MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 731-742

ISSN 0340 - 6253

A New Graph Theoretical Method for Analyzing DNA Sequences Based on Genetic Codes

Nafiseh Jafarzadeh, Ali Iranmanesh*

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

(Received November 2, 2015)

Abstract

In this work, we use some concepts of graph theory to propose an alignment-free method for DNA sequence similarity analysis based on genetic codes. Our approach gives an effective and unique representation for each DNA sequences. In addition, according to our method, we give a numerical characterization of DNA sequences. This characterization facilitates quantitative comparisons of similarities/dissimilarities analysis of DNA sequences based on codons.

1. Introduction

With the exponential growth of biological sequence data, DNA sequence analysis has become an essential task for biologist to understand the features, functions, structures, and evolution of species. Determination of sequence similarity is one of the major steps in computational phylogenetic studies. Analyzing DNA sequences is a fundamental starting point for understanding biological functions. However, we know that it is very difficult to obtain biological information directly from large DNA sequences. Due to the level of complexity, mathematical analysis of large volumes of sequence data is at present a challenge

* Corresponding author
E-mail address: iranmanesh@modares.ac.ir

for bio-scientist. DNA is a nucleic acid that contains the genetic instructions used during the development and functioning of all known living organisms. DNA is a polymer, the monomer units of DNA are nucleotides. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one-letter abbreviations as shorthand for the four bases: A is for adenine, G is for guanine, C is for cytosine, and T is for thymine.

Nowadays, one of the most optimistic fields of mathematics is biomathematics. There are several biological problems that can be treated with mathematical methods. One of the most important problems is sequence (DNA or protein) analysis and comparison. For example in [1–31] you can see some mathematical methods as graphical and numerical representations for similarity analysis of DNA sequences. In bioinformatics, the most popular tools for comparing sequences are alignment methods. A sequence alignment is a way of arranging the sequence of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Alignment-free sequence comparison is frequently used to compare genomic sequences and in particular, gene regulatory regions. Gene regulatory regions are generally not highly conserved making alignment based methods for the identification of gene regulatory regions less efficient [32]. Alignment-free sequence comparison has a relatively long history starting in the mid-1980s [33], see for example the review in [34]. Most of the alignment-based methods compared with alignment-free methods take more computational time. Moreover, another advantage of alignment-free methods to be noted is their sensitivity against short or partial sequences [35].

In 2008, Pesek [31] has presented a new numerical characterization of DNA sequences that was based on the modified graphical representation proposed by Hamori [14], they have used analogous embedding into the strong product of the graphs K_4 and P_n. Although, this approach has merits but this product do not give a unique graph for each sequences. In this paper, we will give a new graph theoretical approach for comparing DNA sequences based on codons. Our method is an alignment-free method and compared with alignment-based is more controllable and it is computationally convenient with large biological databases. Although the graphs in this paper are large, but we extract simple paths from large graphs and give a numerical representation. Also, comparing with other models, our approach has the following preferences:

(a) This method give a representation as a directed and weighted graph for DNA sequences which we will prove that these graphs are unique for each sequence.

(b) We will show that our method is useful for analyzing DNA sequences with having a multiset of all 3- long oligonucleotides of a DNA sequence

2. New graph theoretical method

In this section, at first, we give the definition of DNA graphs and we construct a particular subgraph of these graphs. Then we give the definition of the lexicographic product of two graphs and we discuss about the detail of our method.

Definition 2.1. [36]. 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) $li(x) \in \{A,C,T,G\}$, 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 that 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 instance, 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:

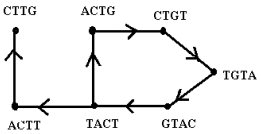

Figure 1. DNA graph of "S" with k = 4

Now, we construct a new DNA graph by another approach that presented by Pevzner [36] 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:

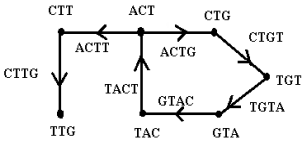

Figure 2. The new DNA graph of Figure 1

Easily seen that, line digraph of this new DNA graph is the DNA graph that is made by the previous approach which is shown in Figure1. In this paper, when we say DNA graph, we mean the DNA graph which is made by second approach.

Let $S = s_1\ s_2\ \ldots.\ s_n$ be a DNA sequence with "*n*" nucleotides. In the following, we wish to construct a particular subgraph of DNA graph according to "S" with $k = 3$ and we call it G_s. In fact the vertices of G_s will be dinucleotides.

