

DNA sequencing by hybridization with additional information available

Piotr Formanowicz

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

*Institute of Bioorganic Chemistry, Polish Academy of Sciences
Noskowskiego 12/14, 61-704 Poznań, Poland*

(Rec. 24 January 2005)

Abstract: In classical DNA sequencing by hybridization it is assumed that the information obtained in the biochemical stage of the method is a set of the l -tuples composing the target sequence. It means that the information concerning the number of the repeated l -tuples is not available. Such an assumption was justified by the DNA chip technology constraints. However, nowadays some approximate information about l -tuple multiplicities can be obtained in the experiments, where DNA chips are used. It was a motivation for formulating combinatorial problems which arise when such additional information is taken into account. The goal of this paper is to formulate and classify these problems, what should establish a good starting point for further research concerning algorithmic methods solving DNA sequencing problems with multiplicity information. Moreover, the computational complexity of the new problems is determined, which in most cases is analogous to the complexity of their classical counterparts.

Key words: DNA sequencing, l -tuple multiplicity, combinatorial problems, computational complexity

1. INTRODUCTION

Reading DNA sequences still remains one of the most important problems in molecular and computational biology. Despite the impressive progress in sequencing genomes of many species there is a need for an efficient and not expensive method for DNA sequencing.

The two main approaches for determining the sequences of DNA are shotgun sequencing and sequencing by hybridization (SBH) [5, 8, 11, 13].

In the latter method a set of l -long substrings composing the analyzed DNA string (*i.e.* the sequence of nucleotides) is determined. The procedure employs one of the fundamental properties of single stranded DNA molecules, *i.e.* their ability to join to complementary strands [14]. As it is known for over 50 years, a single stranded DNA molecule is a sequence (note, that from the computer science point of view it is rather string than sequence) of elementary building blocks called nucleotides, denoted by A, C, G, and T. According to the Watson-Crick complementarity rule nucleotide A is complementary to T and nucleotide C is complementary to G. From the rule it follows that if two single stranded DNAs contain some substrings being complementary to each other then these DNAs may hybridize (*i.e.* join by hydrogen bonds) creating a double stranded molecule.

The idea of SBH is based on the observation that an *oligonucleotide library*, *i.e.* a collection of short single stranded DNA molecules, when put into a solution of a number of copies of single stranded target DNA, the target molecules will hybridize to those elements of the library

which are complementary to some substring of the examined DNA. Nowadays, the oligonucleotide library is made usually as a *DNA chip* [6, 9, 11]. Such a chip is a kind of matrix divided into a number of cells, called *probes*, each of them containing elements of the library of one type, *i.e.* in each of the cells there is a number of identical short DNA strands.

In the standard version of the SBH method the DNA chip is composed of all oligonucleotides of a given length l . In this case the number of library element types (*i.e.* the number of different oligonucleotide sequences) equals 4^l and it is also a number of the probes on the chip. When such a chip is put into solution of the single stranded DNA, the molecules will join to those probes which contain complementary oligonucleotides. If the examined molecules were radioactively or fluorescently labeled then on the basis of the image of the chip one can get the information about the l -tuple composition of the target DNA (*i.e.* the chip is a detector of l -tuples composing the examined DNA). This information is an output of the first, biochemical stage of the SBH method.

In the second, computational stage the target sequence is reconstructed on the basis of the information provided by the biochemical stage. In the ideal case, *i.e.* when no errors occur, this stage provides full information about the l -tuple composition of the target DNA. In other words, one gets a set of all types of substrings of length l occurring in the target DNA sequence. The set is called *spectrum* of a given sequence and its cardinality equals to $n - l + 1$, where n is the length of the sequence. As one can notice,

determining spectrum is not equivalent to determining the whole nucleotide sequence of the examined molecules. In order to discover this sequence one has to determine the order of the spectrum elements in which they appear in the examined DNA. This order is determined in the second, computational stage of the SBH method.

As has been mentioned, the described biochemical stage of SBH is in fact its ideal case, where no errors occur. In practice usually two types of errors appear, *i.e.* the *negative* ones and the *positive* ones [1].

There are two kinds of negative errors. They may result from imperfection of the hybridization experiment, *i.e.* it may happen that in the target molecules there is a substring complementary to oligonucleotides composing some probe of the DNA chip, but the molecules do not hybridize to them. In this case the information about this l -tuple is missed, *i.e.* the cardinality of the obtained spectrum is less than $n - l + 1$. Another source of the negative errors are substring repetitions occurring in the target sequence. More precisely, if in the target DNA molecule there is a repetition of a substring of length at least l , it will not be detected in the hybridization experiment, because due to the current technology constraints in this experiment it is possible to discover only the presence of a given l -tuple in the examined DNA, but not the number of its occurrences (it is the case of the classical SBH variant). It means that spectrum is a set but not a multiset. In this case spectrum cardinality is also less than $n - l + 1$. The spectrum contains information about all types of l -tuples (*i.e.* all different substrings of length l) composing the examined DNA molecule but not about all such l -tuples.

