MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **70** (2013) 401-415

ISSN 0340 - 6253

A New Graph Theoretical Approach to DNA Sequencing with Nanopores

Nafiseh Jafarzadeh and Ali Iranmanesh*

Department of Mathematics, Faculty of Mathematical Sciences,
Tarbiat Modares University, P.O. Box: 14115-137,
Tehran, Iran

iranmanesh@modares.ac.ir

(Received February 28, 2012)

Abstract

DNA sequencing with nanopores (nanopore sequencing) is a method for determining the order in which nucleotides occur on a strand of DNA. One particular way to analyze this method is by using concepts from graph theory. In this paper, we propose a new method for reading a DNA sequence with nanopores.

1. Introduction

Since Watson and Crick [1] proposed the helical structure of DNA, many problems about this structure are posed. An important problem is how to read and recognize primary structure of a DNA sequence. One particular way to analyze DNA sequences and their properties is by using the concept of graph theory. Up to now, many papers published about applications of graph theory in analyzing DNA sequences, for example see [2-14]. There are various methods for DNA sequencing which use concepts of graph theory such as hybridization (SBH) and DNA fragment assembly. In 1999, Ludry and Waterman presented an algorithm for DNA sequencing by hybridization (SBH) by using concepts of graph theory [15]. Pevzner presented graph theoretical approaches to DNA sequencing and fragment assembly [16, 17].

*- Corresponding author (Ali Iranmanesh)

Another way to DNA sequencing is using nanopores. Nanopore sequencing is a method under development since 1995 [18, 19] for determining the order in which nucleotides occur on a strand of DNA. Nanopore devices are used in DNA sequencing with the hopes of enabling large and complete strands of DNA to be read completely.

These nanopores are a few atoms in diameter, and single stranded DNA passes through them by optical or electrical means. Through this passage, each individual base of the strand is identified, producing a seemingly endless chain of A, T, G and C. Based on the work of Watson and Crick [20], we know that DNA is a double-stranded structure and with each single strand of DNA there exists a complement, as A matches with T, and C is complementary with G.

There are some key issues in DNA sequencing with nanopores. The first issue lies on the fact that an entire single strand rarely makes it through the nanopore in one piece. Rather, substrings of lengths around 100,000 bases in length are produced. Further, the problem of the coinciding Watson-Crick complement causes confusion regarding whether or not the original strand of its coinciding complements is being passed through the nanopore. Similarly, there is another issue involving the specified orientation of single strand as it passes through. Upon passage and decomposition into substrings, the direction (3′-5′ or 5′-3′) of the DNA is lost [21]. Bokhari and Saure [22] identify and address the problems that occur in this form of sequencing. They use de bruijn graphs representing DNA data upon determining an algorithm which aids the process of DNA sequencing with nanopores, their de bruijn graph G must be constructed with k-long oligonucleotides (k-mers) of complete DNA sequence which k should be selected so that G contain four paths such that the union of all paths is equivalent to G. Further, no two paths have any equivalence in their intersections. Testing these paths with a permutation further determines whether or not that are, in fact complements and reversals of each other. If this holds true, then the strand being examined comes from an authentic sequence of DNA data [21].

Some authors have tried to solve important computational biology problems with relations between DNA sequences and particular graphs. For example see [23, 24].

In 1999, Blazewicz provided definition of DNA graph and some of its properties [25]. After that some authors work by this graph and discuss about its properties and its applications [26-32].

In this paper we give a new method to DNA sequencing with nanopores using concept of DNA graph which is independent on the long of oligonucleotides (k). We will show that this method is simpler of above method proposed by Bokhari and Saure [22].

2. The relation between DNA graph and line digraph

First, we give definition of DNA graph and then discuss about its properties and its role to DNA sequencing with nanopores.

