ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

A BRIEF REVIEW-GRAPH THEORY CONTRIBUTION TO DNA/RNA SEQUENCES AND ITS STRUCTURES

M. Yamuna and K. Karthika
SAS, VIT University, Vellore, Tamilnadu, India, 632 014.
Email:myamuna@vit.ac.in, karthika.k@vit.ac.in

Received on 15-02-2016 *Accepted on 22-03-2016*

Abstract

Graph theory has established itself as a very strong, effective tool in molecular biology, by its simplicity of representation and its compatibility in computation. In this short survey we have presented a brief note on the contribution of graph theory to DNA and RNA sequences and its structures.

Keywords: DNA**,** Graph, Molecular biology, RNA.

Introduction

Ever since the discover of chemical trees by A. Cayley in [1]. The contribution of graph theory to molecular structures is uncountable. Graph theory as established itself as a unique tool in determining various biological properties due to its ease of representing DNA, RNA structures, protein sequences, metabolic networks, protein – protein interaction networks, genotype, geneticists and hexagonal system. Properties of graph theory like vertex/edge connectivity, blocks, planarity, edge removal and vertex merging has turned graph theory into a very skilful technique in molecular biology. The compatibility of graphs as adjacency, distance matrices has paved way to meet computational demands. Even though graph theory has developed tremendously as a mathematics subject, it is also established as a strong tool in biology. In the short survey, we have tried to highlight the contributions of graph theory in DNA/RNA sequences and structures.

There are numerous results relating graph theory and molecular biology. But we restrict our survey only to contribution of graph theory to DNA and RNA sequences and series. Several results are available in this regard, and many have been omitted in this brief survey. We apologize to the authors for the omission. This survey is restricted to presentation of possible results that describe the effects of the graph theory properties on DNA/RNA sequences

Preliminary Note: In this section we provide the basic details required for this survey.

DNA Sequence

A nucleic acid that carries the genetic information in cells and some viruses, consisting of two long chains of nucleotides twisted into a double helix and joined by hydrogen bonds between the complementary bases adenine and thymine or cytosine and guanine. DNA sequences are replicated by the cell prior to cell division and may include genes, intergenic spacers, and regions that bind to regulatory proteins [2].

RNA Sequence

A nucleic acid present in all living cells and many viruses, consisting of a long, usually single-stranded chain of alternating phosphate and ribose units, with one of the bases adenine, guanine, cytosine, or uracil bonded to each ribose molecule. RNA molecules are involved in protein synthesis and sometimes in the transmission of genetic information [3].

Gap Penalty

Gap penalty values are designed to reduce the score when a sequence alignment has been disturbed by indels. Typically the central elements used to measure the score of an alignment have been matches, mismatches and spaces. Another important element to measure alignment scores are gaps.

A gap is a consecutive run of spaces in an alignment and is used to create alignments that are better conformed to underlying biological models and more closely fit patterns that one expects to find in meaningful alignments. Gaps are represented as dashes on a protein/DNA sequence alignment. The length of a gap is scored by the number of indels (insertions/deletions) in the sequence alignment. In protein and DNA sequence matching, two sequences are aligned to determine if they have a segment each that is significantly similar.

A local alignment score is assigned according to the quality of the matches in the alignment subtracted by penalties for gaps present within the alignment. The best gap costs to use with a given substitution matrix are determined empirically.

Gap penalties are used with local alignment that match a contiguous sub-sequence of the first sequence with a contiguous sub-sequence of the second sequence.

When comparing proteins, one uses a similarity matrix which assigns a score to each possible residue. The score should be positive for similar residues and negative for dissimilar residues pair. Gaps are usually penalized using a linear gap function that assigns an initial penalty for a gap opening, and an additional penalty for gap extensions which increase the gap length [4].

Genetic Code

The genetic code is the set of rules by which information encoded within genetic material (DNA or mRNA sequences) is translated into proteins by living cells. Biological decoding is accomplished by the ribosome, which links amino acids in an order specified by mRNA, using transfer RNA (tRNA) molecules to carry amino acids and to read the mRNA three nucleotides at a time. The genetic code is highly similar among all organisms and can be expressed in a simple table with 64 entries [5].

Graph Theory Terminology and Concepts

Let G = (V, E) be a graph with the vertex set V and edge set E. P_n, C_n denotes the path and cycle graph with n vertices respectively. A directed graph (or digraph) is a graph, or set of vertices connected by edges, where the edges have a direction associated with them. A graph is connected when there is a path between every pair of vertices. A weighted graph is a graph in which each edge is given a numerical weight. The line graph of an undirected graph G is another graph L(G) that represents the adjacencies between edges of G.