The positive errors are also results of imperfection of the hybridization experiment. In this case the target DNA molecules hybridize to oligonucleotides composing some probe of the chip which are not 100% complementary to it. In such case spectrum will contain information about some l -tuples which are not part of the examined sequence and its cardinality will be greater than $n - l + 1$.

As has been mentioned, in classical SBH approach it is assumed that the hybridization experiment provides only information about presence or absence of a given l -tuple in the target DNA sequence (what results in a situation where spectrum is a set but not a multiset). This assumption is justified by the fact that the information about l -tuple composition of the sequenced DNA is read from an image of the DNA chip, where the cells corresponding to l -tuples present in the target sequence shine. So, the information is, in some sense, binary, since the shining probe indicate the presence of an l -tuple in the sequence while a not shining one means that the corresponding l -tuple is not present in the examined molecule.

Nevertheless, at least in principle, it is possible to take into account the intensity of the shining of the chip cells. This intensity is correlated with the number of repetitions of a given l -tuple in the target sequence. However, this

correlation is not simple and many factors can affect the intensity of the chip image. Because of the current technology constraints it is rather impossible to determine exact numbers of l -tuple repetitions but it may be possible, at least in the future, to determine approximate multiplicities of l -tuples. An additional justification of taking into account the intensity of the chip image is the fact that it is common practice in the analysis of the data coming from DNA microarrays used for gene expression analysis [12]. Here, the differences among the intensities of the cell images are interpreted as differences among the levels of gene expression which are equivalent to differences among amounts of nucleic acids hybridized to the microarray cells.

It is a motivation for formulating and classifying combinatorial problems which may arise in SBH approach, where such additional information is available. These problems must be solved in the second, computational stage of the SBH method. Taking into account the (partial) multiplicity information contained in the chip image makes the input data for the computational stage of the method more precise than in its standard version, which should have a positive impact on the quality of the sequences produced by the sequencing algorithms.

In the next section the multiplicity information, whose availability is assumed in this paper, will be described. In Sections 3, 4, and 5 three groups of sequencing problems will be formulated. For each of the groups some precision of the available multiplicity information is assumed. Moreover, computational complexity of most of the problems is established. The paper ends with conclusions in Section 6.

2. THE MULTIPLICITY INFORMATION

In what follows it is assumed that multiplicities of l -tuples detected in the hybridization experiment correspond to numbers of occurrences of these l -tuples in the target sequence. The accuracy of this correspondence depends on the type of multiplicity information assumed and on the hybridization errors which may occur (and, obviously, on the chip technology). Moreover, as mentioned above, it is assumed that these multiplicities are in some way proportional to the intensity of chip cell signals (however, this assumption is not important from the point of view of the correctness of the mathematical models). Moreover, it will be assumed that each positive error has multiplicity equal to 1, which is a reasonable assumption, since such an error is an oligonucleotide only partially complementary to the target sequence and should hybridize weakly giving a weak signal (*i.e.* it should not result in a signal stronger than the one generated by an l -tuple occurring once exactly in the target sequence).

The assumption that the number of repetitions is proportional to the intensity of the signal is justified by the fact that the bigger the number of repetitions of a given

oligonucleotide, the bigger the probability that the target sequence will hybridize to this oligonucleotide on the chip.

Let $S(Q)$ denote the spectrum of sequence Q and let $S^{(is)}(Q)$ denote an ideal spectrum of this sequence. The ideal spectrum contains all types of l -tuples distinguishable in the target sequence, but not all of these l -tuples, which are contained in a multispectrum. Let $S^{(im)}(Q)$ denote an ideal multispectrum for sequence Q . Let us also define $S^{(m)}(Q)$, which is a multispectrum for sequence Q . $S^{(m)}$ differs from $S^{(im)}(Q)$ because multiplicity of every sequence being in $S^{(m)}(Q)$ is limited in most of the problems formulated in this paper and the multispectrum can contain some errors, while an ideal multispectrum cannot. For every sequence $s_i \in S(Q)$ let us define parameter m_i which is equal to the number of occurrences of s_i in $S^{(m)}(Q)$.