Definition 2.1. [33]. Let $k \geq 2$ be an integer. We say that a directed graph D with a set of vertices V(D) and a set of ordered pairs of points (directed edges) E(D), is DNA graph if it is possible to assign a label $(l_1(x),\ldots,l_k(x))$ of length k to each vertex x of V(D) such that:

(a) $l_i(x) \in \{A,C,T,G\}$, for every $i \in \{1,\ldots,k\}$;

(b) All labels are different, that is, $(l_1(x),\ldots,l_k(x)) \neq (l_1(y),\ldots,l_k(y))$ if $x \neq y$;

(c) $(x,y) \in E(D)$ if and only if $(l_2(x),\ldots l_k(x)) = (l_1(y),\ldots l_{k-1}(y))$.

For any multiset which consists of some k-long oligonucleotides, a DNA graph is often constructed as follows:

Each k-long oligonucleotide from the multiset becomes a vertex; two vertices are connected by an arc vertex if the k-1 rightmost nucleotides of first vertex overlap with the k-1 leftmost nucleotides of the second one.

For example let S = {ACTG, CTGT, TGTA, GTAC, TACT, ACTT, CTTG} be a multiset of all 4-long oligonucleotides of a DNA sequence, the DNA graph of "S" is made as follows:

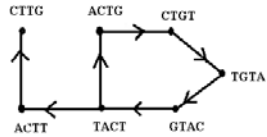

Figure 1. DNA graph of "S"

The above approach, however, leads to an exponential-time algorithm since looking for a Hamiltonian path is in general strongly NP-complete [22].

To obtain the relationship between DNA graphs and line digraphs (directed line graphs), we need some definitions and theorem:

Definition 2.2. [34]. A graph G is a p-graph if for any ordered pair x, y of vertices for G, there are at most p parallel arcs from x to y.

Definition 2.3. [27**].** The adjoint G' = (V,U) of a graph G = (X,V) is the 1-graph with vertex set V and such that there is an arc from a vertex x to a vertex y in G' if and only if the terminal endpoint of the arc x in G is the initial endpoint of arc y in G.

Definition 2.4. [27**].** A graph is a directed line-graph (line digraph) if and only if it is the adjoint of a 1-graph.

Theorem 2.5. [27**].** Let H be the adjoint of graph G. then there is an Eulerian path/circuit in G if and only if there is a Hamiltonian path/circuit in H.

Since line digraphs are especial cases of adjoints, we get the following corollary:

Corollary 2.6. [27**].** Let H be the line digraph of 1-graph G. then there is an Eulerian path/circuit in G if and only if there is a Hamiltonian path/circuit in H.

Now we construct a new DNA graph by another approach which presented by Pevzner [35] as follows:

Each k-long oligonucleotide from the multiset becomes an arc which its initial end point is k-1 rightmost nucleotides of arc and its terminal end point is k-1 leftmost nucleotides.

For example, the new DNA graph of the graph in Figure 1 according to above approach is made as follows:

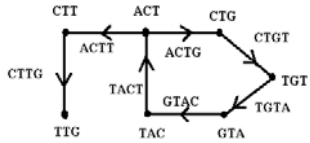

Figure 2. The new DNA graph of Figure 1

Easily seen that, line digraph of this new DNA graph is the DNA graph which is made by the previous approach which is shown in Figure1. Then using corollary 2.6, we find that there is an Eulerian path in this new DNA graph, and it is known that finding an Eulerian path can be done in polynomial time.

3. A new method for DNA sequencing

In this section, by creating and analyzing DNA graph with the approach which discuss in the previous section, we will propose a new method to read a DNA sequence with nanopores. This method is simpler than the method proposed by Bokhari and Sauer in [22], since this is independent on the long of oligonucleotides (k).

There are two certain key issues in DNA sequencing with nanopores as follows:

1. Each DNA helix would be split up into the original and its Watson-Crick complement. Since these would be further broken up into smaller pieces, there is no way for the nanosequencer's signals to reveal which of these two is passing through at any time.

2. There is no control on the orientation in which the strings pass through the nanopore. DNA has well defined orientations (the 3′-5′ direction and vice versa). Since the single strand DNA is broken into subsequences, all information on orientation is lost.

