CMST 10(1) 39-46 (2004) DOI:10.12921/cmst.2004.10.01.39-46

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **10**, 39-47 (2004)

ON THE LINK BETWEEN DNA SEQUENCING AND GRAPH THEORY

MARTA KASPRZAK[1,2]

[1] *Institute of Computing Science, Poznań University of Technology,*
Piotrowo 3A, 60-965 Poznań, Poland

[2] *Institute of Bioorganic Chemistry, Polish Academy of Sciences,*
Noskowskiego 12, 61-704Poznań, Poland

marta@cs.put.poznan.pl

(Rec. 20 November 2003)

Abstract: The methods cited in this paper solve the combinatorial part of DNA sequencing by hybridization, basing on known approaches from graph theory. It is assumed here, that the length of oligonucleotides used in the hybridization experiment is constant within a library.

1. INTRODUCTION

The *DNA sequencing by hybridization,* one of important problems from the computational molecular biology domain, consists in determining a sequence of nucleotides of an unknown DNA fragment [21, 18, 14]. Its input data come from a biochemical *hybridization experiment,* and they can be viewed as a set (called *spectrum*) of words (*oligonucleotides*) over the alphabet {A, C, G, T}, being short subsequences of the studied DNA fragment. The aim is to reconstruct the original DNA sequence of a known length n on the basis of these overlapping words. The spectrum may contain *positive errors,* i.e. oligonucleotides present in the spectrum but absent in the original sequence, and *negative errors,* i.e. oligonucleotides not present in the spectrum, but possible to distinguish in the original sequence. Since the spectrum is a set, repetitions of oligonucleotides in the sequence are also treated as negative errors. The spectrum without any errors is called the *ideal* one.

In the standard approach to the DNA sequencing the oligonucleotide library used in the hybridization experiment contains all possible oligonucleotides of a given constant length. The spectrum being output of the experiment is a subset S of the library, i.e. the set of words of equal length l composing the original sequence. For the standard DNA sequencing, the computational complexity of several variants of the problem is already known. The variant with no errors in the spectrum is polynomially solvable [13], while the variants assuming presence of errors in the data (negative ones, positive ones, or both) are all strongly NP-hard [6]. In this paper, several methods for the DNA sequencing problem with constant-length oliganucleotide library are drawn, the ones basing on approaches from graph theory.

2. METHODS

The first algorithm reconstructing the original sequence on the base of a spectrum was proposed in [2], The spectrum used there does not contain any errors. This branch-and-cut algorithm builds a search tree, where spectrum elements correspond to nodes and two nodes are connected by an arc if last l - 1 letters of the predecessor cover first l - 1 letters of the successor. The element once included into the current path cannot be visited the second time. As the root this oligonucleotide is taken, which begins the original sequence. If we do not know it, the algorithm must construct |S| search trees differing in their roots. The solution is the path from the root to a leaf, which contains all spectrum elements (Example 1).

The second approach to the DNA sequencing problem [12] refers to a well-known problem from graph theory. In a directed graph, built on the base of an ideal spectrum, the Hamiltonian path is looked for. Each vertex in the graph corresponds to other element of the spectrum. Two vertices u and v are connected by the arc (u, v) if last l - 1 letters of the label (oligonucleotide) of u cover first l- 1 letters of the label of v (Example 1).

In [7] an approach similar to the one from [2] was presented, but it avoids excess steps. Subpaths between points of branching in the tree are combined into words of length l or greater. The search tree is built on the base of these words, and two nodes are connected if they overlap on l - 1 letters. Similarly, the solution is the path from the root to a leaf, containing all oligonucleotides from the spectrum (Example 1).

Example 1

Let us assume, that for the original sequence ACTCTGG, the hybridization experiment has been performed without errors and the ideal spectrum has been generated: S = {ACT, CTC, CTG, TCT, TGG}. We see, that |S| = n - l + 1, where n = 7 and l = 3. Assume, that we know the first oligonucleotide of the original sequence. The method from [2] will search the tree from Fig. 1. The lower path traverses through all elements of the spectrum, so it is the solution for the problem. The original sequence can be reconstructed by reading the labels of the vertices from the root to the leaf.

