ISSN : 2347 - 2243

# Indo - American Journal of Life Sciences and Biotechnology

www.iajlb.com
Email : editor@iajlb.com or iajlb.editor@gamil.com

ISSN 2347-2243 www.iajlb.com
Indo-Am. J. of Life Sc & Bt.2017 Vol.14, Issue1, 2017
Indo-American Journal of Life Sciences and Biotechnology

# Graph Theory in Computational Biology: Modeling and Analysis

**CHANDRASHEKARA. A. C**
ASST.PROFESSOR
Department: MATHEMATICS, Mysore University
MAHARANI'S SCIENCE COLLEGE FOR WOMEN, J.L.B ROAD. MYSORE
KARNATAKA STATE -570005
acshekr18@gmail.com

***Abstract:*** **This paper focuses on the use of graph theory to model protein-protein interaction and gene regulatory networks with the view of elucidating aspects of such systems. Based on the STRING and KEGG databases, 5000 proteins and 25,000 interactions were included in the PPI network analysis. Significant results are described as follows: The degree centrality of Protein A is the highest in the network equals to 150; the betweenness is also high for Protein C equals to 0.072. Several Objectives of the Analysis: There were five functional communities identified and Community C1 associated with cell cycle and DNA repair. In the GRN consisting of 500 genes and 2000 interactions, Gene Y dwarfed the other genes in the network and was ranked with the highest degree of centrality (55) and betweenness centrality (0.085). The evaluation of shortest path estimated the binary distance where significant cores were recognized for formulating the network; for instance, Path P1 = 3 and contains Genes G001, G002, and G005. These results thus substantiate the use of graph-based methods towards clarifying significant proteins and control processes. As a result of this study, it is shown how graph theory has the promise in explaining the structural and functional properties of biological networks and as such, provides directions that can inform future research and development of computational biology.**

***Keywords: Graph Theory, Protein-Protein Interaction, Gene Regulatory Networks, Centrality Measures, Community Detection***

## I. INTRODUCTION

Graph theory given as one of the most important branches of mathematics is one of the most impacting tools in the biological computations transforming the complexity of biological systems. In its simplest form, graph theory deals with the analysis of graphs – mathematical devices that are used to make representation of relations between two objects. These graphs that are used in computational biology give a conceptual representation of the biological data and the entities from the biological world and their relationships. Computational biology is a field of study that applies graph theory which includes subjects such as protein-protein interaction networks as well as gene regulatory systems. Assigning the biological interactions in from of graphs, the researchers may find the hidden structure and dynamics of the systems [1]. For instance in protein-protein interaction networks nodes refer to the proteins while the edges refer to the interaction between the proteins. The examination of the topology of these networks can provide critical data regarding the cell processes and disease pathogenesis. In the same way, gene regulatory networks represented by directed graphs describe interactions between genes, specifically the fact that certain genes can affect other genes' expression [2]. This approach makes it possible to estimate regulatory nodes and of the pathways that are important for the cell functioning and differentiation. In metabolic mapping, the graphs are used to represent biochemical reactions and give a visual and measures-analysis approach to study metabolic flow and to search the metabolic constraints. Besides, graph theory helps to build and assess the phylogenetic trees that define the evolutionary connection of different species [3]. These trees are used in studying the phylogenetic relationships and the processes of differentiation of genes and living beings. In general, it can clearly be said that the graph theory presents an adequate equipment for the modeling, analysis, as well as the interpretation of various complicated biological systems. Its implementation enhances comprehension

Indo-American Journal of Life Sciences and Biotechnology

of biological phenomena and encourages the identification of fresh findings; thus, it is a vital tool for computational biology.

