STUDIA UBB CHEMIA, LXI, 1, 2016 (p. 9-16)
(RECOMMENDED CITATION)

Dedicated to Professor Mircea Diudea on the Occasion of His 65th Anniversary

APPLICATION OF GRAPH THEORY TO BIOLOGICAL PROBLEMS

NAFISEH JAFARZADEH[a] AND ALI IRANMANESH[a,*]

ABSTRACT. In this paper, we investigate application of graph theory to some biological problems, specially reconstructing strings based on information about their substrings and sequence comparison by using overlap graphs and also weighted directed graph.

Keywords: *Fragment assembly, Overlap graphs, sequence comparison, Alignment-free method, Weighted directed graph.*

INTRODUCTION

Since the helical structure of DNA was proposed, many problems about this structure are posed. One of the important problems is how to read and recognize primary structure of a DNA sequence. DNA fragment assembly is a newly explored method of determining whether or not a reassembled strand of DNA, matches the original strand. One particular way to perform this method is by using concepts from graph theory. For instance see [1-4]. In order to begin the graph theoretical phase, one needs a directed graph which is built from some k-long oligonucleotides. Fragment assembly is the reconstructing of a string by using a subset of its substrings. In this method, a given piece of DNA (or rather many identical copies of it) is broken into several smaller fragments. The goal is to reconstruct the original DNA string based on the fragments. Therefore we have the following problem: Given a multiset of fragments, construct the best string that contains each

[a] *Department of Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, P.O. Box: 14115-137, Tehran, Iran*
** Corresponding Author: Iranmanesh@modares.ac.ir*

of the strings in the multiset. A string with that property is called a superstring of multiset. Here we want to find a superstring of minimal length, called the shortest common superstring. A mathematical tool to model the problem is a graph that represents the overlap between the strings in the set F. The overlap graph of the multiset contains a node for each string in it. The solution of this problem leads to find a Hamiltonian path. Note this is an instance of the Traveling Salesman Problem, which in its general form is NP complete. Even the existence of a Hamiltonian path in a graph is an NP complete problem that there are no efficient algorithms to solve the problem. But if we can consider fragments as edges of an overlap graph, this solution leads to find an Eulerian path and as we know, solving this problem is not NP complete and there are some efficient algorithms to solve the problem. Some overlap graphs have been proposed for reconstructing DNA sequences. In this paper, by the concept of DNA graph we propose an approach of creating and analyzing a DNA graph and its role in aiding DNA fragment assembly and determining the DNA structure. Graph theory and its concepts are used for application in many fields of study, for example see [5-8]. Also it will be studied how DNA graphs and Eulerian circuits are useful to resolve some problems in DNA fragment assembly.

Biological sequence analysis and comparison is another important problem in bioinformatics which replies to the emergence and need for the analysis of different types of data generated through biological research. Molecular sequence and structure data of DNA, RNA and proteins, gene expression profiles or micro array data, metabolic pathway data are some of the major types of data being analyzed in bioinformatics. Among them, sequence data are increasing at the exponential rate due to advent of next-generation sequencing technologies. Since the origin of bioinformatics, sequence analysis has remained the major area of research with wide range of applications in database searching, genome annotation, comparative genomics, molecular phylogeny and gene prediction. The pioneering approaches for sequence analysis were based on sequence alignment, global or local, pairwise or multiple sequence alignment. Alignment-based approaches generally give excellent results when the sequences under study are closely related and can be reliably aligned, but when the sequences are divergent, a reliable alignment cannot be obtained and hence the applications of sequence alignment are limited. Another limitation of alignment-based approaches is their computational complexity when dealing with large-scale sequence data. The advent of next generation sequencing technologies has resulted in generation of voluminous sequencing data. The size of this sequence data poses challenges on alignment-based algorithms in their assembly, annotation and comparative studies. Thus, alignment-free sequence analysis approaches provide attractive

alternatives over alignment-based approaches. In this paper, we will discuss about applications of product of graphs and overlap graphs to compare DNA sequences based on an alignment-free method.