There are strict correlations among an ideal multispectrum, a multispectrum, an ideal spectrum and a spectrum. Indeed, an ideal multispectrum is a multiset of all l -tuples which compose the target sequence. A multispectrum is a multiset which may contain only part of the l -tuples being elements of an ideal multispectrum and, in addition, it can contain some l -tuples not present in the ideal multispectrum. So, a multispectrum corresponds to the result of a real hybridization experiment, where some information about repetitions of l -tuples is also available. Finally, a spectrum is a set obtained from a multispectrum by discarding multiplicity of l -tuples being elements of this multiset while an ideal spectrum is obtained in the same way from the ideal multispectrum.

In the following sections three groups of problems will be formulated. For each of the groups some precision of the available multiplicity information is assumed. These assumptions are made in order to fit the models to the possible real world hybridization experiments as well as possible. The difficulties of obtaining exact multiplicity information are connected with the nature of the correlation between the intensity of the chip image and the amount of nucleic acids hybridized to the chip cells. Using current chip technology this correlation may be determined only approximately.

It should be noted that the decision versions of all of the problems formulated in this chapter are in class **P** (which is also the case of the classical SBH problems, see [3]). However, the search versions of some of them, being the most interesting from the theoretical, as well as practical point of view, are intractable.

3. INFORMATION ABOUT MULTIPLICITY OF THE TYPE "ONE AND MANY"

3.1. An outline

In this case any oligonucleotide may appear in a multispectrum 0, 1 or 2 times. So, for any sequence s_i ,

$m_i \in \{1, 2\}$. The case $m_i = 2$ is interpreted as at least two occurrences of s_i in the target sequence, if there are no positive errors, or at least one occurrence of s_i and one positive error, when such errors are taken into account.

This type of multiplicity information corresponds to a rather simplified situation, where the analysis of the chip image allows for distinguishing those l -tuples which are repeated in the target sequence, but the number of repetitions remains unknown. Although, such information is not very precise, it may be very useful for the target sequence reconstruction, especially in comparison to the standard models and algorithms, where no multiplicity information is exploited. Moreover, information of this type should be easy to obtain using current DNA chip technology.

3.2. Problem without errors

If there are no errors then $S(Q) = S^{(is)}(Q) = S^{(m)}(Q) = S^{(im)}(Q)$ and the problem can be formulated as follows:

Problem 1

INSTANCE: set $S(Q) = S^{(is)}(Q)$, length n of sequence Q .

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$.

Let us observe that in this case the fact that there are no errors means that there cannot be any repetitions of length, at least l , in the target sequence. Otherwise, if some string s_i were repeated, parameter m_i would have value 2, which could not be unambiguously interpreted as two occurrences of s_i .

This problem is identical to the classical variant of SBH without errors, hence it can be solved in polynomial time using the approach proposed by Pevzner [10].

3.3. Problem with negative errors resulting from repetitions

In this case $S(Q) = S^{(is)}(Q) \subset S^{(m)}(Q) \subseteq S^{(im)}(Q)$ and the problem is formulated as follows:

Problem 2

INSTANCE: set $S(Q) \subset S^{(im)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$, where sequence $s_i \in S(Q)$ appears once in Q' if $m_i = 1$ and it appears at least two times if $m_i = 2$.

Let us observe that if $\forall_{s_i \in S(Q)} m_i = 1$ then the problem is reduced to problem 1, *i.e.* to the problem without repetitions and can be solved in polynomial time.

On the other hand, if $\forall_{s_i \in S(Q)} m_i = 2$, every element of $S(Q)$ is repeated in Q and this subproblem is not equivalent to the classical variant of SBH problem with negative

errors resulting from repetitions only, because in the latter one it is not known which spectrum elements are repeated. However, these two problems are similar to each other, since in the case where $\forall_{s_i \in S(Q)} m_i = 2$ it is known that every s_i is repeated, but the number of repetitions remains unknown.

It is interesting that the complexity status of the classical problem with negative errors resulting from repetitions is an open question and it has been shown that it is polynomially equivalent to some other well known combinatorial problems, (*i.e.* exact Eulerian cycles, exact bipartite matching and restricted exact matching) [2]. However, this problem is not a subproblem of problem 2.

3.4. Problem with negative errors resulting from hybridization

In case of this problem $S(Q) = S^{(m)}(Q) \subset S^{(is)}(Q) = S^{(im)}(Q)$. The problem formulation is as follows:

Problem 3

INSTANCE: set $S(Q) \subset S^{(is)}(Q)$, length n of sequence Q .

ANSWER: sequence Q' of length n containing every element from $S(Q)$ exactly once and, in addition, containing some l -tuples not present in $S(Q)$.

Let us observe that this problem is identical with its classical counterpart since in both of them it is assumed that in the target sequence there are no repetitions of substrings of length l . Hence, the information about l -tuples multiplicity is useless in either of the problems.

