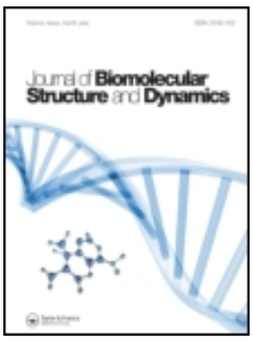


Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

1-Tuple DNA Sequencing: Computer Analysis

Pavel A. Pevzner

To cite this article: Pavel A. Pevzner (1989) 1-Tuple DNA Sequencing: Computer Analysis, Journal of Biomolecular Structure and Dynamics, 7:1, 63-73

To link to this article: http://dx.doi.org/10.1080/07391102.1989.10507752

Published online: 21 May 2012.

Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 7, Issue Number 1 (1989), ©Adenine Press (1989).

l-Tuple DNA Sequencing: Computer Analysis

Pavel A. Pevzner
Laboratory of Mathematical Methods
Institute of Genetics of Microorganisms
1-st Dorozhny 1, Moscow, 113545, USSR

Abstract

A new method of DNA reading was proposed at the end of 1988 by Lysov *et al.* According to the authors' claims it has certain advantages as compared to the Maxam-Gilbert and Sanger methods, which are revealed by automation and rapidity of DNA sequencing. Nevertheless its employment is hampered by a number of biological and mathematical problems. The present study proposes an algorithm that allows to overcome the computational difficulties occurring in the course of the method during reconstruction of the DNA sequence by its l-tuple composition. It is shown also that the biochemical problems connected with the loss of information about the l-tuple DNA composition during hybridization are not crucial and can be overcome by finding the maximal flow of minimal cost in the special graph.

Introduction

Lysov *et al.* (1) suggested a new approach to DNA sequencing based on obtaining information on the l-tuple composition and further reconstruction of the sequences by computer analysis. A similar approach (2) was developed by R. Crkvenjakov at the Genetic Engineering Center, Belgrad, Yugoslavia.

The idea of l-tuple DNA sequencing is connected with the construction of a matrix of immobilized oligonucleotides, that is a membrane on which all oligonucleotides of length l (l-tuple) are fixed at certain points. Lysov *et al.* (1) suppose that the reading of DNA fragments of length $n \approx 200$ can be carried out at $l=8$. Hence the realization of this method requires 65536 preliminary synthesized oligonucleotides. After hybridization of the DNA fragment with the matrix under strict conditions it is possible to determine the l-tuple composition of this fragment. Data concerning the l-tuple composition of the DNA fragment of length n (n-fragment) are entered into the computer (in the case when all l-tuple of the n-fragment are different, the set of $n-l+1$ l-tuple is obtained) and the reconstruction of the DNA sequence is achieved.

Lysov *et al.* (1) note three main problems arising during l-tuple sequencing:

1. Impossibility of sequence reconstruction by the l-tuple composition.

The authors of the method note that there are two possible cases:

- there exists a unique n-fragment corresponding to the experimentally obtained l-tuple composition S (reconstruction of the sequence by the spectrum S is possible on principle).

- there exist several n-fragments corresponding to S (reconstruction is impossible).

In the present paper a necessary and sufficient condition for testing the uniqueness of sequence reconstruction is obtained and an efficient reconstruction algorithm is proposed. Lysov *et al.* give only a weak sufficient condition for such testing. As a result the share of cases not allowing the unique reconstruction was overestimated (according to our statistical experiments - 54% at n=200 and l=8, according to the authors of the method - 20% owing to use of the special recursive procedure for n about 200 and l=8). Our approach allows to decrease the share of unsuccessful attempts of reconstruction to 6% for n=200, l=8. In these cases Lysov *et al.* (1) suggest to use one of the following procedures:

- traditional sequencing methods for non-uniquely reconstructed DNA fragment,

- to carry out an additional definitive experiment with specifically chosen probe,

- to hydrolyze the non-uniquely reconstructed DNA fragment into some smaller fragments,

- to increase l (it is hampered by technical problems connected with the realization of the matrix for 262 144 oligonucleotides).