A planar graph is a graph that can be embedded in the plane, that is it can be drawn on the plane in such a way that its edges intersect only at their endpoints. In other words, it can be drawn in such a way that no edges cross each other. The geometric dual graph of G obtained for a given embedding of G in the plane. A Hamiltonian path is a path in an undirected or directed graph that visits each vertex exactly once. A Hamiltonian cycle (or Hamiltonian circuit) is a Hamiltonian path that is a cycle. An isomorphism of graphs G and H is a bijection between the vertex sets of G and H, f: V (G) $\rightarrow$ V (H) such that any two vertices u and v of G are adjacent in G if and only if f (u) and f (v) are adjacent in H.

Two vertices u and v are said to be identified if they are combined into a single vertex whose neighborhood is the union of the neighborhoods of u and v. The binary operation merge of two graphs G_1 and G_2 forms a new graph G_{uv} by identifying a vertex u in G_1 with a vertex v in G_2.

An adjacency matrix of a graph G with n vertices that are assumed to be ordered from v_1 to v_n is defined by,

$$A = [a_{ij}]_{n\times n} = \begin{cases} 1, \text{ if there exist an edge between } v_i \text{ and } v_j \\ 0, \text{ otherwise.} \end{cases}$$

.

A degree matrix D of a graph G with n vertices is a n × n diagonal matrix defined as

$$d_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

where the degree deg (v_i) of a vertex counts the number of times an edge terminates at that vertex.

Given a simple graph G with n vertices, its Laplacian matrix $L_{n \times n}$ is defined as L = D – A, where D is the degree matrix and A is the adjacency matrix of the graph.

In the case of directed graphs, either the in degree or out degree might be used, depending on the application. The elements of are given by

$$L_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjcent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Where deg (v_i) is degree of the vertex i. For details of on graph theory we refer to [6].

A dominating set, denoted by DS, of G is a set of vertices of G such that every vertex in V – D is adjacent to a vertex in D. The domination number of G, denoted by γ (G), is the minimum cardinality of a DS. The cardinality of any minimum dominating set for G is called the domination number of G and it is denoted by γ (G). γ - set denotes a dominating set for G with minimum cardinality.

A set D is a total dominating set if N (D) = V, if for every vertex v $\in$ V, there is a vertex u $\in$ D, u $\neq$ v, such that u is adjacent to v. The total domination number γ_t (G) equals the minimum cardinality of a total dominating set of G. A dominating set D is called a locating – dominating set if for any two vertices v, w $\in$ V – D, N (v) $\cap$ D $\neq$ N (w) $\cap$ D. Thus, in a locating dominating set, every vertex in V – D is dominated by a distinct subset of the vertices of S. The locating domination number of a graph G is the minimum cardinality among all locating dominating sets in G and is denoted by γ_L (G). A dominating set D is called a differentiating dominating set if for any two vertices v, w $\in$ V, N [v] $\cap$ D $\neq$ N [w] $\cap$ D. The differentiating domination number of a graph G is the minimum cardinality among all differentiating dominating sets in G and is denoted by γ_D (G). The global alliance number of a graph G is the minimum cardinality among all global alliances of G, where a set D is a global alliance if D is a dominating set and for each u $\in$ D, the number of "allies" it has in D are at least as many as it has in V – D. In other words,

D is a dominating set and for each vertex u $\in$ D, it is true that | N [u] $\cap$ D | $\geq$ | N (u) $\cap$ (V – D) |. For details of on domination theory we refer to [7] [8].

D is a dominating set and for each vertex u $\in$ D, it is true that | N [u] $\cap$ D | $\geq$ | N (u) $\cap$ (V – D) |. For details of on domination theory we refer to [7] [8].

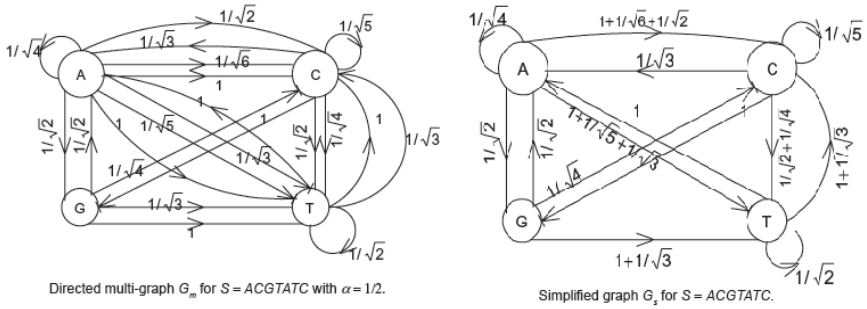
Results and Discussions

In this section, we provide the contribution of graph theory to DNA/RNA sequences and its structures.