Let us also note that in [7] it has been shown that some variant of the shortest common superstring problem is strongly NP-complete. This problem has been used for proving strong NP-hardness of the standard SBH problem with negative errors in [3]. For the complexity status of problem 3 it is important that the transformation shown in [7] is such that in the resulting superstring there are no repetitions of substrings of length equal to the length of the input strings. (In the superstring constructed by the transformation there are no repeated strings of length 3 in case of unbounded alphabet. In case of three-letter alphabet it suffices to encode character “#” by “aa...a” and encode the other characters according to the rule proposed by the authors. Such encoding guarantees that in the resulting superstring there will be no repeated substrings of length equal to the length of input strings – see [7] for details.) From this follows that problem 3 is NP-hard in the strong sense.

3.5. Problem with negative errors of arbitrary types

In this case the negative errors may result from repetitions and from imperfection of the hybridization experiment. The relations between spectra are as follows: $S(Q) \subseteq S^{(is)} \subseteq S^{(im)}(Q)$ and $S(Q) \subseteq S^{(m)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated in the following way:

Problem 4

INSTANCE: set $S(Q) \subseteq S^{(im)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all elements of $S(Q)$, where $s_i \in S(Q)$ appears in Q' once if $m_i = 1$ and at least two times if $m_i = 2$. Moreover, this sequence can contain some l -tuples which are not elements of $S(Q)$.

In case where $\forall_{s_i \in S(Q)} m_i = 1$ the problem is reduced to problem 3, so it is strongly NP-hard. It is worth noting that in case where $\forall_{s_i \in S(Q)} m_i = 2$ the subproblem is not equivalent to the classical SBH problem with negative errors of general type, since here we have precise information which l -tuples are repeated.

This problem can be also formulated in a slightly different way:

Problem 5

INSTANCE: set $S(Q) \subset S^{(im)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all elements of $S(Q)$, where $s_i \in S(Q)$ appears in Q' once if $m_i = 1$ and at least two times if $m_i = 2$. Moreover, this sequence contains at least one l -tuple which is not an element of $S(Q)$.

In this formulation it is assumed that at least one negative error resulting from hybridization imperfection appears in the data. Here we have $S(Q) \subset S^{(is)} \subseteq S^{(im)}(Q)$ and $S(Q) \subseteq S^{(m)}(Q) \subseteq S^{(im)}(Q)$. This modified problem corresponds to the situation, where it is known that errors resulting from hybridization imperfection not only can appear but appear really. Obviously, this problem is also strongly NP-hard.

3.6. Problem with positive errors

In this case $S^{(is)}(Q) \subset S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) = S^{(im)}(Q)$. The problem is formulated in the following way:

Problem 6

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subset S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple occurring in it is an element of $S(Q)$, Q' contains every $s_i \in S(Q)$ for which $m_i = 2$ and there are no repeated l -tuples in Q' .

In case where $\forall_{s_i \in S(Q)} m_i = 1$ the problem becomes equivalent to the classical one with positive errors, since in this case it is unknown which elements of $S(Q)$ are positive errors – like in the classical case. Hence, the classical counterpart being a strongly NP-hard problem [3] is a subproblem of problem 6, from which follows that the latter one is also strongly NP-hard.

On the other hand, if $\forall_{s_i \in S(Q)} m_i = 2$, it is known that all $S(Q)$ elements are present in the target sequence and at the same time they are also positive errors (which makes sense only in the case where multiplicity information is available), so the problem is reduced to the one without errors and can be solved in polynomial time.

3.7. Problem with positive errors and negative ones resulting from repetitions

The motivation for defining such a problem results from the fact that it is possible to set biochemical conditions of the hybridization experiment in such a way that the negative hybridization error rate may be considerably reduced. Such a setting usually results in increasing the positive error rate. Moreover, in many cases it may be impossible to guess if there are some l -tuple repetitions in the examined DNA sequence.

In this case the inclusion relations between spectra are as follows: $S^{(is)}(Q) \subseteq S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated in the following way:

Problem 7

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subseteq S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple appearing in it is an element of $S(Q)$ and Q' contains every $s_i \in S(Q)$ for which $m_i = 2$.

As one can notice, if $\forall_{s_i \in S(Q)} m_i = 1$ then the problem is reduced to the classical SBH problem with positive errors, hence, it is strongly NP-hard.

3.8. Problem with errors of arbitrary types

Here we have $S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated as follows:

Problem 8

INSTANCE: set $S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every sequence $s_i \in S(Q)$ appears in Q' once at the most if $m_i = 1$ and it appears in Q' at least once if $m_i = 2$. Moreover, this sequence can contain some l -tuples which are not in $S(Q)$.

Let us observe that in case where $\forall_{s_i \in S(Q)} m_i = 1$ we have the classical SBH problem with negative and positive errors without repetitions of the l -tuples. Since this problem contains the classical problems with only positive or only negative errors following from hybridization imperfection as subproblems, it is strongly NP-hard.