We will try to solve these problems by our method. Since in DNA sequencing with nanopores, we use the complete DNA sequence, there are four different strings to be recognized. These are the original 3′-5′ string, the reversed 5′-3′ string and the two Watson-Crick complements of these. Our approach is to identify all four of these. Now we give our new method:

Let M be a multiset of all k-long oligonucleotides of a complete DNA sequence, we construct a graph G with these oligonucleotides as follows:

Each oligonucleotide becomes an arc which its initial end point is k-1 rightmost nucleotides of arc and its terminal end point is k-1 leftmost nucleotides. This graph (G) may be connected or disconnected, but this graph includes four connected DNA graphs :G_1, G_2, G_3, G_4 , which G_1 is DNA graph of all k- long oligonucleotides from the original 3′-5′ string, G_2 is DNA graph of all k- long oligonucleotides from the reversed 5′-3′ string, G_3 is DNA graph of all k-

long oligonucleotides from the Watson-Crick complements of the original 3′-5′ string and G_4 is DNA graph of all k- long oligonucleotides from the Watson-Crick complements of the reversed 5′-3′ string. According to the corollary 2.6, each G_i, i = 1,…,4 includes an Eulerian path. These paths determine structure of primary DNA sequence. Our goal is finding these paths.

Suppose that A is a set of all subgraphs of G which include an Eulerian path. Now we define the set B as follows:

$$B = \{\{G_1, G_2, G_3, G_4\} \mid G_i \in A \ \ \& \ \ G_1 \cup G_2 \cup G_3 \cup G_4 = G \ \ \& \ E(G_i) \cap E(G_j) = \emptyset, \forall, i,j, i \neq j \}$$

Let $R(p)$ represent the reversal of a sequence p and $C(p)$ represent the Watson-Crick complement of p. We define a property for the members of the set B.

τ-property: Let p_1, p_2, p_3, p_4 be four paths ,we say these paths satisfy τ-property if ones can find a permutation $\tau(p_i)$ for every i=1,…,4, such that $\tau(p_1) = R(\tau\ (p_2)) = C(\tau\ (p_3)) = R(C(\tau(p_4)))$.

Assume that $\{G_1, G_2, G_3, G_4\} \in B$ and p_i is an Eulerian path of G_i, for i = 1,2,3,4. If this set of subgraphs of G satisfies the τ-property, we can get result that p_1 is single string of DNA sequence. Therefor by this method, we can read a DNA sequence.

Now we give an example for our method.

Example: Let M = { CTG, TCT, CTG, TGA, GAC, ACT, GAG, AGA, GAC, ACT, CTG, TGA, TCA, CAG, AGT, GTC, TCT, CTC, GTC, TCA, CAG, AGA, GAG} be a multiset of all 3-long oligonucleotides of a complete DNA sequence, we construct a graph G with these oligonucleotides as follows:

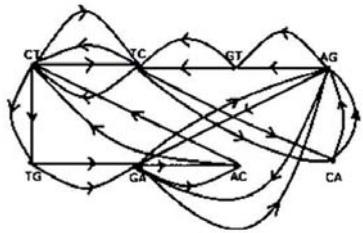

Figure 3. DNA graph of M

Then we obtain four subgarphs of G according to section 3 as follows:

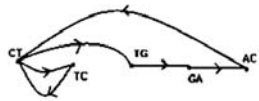

Figure 4. Subgraph G_1 of graph G

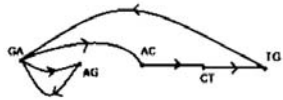

Figure 5. Subgraph G_2 of graph G

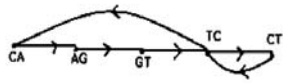

Figure 6. Subgraph G_3 of graph G

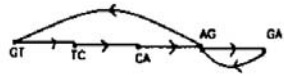

Figure 7. Subgraph G_4 of graph G

We see that G_i , i = 1,…,4 include an Eulerian path so that $G_1 \cup G_2 \cup G_3 \cup G_4 = G$ and $E(G_i) \cap E(G_j) = \emptyset \ \ \forall, i,j, i \neq j$. As you see in figures 4,5,6,7, for each subgraph G_i , we have an Eulerian path p_i , i = 1,…,4, as follows:

p_1 = CTGACTCT

p_2 = TCTCAGTC

p_3 = GAGACTGA

p_4 = AGTCAGAG

Take $\tau(p_1)$= CTCTGACT, $\tau(p_2)$ = TCAGTCTC, $\tau(p_3)$= GAGACTGA, $\tau(p_4)$= AGTCAGAG

Then we have $\tau(p_1) = R(\tau(p_2)) = C(\tau(p_3)) = R(C(\tau(p_4)))$. Therefore p_1 is original single strand of DNA sequence.