Graph Theory in DNA

A DNA sequence has four nucleotides A, T, G, C. Surprisingly researches have independently used these nucleotides as the vertices of graphs constructed by them for various purposes.

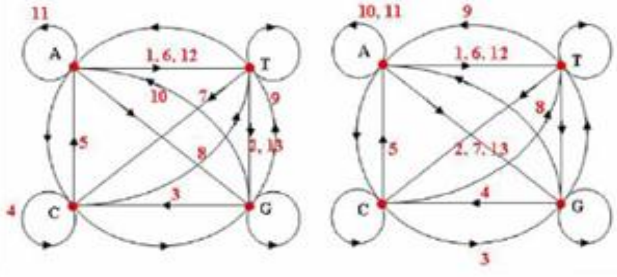
In [9], a weighted directed graph with these four vertices is constructed edges for this graphs is defined as (1 / (j – i)$^{\alpha}$), where $\alpha > 0$. The Snapshot 1 is a sample of the constructed graph when $\alpha = 1 / 2$.



Snapshot 1

Xingqin Qi et al have further devised a simplified weighted directed graph. This graph is and then converted into a 4 × 4 adjacency matrix. This method is then used for characterization of DNA sequences. The results obtained here are consistent with previous studies and biological classification.

In [10], the same vertex set is used for determining DNA gap penalty. Here also weighted graph is constructed. For edge determination a random DNA sequence is considered. This sequence is split into segments of length two and assigned values 1, 2, …, k – 1, if the given sequence of length k. Draw directed edges between these pair of vertices and assigned edge weight as 1, 2, …, k – 1. To determine gap penalty, the corresponding edge weights are compared between sequences. Same weights indicate same pairs. Snapshot 2 provides the gap penalty generated using this method. The penalty matches with the existing classical methods.

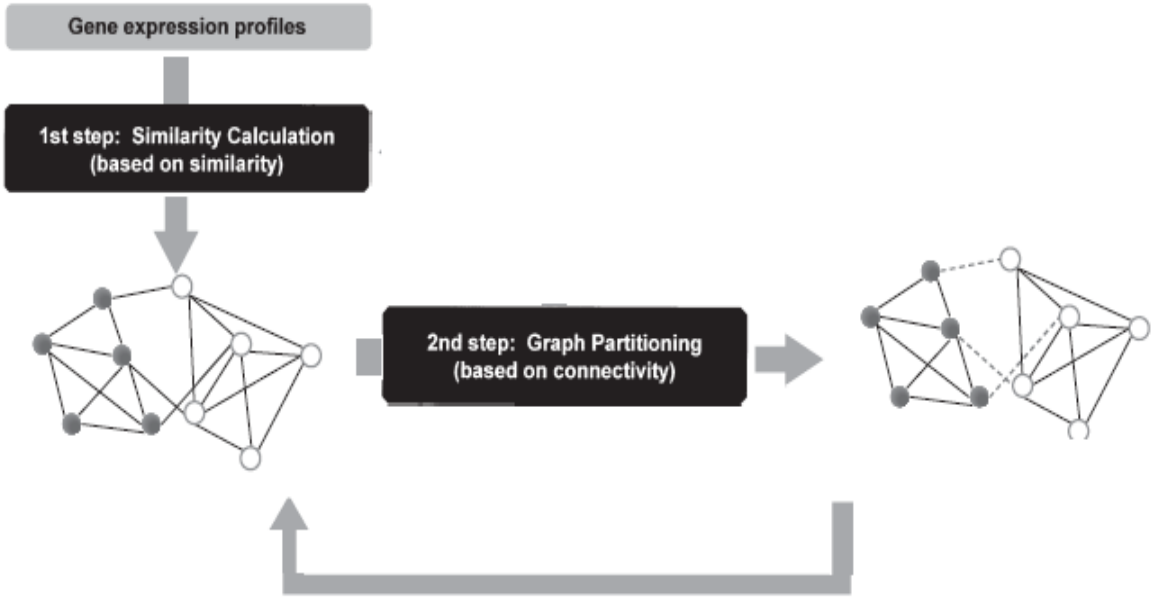
S. No	Sequences	digraphs	Matching weight List	Gap Penalty
1	ATGCCATCTGAATG ATCGCATCTAAATC		1,5,6,7,11,12	5

Snapshot-2

In [11], Takashi Kawamura et al have presented a graph based clustering model for analyzing publicly available micro array data sets. Each data set was divided into model sample data and behind sample data. Initially the graph was constructed based on the similarity of gene expression pattern. Pearson's product moment correlation coefficient was determined between the gene expression patterns. Later graph is partitioned based on connectivity (edge and vertex connectivity). Graph partition is as seen in Snapshot 3.