FRAGMENT ASSEMBLY

In graph theory, an *n*-dimensional De Bruijn graph of *m* symbols is a directed graph representing overlaps between sequences of symbols. It has m^n vertices, consisting of all possible length-*n* sequences of the given symbols; the same symbol may appear multiple times in a sequence. If we have the set of *m* symbols $S = \{s_1, \ldots, s_m\}$ then the set of vertices is:

$$V = S^m = \{(s_1,\ldots, s_1, s_1),(s_1,\ldots, s_1, s_2),\ldots,(s_1,\ldots, s_1, s_m),(s_1, \ldots, s_2, s_1),\ldots,(s_m, \ldots, s_m, s_m)\}$$

If one of the vertices can be expressed as another vertex by shifting all its symbols by one place to the left and adding a new symbol at the end of this vertex, then the latter has a directed edge to the former vertex. Thus the set of arcs (directed edges) is:

$$E = \{((v_1, v_2,\ldots, v_n),(v_2,\ldots, v_n, v_i)) : i = 1,\ldots,m\}.$$

If we consider a set of 4 symbols {A, T, C, G} we get the definition of DNA graph:

Definition 1. [9]. 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 a 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 $(l_2(x),\ldots l_k(x)) = (l_1(y),\ldots l_{k-1}(y))$.

Now, we construct a DNA graph by the approach which was presented by Pevzner [2] 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, a DNA graph with k = 4 is shown in Figure 1. To find the primary sequence, we need to find an Eulerian path.

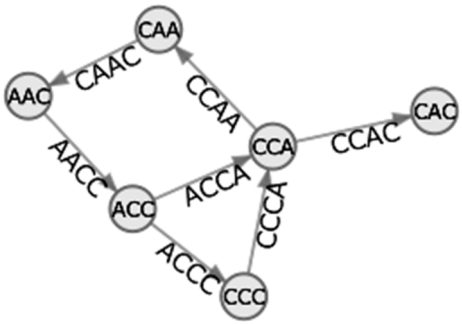

Figure 1. The DNA graph for the sequence “ACCCAACCAC”

Euler’s theorem for directed graphs gives conditions for the existence of an Eulerian path. Define for vertex *v, in(v)* = |{u: uv ∈ E}|, *and out(v)* = *|{w: vw ∈ E}|*.

Label the starting vertex *s* and the terminal vertex *t.* There is an Eulerian path starting at *s* and ending at *t,* if and only if

$$in(v) = out(v) \quad \forall v \neq s, t, \; out(s) - in(s) = 1 \text{ and } out(t) - in(t) = -1$$

Since, each Eulerian path corresponds to a different DNA sequence, we can infer the sequence unambiguously if and only if the number of Eulerian paths in G is exactly one. We can reduce the problem of determining the number of Eulerian paths from *s* to *t* in G to the problem of determining the number of Eulerian cycles in the graph *G* U *ts*, which has a simple solution described below. We define an intersection graph on the cycles of G U *ts* as follows. First decompose *G* U *ts* into simple cycles: v_{i1}, v_{i2}, ... , $v_{ik} = v_{i1}$, where no $v_i = v_j$ except for $v_{ik} = v_{i1}$. An edge can be used in at most one cycle C but a vertex can be used in arbitrarily many cycles. For these cycles, define the intersection graph G_I .of the cycles C_1, C_2, ... , C_m , where cycles correspond to vertices, and if cycles C_i and C_j have *I* vertices in common, we connect them by *I* edges in G_I. The following theorem gives the necessary and sufficient conditions for general graphs.

Theorem 1 [10]**:** There is a unique Eulerian cycle in G if and only if the intersection graph G_I of simple cycles from G is a tree.

Also, there exists a formula for computation of Eulerian circuits in a directed graph. Named after its inventors, de Bruijn, van Aardenne-Ehrenfest, Smith, and Tutte, [11-12], the BEST Theorem reads as follows:

Theorem 2 [13]**:** Given a connected directed graph G and a set of vertices $V(G) = \{v_1,\ldots,v_n\}$ all of even degree, the number of Eulerian circuits $|s(G)|$ is expressed as the following, where $|t_i(G)|$ is the number of spanning trees rooted towards any vertex v_j in G and $d^+(v_j)$ is in-degree of v_j:

$$|s(G)| = |t_i(G)| \prod_{j=1}^{n} (d^+(v_j) - 1)!.$$

There is a theorem about the number of spanning trees in a graph, showing that this number can be computed in polynomial time as the determinant of a matrix derived from the graph. This theorem is named "The matrix tree Theorem":

Theorem 3 [13]**:** Given a directed graph G with the set of vertices $V(G) = \{v_1,\ldots,v_n\}$ and a set of spanning trees $t_i(G)$ oriented towards the vertex v_i, then $|t_i(G)|$ is equal to the cofactor of L(G) (Laplacian matrix of G) on the i-th row and i-th column.