2. Incomplete hybridization - the loss of information about the l-tuples from the spectrum.

Because of the secondary DNA structure there is a possibility to lose information concerning some l-tuples in the course of hybridization (incomplete hybridization of the DNA with probes as a result of self-hybridization of probes can take place also). The problem gets complicated by the fact that different oligonucleotides require varying conditions for hybridization. Thus the spectrum can contain not $n-l+1$ l-tuples, as in the case of "ideal" hybridization, but $n-l+1-k$ ones, where k is called the defect of the experiment (the non-zero defect can be stipulated by the presence of repeats of length l in the n-fragment). Lysov *et al.* (1) suggest no method of DNA reconstruction in this case. The present study suggests an algorithm for identification of the DNA sequences of spectrum with defects ($k>0$) and shows that for small k the algorithm resolution is only slightly reduced compared with the "ideal" spectrum ($k=0$). This algorithm allows to overcome the computational difficulties by involving the discrete optimization, in particular, the transportation problem.

3. Non-specific hybridization - the spectrum comprises l-tuples absent in the original DNA.

Apparently the solution of this problem requires mainly biochemical methods than

mathematical ones. (Lysov *et al.*, 1988 announced the work on choosing conditions permitting to exlcude non-specific hybridization). Nevertheless it is worth to note that even in the case of non-specific hybridization one can try to construct all the DNA sequences with the spectra "similar" in S, and after this to reveal the true sequence in additional experiments. An analogous approach based on the mass-spectrometry ESIAP was suggested recently by M.D. Frank-Kamenetskii, M.A. Grachev, Yu.A. Kusner, O.A. Mirgorodskaya and P.A. Pevzner for express-sequencing of amino acids sequences (3-5).

l-Tuple DNA Identification. Full Spectrum Case (k=0)

The problem of l-tuple DNA sequencing is a particular case of the problem about minimal superword (6): for a given set of words S one has to find the word of minimal length containing all words of the set S. Gallant *et al.* (7) proved the NP-completeness of this problem. Hence one can hope to obtain the efficient algorithms for its solution only in particular cases. For example the problem of the choice of optimal oligonucleotide linkers (in this case S is the set of the restriction recognition sites) was considered in (8). Following this paper we shall formulate the minimal superword problem in terms of graph theory.

The approach of Lysov *et al.* (1) to DNA identification can be formulated (in the graph theory language) as following: According to the spectrum S the digraph H on n−l+1 vertices is constructed (each l-tuple of spectrum S has a corresponding vertex of H). The vertex v of graph H is joined by an arc (v,w) with the vertex w if the last (l−1) nucleotides of v coincide with the first (l−1) nucleotides of w. We say that "i-shift of w in respect to v" is admissible if the last l−i nucleotides of v coincide with the first l−i nucleotides of w (the influence of the shifts between words upon the characteristics of genetic text was studied in (9)). Hence H contains the arc (v,w) if and only if l−shift w in respect to v is admissible. Figure 1 presents the graph H obtained from the

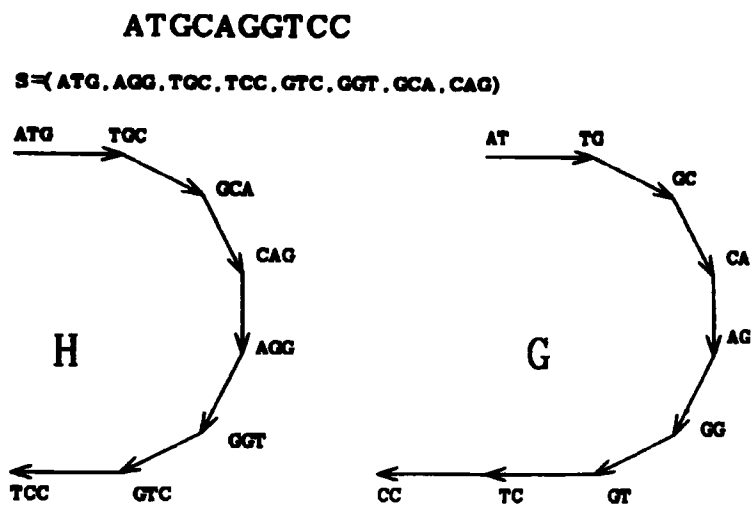


Figure 1: Graphs H and G for ATGCAGGTCC sequence spectrum (l=3).

