A CASE STUDY ON APPLICATIONS OF GRAPH THEORY IN BIOLOGICAL NETWORKS

Mallikarjun S Biradar
Dept. of Mathematics, Govt. First Grade College, Basavakalyan-585327
mallikarjunbiradar54@gmail.com

Abstract

Graph theory is a pivotal branch of mathematics with extensive applications across various disciplines, including computer science, biology, social sciences and more. Graph Theory is used in many areas of Biology. It can be used in drug target identification, determining a protein's function, gene's function. It is also used in studying the structures of DNA and RNA. In this case study, we concentrate on the features of biological networks. We demonstrate approaches, models and methods from the graph theory universe and we discuss ways in which they can be used to reveal hidden properties and features of a network. This network profiling combined with knowledge extraction will help us to better understand the biological significance of the system.

1. Introduction

In the present research study of Graph theory, a prominent and crucial branch of mathematics, explores the study of graphs, which are mathematical structures used to model pairwise relations between objects. Originating from Euler's solution to the Konigsberg bridge problem in 1736, graph theory has since evolved into a vital field of study with applications spanning computer science, biology, social sciences, and more.

In biology, Graph theory can model and analyze the spread of diseases or information within biological networks, helping in designing effective vaccination strategies or understanding ecological interactions.

This work aims to provide a comprehensive understanding of mathematical graph theory with a specific focus on biological networks. We will delve into the foundational concepts of graphs, including vertices, edges, paths, and cycles before progressing to the intricacies of biological networks.

Graphs are used to represent relationships among species on different physical and micro-biological criteria. For example, the evolutionary relationship among the existing species is expressed in a tree structure called phylogenetic tree [20].

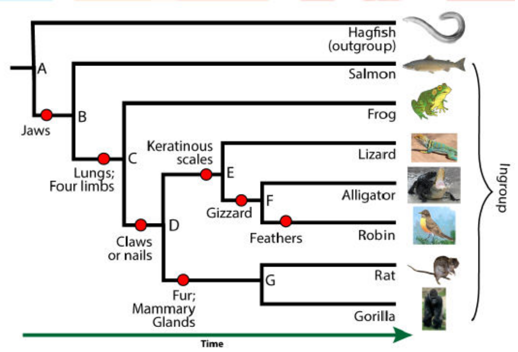


Figure: A Phylogenetic Tree

Graph is also used in analyzing biological data. Ecological landscapes can be modelled using graphs. Habitat patches are represented as vertices and the movement between the patches is represented as edges when ecological landscapes are modelled as graphs.

Similarly, Graph theory is useful in conservation efforts where a vertex represents region where certain species exist and the edges represent migration paths or movement between the regions. This information is important when tracking the spread of diseases, parasites and how the changes in the movement can affect other species.

2. Graph Theory and Definitions:

To introduce the basic concepts of graph theory, we give both the empirical and the mathematical description of graphs that represent networks as they are originally defined in the literature [23].

2.1 Undirected single graph

This is an important feature since there are networks such as protein-protein interaction networks in which two proteins might be evolutionary related, co-occur in the literature or co-express in some experiments, resulting by this way in three different connections, each one with a different meaning. An example of PPI database that takes into account the different types of interactions between proteins is String [12].

2.2 Directed graph

A directed graph is defined as an ordered triple G = (V, E, f), where f is a function that maps each element in E to an ordered pair of vertices in V. The ordered pairs of vertices are called directed edges, arcs or arrows. Directed graphs are mostly suitable for the representation of schemas describing biological pathways or procedures which show the sequential interaction of elements at one or multiple time points and the flow of information throughout the network. These are mainly metabolic, signal transduction or regulatory networks [9].

2.3 Weighted graph

A weighted graph is defined as a graph G = (V, E) where V is a set of vertices and E is a set of edges between the vertices E = {(u, v) | u, v Î V} associated with it a weight function w: E®R, where R denotes the set of all real numbers.

Weighted graphs are currently the most widely used networks throughout the field of bioinformatics. As an example, relations whose importance varies are frequently assigned to biological data to capture the relevance of co-occurrences identified by text mining, sequence or structural similarities between proteins or co-expression of genes [19, 22].

2.4 Bipartite graph is an undirected graph G = (V, E) in which V can be partitioned into two sets V_1 and V_2 such that (u,v) Î E implies either u Î V_1 and v Î V_2 OR v Î V_1 and u Î V_2. Applications of this type of graph to visualization or modelling of biological networks range from representation of enzyme-reaction links in metabolic pathways to ontologies or ecological connections.