4. INFORMATION ABOUT MULTIPLICITY OF THE TYPE "ONE, TWO AND MANY"

4.1. An outline

In this case each oligonucleotide may appear in a multi-spectrum 0, 1, 2 or 3 times. It means that for any sequence

s_i , $m_i \in \{1, 2, 3\}$. The case where $m_i = 3$ means that s_i appears in the target sequence at least three times if there are no positive errors, or it appears at least two times and, in addition, there is one positive error (in the problems with positive errors it is not known which of the two cases takes place). If $m_i = 2$ it means that the target sequence contains exactly two occurrences of s_i in case without positive errors. If there are positive errors then $m_i = 2$ means that the target sequence contains two occurrences of s_i or one its occurrence and one positive error (again, these two cases are indistinguishable). If $m_i = 1$ then s_i appears exactly once in the target sequence in case without positive errors. If there may be positive errors in the hybridization data, $m_i = 1$ means that the target sequence contains exactly one occurrence of s_i or there are no s_i in the target sequence, but there is a positive error in the data (again, these situations are indistinguishable).

The multiplicity information considered in this section is much more precise than the one assumed in the previous section. However, here it is still assumed that only approximate multiplicity information can be deduced from the chip image. The reason for distinguishing between one and two occurrences of an l -tuple in the examined DNA sequence is that in the chip image it should be relatively easily to distinguish between light intensities corresponding to the amount of nucleic acids coming from one and two occurrences of an l -tuple in the target sequence. On the other hand, it may be difficult to distinguish the intensities corresponding to, for example, six and seven repetitions, since the amount of DNA hybridized to a given probe is not exactly proportional to the number of repetitions.

4.2. Problem without errors

Here, like in the analogous problem from the previous section, we have $S(Q) = S^{(is)}(Q) = S^{(m)}(Q) = S^{(im)}(Q)$, and the problem is defined as follows:

Problem 9

INSTANCE: set $S(Q) = S^{(is)}(Q)$, length n of sequence Q .

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$.

Obviously, this problem is identical to the classical problem without errors, so it can be solved in polynomial time.

Let us note that if the multiplicity information available is of the type "one, two and many" it is possible to formulate another problem without errors, where a limited number of repetitions is allowed:

Problem 10

INSTANCE: set $S(Q) = S^{(is)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$, where every $s_i \in S(Q)$ appears in Q' exactly m_i times.

Here, if at the most two occurrences of an l -tuple are present in the target sequence, the multispectrum contains full information about the l -tuple composition of the sequence. The relations between spectra are as follows: $S(Q) = S^{(is)}(Q) \subseteq S^{(m)}(Q) = S^{(im)}(Q)$.

This problem can be solved in polynomial time using a modification of a standard algorithm for finding a directed Eulerian trail (see [4] for a version for undirected graphs which can be easily adopted to directed ones). The graph corresponding to this problem is almost the same as Pevzner graph for the classical problem without errors [10], except that each arc is labeled by the value of corresponding m_i parameter. The modification of the standard algorithm is that after traversing an arc the value of its label is decreased by one and the arc is removed from the graph when the value becomes zero.

4.3. Problem with negative errors resulting from repetitions

In this case we have $S(Q) = S^{(is)}(Q) \subset S^{(m)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated as follows:

Problem 11

INSTANCE: set $S(Q) \subset S^{(im)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, 3\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$, where sequence $s_i \in S(Q)$ appears exactly once in Q' if $m_i = 1$, it appears exactly two times in Q' if $m_i = 2$ and it appears at least three times in Q' if $m_i = 3$.

Let us observe that if $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the one without repetitions. If $\forall_{s_i \in S(Q)} m_i = \{1, 2\}$ we have got problem 10. If $\forall_{s_i \in S(Q)} m_i = 2$ we have a variant of problem 10 where each spectrum element appears two times in the target sequence. All these subproblems are solvable in polynomial time. If $\forall_{s_i \in S(Q)} m_i = 3$ it is known that all spectrum elements have to appear in Q' at least three times.

4.4. Problem with negative errors resulting from hybridization

In this case the relations between spectra are as follows $S(Q) = S^{(m)}(Q) \subset S^{(is)}(Q) = S^{(im)}(Q)$. The problem is formulated in the following way:

Problem 12

INSTANCE: set $S(Q) \subset S^{(is)}(Q)$, length n of sequence Q .

ANSWER: sequence Q' of length n containing every element from $S(Q)$ exactly once and, in addition, containing some l -tuples not present in $S(Q)$.

This problem is identical to its classical counterpart (and to problem 3 so, it is strongly NP-hard.)