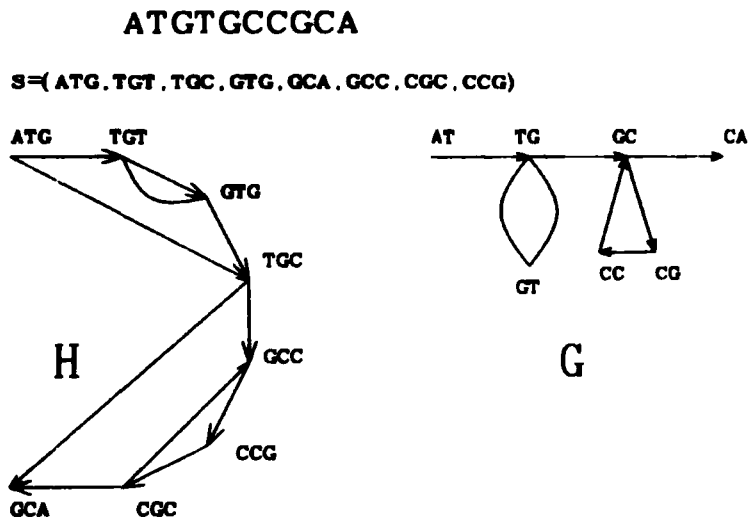


Figure 2: Graphs H and G for ATGTGCCGCA sequence spectrum (l=3).

ATGCAGGTCC sequence spectrum; Figure 2 - from the ATGTGCCGCA sequence spectrum (n=10, l=3).

It is easy to see that the l-tuple DNA sequencing problem is equivalent to the search of the Hamiltonian path in H. Lysov *et al.* (1) formulate the following sufficient condition for the existence of a Hamiltonian path in H between the 5′ end (vertex s) and 3′ end (vertex t):

in(v)=out(v)=1	for all vertices of H, different from s and t	
in(s)=0, out(s)=1	for vertex s	[1]
in(t)=1, out(t)=0	for vertex t	

(here in(v) is the number of arcs entering vertex v, out(v) is the number of arcs leaving vertex v). Condition [1] means that graph H is a directed path without self-intersections (linear graph).

However even for the graph on Figure 2 such a sufficient condition is too weak - in this case criterion [1] does not allow to reconstruct the DNA sequence (although the unique reconstruction is possible). Of course one can try to find Hamiltonian paths in the graphs other than linear, but the problem of searching Hamiltonian path is NP-complete (effective algorithm for its solution is unknown).

In the present paper the problem of DNA identification is reduced to the search of not Hamiltonian, but Eulerian paths in graphs. The existence of an efficient algorithm for finding Eulerian paths allows to avoid computational difficulties and to formulate necessary and sufficient condition for DNA identification from the l-tuple spectrum. In our method the digraph G is constructed on the total set of (l−1)-tuples; the (l−1)-tuple v is joined by an arc with the (l−1)-tuple w if the spectrum S contains l-tuple for

which the first (l−1) nucleotides coincide with v and the last (l−1) nucleotides coincide with w. In the graph G to each l-tuple from S corresponds an arc, but not a vertex (as in H). While graph G on Figure 1 only slightly differs from H, it is evident that G on Figure 2 is significantly "simplified" as compared with H. This approach makes the l-tuple sequencing problem eqivalent to the search of not Hamiltonian, but Eulerian paths in G.