For constructing G_s, at first, we find all types of dinucleotides which exist in "S" and put them as the vertices of this graph. In other words, the vertices of G_s are all of the dinucleotides of "S" without repetition, thus the maximum number of vertices in G_s is 16. After that, according to sequence S, we connect vertex $s_i s_{i+1}$ to vertex $s_{i+1}s_{i+2}$by an arc, for each $1 \leq i \leq n\text{-}2$.

Now, we give an example:

Suppose S = TGTGCA, the G_s of this sequence is presented in the following.

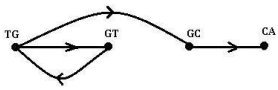

Figure 3. The G_s of S = TGTGCA

In continue, we need the definition of lexicographic product of graphs.

Definition 2.2. [37].The lexicographic product G ∘ H of two graphs G and H is defined on V (G ∘ H) = V (G) × V (H), two vertices (u, x) and (v, y) of G ∘ H being adjacent whenever uv∈ E(G), or u = v and xy ∈ E(H).

In the following, according to the above definition, we propose a weighted and directed graph for each DNA sequence based on triplets that we call it, T-graph (as all of the vertices of this graph are triplets).

Let S = s_1 s_2 …. s_n be a DNA sequence which s_i is a nucleotide for every $1 \leq i \leq n$. Now according to the sequence "S", we consider the graph G_s and also the complete and directed graph K_4 on four vertices A,C,T and G, then we obtain the lexicographic product of these two graphs (i.e., $G_s \circ K_4$).Note that since G_s and $K_4$are directed graphs, then the result of $G_s \circ K_4$ is a directed graph too. For instance, the T-graph of sequence S is presented in Figure 4.

Just as observe in Figure 4, each vertex of T-graph is a triplet; in fact, each vertex is a particular combination of a dinucleotide from G_s and a nucleotide from K_4.

Now, for making the T-graph as a weighted graph, using our previous work [30], we assign a 3D coordinate point to each vertex of T-graph. It has two cases:

1. If the triplet corresponding to a vertex of T-graph is the *i*th codon in the sequence, the coordinate of this vertex is (x,y,*i*),and
2. If the triplet corresponding to the vertex is not a codon in sequence, the coordinate of vertex is (x,y,0), which (x,y) is the 2D coordinate of this triplet as introduced in [29] and Figure 5.

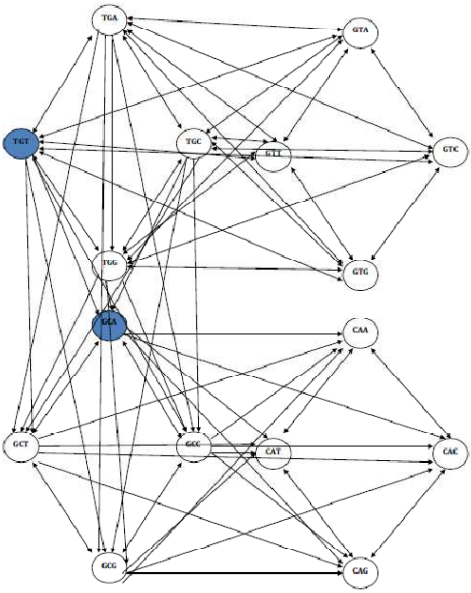

Figure 4. TheT-graph of S = TGTGCA

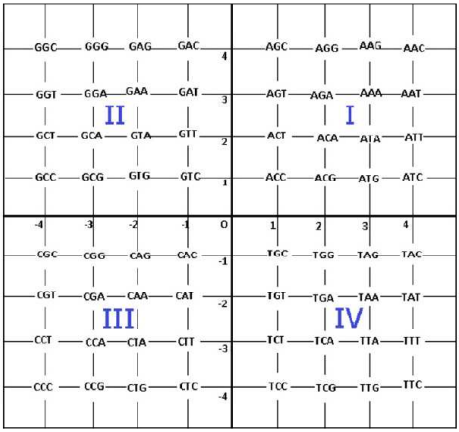

Figure 5. Sixty-four kinds of triplets distributed in Cartesian 2D coordinate

Then we assign a weight to each edge of T-graph by calculating the Euclidean distance between two points corresponding to the vertices of an edge.

Now, we want to discuss about an important property of the T-graph, which made our method powerful, and without error. At first, we give a theorem about lexicographic product and by using this theorem, we prove a particular theorem for T-graphs.