After creating a DNA graph for a multiset of DNA fragments, we need to examine the Eulerian path of this graph. By using the Eulerian graph theorem we make a directed Eulerian graph from DNA graph. Then we use the BEST theorem and the matrix tree theorem for counting Eulerian circuits in DNA graph. It leads us to determine primary DNA sequence.

Now, let $S = n_1\ n_2\ \ldots.\ n_L$ be a DNA sequence and *M* be a multiset of all *k*-long oligonucleotides of a this sequence, then we construct a DNA graph for S and call it G^k_s. The vertices of this graph will be (K -1) oligonucleotides and according to the Pevzner's approach [2], there is an Eulerian path which reveals the primary sequence S and we call P^k_s.

According to the graph theoretical method in [4], in whole genome sequencing with fragment assembly, there are two different strings to be recognized. These are the original 3'-5' string and its complement, in this method, assuming that M be a multiset of all *k*-long oligonucleotides of a complete DNA sequence (S), then according to above approach, G^k_s have been constructed, may be disconnected, but this graph includes two connected DNA graphs belonging to the original 3'-5' string and its complement and each graph includes an Eulerian path and these paths determine the structure of primary DNA sequence.

SEQUENCE COMPARISON

In this section we discuss about the application of graph theory to compare DNA sequences. Let us to mention some concepts in graph theory. By a *graph* we mean a set *V*(*G*) of vertices, together with a set *E*(*G*) of edges. A

graph is the *complete graph* K_n if any two of its distinct vertices are adjacent, and a graph is the *path* P_n if it is isomorphic to a graph on n distinct vertices $v_1, v_2, \ldots, v_n$ and $n-1$ edges v_i, v_{i+1}, $1 \le i < n$.

Definition 2. [14]. The strong product $G \boxtimes H$ of graphs G and H is a graph such that

- the vertex set of $G \boxtimes H$ is the Cartesian product $V(G) \times V(H)$; and
- any two distinct vertices (u,u') and (v,v') are adjacent in $G \times H$ if and only if:
    - u is adjacent to v and $u'=v'$, or
    - $u=v$ and u' is adjacent to v', or
    - u is adjacent to v and u' is adjacent to v'

In [5], Pesek presented a new numerical characterization of DNA sequences that is based on the modified graphical representation proposed by Hamori [15]. He used analogous embedding into the strong product of graphs, $K_4 \boxtimes P_n$, with weighted edges. Based on this representation, a novel numerical characterization is proposed which is based on the products of ten eigenvalues from the start and the end of the descending ordered list of the eigenvalues of the *L/L* matrices associated with DNA.

Now, we can give a new approach to sequence comparison by using pesek's approach. According to the pervious section, let k >1 be an integer and $S = n_1\ n_2\ \ldots.\ n_L$ be a DNA sequence and M be a multiset of all k-long oligonucleotides of a this sequence; we consider $K_4 \boxtimes P^k{}_s$ as a new graph which just is made by multiset of all k-long oligonucleotides of a sequence and we can continue the details of pesek's method for comparing sequences.

Another approach that applies concepts of graph theory for sequence comparison is proposed by Qi et al, [16]. They constructed novel mathematical descriptors based on graph theory, for each DNA sequence, they sat up a weighted directed graph. The adjacency matrix of the directed graph will be used to induce a representative vector for DNA sequence. This new approach measures similarity based on both ordering and frequency of nucleotides so that much more information is involved.

They have shown how to construct the weighted directed multi-graph for $S = s_1\ s_2\ \ldots\ s_n$, which is denoted by $G_m = (V(G_m), A(G_m))$. The vertex set $V(G_m) = \{A, C, G, T\}$. For each pair of nucleotides s_i and s_j in S with i , j, put an arc from s_i to s_j , and define a special weight of that.