The criterion of existence of an Eulerian path in G is given by the Eurler theorem for digraphs (10):

$$\begin{array}{ll} in(v)=out(v) & v\neq s,t \\ out(s)-in(s)=1 & \text{for 5' end s} \\ out(t)-in(t)=-1 & \text{for 3' end t} \end{array}$$

The number of Eulerian paths in given by deBruijn - vanAardenne-Ehrenfest - Smith - Tutte theorem (10):

$$E = C \prod_{v \varepsilon VG} (d(v)-1)! \qquad [2]$$

where E is the number of Eulerian paths in G, VG is the vertex set of G, d(v)=in(v)=out(v) is the semi-degree of the vertex v, C is the common value of all signed minors of the Kirchhoff matrix (we added a fictious arc (t,s) to the graph G in order to make it Eulerian and to formulate the deBruijn - vanAardenne-Ehrenfest - Smith - Tutte theorem in its original version).

Condition E=1 is fulfilled if and only if C=1 and $d(v)\leqslant 2$ for all $v\varepsilon V$. Statistically these conditions are fulfilled significantly more often than criterion [1]. To compare criteria [1] and [2] the "C" program EULER was written. Statistical experiments show that condition [2] permits to increase the share of successful identification for n=200 and l=8 from 46% to 94% (Figure 3). While implementing the EULER program we avoid the tedious computation of signed minors of the Kirchhoff matrix (there is no use of it since we were interested not in the precise C value, but simply testing condition C=1). For testing condition C=1 we suggest the following algorithm:

- define an arbitrary partition of graph G into simple cycles,

- construct the intersection graph GC of simple cycles

(vertices of GC correspond to the simple cycles of the chosen partition) according to the rule: vertices corresponding to simple cycles c_1 and c_2 of GC are connected by k edges if c_1 and c_2 have k common vertices. After the construction of the graph GC is completed the problem of DNA identification is easily solved due to the following lemma:

Lemma. C equals 1 if GC is a tree.

Note that for the example of ambiguous DNA reconstruction described in (1) the

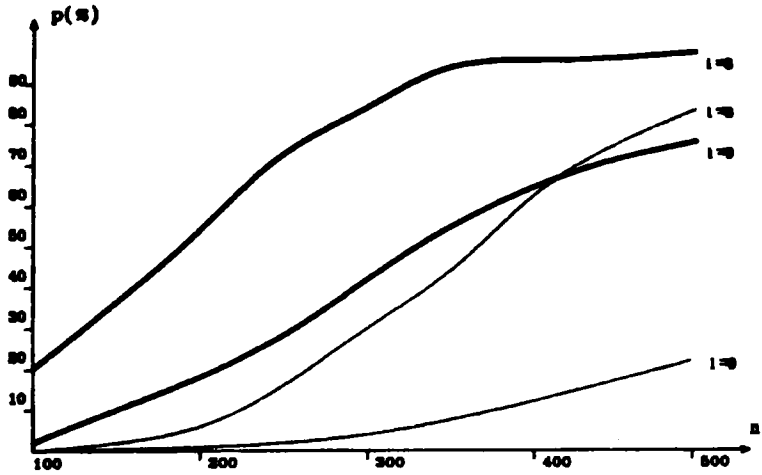


Figure 3: The dependence of "unsuccessful" attempts of DNA sequence reconstruction (p) on the length of sequence (n): ------ by criterion (1); —— by criterion (2).

case C>1 takes place, and the corresponding graph GC is the cycle of length 2.

l-Tuple DNA Identification - Incomplete Spectrum Case (k>0)

Consider a case when as a result of experiment for n-fragment a spectrum with not n−l+1 l-tuples (as for the ideal cases), but n−l+1−k ones is obtained (here k is the defect of the experiment). Note that k can exceed 0 not only because DNA did not hybridize with some probes, but also due to DNA repeats of the length l. Consider for example the spectrum on Figure 1, of which three probes were "lost" - TGC,CAG and GTC. In this case graph G involves not a single path, but four ones (Figure 4). In order to reconstruct the original DNA sequence it is necessary to add arcs to the graph G to join these four paths into one. However besides the "correct" joint (scheme G_1) other variants are possible (scheme G_2), and it seems that identification is impossible. We suggest an algorithm for DNA sequence reconstruction from the spectrum with defects. It is shown that the probability of unique identification in the case of small defect is only slightly reduced as compared with the ideal spectra.