Theorem 2.1. [37]. Let X, Y, A and B be graphs. If $X \circ Y \cong A \circ B$ and $|Y| = |B|$, then $Y \cong B$ and $X \cong A$.

Theorem 2.2. The T-graph for each DNA sequence is unique.

Proof. Let S and S´ be DNA sequences which have the same T-graph, then we will show that S=S´. The T-graph of S is $G_s \circ K_4$ and the T-graph of S´ is $G_{s'} \circ K_4$, then $G_s \circ K_4 = G_{s'} \circ K_4$, it means that; $G_s \circ K_4 \cong G_{s'} \circ K_4$and Both of $G_s \circ K_4$ and $G_{s'} \circ K_4$ have the same labeling for their vertices and also the same directs of their edges. Therefore, by Theorem 1.2, we conclude that $G_s \cong G_{s'}$ and also G_s and $G_{s'}$ have the same labeling, the same directs and the same edges weight, accordingly, we have $G_s = G_{s'}$. Thus, according to the structure of these two overlapping graphs(G_s and $G_{s'}$) as a result S = S´. ■

In [38], we proposed a new approach for DNA sequencing by concepts of graph theory. In that method when we choose k = 3, in fact, the Eulerian path from DNA graph of all 3-long oligonucleotides from the original string of sequence S is the graph G_s. Therefore, according to our previous work [38], we can use our new method for analyzing DNA sequences, which the sequence of them is not determined.

3. Numerical characterization of DNA sequences

In this section, we give a numerical representation according to T-graph. For this purpose, we extract a particular directed path from T-graph of sequence S and we call it C-path. In fact, C-path is shortest path according to the weights of edges that include all the codons of sequence S. Note that in C-path, we have consider the position of codons in succession.

For instance, Figures 6, 7, and 8 show the C-path extracted from T-graph of sequences S_1 =ATGGTGCACCTG,S_2=ATGGTGCACCGT and S_3 =ATGCTGCAGTTG.