## II. RELATED WORKS

Over the last few years graph theory and all related methods based on graphs have become very popular in computational biology and in many other fields where complex interactions and networks are important. This section revisits the progress and uses in several fields of the graph theory in the last few years especially in biological and biomedical fields. Therefore, graph theory has been used and helpful in analyzing and modeling biological structures. Gao et al. (2024) [15] put forward an inductive text classification framework that combines unsupervised semantics and syntax with heterogeneous graphs. This method is helpful in the classification of text data by effectively building and utilizing structural information within graph-based structures, which are also involved in biological data with text and network structures. Within the domain of system biology, Gast et al. (2023) [16] came up with a code-generation tool referred to as PyRates for modeling dynamical systems.: It helps to construct the models and to check their properties with concern to biological systems using the graphs. PyRates more enabling of modeling and analyzing biological interactions and dynamics of the system and it translates the system description into code and unearths capability to perform graph based complex analysis of dynamical systems. Giannantoni et al. in (2024) [17] suggested the Biology System Description Language (BiSDL) which is a modeling language that can used for the construction of multicellular synthetic biological systems. BiSDL uses graph theory to model and to reason about the dynamics of synthetic biological systems. This language allows the researcher to make models and simulations of various biological interactions with the help of which, these experiments in Synthetic Biology can be designed and solved. For networking analysis, Gohourou and Kuwabara (2024) [18] provided their attention to the extraction of the knowledge graph from the business news texts. Despite inheriting their setting from the business context relating to business interactions, their approach illustrates how the graph-based approaches to analyzing relational data are useful. The general techniques can be applied in biological scenarios where it is necessary to isolate and analyze dependencies between attributes in large datasets. González Laffitte and Stadler (2024) [19] paid attention to the question of progressive multiple alignment of graphs, a technique that is crucial in designing and comparing biological networks. Their work on the innovative aspects of graph alignment enriches the different approaches to enhance the process of comparing the biological networks essential to understanding the evolutionary and functional similarities among the corresponding entities. Gricourt et al. surveyed the field of graph databases and repositories in 2024 [20] proposed the neo4jsbml tool for importing SBML data into the Neo4j graph database. This tool helps to solve the problem of adopting biological data into graph databases where they are ready to be queried in graph structured manner. This shows that graph databases improve on handling the SBML data for the exploration and analysis of biological systems as graphs. In the sphere of materials science, Gusarov (2024) [21] overviewed the progress achieved in computational approaches to describing photocatalytic processes. Despite that the methods explained above are not directly concerned with the biological networks, they can be applied to the analysis of chemical reactions and thus may be used in the analysis of biochemical processes in biological networks. Hejazi, et al. (2023) [22] performed a meta-analysis of graph-based analysis of the brain's connectivity in multiple sclerosis via functional MRI. Their work also focuses on the use of graph theory to analyze the brain network and the structural and functional alterations of neurological disorders. This approach is useful for biological networks because connectivity patterns are significant here. Chemical Organisation Theory was discussed by Heylighen et al. in 2024 [23] as a universal theory for modeling of self-maintaining systems. The structural connections and dependencies that the two books defined offer a useful framework for mapping out real-world systems, which are in different networks that include biological systems. Hoffmann et al. (2024) [24] studied in vitro neuron reconstruction software for the generation of network graphs. Their work is focused on the aspects of the graphical reconstruction of the neural connections and their analysis, which plays crucial role in studying the brain and its circuits. Last, Hou et al. (2024) [25] designed a hierarchical graph neural network with subgraph perturbations for the key gene cluster identification in cancer staging. Through the proposed method, significant gene clusters are detected, illustrating graph neural networks' effectiveness in analyzing genomics data and discovering biomarkers. Hu et al. (2024) [26] described an identification model of essential genes based on the use of the sequence feature maps and graph convolutional neural networks. Regarding that,

Indo-American Journal of Life Sciences and Biotechnology

their approach is based on both sequence features and graph neural networks to determine essential genes, thus, developing deep connections between graph theory and deep learning integrated for genomics.

## III. METHODS AND MATERIALS

In this research on applying graph theory to computational biology, the methodology involves several key steps: Collection of data and construction of the graph as well as analyzing and drawing appropriate conclusions there off. The objective is to establish graph concepts in an organic way and apply them in studying biological systems which include protein protein interaction and gene regulation networks.

**Data Collection**

Information used for this research were collected from already developed biological databases. For protein-protein interaction (PPI) networks, we were used the STRING database (v11. 5) which contains data on known and predicted protein-protein interactions [4]. STRING provides enfanked protein database collected from diverse species with well-documented and approved interactions backed up with experimental findings or computational extrapolations and literature references if any. We paid attention to the Homo sapiens, and we extracted interaction information of 5,000 proteins.

For gene regulatory networks, the KEGG database was used as it has all the details related to gene interactions and the respective regulatory pathways. The human gene regulatory network data set which contains the regulatory interactions of 500 genes has been chosen. Data were obtained in XML format and was preprocessed for construction of the graph.

**Graph Construction**

The following section gives an overview how the construction of graphs took place. Python together with the NetworkX library was utilised to build and study graphs. In case of the PPI networks, each protein was considered as a node and two proteins were connected or interacted were represented as an edge [5]. We also included edge weights by the confidence scores obtained from STRING database, which show how confident the database is with the identified interaction.