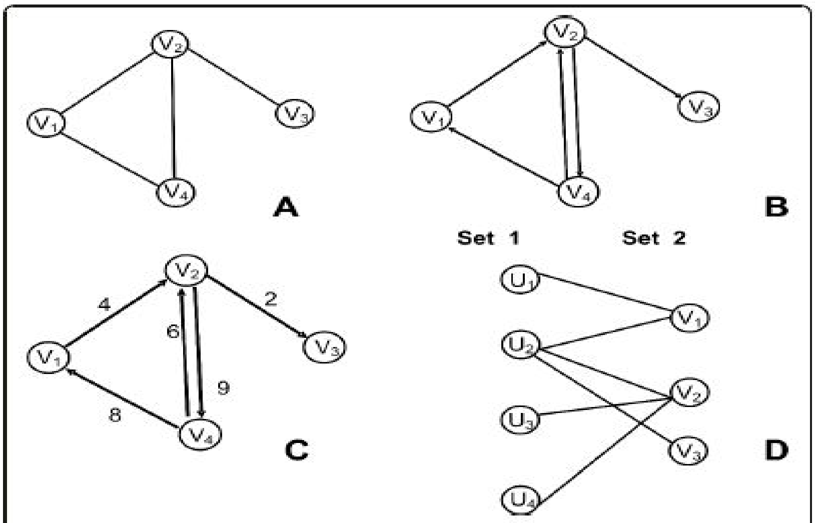


Figure 1 Undirected, Directed, Weighted, Bipartite graphs. A. Undirected Graph: $V = \{V_1, V_2, V_3, V_4\}$, $|V| = 4$, $E = \{(V_1, V_2), (V_2, V_3), (V_2, V_4), (V_4, V_1)\}$, $|E| = 4$. B. Directed Graph: $V = \{V_1, V_2, V_3, V_4\}$, $|V| = 4$, $E = \{(V_1, V_2), (V_2, V_3), (V_2, V_4), (V_4, V_1), (V_4, V_2)\}$, $|E| = 5$. C. Weighted Graph: $V = \{V_1, V_2, V_3, V_4\}$, $|V| = 4$, $E = \{(V_1, V_2, V_4), (V_2, V_3, V_2), (V_2, V_4, V_9), (V_4, V_1, V_8), (V_4, V_2, V_6)\}$, $|E| = 5$. D. Bipartite graph: $V = \{U_1, U_2, U_3, U_4, V_1, V_2, V_3\}$, $|V| = 7$, $E = \{(U_1, V_1), (U_2, V_1), (U_2, V_2), (U_2, V_3), (U_3, V_2), (U_4, V_2)\}$, $|E| = 6$.

Figure: Undirected, Directed, Weighted, Bipartite Graph

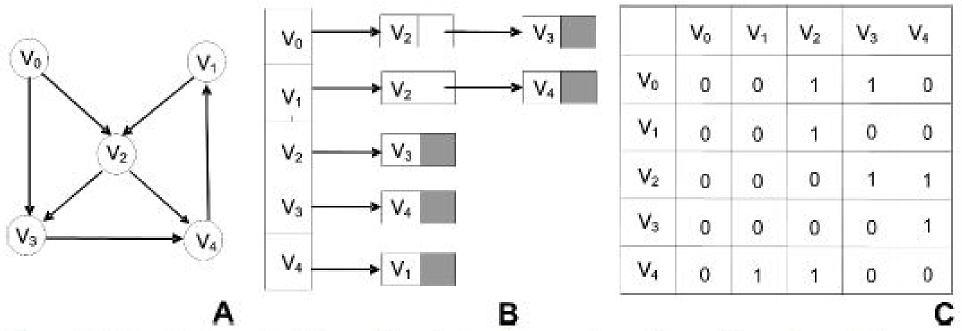
The total connectivity of a network is defined as $C = \dfrac{E}{N(N-1)}$ where E is the number of edges and N the total number of nodes. The connectivity structure of biological networks is often informative with respect to reaction interplay and reversibility, compounds that structure the network, like in metabolism, or trophic relationships, like in food-web networks. Such connectivity profiles can be detected based on mixture models using software like MixNet [6, 24].