The classification model was constructed for four micro array datasets, leukemia, breast cancer, prostate cancer and colon cancer and the accuracies of classification with k – nearest neighbor were all more than 80%.

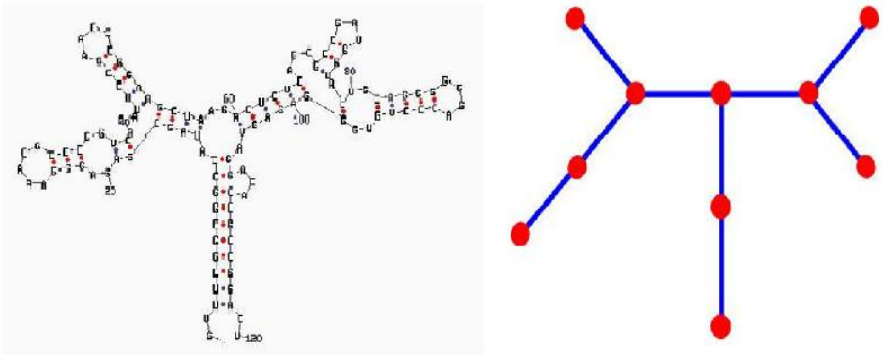
Any graph theorist knows that searching of Hamiltonian cycles in a digraph is strongly NP – complete. To determine the DNA sequence assembly J. Blazewicz et. al devised a method on building a multigrpah with vertices corresponding to input sequences not contained in others. Edges between pair of vertices correspond to possible overlaps of the sequences observing the assumed error bound. Procedure based on edge removal in this constructed graph is designed indetermining heuristic algorithm reducing graphs towards simplifying the Hamiltonian cycle problem, without losing any feasible solution [12].



Snapshot-3

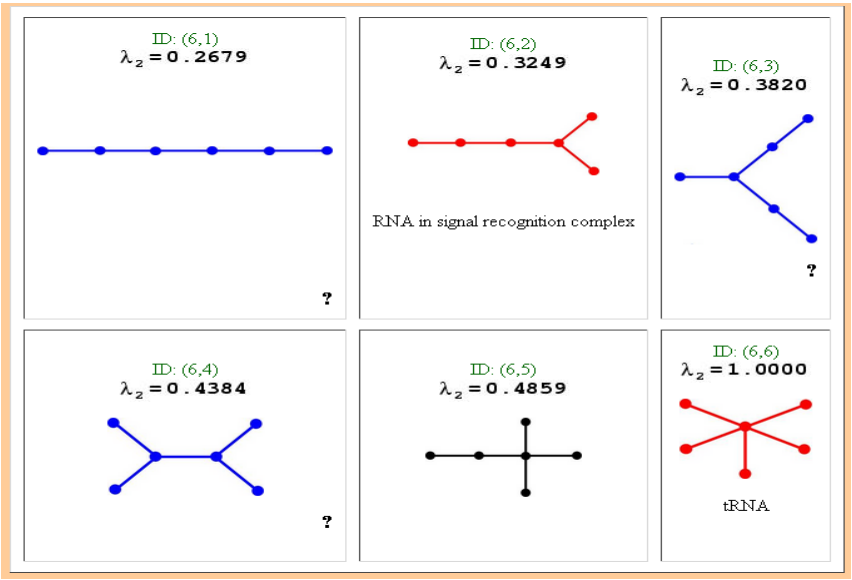
Graph Theory in RNA

In [13], Tamar Schlick et al have discussed a novel method of representing RNA tree graph. This RNA tree is developed a nucleotide bulge, junctions, hairpin loop, or internal loop is considered as a vertex. An RNA stem with more than one complementary base pair is considered an edge. A secondary RNA structure and resulting tree from RAG is provided in Snapshot 4.