Let us explain how to choose between G_1 and G_2 on Figure 4, and then describe the algorithm itself. At first note that thick arcs of G_1 and G_2 on Figure 4 show only the way in which the components of G are joined together; therefore these arcs can join not only neighboring vertices (i.e., the pairs of vertices for which 1-shift is admissible), but also those lying at a distance greater than 1 (e.g., arc TG→AG on scheme G_2). Put div(v)=in(v)−out(v). Schemes G_1 and G_2 both have 4 pairs of vertices with div(v)≠0. For TG, CA, GT, CC div(v) equals 1 and for GC, AG, TC, AT div(v) equals −1. However it remains unclear which of the vertices with div(v)=−1 is DNA 5′ terminal and which of the vertices with div(v)=1 is DNA 3′ terminal. On the scheme G_1 3 new paths (corresponding to the thick arcs) are added in G: TG→GC, CA→AG and GT→TC. Each of them can be made of one arc since pairs (TG, GC), (CA, AG), (GT, TC)

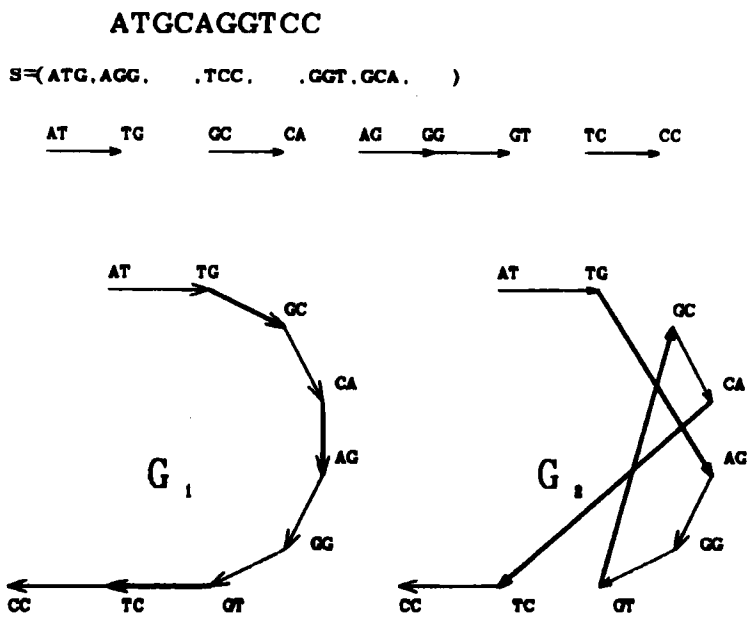


Figure 4: Distinct schemes of joint of components G into one directed path.

admit an 1-shift. Thus in order to join the four G components into one directed path 1+1+1=3 arcs are necessary (in the case of scheme G_1). For the scheme G_2 the path TG→TC - not less than 2 arcs (minimal shift of AG in respect to TG equals 2), CT→GC - not less than 2 arcs, CA→TC - not less than 2 arcs. Hence in the case of G_2 in order to join four components of G into one directed path at least 2+2+2=6 arcs are necessary. It means that to join 4 paths of graph G in the order given by G_2 at least 6 arcs are necessary. Taking into account that the defect of the spectrum equals only 3 we obtain the contradiction and eliminate the variant of joint given on shceme G_2.

Thus, the idea of reconstructing DNA from the spectrum with defects consists of the construction of a shceme G_{opt} in which the number of arcs necessary to join the fragments of G into a directed path, would not exceed k. In the case when such a graph is unique, identification is possible; in other cases it is impossible. Statistical experiments allowed to reveal that within values n and l suggested (1) the share of unsolvable cases is small for low values of k (Figure 5), below an algorithm for finding a scheme G_{opt} is suggested. It is based on the reduction to the transportation problem.