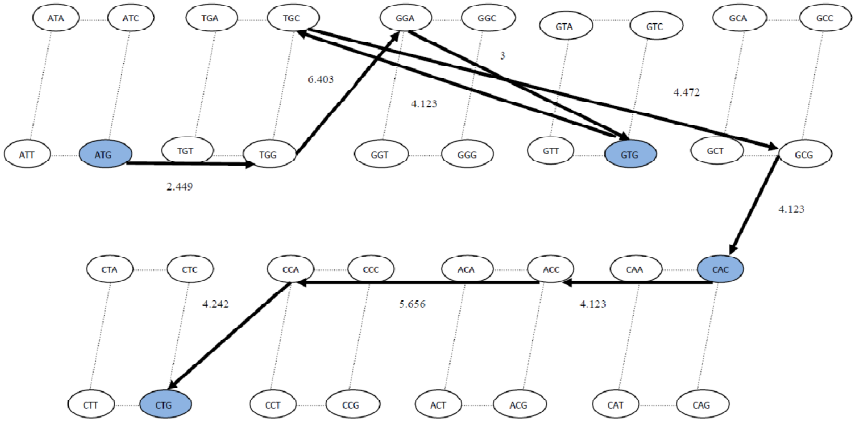

Figure 6. The C-path of S_1 = ATGGTGCACCTG

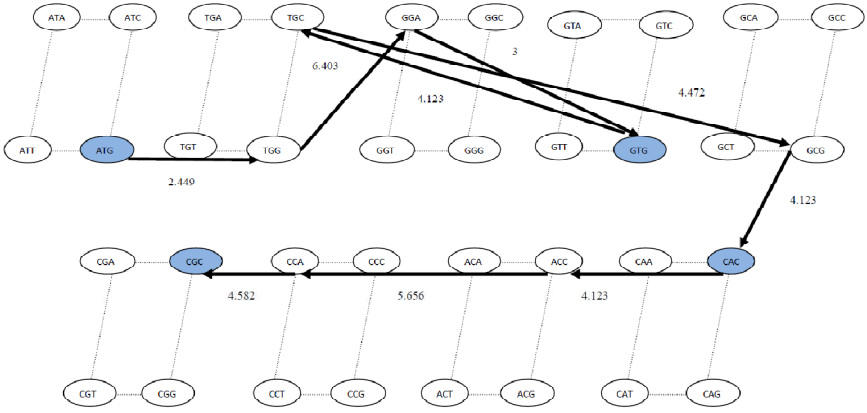

Figure 7. The C-path of S_2 = ATGGTGCACCGT

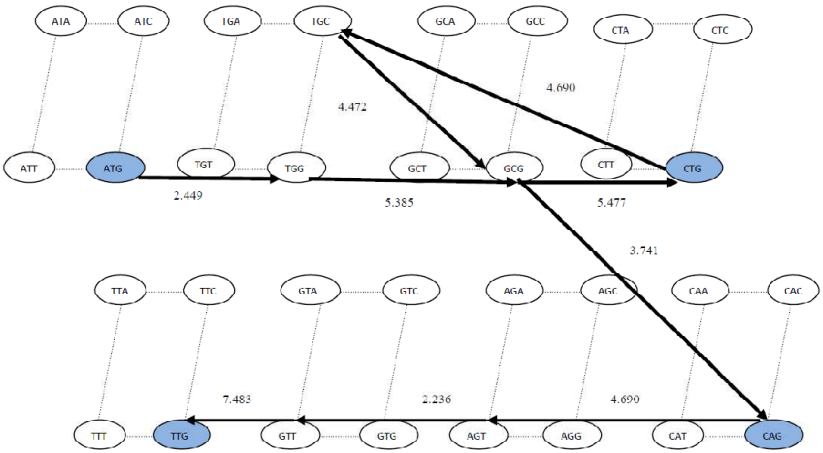

Figure 8. The C-path of S_3 = ATGCTGCAGTTG

In order to numerically characterize of a DNA sequence based on our approach, one can associate with a corresponding C-path, a matrix and consider a matrix invariant. One of the possible matrices is the D/D matrix (distance/distance matrix) whose elements are defined as the quotient of the Euclidean distance and the topological distance between a pair of the vertices of a graph. Here, we use analogous matrix based on the weighted graph, i.e., instead of Euclidean distance between a pair of vertices, we put the sum of weights between them.

One of the important invariants for DNA sequences is leading eigenvalue (λ) and another invariant is ALE-index (χ). ALE-index is defined by Li and Wang [22], which is an invariant of DNA sequences:

Let $M = (a_{ij})_{n\times n}$ be such a matrix, with the following property:

$a_{ij} \geq 0$, $a_{ij} = a_{ji}$, and $a_{ii} = 0$ for i,j = 1,2,...,n.

The ALE-index of M is defined as follows:

$\chi = \chi(M) = \frac{1}{2}\left(\frac{1}{n}\|M\|_{m1} + \sqrt{\frac{n-1}{n}}\|M\|_F\right)$, where $\|.\|_{m1}$ and $\|.\|_F$ are the m_1- and F-norms of a matrix, respectively.

According to [29], clearly we can show that the ALE-index is compatible as a sequence invariant for the D/D matrix of C-path too. Therefore, by using this invariant, we obtain a numerical representation for our graph theoretical method.

Finally, we exemplify our approach by the above three sequences S_1, S_2 and S_3.

Table 1. The similarity/dissimilarity matrix for the sequences S_1, S_2 and S_3

	Sequence	S_1	S_2	S_3
S_1	0.85731.367			
S_2	0	30.510		
S_3	0			

In Table 1, we give the similarity/dissimilarity matrix. The smallest entry in this table is associated with the pair (S_1,S_2) and the largest entry in the matrix appear in column belonging to S_3.

4. Conclusion

In this paper, we proposed a new graph theoretical method for DNA sequence similarity analysis based on codons. The basis of our method is using the lexicographic product of an overlapping graph, which its vertices are dinucleotides, and a complete graph that its vertices are nucleotides. The result of this product is a unique weighted graph for each DNA sequence, which its vertices are triplets. Furthermore, we discussed why we choose the lexicographic product to build a new graph. Then, according to this graph, we extracted a directed path for each sequence based on codons and we gave a numerical characterization of DNA sequences. Our method in this paper is an alignment-free method and compared with alignment-based is more controllable. In addition, comparing with other models, our approach has the following preferences:

(a) This method gave a representation as a directed and weighted graph for DNA sequences, which we proved that this graphs are unique for each sequence.

(b) We showed that our method is useful for analyzing DNA sequences with having a multiset of all 3- long oligonucleotides of a DNA sequence.

Acknowledgment: The authors would like to thank to referee for the valuable comments. This research is partially supported by Iran National Science Foundation (INSF) (Grant No. 93036169).