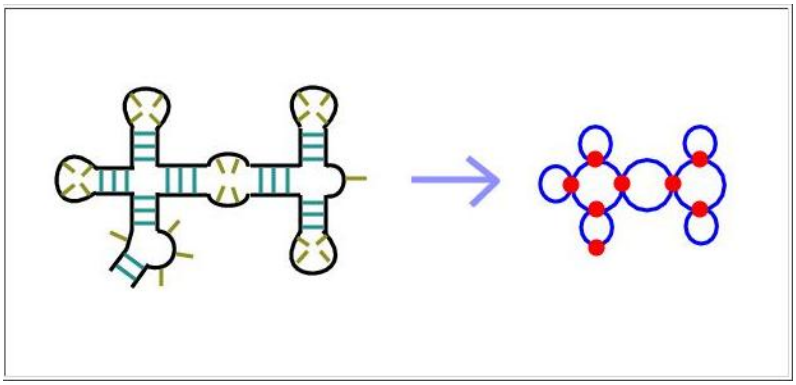
Snapshot-4

Tamar Schlick et al defined a Laplacian matrix, calculated the eigen value for this matrix and consider the second smallest eigen value. Using this RNA tree motifs which match with the RNA structures found a nature and not found a native nature are characterized. The Snapshot 5 is a sample of RNA trees with six vertices.



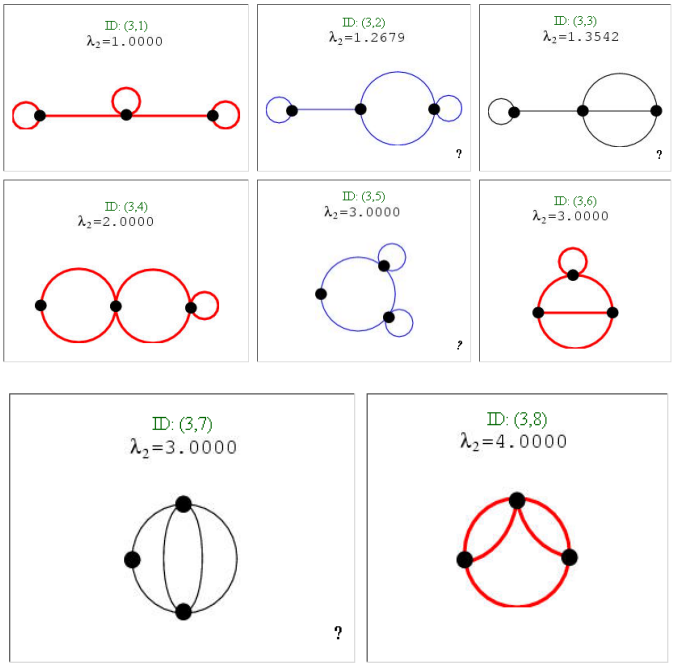
Snapshot-5

They have further improved this and extended then to dual graphs, where nucleotide bulges, hairpin loops, or internal loops with more than one non – complementary base pair or more than three unmatched nucleotides are represented by circular edges. A secondary RNA structure and resulting dual graph from RAG is seen in Snapshot 6.



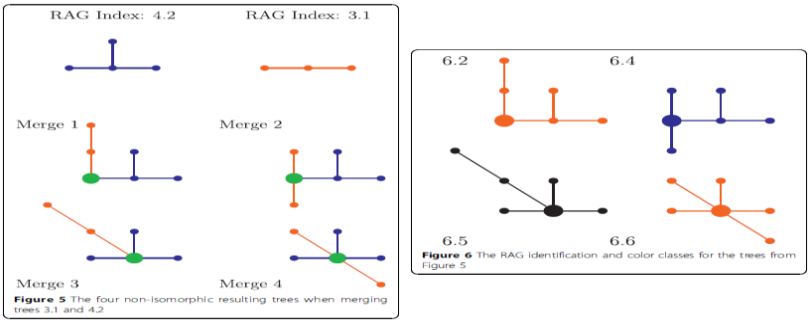
Snapshot-6

Similar to RNA trees, the Laplacian matrix is constructed, the second smallest eigen value is calculated and RNA graphs existing and not existing in nature are classified. The Snapshot 7 is a sample of RNA dual graphs with 3 vertices.



