds,* Lecture Notes in Computer Science, Algorithms - ESA 2003 **2832**, 506–516 (2003).

[13] S. Broude, T. Sano, C. Smith, C. Cantor, *Enhanced DNA sequencing by hybridization,* Proceedings of National Academy of Science USA **91**, 3072–3076 (1994).

[14] S. Hannenhalli, P. Pevzner, H. Levis, S. Skiena, *Positional sequencing by hybridization,* Computer Applications in Biosciences **12**, 19–24 (1996).

[15] A. Ben-Dor, I. Pe'er, R. Shamir, R. Sharan, *On the complexity of positional sequencing by hybridization,* Journal of Computational Biology **8**, 361–371 (2001).

[16] J.-H. Zhang, L.-Y. Wu, Y.-Y. Zhao, X.-S. Zhang, *An optimization approach to the reconstruction of positional* DNA *sequencing by hybridization with errors,* European Journal of Operational Research **182**(1), 413–427 (2007).

[17] J. Błażewicz, F. Glover, M. Kasprzak, *Evolutionary approaches to* DNA *sequencing with errors,* Annals of Operations Reasearch **138**(1), 67–78 (2005).

[18] J. Błażewicz, M. Kasprzak, W. Kuroczycki, *Hybrid genetic algorithm for* DNA *sequencing with errors,* Journal of Heuristics **8**(5), 495–502 (2002).

[19] J. Błażewicz, F. Glover, M. Kasprzak, DNA *sequencing – tabu and scatter search combined,* INFORMS Journal on Computing **16**(3), 232–240 (2004).

[20] J. Błażewicz, F. Glover, M. Kasprzak, W.T. Markiewicz, C. Oguz, D. Rebholz-Schuhmann, A. Świercz, *Dealing with repetitions in sequencing by hybridization,* Computational Biology and Chemistry **30**(5), 313–320 (2006).

[21] J. Błażewicz, C. Oguz, A. Świercz, J. Węglarz, DNA *sequencing by hybridization via genetic search,* Operations Research **54**(6), 1185–1192 (2006).

[22] J.-H. Zhang, L.-Y. Wu, X.-S. Zhang, *Reconstruction of* DNA *sequencing by hybridization,* Bioinformatics **19**(1), 14–21 (2003).

[23] P. Formanowicz, DNA *sequencing by hybridization with additional information available,* Computational Methods in Science and Technology **11**(1), 21–29 (2005).

[24] M. Schena, *Microarray Analysis*. Hoboken, New Jersey: Wiley-Liss, 2003.

[25] P. Formanowicz, *Isothermic sequencing by hybridization problems with information about repetitions,* Electrical Review **9**, 103–107 (2008).

[26] K. Kwarciak, M. Radom, P. Formanowicz, DNA *sequencing with negative errors and information about repetitions,* Zeszyty Naukowe Politechniki Śląskiej: Automatyka **151**, 215–222 (2008).

[27] K. Kwarciak and P. Formanowicz, *A greedy algorithm for the* DNA *sequencing by hybridization with positive and negative errors and information about repetitions,* Bulletin of the Polish Academy od Sciences, Technical Sciences **51**(1), 111–115 (2011).

[28] K. Kwarciak and P. Formanowicz, *Tabu search algorithm for* DNA *sequencing by hybridization with multiplicity information available,* Computers and Operations Research **47**, 1–10 (2014).

[29] K. Kwarciak, M. Radom, P. Formanowicz, *A multilevel ant colony optimization algorithm for* DNA *sequencing by hybridization with multiplicity information available.* in press, 2015.

[30] C. Blum, M.Y. Vallès, M.J. Blesa, *An ant colony optimization algorithm for* DNA *sequencing by hybridization,* Computers and Operations Research **35**(11), 3620–3635 (2008).

[31] The Chimpanzee Sequencing and Analysis Consortium, *Initial sequence of the chimpanzee genome and comparison with the human genome,* Nature **437**, 69–97 (2005).

[32] I. Pe'er and R. Shamir, *Spectrum alignment: Efficient resequencing by hybridization,* in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 260–268, AAAI Press, 2000.

[33] I. Pe'er, N. Arbili, R. Shamir, *A computational method for resequencing long* DNA *targets by universal oligonucleotide arrays, Proceedings of the National Academy of Sciences*, vol. 99, no. 24, pp. 15492–15496, 2002.

[34] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley, E.S. Lander, *Characterization of single-nucleotide polymorphisms in coding regions of human genes,* Nature Genetics **22**, 231–238 (1999).

[35] J.G. Hacia, *Resequencing and mutational analysis using oligonucleotide microarrays,* Nature Genetics **21**, 42–47 (1999).