2.5 Data Structures

The two main data structures used to store network graph representations are described. This data structure is more efficient for cluttered networks, where the density of the connections between elements is relatively high. In the case of a fully connected graph where all nodes are connected with each other, adjacency matrices are highly suggested.

2.6 Adjacency list

Given a graph G = (V, E) the adjacency list representation consists of an array Adj of |E| elements where for each eÎE Adj(0, e) = i ÎV. Adjacency lists require space Θ (|V| +|E|) and are preferable for sparse graphs with a low density of connections. An example of how these data structures represent a graph is given in Figure 1.2.



	V0	V1	V2	V3	V4
V0	0	0	1	1	0
V1	0	0	1	0	0
V2	0	0	0	1	1
V3	0	0	0	0	1
V4	0	1	1	0	0

Figure 1.2 **Data structures**. A. A Directed Graph: A random graph consisting of five nodes and six directed edges. B. Adjacency List: The data structure which represents the directed graph using lists. C. Adjacency Matrix: The data structure which represents the directed graph using a 2D matrix. The zeros represent the absence of the connection whereas the ones represent the existence of the connection between two nodes. The matrix is not symmetric since the graph is directed.

Figure: Data Structure

2.7 Network Properties

Looking at different network properties can provide valuable insight into the internal organization of a biological network, the repartition of molecules among cellular processes, as well as the evolutionary constraints that have shaped an organism's protein, metabolic or regulatory network into a functional, feasible structure. In the following, we give a short description of the main properties that are commonly analyzed in networks.

2.8 Graph Isomorphism

Let G1= (V1, E1) and G2= (V2, E2) be two undirected graphs. A function f: V1 >V2 is called isomorphism if f is an edge-preserving bisection, such that for all a, bÎV1, (a, b) ÎE1 if and only if (f(a), f(b)) Î E2. When such function exists, then G1 and G2 are called isomorphic. An example is shown in Figure 2.1.

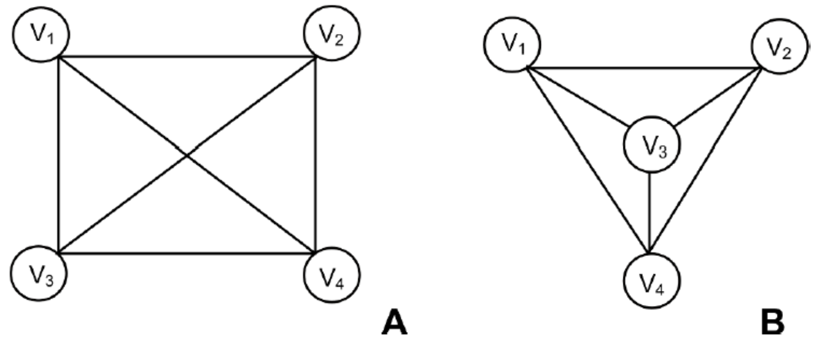


Figure 2.1 **Graph Isomorphism**. $V = \{V_1, V_2, V_3, V_4\}$, $|V| = 4$, $E = \{(V_1, V_2), (V_1, V_3), (V_1, V_4), (V_2, V_3), (V_2, V_4), (V_3, V_4)\}$, $|E| = 6$. Graphs *A* and *B* have different topology but they are isomorphs. The graph is fully connected and every node is connected to any other so that it forms a fully connected clique.

Figure: Graph Isomorphism

A walk is a pass through a specific sequence of nodes (v1, v2,..., vL) such that {(v1,v2), (v2, v3),..., (vL-1, vL)} ⊆ E. A simple path is a walk with no repeated nodes. A cycle is a walk (v1, v2,..., vL) where v1= vL with no other nodes repeated and L >3, such that the last node is the same with the first one. A trail is a path where no edge can be repeated. A graph is called cyclic if it contains a cycle. In any other case it is called acyclic. All of the aforementioned can be found as an example in Figure 4. A complete graph is a graph in which every pair of nodes is adjacent. If (i, j) is an edge in a graph G between nodes i and j, we say that the vertex i is adjacent to the vertex j. An undirected graph is connected if one can get from any node to any other node by following a sequence of edges.