Analysis of Spectra with Defects and the Transportation Problem

Consider the vertices of graph G in which Eulerian condition is violated. Define

$$X = \{v \varepsilon VG: \mathrm{div}(v)<0\}$$
$$Y = \{v \varepsilon VG: \mathrm{div}(v)<0\}$$
$$m = -\sum_{v \varepsilon X} \mathrm{div}(v) = \sum_{v \varepsilon Y} \mathrm{div}(v)$$

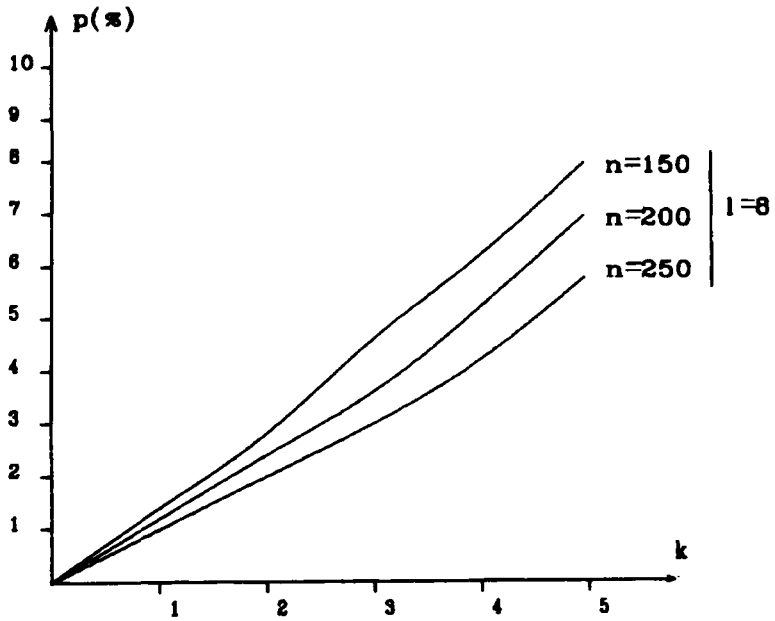


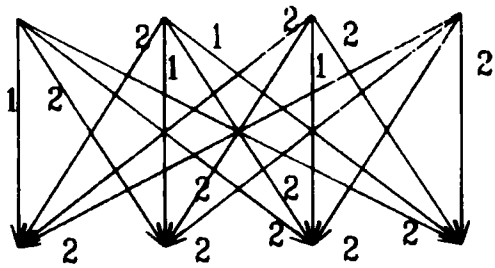
Figure 5: The dependence of the share of "unsuccessful" attempts (p) upon reconstructing the complete spectrum from the spectrum with defects, on the value of the defect (k).

Usually within the defect k each of the sets X and Y contains k+1 vertices. Apparently, $m \leqslant k+1$. Construct now complete bipartite graph $K_{m,m}$ (Figure 6), such that vertices of the first part correspond to the set X, and the vertices of the second part - to the set Y (the vertices with $|div(v)|>1$ duplicate $|div(v)|$ times). Each arc (v,w) of the graph $K_{m,m}$ is assigned with a cost c(v,w) equal to the value of the minimal shift of w in respect to v (i.e., c(v,w) is the minimal number of arcs necessary for $v \rightarrow w$ transition in the graph G).

To find which arcs should be added to G to ensure the existence of the Eulerian path one should find a certain matching (11) in $K_{m,m}$ wth m−1 arcs. These arcs must join the separate paths of G into the Eulerian path, and their total cost should not exceed k. Add tothe graph $K_{m,m}$ four vertices s_1,s_2,t_1,t_2 and 2m+2 arcs (Figure 6):

- (s_1,s_2) with capacity m−1,
- (t_1,t_2) with capacity m−1,
- m arcs of type (s_2,x), where $x \varepsilon X$ with capacity 1,
- m arcs of type (y,t_2), where $y \varepsilon Y$ with capacity 1.