Finally for illustrating the utility of this method, we use this method for sequencing of a complete segment of the first exon of β-globin human gene [12]. First we take k = 4 and use our new method to sequencing this gene and then we take k = 8 and then we will compare results.

Let M' = { ATGG, TGGT, GGTG, GTGC, TGCA, GCAC, CACC, ACCT, TACC, ACCA, CCAC, CACG, ACGT, CGTG, GTGG, TGGA, GGTA, TGGT, GTGG, CGTG, ACGT, CACG, CCAC, TCCA, CCAT, ACCA, CACC, GCAC, TGCA, GTGC, GGTG, AGGT} be a multiset of all 4-long oligonucleotides of a complete segment of the first exon of β-globin human gene, we construct a graph G' with these oligonucleotides as follows:

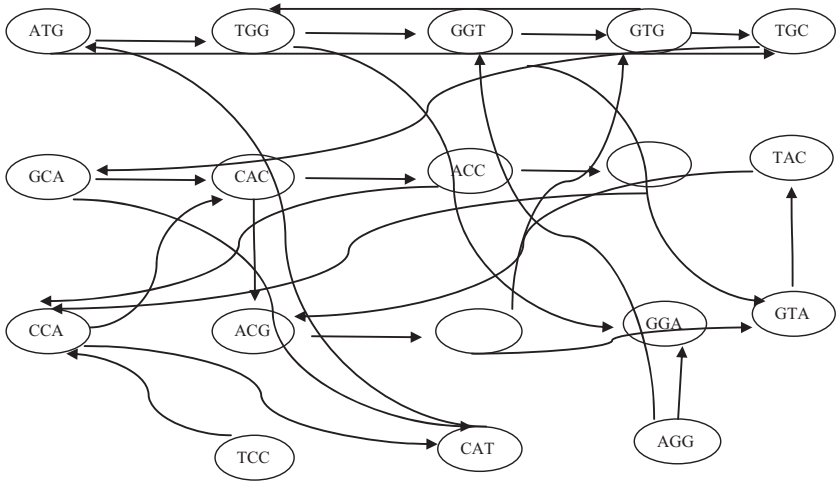

Figure 8. DNA graph of M′

Then we obtain four subgarphs of G' according to section 3 as follows:

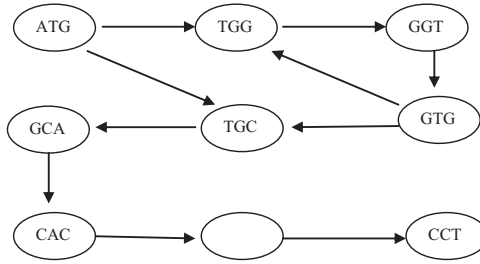

Figure 9. Subgraph G'_1 of graph G'

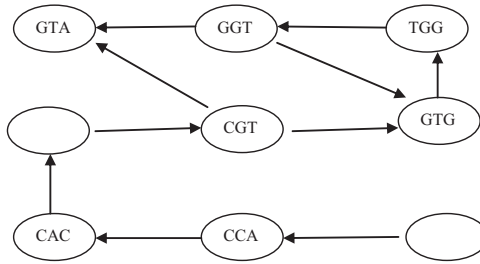

Figure 10. Subgraph G'_2 of graph G'

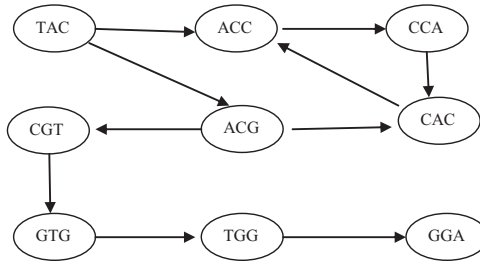

Figure 11. Subgraph G'_3 of graph G'

