BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 59, No. 1, 2011
DOI: 10.2478/v10175-011-0015-0

A greedy algorithm for the DNA sequencing by hybridization with positive and negative errors and information about repetitions

K. KWARCIAK[1]* and P. FORMANOWICZ[1,2]

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

Abstract. In this paper a greedy algorithm for some variants of the sequencing by hybridization method is presented. In the standard version of the method information about repetitions is not available. In the paper it is assumed that a partial information of this type is a part of the problem instance. Here two simple but realistic models of this information are taken into consideration. The first one assumes it is known if a given element of a spectrum appears in the target sequence once or more than once. The second model uses the knowledge if a given element of a spectrum occurs in the analyzed sequence once, twice or at least three times. The proposed greedy algorithm solves the variant of the problem with positive and negative errors. Results of a computational experiment are reported which, among others, confirm that the additional information leads to the improvement of the obtained solutions. They also show that the more precise model of information increases the quality of reconstructed sequences.

Key words: DNA sequencing, l-mer multiplicity, greedy algorithm, combinatorial problems.

1. Introduction

The DNA sequencing is one of the most important problems in computational and molecular biology. The goal is to determine the sequence of nucleotides DNA consists of. There are many methods to solve it. One of them is sequencing by hybridization (SBH) [1, 2]. This approach is comprised of two stages: a biochemical one and a computational one. In the first phase a set of l-long oligonucleotides (called also l-mers) composing the original DNA sequence is determined. In the second stage a combinatorial problem must be solved. The information about l-mers composition is used to reconstruct the target sequence.

The biochemical stage utilizes a DNA chip [3, 4] containing a full l-long oligonucleotide library. It is a kind of matrix divided into cells. Each cell contains a number of identical l-mers representing one of the oligonucleotides from the library. When such a chip is put into a solution of a number of copies of the single stranded target DNA sequence, some fragments of the examined DNA hybridize to complementary l-mers on the chip. If the copies of the analyzed sequence are radioactively or fluoroscently labelled then one can obtain an image of the DNA chip corresponding to the set of l-mers composing the target DNA. This set of l-long oligonucleotides is called spectrum.

In the ideal case the biochemical experiment provides full and proper information about l-mers present in the target sequence. However, during the experiment some errors may occur. There are two types of errors: positive ones and negative ones. The positive error occurs when the target sequence hybridizes to an oligonucleotide on the chip which is not perfectly complementary to it. As a result a spectrum containing additional l-mers that are not a part of the target sequence is obtained. An opposite situation may also take place. The target sequence do not need to hybridize to the complementary l-mer on the chip. In this case some information about l-mer composition of the examined DNA sequence is missed. The missing l-mers are negative errors. Another source of negative errors are repetitions in the target sequence of length equal or greater than l. Since in the classical SBH method spectrum is a set, not a multiset, the information about repetitions is lost.

In the classical SBH approach the output of the biochemical phase is the binary information about oligonucleotide presence in the target sequence (i.e. a given l-mer is or is not present in the target DNA). If the examined DNA sequence is repetitive then negative errors resulting from repetitions occur. However, the development of the DNA chip technology enables to take into account an information about an intensity of the chip signals. This intensity can be, at least to some extent, correlated with the multiplicity of a given l-mer in the target sequence. Unfortunately, the precision of this information decreases when the number of occurrences of an l-mer in the sequence increases. It is possible to easily distinguish the signal coming from one occurrence and many occurrences, but differentiating the shining of, for example, seven and eight occurrences would be very hard (or even impossible). Nevertheless, even partial information about repetitions can be very useful [5, 6].

In this paper two realistic models of the multiplicity information are considered. In the first of them, called “one and many”, it is assumed that the information coming from the hy-

*e-mail: kamil.kwarciak@cs.put.poznan.pl bridization experiment allows for distinguishing between one and more than one occurrences of any l-mer in the analyzed DNA sequence. According to the second model, called “one, two and many”, it is possible to distinguish between one, two and more than two occurrences of an l-mer. The assumptions are well justified by the current DNA chip technology.

2. Problem definition

In this paper we consider DNA sequencing by hybridization with positive and negative errors. The source of an error (repetitious sequence or imperfect experiment hybridization) is irrelevant. In order to precisely define various variants of SBH problems with additional information it is necessary to introduce four types of spectra [5].