It is possible to show that if the constructed graph has the unique maximal flow of cost k than the problem of DNA sequence reconstruction from the l-tuple composition with defect $k>0$ has unique solution. The existence of the fast algorithms for finding maximal flow of minimal cost (11) allows to reconstruct the DNA from the spectrum with defects efficiently.



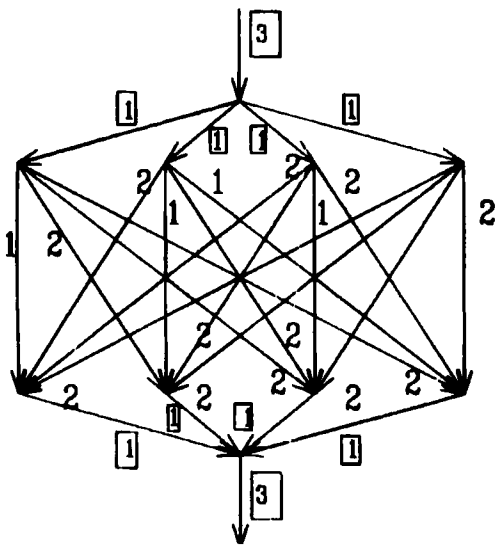


Figure 6: Graph $K_{m,m}$ for spectrum with defects given on Figure 4 and transportation network constructed for DNA sequence reconstruction (the capacities of arcs are shown in the boxes).

Remarks

1. In the case when $|div(v)|>1$ it is unnecessary to duplicate the vertices - in this case the problem of maximal flow of minimal cost can be solved by defining arc capacity of (s_2,v) or (v,t_1) as $|div(v)|$ (the choice between these two arcs is determined by the sign of div(v)).

2. The proposed approach requires to solve the problem of the uniqueness of maximal flow of minimal cost. This problem in its simplest case can be obtained by solving $m-1$ transportation problems (prohibiting subsequently $m-1$ arcs of optimal plan). We suggest to reduce this problem to the search of the minimal cycle in the graph on m vertices, obtianed from $K_{m,m}$ by compressing the arcs of the optimal plan with the specially introduced cost function.

3. In the course of implementation of the method arises the question concerned with the probability distribution of the value of the minimal shift for random words v and w. It occurred that the average value of the minimal shift is very large (approximately to $l-1$) as compared with shift 1 between the neighboring vertices. This circumstance is the reason of uniqueness of the optimal solution of the transportation problem in the majority of cases.

4. Upon solving the transportation problem it is necessary to consider whether the components of G are joined into one connected path. The variants for which the joint leads to the non-connected graph should be neglected.

5. For the graph, given on Figure 7 from the deBruijn - vanAardenne-Ehrenfest - Smith - Tutte theorem point of view two Eulerian paths are possible (depending on which of the two multiple arcs is going first in the Eulerian path). However the same DNA sequence corresponds to both of these paths. We test such situations and reconstruct DNA uniqueness even in these cases.

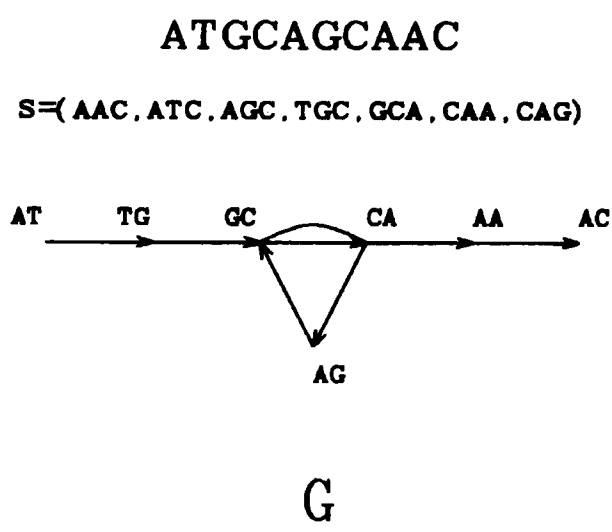


Figure 7: Two Eulerian paths are possible for graph G depending on which of the multiple arcs (GC→CA) is the first in the path. Both paths correspond to the same ATGCAGCAAC DNA sequence.