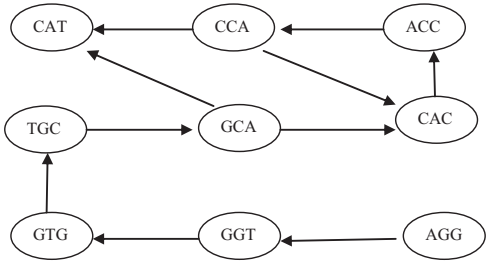

Figure 12. Subgraph G'_4 of graph G'

As you see in figures 9, 10, 11 and 12, $\{G'_1, G'_2, G'_3, G'_4\} \in B$ and for each subgraph G'_i we have an Eulerian path p'_i , $i = 1,\ldots,4$, as follows:

p'_1 = ATGGTGCACCT

p'_2 = TCCACGTGGTA

p'_3 = TACCACGTGGA

p'_4 = AGGTGCACCAT

Take τ = identity permutation, we have $\tau(p'_1)$= ATGGTGCACCT, $\tau(p'_2)$ = TCCACGTGGTA

, $\tau(p'_3)$= TACCACGTGGA, $\tau(p'_4)$= AGGTGCACCAT

Then we have $\tau(p'_1) = R(\tau\,(p'_2)) = C(\tau\,(p'_3)) = R(C(\tau(p'_4)))$. Therefore p'_1 = ATGGTGCACCT is original single strand of DNA sequence.

Now let M'' = { ATGGTGCA, TGGTGCAC, GGTGCACC, GTGCACCT, ACGTGGTA, CACGTGGT, CCACGTGG, TCCACGTG, TACCACGT, ACCACGTG, CCACGTGG, CACGTGGA, TGCACCAT, GTGCACCA, GGTGCACC, AGGTGCAC} be a multiset of all 8-long oligonucleotides of a complete segment of the first exon of β-globin human gene, we construct a graph G'' with these oligonucleotides as follows:

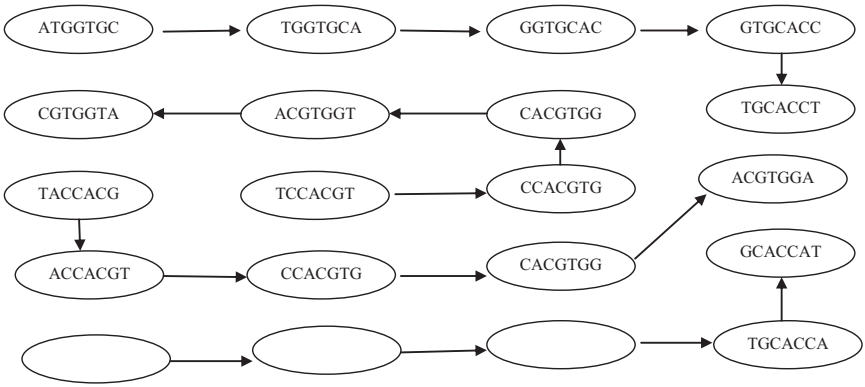

Figure 13. DNA graph of M″

Then we obtain four subgarphs of G'' according to section 3 as follows:

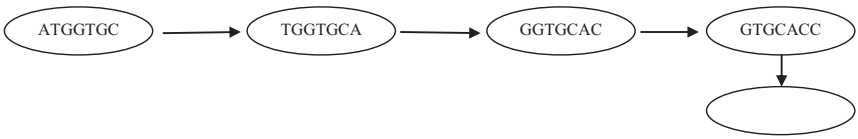

Figure 14. Subgraph G''_1 of graph G''

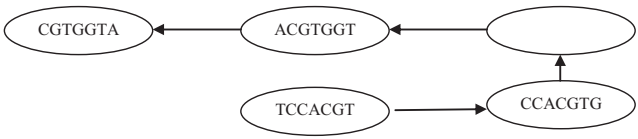

Figure 15. Subgraph G''_2 of graph G''