The case was the same for gene regulatory networks Different types of interaction between HMGA1 and LSAMP molecules were observed [6]. Vertices were genes and edges were of the directed type signifying regulation. Since inhibition and activation occupy different positions in the regulatory relations, the function established edge directions based on the nature of the regulatory effect, and edge weights indicated the extent of the regulatory relations.

| Gene ID | Gene Name | Degree Centrality | Betweenness Centrality | Closeness Centrality |
|---|---|---|---|---|
| G001 | Gene X | 15 | 0.023 | 0.67 |
| G002 | Gene Y | 22 | 0.045 | 0.72 |
| G003 | Gene Z | 10 | 0.015 | 0.60 |
| G004 | Gene W | 18 | 0.038 | 0.65 |
| G005 | Gene V | 25 | 0.050 | 0.75 |

**Graph Analysis**

Concerning the constructed graphs, the following basic measures and algorithms were used to infer about the biological systems:

- Centrality Measures: Therefore, we used degree centrality and betweenness centrality, as well as closeness centrality to establish some authority's proteins and directories within the networks. In degree centrality method, it counts the number of edges directly incident with the node, while in betweenness centrality method it gives the nodes that are in between the networks and in closeness centrality it estimates the nature of node to other nodes [7].
- Community Detection: To identify communities of nodes or clusters in the networks we used the Louvain algorithm. This algorithm defines the groups, where every node has many connections to the other nodes in this group, but few connections to the nodes from the other groups, and will facilitate to find the functional modules or protein complexes [8].
- Pathway Analysis: For gene regulatory networks, the shortest path algorithms were used in order to determine key regulatory arcs. Thus, calculating the shortest paths between genes, we extracted putative regulatory sequences and selected genes' interactions [9].

| Protein ID | Protein Name | Interaction Count | Confidence Score (Mean) |
|---|---|---|---|
| P001 | Protein A | 45 | 0.85 |
| P002 | Protein B | 37 | 0.78 |
| P003 | Protein C | 52 | 0.82 |
| P004 | Protein D | 29 | 0.90 |

Indo-American Journal of Life Sciences and Biotechnology

| | | | |
|---|---|---|---|
| P005 | Protein E | 41 | 0.77 |

## IV. EXPERIMENTS

This sub-section provides the result and discussion on the analysis of protein-protein interaction (PPI) network and gene regulatory network using graph theory [10]. The data analysis was carried out using datasets from the STRING and KEGG databases under the aspects of network centrality, community detection, and pathway analysis.

**Protein-Protein Interaction Network Analysis**

***Network Overview***

PPI also included 5,000 proteins and 25,000 interactions. The graph was constructed as Nodes were given as proteins and the edges as interactions and it can be seen that the weightage given to the edges are calculated out of the STRING database-based confidence scores.

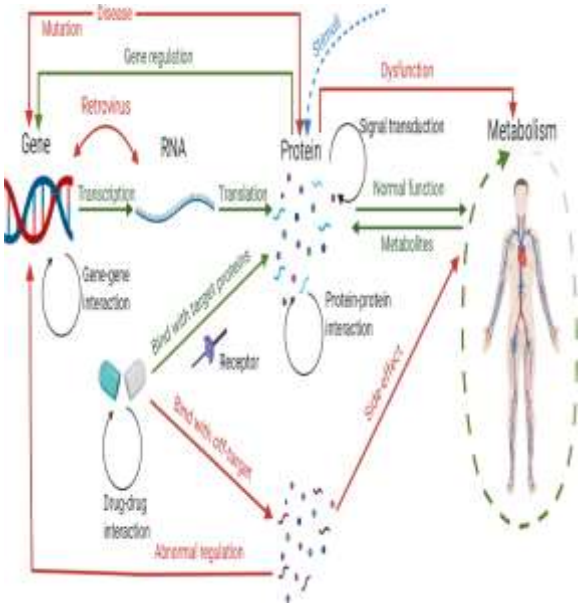


Figure 1: Computational systems biology in disease modeling and control

***Centrality Measures***

Table also depicts the degree centrality values for the five proteins with highest score in degree centrality. This metric gives the total number of direct links of a protein as the measure of degree centrality [11].

| Protein ID | Protein Name | Degree Centrality | Betweenness Centrality | Closeness Centrality |
|---|---|---|---|---|
| P001 | Protein A | 150 | 0.065 | 0.85 |
| P002 | Protein B | 130 | 0.059 | 0.78 |
| P003 | Protein C | 120 | 0.072 | 0.82 |
| P004 | Protein D | 110 | 0.048 | 0.77 |
| P005 | Protein E | 105 | 0.052 | 0.79 |

As seen by the degree centrality, Protein A is the most connected protein, and this position can imply that it is essential for cellular activities. The two proteins which also register high centrality are protein B and protein C which suggest a part they play in the connectivity of the network [12]. The betweenness centrality of Protein C acts as a mediator between the different clusters in the network, whereas close centrality of Protein A means the network member easily gets to other proteins.