Let $S(Q)$ denote a spectrum of sequence Q and let $S^{(is)}(Q)$ denote an ideal spectrum of this sequence. The ideal spectrum consist of all and only those types of l-mers but not all of these l-mers, which compose the target sequence Q. All of these l-mers compose an ideal multispectrum of sequence Q, which will be denoted by $S^{(im)}(Q)$. Note that the number of occurrences of any l-mer in the ideal multispectrum is equal to the number of repetitions of this l-mer in sequence Q. Let $S^{(m)}(Q)$ be a multispectrum of sequence Q. This spectrum may not contain full information about oligonucleotide repetitions and in addition it can contain some positive errors (l-mers which are not a part of the target sequence) or negative errors (some l-mers can be missed). For every sequence (l-mer) $s_i \in S(Q)$ let m_i be the number of occurrences of s_i in $S^{(m)}(Q)$.

Assuming there does not exist the additional multiplicity information the combinatorial problem may be defined as follows (cf. [7]):

Problem 1. Lack of the multiplicity information
Instance: set $S(Q)$, length n of sequence Q
Answer: sequence Q' of length n containing the maximum number of elements of $S(Q)$. Moreover, Q' can contain some l-mers which are not elements of $S(Q)$.

Let us assume that it is possible to obtain from the biochemical experiment the partial multiplicity information of type “one and many”. Then the sequencing problem may be stated as follows [5]:

Problem 2. Multiplicity information of the type “one and many”
Instance: set $S(Q)$, length n of sequence Q, parameter $m_i \in \{1, 2\}$ for every $s_i \in S(Q)$
Answer: sequence Q' of length 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 $S(Q)$.

Finally, let us assume there exists approximate multiplicity information of type “one, two and many”. In this case the sequencing problem may be defined is as follows [5]:

Problem 3. Multiplicity information of the type “one, two and many”
Instance: set $S(Q)$, length n of sequence Q, parameter $m_i \in \{1, 2, 3\}$ for every $s_i \in S(Q)$
Answer: sequence Q' of length 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 $S(Q)$.

It is possible to transform any problem stated above into a variant of the travelling salesman problem (TSP). The classical TSP has a directed or undirected graph as an input. Each arc or edge has assigned a weight (cost). The goal is to find the minimal cost Hamiltonian cycle (cycle which contains all vertices).

One should apply the following modification on the travelling salesman problem to obtain a problem which is equivalent to the sequencing by hybridization. Firstly, the goal should be to find a path not a cycle. The cost of the path is constrained by the length n of the target sequence. Moreover, the first vertex on the path should correspond to the first l-mer in the sequence. Finally, acceptable solutions need not to visit all vertices and some vertices may be visited more than once (i.e. according to the value of m_i).

If the vertices in the directed input graph represent oligonucleotides from the spectrum then the travelling salesman problem customized as mentioned above corresponds to the sequencing by hybridization problem. The cost of an arc is related to how the two l-mers overlap each other. The cost of the arc is equal to the oligonucleotides length l minus the length of the common fragment. For example, the cost of an arc from the vertex representing oligonucleotide CGCTTA to the vertex representing GCTTAT is equal to 1 because they have an common subsequence GCTTA of length equal to 5. Note that any two oligonucleotides may have more than one common subsequence, so the input graph is in fact a multigraph, where each arc represents one possible l-mers' overlapping.

The traveling salesman problem is strongly NP-hard, so there does not exist a polynomial time algorithm to solve it (assuming $P \neq NP$). The sequencing by hybridization is also an intractable problem. The strong NP-hardness of the classical SBH approach has been proved in [7] and of the above variants with the additional information about oligonucleotides multiplicity in [5] and [6].

3. Algorithm

The time of solving optimally an intractable problem increases exponentially compared to the instance size and exact algorithms for such problems have limited application in practice. This justifies development of heuristic algorithms which enable to obtain an approximate solution in polynomial time. This is a compromise between the quality of a solution and the time needed to obtain it.

There exist many types of heuristic algorithms. One of them is a greedy algorithm [8]. This approach constructs the solution iteratively. In each step the current partial solution is extended on the basis of the locally optimal choice. There is no guarantee the optimal solution will be found but it reduces the time of computation.

The computational complexity of the sequencing by hybridization was a motivation to develop a greedy algorithm which could be useful in practice to obtain an approximate sequence of nucleotides in the target sequence of DNA. The greedy algorithm presented in this paper is a heuristic based on the greedy algorithm defined in [9] for the SBH problem with positive and negative errors. The original algorithm has been extended to take into consideration the information about l-mers multiplicity.

The algorithm starts at an initial oligonucleotide and adds successive l-mers. This process ends if adding another l-mer will violate the maximum length constraint. The obtained sequence cannot be longer than the target sequence of length n.

To make the local choice, the cost of appending all of the remaining (not used yet) l-mers to the current vertex is verified and the best option is chosen. The criterion of a new nucleotide selection is the cost of overlapping of a current l-mer and the new one plus the smallest cost of overlapping of the new one and one of its possible successors.