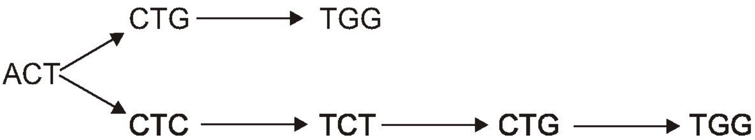
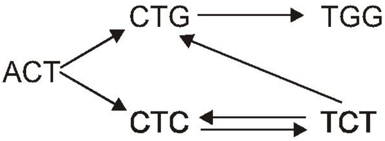


Fig. 1. A search tree from the method of Bains and Smith [2]

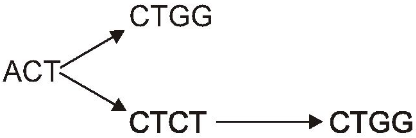
The graph from the method from [12], built on the base of the same spectrum, is presented in Fig. 2. In the graph exactly one Hamiltonian path exists, and it corresponds to a unique sequence of length n possible to build from the spectrum.

Fig. 2, A graph from the method of Lysov *et al.* [12]



The method from [7] search a similar tree as in Fig. 1, but subpaths between branching nodes are combined into vertices (Fig. 3). Here, the solution is path (ACT, CTCT, CTGG) corresponding to sequence ACTCTGG.

Fig. 3. A search tree from the method of Drmanac *et al.* [7]



All the above approaches accept as the input data only the ideal spectrum, nevertheless their computational complexity is exponential in time. The first and only polynomial-time algorithm solving the case of errorless, constant-length spectrum was presented in [13], what proved the membership of this variant of the DNA sequencing to the class of easy solvable problems. In the algorithm, the Eulerian path is looked for in a directed graph, and now elements of the spectrum are associated with arcs in the graph. An arc labeled by oligonucleotide *o* of length *l* leaves the vertex labeled by first *l* - 1 letters of *o* and enters the vertex labeled by last *l* - 1 letters of o (Example 2). Thus, the number of vertices is the same as the number of different *l* -1 letters substrings in the spectrum.

This interesting transformation of the graph, in which the Hamiltonian path is looked for, to the graph, in which the Eulerian path is searched, i.e. changing the computational complexity of the problem, was not commented by Pevzner. The class of graphs, for which such transformation is possible *(labeled graphs),* was widely examined in [5]. The graphs built on the base of the spectrum, called *DNA graphs,* belong to this class. The graph constructed by the method of Lysov *et al.* is the line graph of the graph by Pevzner for the same spectrum. For that pair of directed graphs the two problems of looking for the Hamiltonian and Eulerian paths are equivalent.

In [13] also a variant of the method for spectrum with negative errors was proposed. It has a polynomial-time complexity but it not always finds a solution [4], thus it may be treated only as a heuristic. Every negative error in the spectrum causes the lack of one arc in the graph. The method looks for these missing arcs, transforming the problem to the one of searching for a flow in a network based on a bipartite graph $K_{m,n}$. The bipartite graph is constructed from the vertices of the base graph having different numbers of incoming and outgoing arcs. The vertex for which this difference is greater than 1 is multiplied in the bipartite graph. Arcs in this graph are drawn from the vertices having a greater indegree in the base graph to the ones of a greater outdegree. With every arc a cost is assigned, equal to the minimal shift of vertex labels in their overlap. Therefore, the arc between vertices representing labels of themaximum possible overlap (i.e. on l - 2 letters) has the cost equal to 1, while the arc between vertices of labels having no common part has its cost equal to l - 1. The latter arc corresponds in the base graph to a path between these two vertices, composed of l - 1 arcs, what results in addition of l- 1 new elements to the spectrum. To this bipartite graph the source s and the sink t must be added, and the arcs from s to all vertices of the bipartite graph with zero indegree, and from all vertices with zero outdegree to t. All arcs in the network have its capacity equal to 1. In this network we look for the flow of value m - 1 and of the minimum cost. If the cost appears to be equal n-l+1-|S|, the base graph will be completed by arcs (paths) composing the flow. Then, if the new base graph is connected, we can search in it for the Eulerian path, corresponding to the original DNA sequence (Example 2).

Example 2

For the ideal spectrum from Example 1: {ACT, CTC, CTG, TCT, TGG}, the graph constructed in the Pevzner's method is shown on Fig. 4.

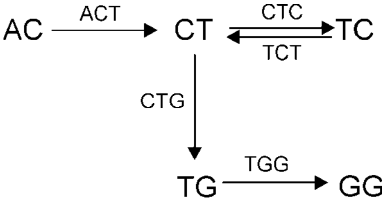


Fig. 4. A graph from the method of Pevzner [13]

The Eulerian path in this graph corresponds to the original sequence ACTCTGG. Let us assume now, that as a result of an experimental error we lose one element: CTG. In order to make the base graph an Eulerian one, we should find a flow of cost equal to n-l + 1 -|S|=1 in a network constructed on the base of this graph (Fig. 5).