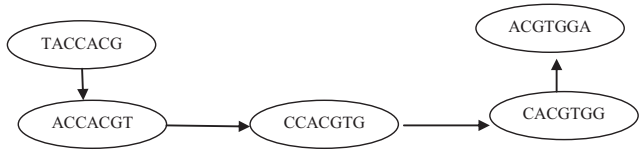

Figure 16. Subgraph G''_3 of graph G''

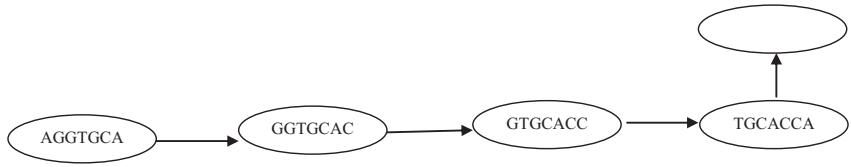

Figure 17. Subgraph G''_4 of graph G''

As you see in figures 14-17, $\{G''_1, G''_2, G''_3, G''_4\} \in B$ and for each subgraph G''_i, we have an Eulerian path p''_i , $i = 1,\ldots,4$, as follows:

p''_1 = ATGGTGCACCT

p''_2 = TCCACGTGGTA

p''_3 = TACCACGTGGA

p''_4 = AGGTGCACCAT

Same as above take τ = identity permutation, we have $\tau(p''_1)$= ATGGTGCACCT, $\tau(p''_2)$ = TCCACGTGGTA, $\tau(p''_3)$= TACCACGTGGA, $\tau(p''_4)$= AGGTGCACCAT

Then we have $\tau(p''_1) = R(\tau\ (p''_2)) = C(\tau\ (p''_3)) = R(C(\tau(p''_4)))$. Therefore p''_1 = ATGGTGCACCT is original single strand of DNA sequence. Comparing the above two cases, can be easily found that the results are similar, in other words $p'_1 = p''_1$, $p'_2 = p''_2$, $p'_3 = p''_3$ and $p'_4 = p''_4$.

As seen in the example above, this method gives the same result for each k.

4. Conclusion

We give a new method to read a DNA sequence using concepts of graph theory, which work in DNA sequencing with nanopores. Compared with the method proposed by Bokhari [22] our method is simpler. In their method, first we are supposed to find a suitable k and then construct a de bruijn graph with k-oligonucleotides of a DNA sequence, but our method is independent of the length of oligonucleotides (k), in other words, this method does not require a specific amount for k. on the other hand steps of the algorithm according to this method will be less than the previous methods. These advantages make our method simpler and more useful.

Acknowledgement

The authors would like to thank the referee for the valuable comments.

References

[1] J. D. Watson, F. H. C. Crick, Molecular structure of nucleic acids - A structure for deoxyribose nucleic acid, *Nature* **171** (1953) 737–738.

[2] R. Wu, Q. Hu, R. Li, G. Yue, A novel composition coding method of DNA sequence and its application, *MATCH Commun. Math. Comput. Chem.* **67** (2012) 269–276.

[3] X. Zhou, K. Li, M. Goodman, A. Sallam, A novel approach for the classical Ramsey number problem on DNA-based supercomputing, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 347–370.

[4] Q. Zhang, B. Wang, On the bounds of DNA coding with H-distance, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 371–380.

[5] Q. Zhang, B. Wang, X. Wei, Evaluating the different combinatorial constraints in DNA computing based on minimum free energy, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 291–308.

[6] Y. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 477–488.

[7] R. Wu, R. Li, B. Liao, G. Yue, A novel method for visualizing and analyzing DNA sequences, *MATCH Commun. Math. Comput. Chem*. **63** (2010) 679–690.

[8] W. Chen, B. Liao, Y. Liu, W. Zhu, Z. Su, A numerical representation of DNA sequences and its applications, *MATCH Commun. Math. Comput. Chem*. **60** (2008) 291–300.

[9] V. Aram, A. Iranmanesh, 3D-dynamic representation of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **67** (2012) 809–816.

[10] J. Pesek, A. Zerovnik, Numerical characterization of modified Hamori curve representation of DNA sequences, *MATCH Commun. Math. Comput. Chem*. **60** (2008) 301–312.