For each l-mer (s_i) the number of occurrences in the current solution is stored. If this number reaches the maximum value according to the parameter m_i then the oligonucleotide is not considered any more as a possible successor of the last vertex in the current solution.

The above algorithm uses as a part of the input data information about the first l-mer of the analyzed sequence. The assumption that the first l-mer is known is well justified since the PCR commonly used to obtain a large number of copies of the target requires the knowledge of the first oligonucleotide.

4. Computational experiment

The algorithm described in the previous section has been tested on the real DNA sequences obtained from GenBank. The human DNA has been used. The sequences have been divided into fragments of length 109, 209, 309, 409 and 509 nucleotides. For each sequence length an independent test set with 1, 2, 3, 4 and 5 percent of negative errors coming from repetitions has been prepared. Each test set contained 50 various sequences.

Spectra have been prepared for oligonucleotides of length equal to 10. Hybridization errors in biochemical experiment have been simulated by deletion of 15% random l-mers and insertion of 15% random l-mers.

The similarity of an obtained sequence and the original one has been measured using the Smith-Waterman algorithm for calculating the local alignment [10]. The following values of parameters have been used: match (the same nucleotides) $+1$, mismatch (different nucleotides) -1, insertion (a nucleotide in one sequence and the blank position in the second sequence) -1. A solution, which is the same as the target, obtains the highest score (similarity) equals to the number of nucleotides in the sequence.

The average local alignment of the obtained results has been shown in Table 1 (lack of multiplicity information), Table 2 (multiplicity information of the type "one and many") and Table 3 (multiplicity information of the type "one, two and many"). The results are presented as a function of the sequence length and the percent of negative errors resulting from repetitions. Each value is an average computed on the basis of 50 sequences and for each sequence 20 independent tests have been performed. Hence, each entry represents 1000 solutions. There are two values in each cell of the tables. The first one is the score of the local alignment calculated using the Smith-Waterman algorithm. The second one is a percent value related to the maximum similarity score (100% if the obtained solution is identical to the target sequence).

Table 1
Similarity – lack of the multiplicity information

Repetitions count	Spectrum size				
	100	200	300	400	500
1%	100.64	180.23	244.57	293.49	317.59
	92.33%	86.23%	79.15%	71.76%	62.39%
2%	100.83	172.93	238.01	278.68	320.33
	92.50%	82.74%	77.03%	68.14%	62.93%
3%	96.32	171.05	223.75	261.88	296.96
	88.37%	81.84%	72.41%	64.03%	58.34%
4%	92.17	161.12	220.86	257.80	290.33
	84.56%	77.09%	71.47%	63.03%	57.04%
5%	91.91	151.46	211.36	248.95	275.94
	84.32%	72.47%	68.40%	60.87%	54.21%

Table 2
Similarity – multiplicity information of the type "one, and many"

Repetitions count	Spectrum size				
	100	200	300	400	500
1%	102.08	183.24	241.30	297.55	319.09
	93.65%	87.68%	78.09%	72.75%	62.69%
2%	102.45	179.24	240.28	277.08	325.59
	93.99%	85.76%	77.76%	67.75%	63.97%
3%	98.67	177.51	231.36	271.35	302.38
	90.52%	84.93%	74.87%	66.34%	59.41%
4%	96.51	166.63	230.29	266.18	297.16
	88.54%	79.73%	74.53%	65.08%	58.38%
5%	95.15	159.18	220.96	258.17	295.79
	87.29%	76.16%	71.51%	63.12%	58.11%

Table 3
Similarity – multiplicity information of the type "one, two and many"

Repetitions count	Spectrum size				
	100	200	300	400	500
1%	102.04	183.95	243.9	298.98	311.94
	93.61%	88.01%	78.93%	73.10%	61.28%
2%	104.3	177.54	242.85	274.25	326.32
	95.69%	84.95%	78.59%	67.05%	64.11%
3%	99.79	177.51	226.73	266.34	306.15
	91.55%	84.93%	73.38%	65.12%	60.15%
4%	97.65	170.26	229.5	264.19	295.29
	89.59%	81.46%	74.27%	64.59%	58.01%
5%	96.26	161.24	218.29	261.76	291.6
	88.31%	77.15%	70.64%	64.00%	57.29%

The results have shown that taking into consideration the additional information about repetitions increases the quality (similarity) of the obtained solutions. Additionally, one may suppose that the negative errors resulting from the repetitions have much greater impact on the results quality than the positive and negative errors resulting from biochemical experiment (at least in the case of the greedy algorithm).