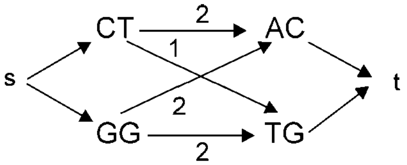


Fig. 5. A network from the method of Pevzner [13]

After source s we have the vertices of a greater indegree in the base graph (see Fig. 4 after removing arc CTG), before sink t we have vertices of a greater outdegree. Costs of arcs are equal to shifts of vertex labels in their maximum possible overlap. The arcs of a cost greater than 1 represent in fact sequences of arcs and vertices not necessarily present in the base graph. Capacities of all arcs are equal to 1. There exists one flow of value m - 1 = 1 and of cost 1, it contains arc (CT, TG) corresponding to the missing oligonucleotide CTG. After the addition of this arc to the base graph, we may there look for the Eulerian path.

The next method [11] accepts the data with negative errors and also with positive ones matching a restricted scheme. It is assumed there, that a positive error comes only from a misreading an existing oligonucleotide, which then cannot appear in the spectrum in its correct form, i.e. it is always associated with a negative error. Moreover, the misreading can concern only terminal bases of the oligonucleotide. The method requires the additional information about percentage of errors present in the spectrum, both negative and positive. The method was derived from the Pevzner's method for negative errors (see above), thus it copies faults of that approach, especially a great probability of obtaining a disconnected solution. A bipartite digraph representing the whole spectrum is constructed, and a probability of the existence of an oligonucleotide within the original sequence is associated with every arc. Next, the Hungarian method searching for the best assignment in this graph is called, and the result is used to update the probability values of oligonucleotides. These steps are repeated until the successive probability values converge.

The algorithm from [10] accepts any negative errors, but its memory usage grows exponentially. In a directed graph built on 4^{l-1} vertices, where arcs correspond to real and hypothetical spectrum elements, a path is looked for by a stochastic algorithm. An additional knowledge is required to assign to arcs a rough number of occurrences of oligonucleotides within the original sequence.

The first algorithm accepting any negative and positive errors in a spectrum was proposed in [3]. The problem of finding the original DNA sequence was formulated there as a variant of Selective Traveling Salesman Problem. The complete directed graph is constructed with oligonucleotides from the spectrum associated with vertices. To every vertex a profit equal to 1 is assigned, and to every arc a cost equal to the shift of oligonucleotides in their best possible overlap. In this graph, a simple path of a maximum total profit and of a total cost not greater than $n - l$ is looked for, what is equivalent to a sequence of a length not greater than n, composed of as many spectrum elements as possible (see Example 3).

Example 3

Let our spectrum with negative and positive errors, for sequence ACTCTGG, be {ACT, CTC, GCC, TCT, TGG} (i.e. the negative error is CTG and the positive one is GCC). The complete graph constructed in the method from [3] would have the costs of arcs as in Table 1. All profits of vertices would equal 1. Thus, the solution would be path (ACT, CTC, TCT, TGG), since it is the only one visiting four vertices and having the total cost not greater than $n - l = 4$. The original sequence can be reconstructed by overlapping the labels of visited vertices with shifts equal to the costs of arcs.

In [16] an interactive approach to the reconstructing of a DNA sequence, on the base of a spectrum with negative and positive errors, was proposed. The traditional hybridization experiment is supplemented there with a series of comparisons of short probes against the original sequence. The length of the probes increases starting from l, what improves the reconstruction phase by resolving ambiguities. The graph in this method is constructed in the same way as in [12]: oligonucleotides correspond to vertices and a pair of them is connected by an arc if the oligonucleotides overlap with shift equal to 1. The series of additional queries helps to remove from the graph positive errors and introduce to the graph missing vertices filling gaps between the existing ones. The querying stops when all branching points in the graph are resolved.

Table 1. Arc costs in a graph from the method of Błażewicz *et al.* [3]

	ACT	CTC	GCC	TCT	TGG
ACT	–	1	3	2	2
CTC	3	–	3	1	3
GCC	3	2	–	3	3
TCT	3	1	3	–	2
TGG	3	3	2	3	–

A similar approach was described in [9], but errorless spectrum was assumed there together with additional knowledge what oligonucleotides occur more than once in the original sequence. First, a directed graph is constructed as in the Pevzner's method [13]: oligonucleotides correspond to arcs. Because we know here, which oligonucleotides are repeated, solving ambiguities in this graph is restricted to these arcs. Then, additional longer queries are costructed, corresponding to all possible subpaths covering the strings of the repeatable oligonucleotides.