[11] Y. Zhang, W. Chen, New invariant of DNA sequences*, MATCH Commun. Math. Comput. Chem*. **58** (2007) 197–208.

[12] N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons, *MATCH Commun. Math. Comput. Chem*. **68** (2012) 611–620.

[13] J. Yu, J. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation*, MATCH Commun. Math. Comput. Chem*. **63** (2010) 493–512.

[14] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, *MATCH Commun. Math. Comput. Chem*. **56** (2006) 209–216.

[15] R. M. Ldury, M. S. Waterman, A new algorithm for DNA sequencing assembly, *J. Comput. Biol.* **2** (1995) 291–306.

[16] P. A. Pevzner, DNA physical mapping and alternating Eulerian cycles in colored graphs, *Algorithmica* **13** (1995) 77–105.

[17] P. A. Pevzner, H. Tang, M. S, Waterman, A new approach to fragment assembly in DNA sequencing, *RECOMB* **1** (2001) 256–267.

[18] G. M. Church, D. W. Deame, D. Branton, R. Baldarelli, J. Kasianowicz, Characterization of individual polymer molecules based on monomer-interface interactions, *U. S. Patent* (1998) 5795782.

[19] J. Kasianowicz, E. Brandin, D. Branton, D. W. Deamer, Characterization of individual polynucleotide molecules using a membrane channel, *Proc Natl Acad Sci USA* **93** (1996) 13770–13773.

[20] J. D. Watson, F. H. C. Crick, Molecular structure of nucleic acids, *Am. J. Psych.* **160** (2003) 623–624.

[21] J. Kaptcianos, A graph theoretical approach to fragment assembly, *Am. J. Undergrad. Res.* **7** (2008) 311–329.

[22] S. H .Bokhari, J. R. Sauer, A parallel graph decomposition algorithm for DNA sequencing with nanopores, *J. Bioinform.* **21** (2005) 889–896.

[23] C. Li, N. Tang, J. Wang, Directed graphs of DNA sequences and their numerical characterization, *J. Theor. Biol*. **241** (2006) 173–177.

[24] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem*. **63** (2010) 493–512.

[25] J. Blazewicz, A. Hertz, D. Kobler, On some properties of DNA graphs, *Discr. Appl. Math*. **98** (1999) 1–19.

[26] J. Wang, C. Xu, A further study on DNA graphs and DNA labeling graphs, *Acta Math. Appl. Sin.* **33** (2010) 982–989 (in Chinese).

[27] J. Hao, The adjoints of DNA graphs, *J. Math. Chem.* **37** (2005) 333–346.

[28] X. Li, H. Zhang, Characterizations for some types of DNA graphs, *J. Math. Chem.* **42** (2007) 65–79.

[29] S. Wang, J. Yuan, DNA computing of directed line-graphs, *MATCH Commun. Math. Comput. Chem*. **56** (2006) 479–484.

[30] P. Sa-Ardyen, N. Jonoska, Self-assembling DNA graphs, *Nat. Comput.* **2** (2003) 427–438.

[31] R. Pendavingh, P. Schuurman, G. Woeginger, Recognizing DNA graphs is difficult, *Discr. Appl. Math.* **127** (2003) 85–94.

[32] J. Blazewicz, M. Bryja, M. Figlerowicz, P. Gawron, M. Kasprzak, E. Kirton, D. Platt, J. Przybytek, A. Swiercz, L. Szajkowski, Whole genome assembly from 454 sequencing output via modified DNA graph concept, *Comput. Biol. Chem.* **33** (2009) 224–230.

[33] S. Y. Wang, J. Yuan, S. Lin, DNA labelled graphs with DNA compuing, *Sci. China Ser. A: Math.* **51** (2008) 437–452.

[34] G. Chartrand, L. Lesniak, *Graphs and Digraphs*, Wadsworth & Brooks, Monterey, 1986.

[35] P. A. Pevzner, H. Tang, M. S. Waterman, An Eulerian path approach to DNA fragment assembly, *Proc. Nat. Acad. Sci.* **98** (2001) 9748–9753.