A directed graph is strongly connected if there is a directed path from any node to any other node. This does not require an all-against combination. The distance δ(i, j) from i to j is the length of the shortest path from i to j in G. If no such path exists, then we set δ(i, j) = ∞ assuming that the nodes are so far between each other so they are not connected. Practically, for the distance δ(i, j) = ∞ we can. use the maximum weight of the graph by adding one. Thus δ(i, j) = ∞ = (maxd(i, j)+1). To define the shortest path problem we can briefly say that it is the methodology of finding a path between two nodes such that the sum of the weights of its constituent edges is minimized.

The average path length and the diameter of a graph G are defined to be the average and maximum value of δ(i, j) taken over all pairs of distinct nodes, i, j ÎV(G) which are connected by at least one path. More specifically, the average path length of a network is the average number of edges or connections between nodes, which must be crossed in the shortest path between any two nodes. It is calculated as $\delta = \frac{2}{N(N1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{\min}(i,j)$ where $\delta_{\min}$(i, j) is the minimum distance between nodes i and j. The diameter of a network is the longest shortest path within a network. The diameter is defined as D = max i,j $\delta_{\min}$(i, j). The most common algorithms for calculating the shortest paths are Dijkstra's greedy algorithm and Floyd's dynamic algorithm. Dijkstra's algorithm has running time complexity O(N2) where N is the number of vertices and returns the shortest path between a source vertex i and all other vertices in the network. Floyd's algorithm has running time complexity O(N3) and requires an all-against-all matrix that contains the distances of every node in the network to every other node in the network. [7]

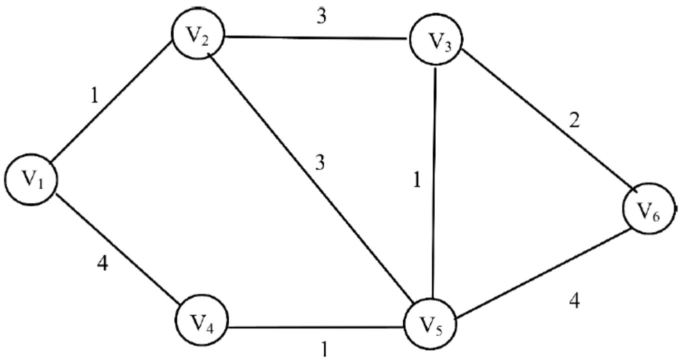


Figure: Walks, simple paths trails and cycles in graphs

3. Applications of Graph Theory in Biological system:

Biological networks come in a variety of forms. Nodes in biological networks represent bimolecular such as genes, proteins or metabolites, and edges connection these nodes indicate functional, physical or chemical interactions between the corresponding bimolecular. Understanding these complex biological systems has become an important problem that has lead to intensive research in network analyses, modelling and function and disease gene identification and prediction. The hope is that utilizing such system-level approaches to analyzing and modelling complex biological systems will provide insights into the inner working of the cell, biological function, and disease.

4. Transcriptional regulation networks:

In transcriptional regulation networks, nodes represent genes and edges are directed from a gene that encodes for a transcription factor protein to a gene transcriptional regulated by that transcription factor 9, see figure). Thus, the network structure in an abstraction of the system's biochemical dynamics that is responsible for regulating the expression of genes in cells. The two best characterized transcriptional regulation networks are those of a eukaryote, the yeast and a bacterium.

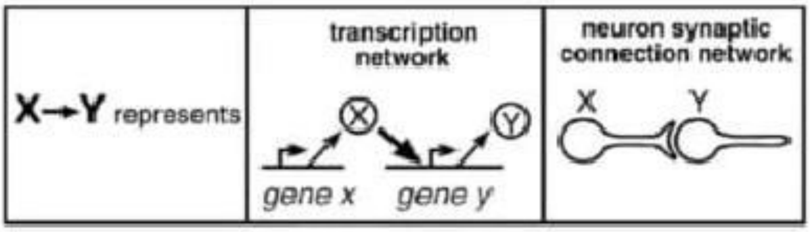


Figure B Two examples of biological network

5. Metabolic Networks

One of the most important life processes is the metabolism of an organism, the basic chemical system that generates essential components such as amino acids, sugars and lipids, and the energy required to synthesize them and to use them in creating proteins and cellular structures. A metabolic network represents this system of connected chemical reactions, i.e., the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. Metabolism network reconstruction breaks down metabolism pathways into their respective reactions and enzymes. Thus, in these networks, small molecule substrates can be envisioned as nodes and the links as the enzyme-catalyzed reactions that transform one metabolite into another, with the sequencing of complete genomes, it is now possible to reconstruct the network o biochemical reactions in many organisms, from bacteria to human. These networks are available in several databases, such as Kyoto Encyclopaedia of Genes and Genomes (KEGG).