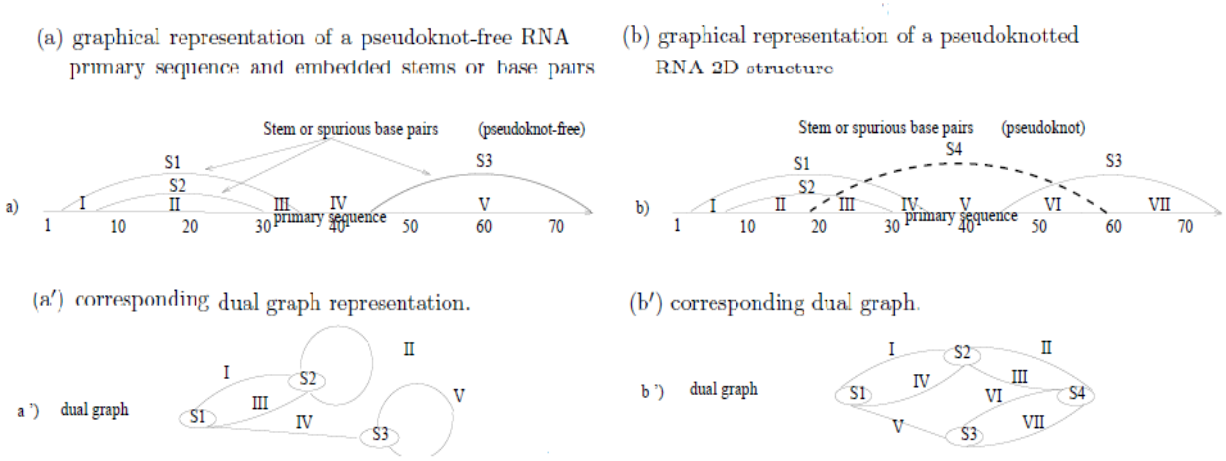
Snapshot-7

In Snapshot 5 and 7, Red graphs denote those structures found in nature and black graphs denote those that have either not yet been found or do not exist (a question mark is placed for these in the lower right corner). Blue graphs represent candidate RNA tree motifs (which have not yet been found in nature). Adopting the tree and dual structures used [13] in [14], by Samuela Pasquali et al. They have devised a technique for the occurrence frequency of the topological tree which has an advantage over sequence alignment since functionally related RNA of lack sequence similarity. In [15], D. R. Koessler et al have slightly modified the original procedure. Two trees are combined together using vertex merging. The resulting new tree is compare and identified with already existing trees classified by [13]. Four non – isomorphic trees and official RAG identification and color classes for the trees is seen in Snapshot 8.



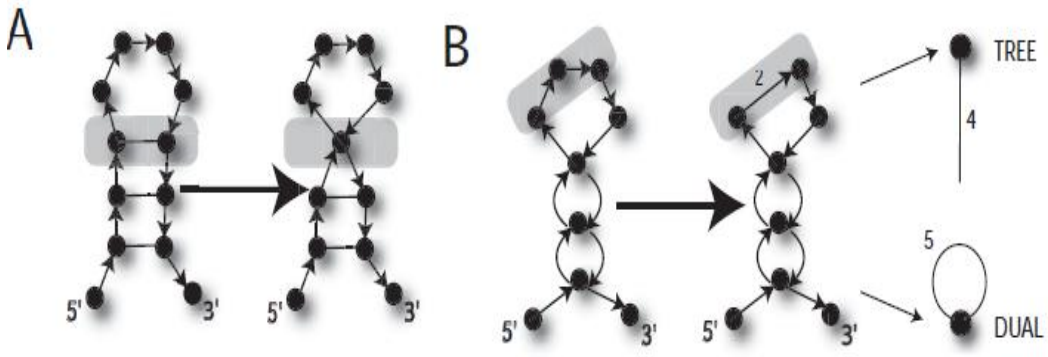
Snapshot-8

Contribution of graph theory to DNA, RNA is not restricted to just trees. Planar graphs and its dual have their contribution to model RNA secondary structures. A graph and dual graph representation of an RNA 2D – structure is determined in [16] by Louis Petingi et al. Graphical representation of a pseudoknot – free RNA and RNA 2D structure is provided in Snapshot 9.



Snapshot-9

Here Louis Petingi et al have provided a partitioning approach of the dual graph representation if RNA 2D structures into maximal non separable components which they believe could guide the discovery of modular regions of RNA. Graph vertex contraction as also contributed in RNA secondary structure representation. In [17] C. H. Q. Ding et al have provided a method of vertex contraction to represent different secondary structures. RNA contraction graph is seen in Snapshot 10.



Snapshot-10

A. The shaded base pair (left) is shown after contraction (right). Arcs are conserved during contraction along base pairs.

B. The shaded edges in the loop region (left) are shown after contraction (right). Flow conservation yields a weight of two on the resulting edge. Also note that all base pairs in the

helical region have been contracted.

Domination Theory in RNA

Surprisingly a dominating set is used in determining the same second smallest eigen value. For this purpose, the domination, total domination, global alliance, locating – dominating and differentiation domination numbers are used. In [18] by T. Haynes et al. have defined P_1, P_2 and P_2* as follows.

$$P_1 = \frac{\gamma + \gamma_t + \gamma_a}{n};\ P_1 = \frac{\gamma_L + \gamma_D}{n};\ P_2^* = \gamma_L + \gamma_D + n\lambda_2 .$$

Finally they have shown that these combination matches with the RAG database determined in [13].