References

[1] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem*. **258** (1983) 1318–1327.
[2] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **119** (1986) 319–328.
[3] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–313.
[4] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci*. **11** (1995) 503–507.
[5] C. T. Zhang, R. Zhang, H. Y. Ou, The Z-curve databases: a graphic representation of genome sequence, *Bioinformatics* **19** (2003) 593–599.
[6] R. Zhang, C. T. Zhang, Z curve, an intuitive tool for visualizing and analyzing the DNA sequences, *J. Biomol. Struct. Dyn.* **11** (1994) 767–782.
[7] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences, *ARKIVOC* **9** (2006) 211–238.
[8] B. Liao, T. M. Wang, New 2D graphical representation of DNA sequences, *J. Comput. Chem.* **25** (2004) 1364–1368.
[9] B. Liao, W. Zhu, Y. Liu, 3D Graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 209–216.
[10] B. Liao, A 2D graphical representation of DNA sequence, *Chem. Phys. Lett.* **401** (2005) 196–199.
[11] M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.
[12] 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.
[13] M. Randić, A. T. Balaban, On a four-dimensional representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci*. **43** (2003) 532–539.
[14] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem*. **258** (1983) 1318–1327.
[15] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Curr. Sci*. **70** (1996) 661–668.
[16] A. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, *Chem. Phys. Lett.* **368** (2003) 102–107.
[17] X. Guo, X. Nandy, Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, *Chem. Phys. Lett.* **369** (2003) 361–366.
[18] M. Randić, M. Vračko, On the similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci*. **40** (2000) 599–606.

[19] 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.
[20] X. Zhou, K. Li, M. Goodman, A. Sallam, A novel approach for the classical Ramseynumber problem on DNA-based supercomputing, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 347–370.
[21] Q. Zhang, B. Wang, On the bounds of DNA coding with H-distance, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 371–380.
[22] C. Li, J. Wang, New invariant of DNA sequences, *J. Chem. Inf. Model*. **45** (2005) 115–120.
[23] Y. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol*. **242** (2006) 382–388.
[24] J. Feng, Y. Hu, P. Wan, A. Zhang, W. Zhao, New method for comparing DNA primary sequences based on a discrimination measure, *J. Theor. Biol*. **266** (2010) 703–707.
[25] X. Q. Liu, Q. Dai, Z. Xiu, T. Wang, PNN-curve: A new 2D graphical representation of DNA sequences and its application, *J. Theor. Biol*. **243** (2006) 55–56.
[26] P. He, D. Li, Y. Zhang, X. Wang, Y. Yao, A 3D graphical representation of protein sequences based on the Gray code, *J. Theor. Biol*. **304** (2012) 81-87.
[27] J. F. Yu, X. Sun, J. H. Wang, TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, *J. Theor. Biol*. **261** (2009) 459–468.
[28] X. Q. Qi, J. Wen, Z. H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, *J. Theor. Biol*. **249** (2007)681–690.
[29] 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.
[30] N. Jafarzadeh, A. Iranmanesh, C-curve: A novel 3D graphical representation of DNA sequence based on codons, *Math. Biosci*. **241** (2013) 217–224.
[31] I. Pesek, J. Zerovnik, A numerical characterization of modified Hamori curve representation of DNA sequences, *MATCH Commun. Math. Comput. Chem*. **60** (2008) 301–312.
[32] A. Ivan, M. S. Halfon, S. Sinha, Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs, *Genome Biol*. **9** (2008) #1.
[33] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Nat. Acad. Sci. USA* **83** (1986) 5155–5159.
[34] S. Vinga, J. Almeida, Alignment-free sequence comparison: a review, *Bioinformatics* **19** (2003) 513–523.
[35] L. W. Parfrey, J. Grant, Y. I. Tekle, E. Lasek-Nesselquist, H. G. Morrison, M. L. Sogin, D. J. Patterson, L. A. Katz, Broadly sampled multigene analyses yield a well resolved eukaryotic tree of life, *Syst. Biol*. **59** (2010) 518–533.
[36] S. Y. Wang, J. Yuan, S. Lin, DNA labelled graphs with DNA computing, *Sci. Chin. Ser. A: Math*. **51** (2008) 437–452.
[37] W. Imrich, S. Klažvar, *Product Graphs: Structure and Recognition*, Wiley, New York, 2000.
[38] N. Jafarzadeh, A. Iranmanesh, A new graph theoretical approach to DNA sequencing with nanopores, *MATCH Commun. Math. Comput. Chem*. **70** (2013) 401–415.