Metabolic networks are powerful tools for studying and modelling metabolism. However, graph theoretic description of real-world metabolic networks (Figure C a still needs to be established precisely. For example, in the most abstract approach, all interacting is considered equally and the edges between nodes represent reactions that convert one substrate into another Figure C b). However, for many biological applications, it is useful to ignore co-factors, which can result in a completely different type of mapping that connects only the main source metabolites to the main products (Figure C c)

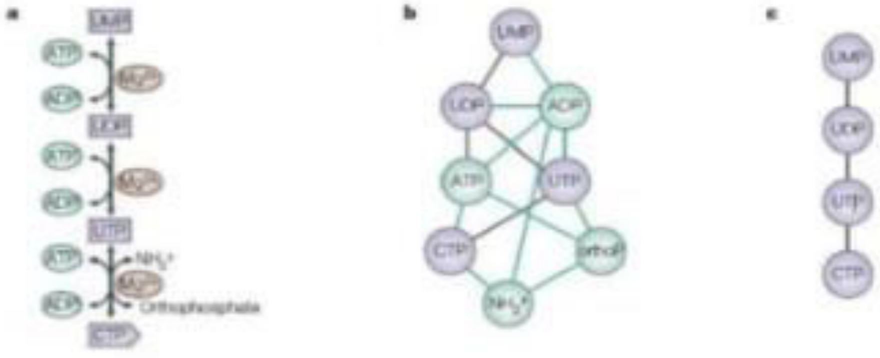


Figure: C Metabolic networks

6. PPI Network

Finally, in protein-protein interaction (PPI) networks, nodes correspond to proteins and undirected edges represent physical interaction amongst them. PPI networks represent an opportunity as well as the challenge. Analyzing these networks may provide useful clues about the function of individual proteins, protein complexes, and larger cellular machines. However, PPI data volume and noisiness is making many algorithms for its analyses intractable. Additionally, graph representation of PPI data with nodes and edges corresponding to proteins and protein interactions, respectively, does not address some of the major properties of protein interaction data. It does not deal with the noisiness of the data, i.e., the large number of false positives and negatives. Moreover, all spatial and temporal information is lost, as well as the information about the conditions of biochemical experiments, confidence of interactions, number of experiments confirming the interactions, etc.

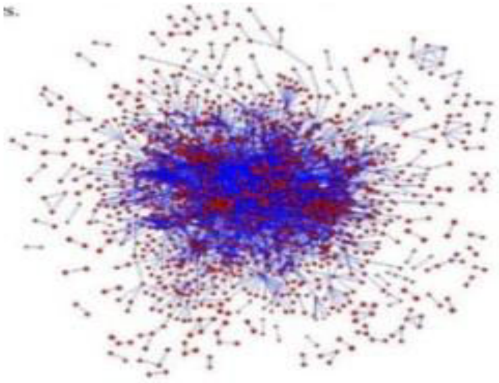


Figure: D PPI Network

7. Characterizing drug-drug target relationship

An assessment of the number of drug targets, i.e., molecular targets that represent an opportunity for therapeutic intervention, as well as their identification, is crucial to the development of post-genomic research strategies within the pharmaceutical industry.

Now that the size of the human genome is known, it is interesting to consider just how many molecular targets this opportunity represents. Additionally, identifying and characterizing the relationships between drugs and their protein targets, as well as between drug targets and disease-gene products in the human protein-protein interaction network still remains a challenge.

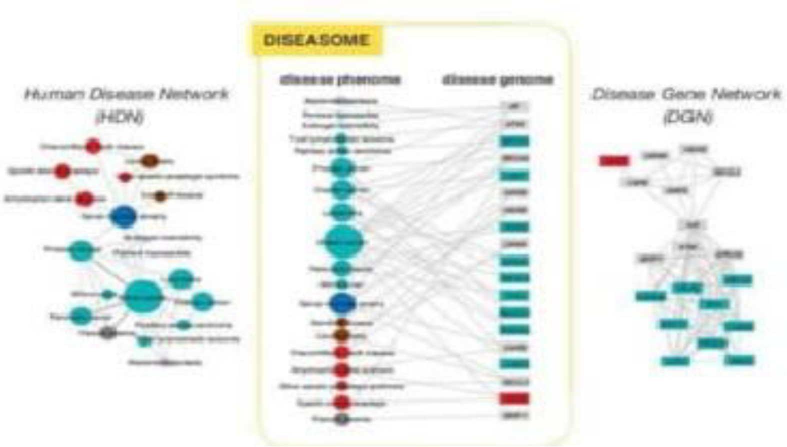


Figure: E Construction of diseasome bipartite network

To study drug-drug target relationship, it will be divided into sub categories to study drug-drug target relationship.