Using the same formula for P_1, P_2 along with the new formula

$P_3 = \frac{\mathrm{diam}(L(T)) + \mathrm{rad}(L(T)) + |B|}{n}$ in [19] Haynes et al have proved that a similar method of classification yields similar results to the statistical analysis in RAG. In P_3, L (T) represents the line graph of the tree and | B | is the number of blocks in the line graph of the tree. Dominating sets have further aided in encrypting DNA sequences. A graph G is said to be domination subdivision stable (DSS), if the γ - value of G does not change by subdividing any edge of G [20]. An approach using DSS graphs is discussed in [21]. Here edge values for binary encryption based on DSS properties are discussed and finally each edge is assigned a binary string value. Using path P_4 a method of encryption and decryption of chemical formulae as a sequence of numbers and DNA sequence is given. Using similar methodology encryption of a medicine name of the chemical formula for the medicine name as a RNA sequence and as a binary numbers is also discussed by the same authors in [22]. The RNA binary conversion table is as shown in Snapshot 11. In [23], genetic code using binary and gray code is devised by Uday Bhaskar et al. They have created RNA/DNA basic coed table, RNA/ DNA gray code table are used for converting a DNA, RNA sequence into a binary string for a safe transfer. This method is independent of dominating sets. The basic code table for RNA is as shown in Snapshot 12.

Binary Conversion Table

	U	C	A	G	
U	UUU – 111111 UUC – 111101 UUA – 111100 UUG – 111110	UCU – 110111 UCC – 110101 UCA – 110100 UCG – 110110	UAU – 110011 UAC – 110001 UAA – 110000 UAG – 110010	UGU – 111011 UGC – 111001 UGA – 111000 UGG – 111010	U C A G
C	CUU – 011111 CUC – 011101 CUA – 011100 CUG – 011110	CCU – 010111 CCC – 010101 CCA – 010100 CCG – 010110	CAU – 010011 CAC – 010001 CAA – 010000 CAG – 010010	CGU – 011011 CGC – 011001 CGA – 011000 CGG – 011010	U C A G
A	AUU – 001111 AUC – 001101 AUA – 001100 AUG – 001110	ACU – 000111 ACC – 000101 ACA – 000100 ACG – 000110	AAU – 000011 AAC – 000001 AAA – 000000 AAG – 000010	AGU – 001011 AGC – 001001 AGA – 001000 AGG – 001010	U C A G
G	GUU – 101111 GUC – 101101 GUA – 101100 GUG – 101110	GCU – 100111 GCC – 100101 GCA – 100100 GCG – 100110	GAU – 100011 GAC – 100001 GAA - 100000 GAG – 100010	GGU – 101011 GGC – 101001 GGA – 101000 GGG – 101010	U C A G

Snapshot-11

Table 4 RNA Basic Code Table

	U		C		A		G		
U	UUU	00.111.11	UCU	00.131.10	UAU	00.101.10	UGU	00.121.10	U
	UUC	00.113.11	UCC	00.133.10	UAC	00.103.10	UGC	00.123.10	C
	UUA	00.110.11	UCA	00.130.10	UAA	00.100.00	UGA	00.120	A
	UUG	00.112.11	UCG	00.132.10	UAG	00.102.00	UGG	00.122.11	G
C	CUU	00.311.11	CCU	00.331.11	CAU	00.301.00	CGU	00.321.00	U
	CUC	00.313.11	CCC	00.333.11	CAC	00.303.00	CGC	00.323.00	C
	CUA	00.310.11	CCA	00.330.11	CAA	00.300.01	CGA	00.320.00	A
	CUG	00.312.11	CCG	00.332.11	CAG	00.302.01	CGG	00.322.00	G
A	AUU	00.011.11	ACU	00.031.10	AAU	00.001.01	AGU	00.021.10	U
	AUC	00.013.11	ACC	00.033.10	AAC	00.003.01	AGC	00.023.10	C
	AUA	00.010.11	ACA	00.030.10	AAA	00.000.00	AGA	00.020.00	A
	AUG	00.012.11	ACG	00.032.10	AAG	00.002.00	AGG	00.022.00	G
G	GUU	00.211.11	GCU	00.231.11	GAU	00.201.01	GGU	00.221.11	U
	GUC	00.213.11	GCC	00.233.11	GAC	00.203.01	GGC	00.223.11	C
	GUA	00.210.11	GCA	00.230.11	GAA	00.200.01	GGA	00.220.11	A
	GUG	00.212.11	GCG	00.232.11	GAG	00.202.01	GGG	00.222.11	G

Snapshot-12

Conclusion

In the course of this survey, it is exciting and surprising to see the traces that graph theory and graph theorists have left behind to molecular biology. Infinite contributions that a small structure could provide in analyzing, determining, deciding and paving way for a new area of research. In this short survey we could manage to provide a very small glimpse of the contribution of graph theory to DNA/ RNA sequences and series.