Theorem 4 [16]. It is a one-to-one mapping between a DNA sequence S and its corresponding weighted directed multi-graph G_m.

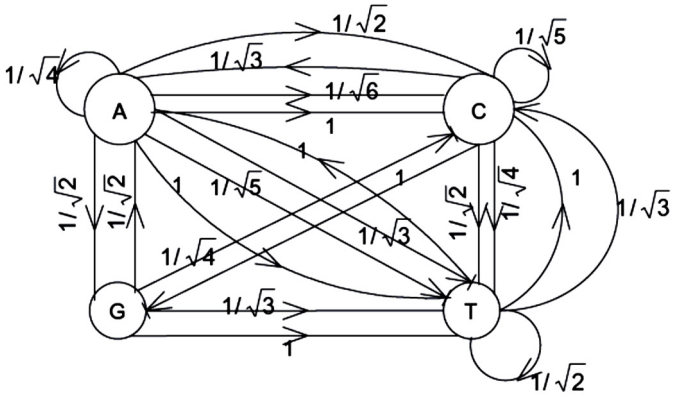

Figure 2. Directed multi-graph Gm for S = ACGTATC

CONCLUSIONS

Our purpose in this paper was to investigate application of some concepts in graph theory to sequencing and sequence comparison. At first, we used a multiset of fragments of a given DNA sequences and applied Pevzner's approach to achieve an overlap graphs for which the finding of superstring of minimal length is not NP-complete. Then we utilized the Eulerian path of this graph to construct a new graph by using strong product of graphs for comparing and analysing DNA sequences.

ACKNOWLEDGMENTS

This research is partially supported by Iran National Science Foundation (INSF) (Grant No. 93036169).

R E F E R E N C E S

1. R.M. Ludry, M.S. Waterman, A new algorithm for DNA sequencing assembly. *J. Comput. Biol.,* **1995**, *2*, 291.
2. P.A. Pevzner, H. Tang, M.S. Waterman, A new approach to fragment assembly in DNA sequencing. *RECOMB*., **2001**, *1*, 256.
3. 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.,* **2009**, *33*, 224.

4. N. Jafarzadeh, A. Iranmanesh, Application of DNA graphs to whole genome sequencing, *International Journal of Green Nanotechnology*, **2014**, *2*, 373.
5. J. Pesek, A. Zerovnik, Numerical Characterization of Modified Hamori Curve Representation of DNA Sequences, *MATCH Commun. Math. Comput. Chem*., **2008**, *60*, 301.
6. Y. Zhang, W. Chen, New Invariant of DNA Sequences*, MATCH Commun. Math. Comput. Chem*., **2007**, *58*, 197.
7. N. Jafarzadeh, A. Iranmanesh, A Novel Graphical and Numerical Representation for Analyzing DNA Sequences Based on Codons, *MATCH Commun. Math. Comput. Chem*., **2012**, *68*, 611.
8. J. Yu, J. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation*, MATH Commun. Math. Comput. Chem*., **2010**, *63*, 493.
9. J. Blazewicz, A. Hertz, D. Kobler, On some properties of DNA graphs, *Discrete Appl Math*., **2003**, *1*, 98.
10. P.A. Pevzner, *l*-tuple DNA sequencing: a computer analysis, *J. Biota. Struct. Dyn.,* **1989**, *7*, 63.
11. T. van Aardenne Ehrenfest, N.G. de Bruijn, Circuits and Trees in Oriented Linear Graphs, *Simon Stevin*., **1951**, *28*, 203.
12. W.T. Tutte, C.A.B. Smith, On Unicursal Paths in a Network of Degree 4, *Amer. Math. Monthly,* **1941**, *48*, 233.
13. J. Kaptcianos, A graph theoretical approach to fragment assembly, *American Journal of Undergraduate Research*, **2008**, *7*, 311.
14. W. Imrich, S. Klavžar, "Product Graphs: Structure and Recognition", John Wiley & Sons, New York, **2000**.
15. E. Hamori, Graphical representation of long DNA sequences by methods of H curves, current results and future aspects, *Biotechniques*, **1989**, *7*, 710.
16. X. Qi, Q. Wu, Y. Zhang, E. Fuller, C. Zhang, A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory, *Evolutionary Bioinformatics,* **2011**, *7*, 149.