Let us observe that the following variant of problem 12 can be formulated:

Problem 13

INSTANCE: set $S(Q) \subset S^{(is)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$, where every $s_i \in S(Q)$ appears in Q' exactly m_i times. Moreover, Q' contains some l -tuples which are not in $S(Q)$.

Here $S(Q) \subset S^{(is)}(Q) \subset S^{(im)}(Q)$ and $S(Q) \subset S^{(m)}(Q) \subseteq S^{(im)}(Q)$. Let us observe that in case where $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to problem 12 so it is strongly NP-hard.

4.5. Problem with negative errors of arbitrary types

In this case $S(Q) \subseteq S^{(is)}(Q) \subseteq S^{(im)}(Q)$ and $S(Q) \subseteq S^{(m)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated as follows:

Problem 14

INSTANCE: set $S(Q) \subseteq S^{(im)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, 3\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all elements of $S(Q)$, where sequence $s_i \in S(Q)$ appears exactly once in Q' if $m_i = 1$, it appears exactly two times in Q' if $m_i = 2$ and it appears at least three times in Q' if $m_i = 3$. Moreover, this sequence can contain some l -tuples which are not elements of $S(Q)$.

In case where $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the classical problem with negative errors following from hybridization imperfectness, so it is strongly NP-hard.

Analogously as in the previous section the problem can be also formulated in a slightly different way:

Problem 15

INSTANCE: set $S(Q) \subset S^{(im)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, 3\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all elements of $S(Q)$, where sequence $s_i \in S(Q)$ appears exactly once in Q' if $m_i = 1$, it appears exactly two times in Q' if $m_i = 2$ and it appears at least three times in Q' if $m_i = 3$. Moreover, this sequence contains at least one l -tuple which is not an element of $S(Q)$.

This formulation corresponds to the situation where it is known that at least one negative error resulting from hybridization appears in the data. In this case $S(Q) \subset S(Q)^{(is)} \subseteq S^{(im)}(Q)$ and $S(Q) \subseteq S^{(m)}(Q) \subset S^{(im)}(Q)$. Obviously, this problem is strongly NP-hard.

4.6. Problem with positive errors

In this case $S^{(is)}(Q) \subset S(Q) \subseteq S^{(m)}(Q)$, $S^{(is)}(Q) = S^{(im)}(Q)$ and the problem is formulated as follows:

Problem 16

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subset S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple appearing in it is an element of $S(Q)$, Q' contains every $s_i \in S(Q)$ for which $m_i = 2$ and there are no repeated l -tuples in Q' .

Let us observe that this problem is identical to problem 6 and it contains as a subproblem the classical problem with positive errors in case where $\forall_{s_i \in S(Q)} m_i = 1$. Hence, the problem is strongly NP-hard.

On the other hand, the problem could be modified in such a way that it would be possible for some $s_i \in S(Q)$ to appear in Q' two times. But in such a case it would be, at least in some sense, a problem with positive errors and negative errors resulting from repetitions. The reason is that $m_i = 2$ would mean that either s_i appears two times in Q' or s_i appears once in Q' and the second occurrence of s_i in $S^{(m)}(Q)$ would be a positive error. In such a situation it would be not known whether there is a repetition of s_i in Q' or not. So, we would have an error resulting from repetition.

Let us observe that the following variant of the problem can be formulated:

Problem 17

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subset S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, 3\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple appearing in it is an element of $S(Q)$ and if for $s_i \in S(Q)$, $m_i = 2$ then Q' contains at least one occurrence of s_i and if $m_i = 3$ then Q' contains two occurrences of s_i .

Here, we have $S^{(is)}(Q) \subset S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$. In case where $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the classical problem with only positive errors, hence, it is strongly NP-hard.

4.7. Problem with positive errors and negative ones resulting from repetitions

In this case the inclusion relations between spectra are as follows: $S^{(is)}(Q) \subseteq S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated in the following way:

Problem 18

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subseteq S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, 3\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple appearing in it is an element of $S(Q)$ and if for $s_i \in S(Q)$, $m_i = 2$ then Q' contains at least one occurrence of s_i and if $m_i = 3$ then Q' contains at least two occurrences of s_i .

If $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the classical SBH problem with positive errors, hence, it is strongly NP-hard.

4.8. Problem with errors of arbitrary types

In this case $S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$ and the problem is formulated in the following way:

Problem 19

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

ANSWER: sequence Q' of length n such that every sequence $s_i \in S(Q)$ appears in Q' once at the most if $m_i = 1$, it appears once or two times in Q' if $m_i = 2$, and s_i appears in Q' at least two times if $m_i = 3$. Moreover, this sequence can contain some l -tuples which are not in $S(Q)$.