References

1. A. Cayley, On the Theory of the Analytical Forms Called Trees, Philosophical Magazine, 1857,Vol 13, pp 172 – 176.
2. http://www.thefreedictionary.com/DNA.
3. http://www.thefreedictionary.com/RNA.
4. F. Harary. Graph Theory, Addison Wesley/ Narosa Publishing House Reprint 1988, 10th reprint, 2001.
5. T.W. Haynes, S.M. Hedetniemi, S.T. Hedetniemi Domination and Independence Subdivision Numbers of Graphs, Discussiones Mathematicae, Graph Theory, 2000, Vol 20, pp 271 – 280.
6. T.W. Haynes, S.M. Hedetniemi, S.T. Hedetniemi, D. P. Jacobs, J. Knisely, L. C. V. D. Merwe Domination Subdivision Number, Discussiones Mathematicae, Graph Theory, 2001, Vol 21, pp 239 – 253.
7. Xingqin Qi, Qin Wu, Yusen Zhang, Eddie Fulle, Cun – Quan Zhang. A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory, Evolutionary Bioinformatics, 2011, Vol 7, pp 149 – 158.
8. M. Yamuna. DNA Gap Penalty Using Directed Graphs, Der Pharmacia Lettre, 2015, Vol 7, pp 392 – 398.

9. Takashi Kawamura, Hironori Mutoh, Yasuyuki Tomita, Ryuji Kato, Hiroyuki Honda. Cancer DNA Microarray Analysis Considering Multi – subclass with Graph – based Clustering Method, Journal of Bioscience and Bioengineering, 2008, Vol 106, pp 442 – 448.
10. J. Blazewicz, M. Kasprzak. Graph Reduction and its Application to DNA Sequence Assembly, Bulletin of the Polish Academy of Sciences, 2008, Vol 56, pp 65 – 70.
11. Samuela Pasquali, Hin Hark Gan, Tamar Schlick. Modular RNA Architecture Revealed by Computational Analysis of Existing Pseudoknots and Ribosomal RNAs, Nucleic Acids Research, 2005, Vol 33, pp 1384 – 1398.
12. Denise R Koessler, Debra J Knisley, Jeff Knisley, Teresa Haynes. A Predictive Model for Secondary RNA Structure Using Graph Theory and a Neural Network, BMC Bioinformatics, 2010, Vol 11, pp 1 – 10.
13. Chris H.Q. Ding, Richard F. Meraz, Xiaofeng He, Stephen R. Holbrook. Contraction Graphs for Representation and Analysis of RNA Secondary Structure, Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004) 0-7695-2194-0/04 $20.00 © 2004 IEEE.
14. T. Haynes, D. Knisley, E. Seier, Yue Zou. A Quantitative Analysis of Secondary RNA Structure Using Domination Based Parameters on Trees, BMC Bioinformatics, 2006, Vol 7, pp 1 – 11.
15. T. Haynes, D. Knisley, J. Knisley. Using a Neural Network to Identify Secondary RNA Structures Quantified by Graphical Invariants, Match Communications in Mathematical and in Computer Chemistry, 2008, Vol 60, pp 277 – 290.
16. M. Yamuna K. Karthika. Domination Subdivision Stable Graphs. International Journal of Mathematical Archive, 2012, Vol 3, pp 1467 – 1471.
17. M. Yamuna, K. Karthika. Chemical Formula: Encryption Using Graph Domination and Molecular Biology, International Journal of ChemTech Research, 2013, Vol 5, pp 2747 – 2756.
18. M. Yamuna, K. Karthika. Medicine Names as a DNA Sequence Using Graph Domination, Der Pharmacia Lettre, 2014, Vol 6, pp 175 – 183.
19. Uday Bhaskar, Adeesh Nagpal, Himanshu Paul, M Yamuna. Genetic Code as Binary BCD and Gray Code, Research Journal of Pharmaceutical, Biological and Chemical Sciences, 2014, Vol 5, pp 438 – 451.

Corresponding Author:
M. Yamuna*,
Email: *myamuna@vit.ac.in*