6. For some spectra with defect k the cost of the corresponding maximal flow occurred to be less than k (e.g., in the case of more than l identical nucleotides running in the DNA fragment). Graph on Figure 5 takes into account such situations.

7. Statistical experiments show that the cases in which the information about l-tuples near to 5′-end or 3′-end is lost give the main contribution to the number of unsuccessful attempts upon reconstructing the full spectrum from the spectrum with k defects (Figure 5).

Discussion

The suggested graph-theoretical algorithm for identification the DNA sequences allows to reduce significantly the share of cases for which indentification is impossible (for n=200 and l=8 it give a 9 times decrease). The question about the theoretical evaluation of the probability of the l-tuple composition coincidence for random n-fragments remains open. A simpler question concerned with the probability characteristics of the statistical distances between texts (statistical distance is the distance between l-tuple compositions when the numbers of entries of each l-tuple is taking into account) was studied in (12). It was shown that the share of sequences with identical l-tuple composition (considering the number of entries of each l-tuple into the spectrum) decreases rapidly while l increases. While going from l=8 to l=9 (Figure 3) the share of such cases becomes negligible, but the technical difficulties in the realization of the method increase significantly. Probably that the realization of l-tuple sequencing by employing of the "classical" methods in "ambiguous" cases will allow to decrease the cost and to increase the sequencing rate.

It is shown that the loss of l-tuples in the spectrum is not fatal for reconstructing: for

small number of defects the share of unsolved cases changes only slightly. The case with erroneous l-tuples in the spectrum is the most complicated. Now our efforts are aimed to avoid the case of "erroneous" l-tuples in the spectrum not by biochemical (as in (1)), but by mathematical means: by using ideas employed in express-sequencing of amino acids by means of ESIAP mass-spectrometry.

Acknowledgments

I wish to thank Dr. A.R. Rubinov for valuable discussions on flows in networks and Drs. L.Ya Lomakina and B.Z. Shapiro for helping in preparing the English version of this manuscript.

References and Footnotes

1. Yu.P. Lysov, V.L. Florent'ev, A.A. Khorlin, Khrapko, V.V. Shik and A.D. Mirzabekov, *Dokl. Acad. Sci. USSR 303*, 1508-1511 (1988).
2. L. Roberts, *Science 242*, 1245 (1988).
3. Yu.V. Gavrilov, A.D. Frank-Kamenetzky and M.D. Frank-Kamenetzky, *Biochimia 31*, 709-804 (1966).
4. M.L. Alexandrov, G.I. Baram, L.N. Gall, M.A. Grachev, V.D. Knorre, N.V. Krasnov, Yu.S. Kusner, O.A. Mirgorodskaya, V.I. Nikolaev and V.A. Shkurov, *Bioorganik. Chem. 11*, 705-708 (1985).
5. P.A. Pevzner, *Combinatorial Methods of Biopolymer Structure Analysis*, Moscow Physical and Technical Institute, p. 24 (in Russian) (1988).
6. M.D. Garey and D.S. Johnson, *Computers and Intractability*, Friedman & Co., N.Y., p. 416 (1979).
7. J. Gallant, D. Maier and J.A. Storer, *J. Comput. Syst. Sci. 20*, 50-58 (1980).
8. P.A. Pevzner and V.P. Veiko, *Molek. Biol. 23*, in press (1989).
9. P.A. Pevzner, M.Yu Borodovskii and A.A. Mironov, *J. Biomol. Struct. Dyn. 6*, 1013-1026 (1989).
10. F. Harary, *Graph Theory*, Addison-Wesley, N.Y., p. 300 (1969).
11. G.M. Adelson-Velsky, E.A. Dinic and A.V. Karzanov, *Flow Algorithms*, Nauka, Moscow, p. 119 (in Russian) (1975).
12. P.A. Pevzner, *Nucl. Acids Res. 17*, submitted (1989).

Date Received: May 10, 1989

Communicated by the Editor Maxim Frank-Kamenetskii