If $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the classical one with negative and positive errors without repetitions (like in case of problem 8) so, it is strongly NP-hard.

5. COMPLETE MULTIPLICITY INFORMATION**5.1. An outline**

Here we assume that full multiplicity information is available. This assumption is rather unrealistic from the viewpoint of current DNA chip technology, but it seems to be possible to distinguish more multiplicities than "one, two and many" considered in the previous section. Moreover, it is reasonable to expect that, as the chip technology evolves, the information possible to extract from the chip image will become more and more precise.

In case of multiplicity information whose availability is assumed in this section each l -tuple can appear in a multispectrum an arbitrary number of times. For every $s_i \in S(Q)$ parameter m_i is defined whose value is equal to a multiplicity of s_i in $S^{(m)}$. Let us also observe that in this case the problem with negative errors resulting from repetitions does not make sense, since a multispectrum contains full information about all occurrences of every l -tuple in the target sequence (if there are no hybridization errors, that is).

5.2. Problem without errors

In this case, like in the previous sections we have $S(Q) = S^{(is)}(Q) = S^{(m)}(Q) = S^{(im)}(Q)$ and the problem can be formulated in the following way:

Problem 20

INSTANCE: set $S(Q) = S^{(is)}(Q)$, length n of sequence Q .

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$.

Obviously, this problem is identical to the classical problem without errors and can be solved in polynomial time. However, if the full multiplicity information is available, the problem with an arbitrary number of repetitions is also the problem without errors and can be defined as follows:

Problem 21

INSTANCE: set $S(Q) = S^{(is)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, \dots, n-l+1\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$, where every $s_i \in S(Q)$ appears in Q' exactly m_i times.

Here we have the relations $S(Q) = S^{(is)}(Q) \subseteq S^{(m)}(Q) = S^{(im)}(Q)$. It is easy to see that problem 21 contains problem 20 as a subproblem in the case where $\forall_{s_i \in S(Q)} m_i = 1$. This more general problem without errors can also be solved in polynomial time using the modified algorithm for finding a directed Eulerian trail proposed to solve problem 10. In this case each arc in the Pevzner graph is labeled by the value of corresponding m_i parameter, which in this case can be in the range $1, 2, \dots, n-l+1$.

5.3. Problem with negative errors resulting from repetitions

As has been mentioned above this problem does not make sense when the full multiplicity information is available.

5.4. Problem with negative errors resulting from hybridization

The inclusion relations between spectra in this case are as follows: $S(Q) = S^{(m)}(Q) \subset S^{(is)}(Q) = S^{(im)}(Q)$ and the problem is defined in the following way:

Problem 22

INSTANCE: set $S(Q) \subset S^{(is)}(Q)$, length n of sequence Q .

ANSWER: sequence Q' of length n containing every element from $S(Q)$ exactly once and, in addition, containing some l -tuples not present in $S(Q)$.

The problem is identical to its classical counterpart, so it is strongly NP-hard. Here, like in the previous section another variant of the problem can be defined:

Problem 23

INSTANCE: set $S(Q) \subset S^{(is)}(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, \dots, n-l+1\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n containing all and only those l -tuples which are elements of $S(Q)$, where every $s_i \in S(Q)$ appears in Q' exactly m_i times. Moreover, Q' contains some l -tuples which are not in $S(Q)$.

In this case $S(Q) \subset S^{(is)}(Q) \subset S^{(im)}(Q)$ and $S(Q) \subset S^{(m)}(Q) \subset S^{(im)}(Q)$. When $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to problem 22, hence, it is strongly NP-hard.

5.5. Problem with negative errors of arbitrary types

Let us observe that in case where the complete multiplicity information is available the problem is identical to the one with negative errors resulting from hybridization, since there cannot be any errors resulting from l -tuple repetitions.

5.6. Problem with positive errors

In this case $S^{(is)}(Q) \subset S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) = S^{(im)}(Q)$ and the problem is formulated in the following way:

Problem 24

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subset S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple appearing in it is an element of $S(Q)$ and Q' contains every $s_i \in S(Q)$ for which $m_i = 2$.

It is easy to note that in case where $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the classical SBH problem with positive errors, hence, it is strongly NP-hard.

Let us also observe that in case of full multiplicity information available a variant of the above problem can be formulated, where some l -tuple repetitions are allowed:

Problem 25

INSTANCE: set $S(Q)$ such that $S^{(is)}(Q) \subset S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, \dots, n-l+1\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every l -tuple appearing in it is an element of $S(Q)$ and every $s_i \in S(Q)$ has to appear in Q' $m_i - 1$ or m_i times.