[36] K.L. Gunderson, X.C. Huang, M.S. Morris, R.J. Lipshutz, D.J. Lockhart, M.S. Chee, *Mutation detection by ligation to complete n-mer* DNA *arrays.,* Genome Res **8**(11), 1142–1153 (1998).

[37] N.J. Haslam, N.E. Whiteford, G. Weber, A. Prugel-Bennet, J.W. Essex, C. Neylon, *Optimal probe length varies for targets with high sequence variation: Implications for probe library design for resequencing highly variable genes, PLOS One* **3**(6), e2500 (2008).

[38] I. Pe'er, N. Arbili, Y. Liu, C. Enck, C.A. Gelfand, R. Shamir, *Advanced computational techniques for re-sequencing* DNA *with polymerase signaling assay arrays,* Nucleic Acids Research **31**(19), 5667–5675 (2003).

[39] M.T. Boyce-jacino, M.B. Addelston, S.R. Head, *Polymerase signaling assay,* http://www.freepatentsonline.com/6872521.html, March 2005.

[40] P. Pevzner and R. Lipshutz, *Towards DNA sequencing chips,* Symposium on Mathematical Foundations of Computer Science, Lecture Notes in Computer Science **841**, 143–158 (1994).

Symposium on Mathematical Foundations of Computer Science **841**, 143–158 (1994).

[41] M. Radom and P. Formanowicz, *Algorithms for sequencing by hybridization problems based on non-classical DNA chips,* Przeglad Elektrotechniczny **10**, 97–100 (2010).

[42] M. Radom and P. Formanowicz, *An algorithm for sequencing by hybridization based on an alternating DNA chip,* Computational Biology and Chemistry **10**(3), 67–78 (2018).

[43] P. Zhang, M. Egholm, N. Paul, M. Pingle, D. Bergstrom, *Peptide nucleic acid*-DNA *duplexes containing the universal base 3-nitropyrrole,* Methods **23**, 132–140 (2001).

[44] K. Too and D. Loakes, *Universal Base Analogues and their Applications to Biotechnology*. Wiley-VCH Verlag GmbH and Co. KGaA, 2008.

[45] A. Frieze, F. Preparata, E. Upfal, *Optimal reconstruction of a sequence from its probes,* Journal of Computational Biology **6**, 361–368 (1999).

[46] F. Preperata and J. Oliver, DNA *sequencing by hybridization using semi-degeneratae bases,* Journal of Computational Biology **11**, 753–765 (2004).

[47] P. Sachadyn and J. Kur, *Reducing the number of microlocations in oligonucleotide microchip matrices by the application of degenerate oligonucleotides,* Journal of Computational Biology **197**, 393–401 (1999).

[48] J. Błażewicz, P. Formanowicz, M. Kasprzak, W. Markiewicz, *Sequencing by hybridization with isothermic oligonucleotide libraries,* Discrete Applied Mathematics **145**, 40–51 (2004).

[49] J. Błażewicz and M. Kasprzak, *Computational complexity of isothermic DNA sequencing by hybridization,* Discrete Applied Mathematics **154**, 718–729 (2006).

[50] J. Błażewicz, P. Formanowicz, M. Kasprzak, W. Markiewicz, A. Świercz, *Tabu search algorithm for* DNA *sequencing by hybridization with isothermic libraries,* Computational Biology and Chemistry **28**, 11–19 (2004).

[51] J. Błażewicz, P. Formanowicz, *Multistage isothermic sequencing by hybridization,* Computational Biology and Chemistry **29**, 69–77 (2005).

Kamil Kwarciak works as Technical Project Manager in Cognifide. He received his PhD in Computer Science from Poznan University of Technology in 2014. His research interests focus on DNA sequencing and heuristic algorithms (i.e. tabu search, ant colony optimization). He is also interested in project management and programming.

Marcin Radom has been working as Assistant Professor in the Faculty of Computing Science, Poznan University of Technology, since 2012. He received his PhD in Computer Science from Poznan University of Technology in 2011. He also works in the Institute of Bioorganic Chemistry, Polish Academy of Sciences. His research focuses on DNA sequencing, precisely the complexity problems connected with various sequencing methods. Other areas of interest include systems biology (i.e., modeling complex biological systems using Petri nets and model analysis issues) and programming in Java and C#.

Piotr Formanowicz received the PhD degree in 2000 and the habilitation qualification in 2005 from Poznan University of Technology. In 2015, he received the title of professor. He is an associate professor at the Institute of Computing Science, Poznan University of Technology, and a full professor at the Institute of Bioorganic Chemistry, Polish Academy of Sciences. His research interests include combinatorial optimization, scheduling theory, computational complexity, graph theory, computational biology and systems biology.

DOI:10.12921/cmst.2018.0000055