However, increasing the precision of the multiplicity information does not clearly imply further quality improvement as one may expect. It may be caused by a low repetitions count for a given l-mer in the analyzed sequences, i.e. if the average multiplicity of the repetitious l-mers in the tested sequences is close to 2. An supplementary experiment has been developed as follows to verify this hypothesis.

An additional test set has been prepared. It contained real human DNA sequences of length from 109 to 509 nucleotides with 1, 2, 3, 4 and 5 repetitions of exactly one l-mer. The length of oligonucleotides in spectra was also 10. Hybridization errors were simulated too. The number of deletions and the number of insertions were the same as in the main experiment, i.e. 15%. There were 25 different sequences of each type (lenght & repetition count). Every sequence was tested 40 times, so results represent also 1000 solutions.

The more precise model of multiplicity information “one, two and many” resulted in better alignment score only for shorter sequences (see results in Table 4). This may be caused by significantly higher hybridization errors rate in longer sequences. The maximum number of repetitions was 5 while for a sequence of length 509 the number of insertions was 75 and the number of deletions was also 75.

Table 4
Using the model “one, two and many” – the influence on the alignment score in comparison to the model “one and many” (insertions: 15%, deletions: 15%)

Repetitions count	Spectrum size				
	100	200	300	400	500
1	−0.08%	1.48%	0.36%	−0.56%	−0.43%
2	1.85%	−0.06%	0.04%	−0.27%	0.28%
3	1.42%	0.98%	−0.19%	0.95%	−0.46%
4	0.96%	0.57%	−0.52%	−0.01%	−0.24%
5	0.36%	0.50%	0.16%	0.21%	0.50%

Table 5
Using the model “one, two and many” - the influence on the alignment score in comparison to the model “one and many” (insertions: 5%, deletions: 5%)

Repetitions count	Spectrum size				
	100	200	300	400	500
1	0.21%	−0.44%	0.39%	−0.16%	−0.90%
2	1.07%	0.44%	−0.99%	0.69%	0.81%
3	1.21%	0.61%	0.07%	1.03%	1.72%
4	1.61%	0.54%	0.06%	−0.01%	−0.17%
5	1.13%	0.58%	0.46%	0.86%	0.68%

The insertions count and the deletions count have been set to 5% and the experiment has been run again. This time the results have shown that the model “one, two and many” leads to the improvement of obtained solutions for longer sequences too (see Table 5). However, it is hard to observe the superiority of any multiplicity information model if the analyzed sequence cointains only 1 repetition. In this case both models are indistinguishable.

5. Conclusions

In this paper has been shown how the partial information about repetitions in an analyzed sequence influences the resequencing by hybridization results. The two simplest, but realistic, models of such information have been taken into consideration. The first one uses the knowledge if a given l-mer occurs in the analyzed sequence once or more than once. The second model contains the information if a given oligonucleotide occurs in the target sequence once, twice or at least three times. This additional information is not very precise, but the results have shown that it leads to the quality improvement of the obtained solutions. Note that the current DNA chip technology has some constraints and cannot provide the precise number of repetitions of a given l-mer. Currently, considering more accurate models of multiplicity information does not have practical applications.

The sequencing by hybridization is an intractable problem (NP-hard in the strong sense) which justifies the implementation of the greedy algorithm. It was used to verify the impact of the partial information about the multiplicity on the solutions quality. However, the greedy heuristic may be used to obtain quickly (i.e. in polynomial time) an approximate sequence of nucleotides in the analyzed DNA sequence. Moreover, it can provide an initial solution for more advanced heuristics, e.g. tabu search. It is the subject of the current authors’ research in this area.

Acknowledgments. This research has been partially supported from the Polish Ministry of Science and Higher Education grant No N N519 314635. The preliminary version of this paper appeared at the 17th KKAPD (State Conference of Discrete Processes Automatization) 2010, Zakopane, Poland.

REFERENCES

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

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

[3] A.C. Pease, D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S.P. Fodor, “Light-generated oligonucleotide arrays for rapid DNA sequence analysis” *Proc. National Academy of Science USA* 91, 5022–5026 (1994).

[4] P.A. Pevzner, *Computational Molecular Biology. An Algorithmic Approach*, The MIT Press, Cambridge, 2000.

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

[6] P. Formanowicz, *Selected Combinatorial Aspects of Biological Sequence Analysis*, Poznań University of Technology Publishing House, Poznań, 2005.

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

[8] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*, Scientific-Technical Publishing House, Warszawa, 2005.

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

[10] T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences", *J. Molecular Biology* 147, 195-–197 (1981).