Here we have $S^{(is)}(Q) \subset S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$. Obviously, this problem reduces to problem 24 if $\forall_{s_i \in S(Q)} m_i = \{1, 2\}$ which results in its strong NP-hardness.

Problem with positive errors and negative ones resulting from repetitions

If the full multiplicity information is available this problem does not make sense.

5.7. Problem with errors of arbitrary types

In this case the inclusion relations are as follows: $S(Q) \subseteq S^{(m)}(Q)$ and $S^{(is)}(Q) \subseteq S^{(im)}(Q)$. The problem is formulated in the following way:

Problem 26

INSTANCE: set $S(Q)$, length n of sequence Q , parameter $m_i \in \{1, 2, \dots, n-l+1\}$ for each $s_i \in S(Q)$.

ANSWER: sequence Q' of length n such that every sequence $s_i \in S(Q)$ appears in Q' $m_i - 1$ or m_i times. Moreover, this sequence can contain some l -tuples which are not in $S(Q)$.

If $\forall_{s_i \in S(Q)} m_i = 1$ the problem reduces to the classical one with negative and positive errors without repetitions (similarly like in the case of problems 8 and 19) so, it is strongly NP-hard.

CONCLUSIONS

In this paper the classical SBH problems with and without errors have been reformulated in such a way that some information about multiplicity of spectrum elements is taken into account. To the best of our knowledge problems of this type have not been considered in the literature so far. In classical approaches to SBH the multiplicity information is assumed not to be available,

which was caused by the DNA chip technology constraints. However, the rapid development of this technology leads to an improvement of the quality of results obtained in the experiments based on DNA chips. In particular, at least an approximate relative amount of nucleic acids hybridized to chip probes can be estimated, which is a common practice in gene expression analysis (where also DNA chips are used).

The computational complexity of the new problems is, in most cases, analogous to the complexity of their classical counterparts. On the other hand, the additional information available should have a positive impact on the quality of results provided by the algorithms solving the new problems (in the sense of a similarity of these results to the real target DNA sequences). Such algorithms are the subject of our future research in this area. As the computational complexity results suggest, they should be heuristic or some efficient in average case enumeration methods, like branch and bound algorithms.

References

- [1] J. Błażewicz, P. Formanowicz, M. Kasprzak, W. T. Markiewicz and J. Węglarz, *DNA sequencing with positive and negative errors*, Journal of Computational Biology **6**, 113-123 (1999).
- [2] J. Błażewicz, P. Formanowicz, M. Kasprzak, P. Schuurman and G. J. Woeginger, *DNA sequencing, Eulerian graphs, and the exact perfect matching problem*, Lecture Notes in Computer Science **2573**, 13-24 (2002).
- [3] J. Błażewicz and M. Kasprzak, *Complexity of DNA sequencing by hybridization*, Theoretical Computer Science **290**, 1459-1473 (2003).
- [4] V. Bryant, *Aspects of Combinatorics. A Wide-Range Introduction*, Cambridge University Press 1993.
- [5] R. Drmanac, I. Labat, I. Brukner and R. Crkvenjakov, *Sequencing of megabase plus DNA by hybridization: theory and the method*, Genomics **4**, 114-128 (1989).
- [6] S. P. A. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. Lu and D. Solas, *Light-directed, spatially addressable parallel chemical synthesis*, Science **251**, 767-773 (1991).
- [7] J. Gallant, D. Maier and J. A. Storer, *On finding minimal length superstrings*, Journal of Computer and System Sciences **20**, 50-58 (1980).
- [8] K. R. Khrapko, Y. P. Lysov, A. A. Khorlin, V. V. Shik, V. L. Florentiev and A. D. Mirzabekov, *An oligonucleotide approach to DNA sequencing*, FEBS Letters **256**, 118-122 (1989).
- [9] A. C. Pease, D. Solas, E. Sullivan, M. Cronin, C. Holmes and S. Fodor, *Light-generated oligonucleotide arrays for rapid DNA sequence analysis*, Proceedings of the National Academy of Sciences of the USA **91**, 5022-5026 (1994).
- [10] P. A. Pevzner, *l-tuple DNA sequencing: computer analysis*, Journal of Biomolecular Structure and Dynamics **7**, 63-73 (1989).
- [11] P. A. Pevzner, *Computational molecular biology. An algorithmic approach*, The MIT Press, Cambridge, Massachusetts 2000.
- [12] M. Schena, *Microarray Analysis*, Wiley-Liss, Hoboken, New Jersey 2003.
- [13] E. M. Southern, U. Maskos and J. K. Elder, *Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models*, Genomics **13**, 1008-1017 (1992).
- [14] J. D. Watson and F. H. C. Crick, *Genetic implications of the structure of deoxyribonucleic acid*, Nature **171**, 964-967 (1953).