A set of "gapped probes", instead of the usual oligonucleotide library of cardinality 4^l, was used in the hybridization experiment in the approach from [17]. The gapped probes are the oligonucleotides composed of the nucleotides A, C, G, and T, and moreover of the universal nucleotide U which is able to hybridize with any other nucleotide. The gapped probes used in this approach match the following scheme. The probe described as (s, r)-probe begins with a sequence of s common nucleotides, followed by r occurrences of the pattern: s - 1 universal nucleotides and 1 common nucleotide. For example, a gapped (3,2)-probe looks like XXXUUXUUX, where X means a common nucleotide. This proposition enables to increase the length of the oligonucleotides from the library without enlarging the whole library, what leads to a substantial increase of the length of original sequences which can be reconstructed unambiguously. The method assumes a spectrum without errors. If we build agraph similarly like in the Pevzner's method [13], two vertices will be joined by the arc if their oligonucleotides overlap with shift 1, universal nucleotides being able to match any other ones. During the construction of the Eulerian path, a choosing of an arc in a point of branching must be confirmed by overlaps of preceding and succeeding oligonucleotides, i.e. the whole neighborhood should match (see Example 4).

Example 4

Let again the original sequence be ACTCTGG. The set of gapped (2,1)-probes without any errors would look like as follows: S = {ACUC, CTUG, CTUT, TCUG}. The graph, in which we can search for the Eulerian path is shown in Fig. 6. There exist two Eulerian paths in this graph: (ACUC, CTUT, TCUG, CTUG) and (ACUC, CTUG, TCUG, CTUT), however, only the first one does not produce conflicts in overlaps and it is the solution for this instance.

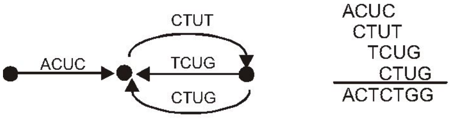


Fig. 6. A graph and the solution of the problem formulated by Preparata *et al.* [17]

3. CONCLUSIONS

There are also many other methods for DNA sequencing by hybridization, assuming spectra with constant or variable oligonucleotide length, with or without errors, however they usually do not base on a clear, known approach from graph theory. The common goal of the presented methods is searching for a path in a directed graph. Since vertices or arcs are labeled by elements of the spectrum, the path can be easily translated to a sequence of nucleotides. A separate problem is the uniqueness of the solution, i.e. whether the result covers the original sequence. This problem was considered in a number of papers (e.g. [20, 15, 8, 1, 19]), the certainty about the obtained result can be achieved only by additional hybridization experiments.

Acknowledgements

The research has been supported by KBN grant.

References

[1] R. Arratia, D. Martin, G. Reinert, and M. S. Waterman, *Poisson process approximation for sequence repeats and sequencing by hybridization,* Journal of Computational Biology **3**, 425-463 (1996).

[2] W. Bains and G. C. Smith, *A novel methodfor nucleic acid sequence determination,* Journal of Theoretical Biology **135**, 303-307 (1988).

[3] 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).

[4] J. Błażewicz, Ł. Gwóźdź, M. Kasprzak, and M. Przysucha, *A comparison of two DNA sequencing methods,* Computational Methods in Science and Technology **2**, 17-32 (1996).

[5] J. Błażewicz, A. Hertz, D. Kobler, and D. de Werra, *On some properties of DNA graphs,* Discrete Applied Mathematics **98**, 1-19 (1999).

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

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

[8] M. Dyer, A. Frieze, and S. Suen, *The probability of unique solution of sequencing by hybridization,* Journal of Computational Biology **1**, 105-110 (1994).

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

[10] J. N. Hagstrom, R. Hagstrom, R. Overbeek, M. Price, and L. Schrage, *Maximum likelihoodgenetic sequence reconstructionfrom oligo content,* Networks **24**, 297-302 (1994).

[11] R. J. Lipshutz, *Likelihood DNA sequencing by hybridization,* Journal of Biomolecular Structure andDynamics **11**, 637-653 (1993).

[12] 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).

[13] P. A. Pevzner, *l-tuple DNA sequencing: computer analysis,* Journal of Biomolecular, Structure and Dynamics **7**, 63-73 (1989).

[14] P. A. Pevzner, *Computational Molecular Biology: an Algorithmic Approach,* MIT Press, Cambridge (2000).

[15] P. A. Pevzner and R. J. Lipshutz, *Towards DNA sequencing chips,* Lecture Notes in Computer Science **841**, 143-158 (1994).

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

[17] F. P. Preparata and E. Upfal, *System and methods for sequencing by hybridization,* United States Patent ApplicationUS 2001/0004728A1 (2001).

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

[19] R. Shamir and D. Tsur, *Large scale sequencing by hybridization,* Journal of Computational Biology **9**, 413-428 (2002).

[20] 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).

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