1) Drug able protein

2) Drug Bank

3) Drug Target Network [15]

8. CONCLUSION

There is a need to develop an integrative, systems-oriented analysis in Biology. While ultimately, we may wish to understand the dynamical processes that take place in living organisms, we first need to understand how the components in biological systems interact with each other and the biological significance of those interactions. Biological network analysis is thus a necessary and highly important aspect of the general systems – driven approach to biology. Recent developments in biology and medicine have led to a clear need for biological network analysis.

Moreover, current techniques for the generation of network data are error-prone. Network analysis techniques can be used to assess the accuracy of such data and to help in obtaining more reliable network maps in the future.

Despite the progress that has been made in the analysis of biological networks, there are many major issues that still need to be addressed. The unreliable quality and incompleteness of existing data sets, is a serious impediment to network research and the development of improved experimental and statistical techniques, to enhance the accuracy of network maps, is of vital importance for future research efforts.

REFERENCES

1) Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

2) Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews genetics, 12(1), 56-68.

3) Burgos, E., Ceva, H., Hernández, L., Perazzo, R. P., Devoto, M., & Medan, D. (2008). Two classes of bipartite networks: nested biological and social systems. Physical Review E-Statistical, Nonlinear, and Soft Matter Physics, 78(4), 046113.

4) Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). Introduction to algorithms. MIT press.

5) D'Andrade, R. G. (1978). U-statistic hierarchical clustering. Psychometrika, 43, 59-67.

6) Dijkstra, E. W. (2022). A note on two problems in connexion with graphs. In Edsger Wybe Dijkstra: his life, work, and legacy (pp. 287-290).

7) Erdős, P., & Rényi, A. (1961). On the strength of connectedness of a random graph. Acta Mathematica Hungarica, 12(1), 261-267.

8) Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

9) Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., ... & von Mering, C. (2009). STRING 8-a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research, 37(suppl_1), D412-D416.

10) Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. Nature Reviews Genetics, 21(7), 428-444.

11) Kholodenko, B. N., Hancock, J. F., &Kolch, W. (2010). Signalling ballet in space and time. Nature reviews Molecular cell biology, 11(6), 414-426.

13) Koschützki, D., & Schreiber, F. (2004). Comparison of centralities for biological networks.

14) Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. Molecular systems biology, 4(1), 213.

15) Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome research, 14(6), 1085-1094.

16) Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., ... & Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. science, 298(5594), 799-804.

17) Lima-Mendez, G., & Van Helden, J. (2009). The powerful law of the power law and other myths in network biology. Molecular BioSystems, 5(12), 1482-1493.

18) Lloyd, C. M., Halstead, M. D., & Nielsen, P. F. (2004). CellML: its future, present and past. Progress in biophysics and molecular biology, 85(2-3), 433-450.

19) Mazurie, A., Bonchev, D., Schwikowski, B., & Buck, G. A. (2010). Evolution of metabolic network organization. BMC Systems Biology, 4, 1-10.

20) Pellegrini, M., Haynor, D., & Johnson, J. M. (2004). Protein interaction networks. Expert review of proteomics, 1(2), 239-249.

21) Picard, F., Miele, V., Daudin, J. J., Cottret, L., & Robin, S. (2009, June). Deciphering the connectivity structure of biological networks using MixNet. In BMC bioinformatics (Vol. 10, pp. 1-11). BioMed Central.

22) Prathik, A., Uma, K., & Anuradha, J. (2016). An Overview of application of Graph theory. International Journal of ChemTech Research, 9(2), 242-248.

23) Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., &Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. science, 297(5586), 1551-1555.

24) Watts, D. J., &Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. nature, 393(6684), 440-442.

25) Zotenko, E., Mestre, J., O'Leary, D. P., &Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: re-examining the connection between the network topology and essentiality. PLoS computational biology, 4(8), e1000140.