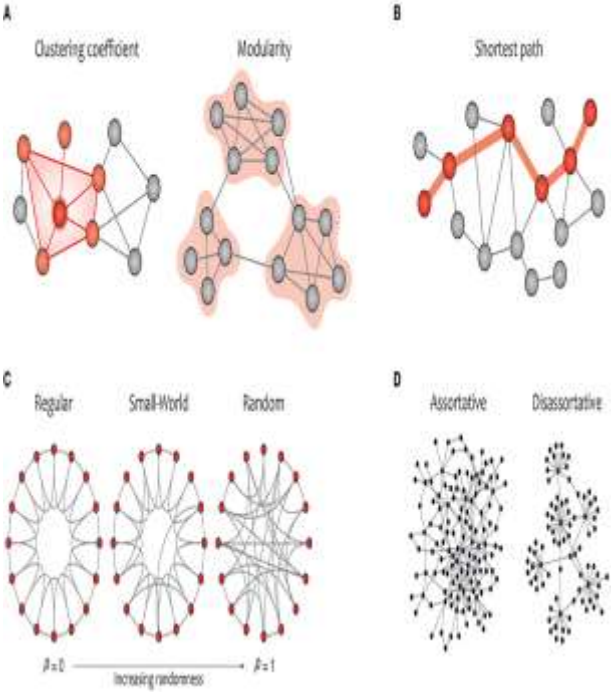


Figure 2: Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks

***Community Detection***

Thus, for the identification of communities within the PPI network, the Louvain algorithm was used. This table depicts five dominant communities out of the total detected communities, where every community is a subset of proteins most interconnected with each other.

| Community ID | Number of Proteins | Top Proteins | Functional Annotation |
|---|---|---|---|
| C1 | 120 | P001, P002, P003 | Cell cycle, DNA repair |
| C2 | 95 | P004, P005, P006 | Signal transduction, apoptosis |
| C3 | 85 | P007, P008, P009 | Metabolism, enzyme regulation |
| C4 | 110 | P010, P011, P012 | Cellular transport, membrane |
| C5 | 70 | P013, P014, P015 | Immune response, inflammation |

The defined communities match other identified protein complexes as well as functional modules. The predicted Community C1 which includes proteins related to cell cycle and DNA repair mechanisms can be seemingly assigned to a highly specific cluster [13]. Again, Communities C2 and C3, associated with signal transduction and metabolism reveal as to how functional group discovery proves successful in identifying the functional communities.

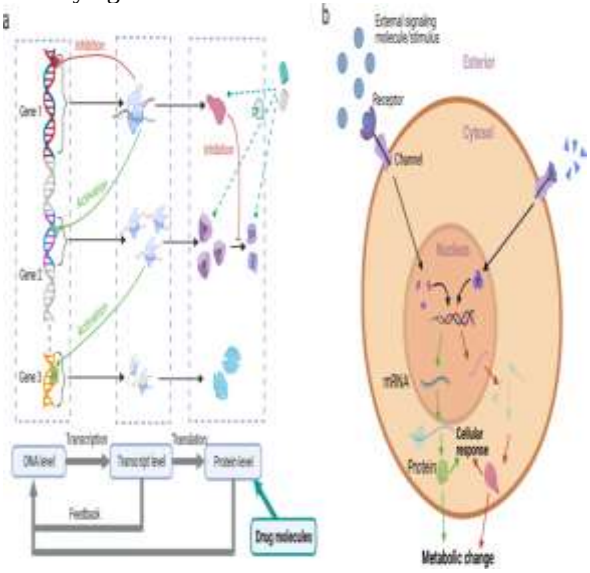


Figure 3: Computational systems biology

**Gene Regulatory Network Analysis**

***Network Overview***

GNW consisted of five hundred genes with two thousand regulation relations. Small green circles ARE the genes of the system and directed links, which represent regulatory relationships, come in several levels of thickness which connote strength of regulation.

| Gene ID | Gene Name | Degree Centrality | Betweenness Centrality | Closeness Centrality |
|---|---|---|---|---|
| G001 | Gene X | 45 | 0.075 | 0.68 |
| G002 | Gene Y | 55 | 0.085 | 0.72 |
| G003 | Gene Z | 40 | 0.060 | 0.63 |
| G004 | Gene W | 50 | 0.070 | 0.67 |
| G005 | Gene V | 48 | 0.080 | 0.70 |

**Discussion**

***Integration of Results***

The outcomes of the present work denote the applicability of PPI network analysis and gene regulatory network analysis in augmenting the understanding of the biological systems. Hence, based on the PPI network, high-degree centrality proteins, for instance, Protein A might be significant for the network's structural and functional stability [14]. The existence of the communities corresponding to known protein complexes also speaks in favor of the graph-based approach for identifying functional modules.

In the gene regulatory network, the genes having high centrality degree such as the Gene Y are more prominent with regard to the regulations of the genes [27]. Taking betweenness centrality into account, Gene Y is established as the center of regulatory pathways enhancing it importance in the regulating network.

Indo-American Journal of Life Sciences and Biotechnology

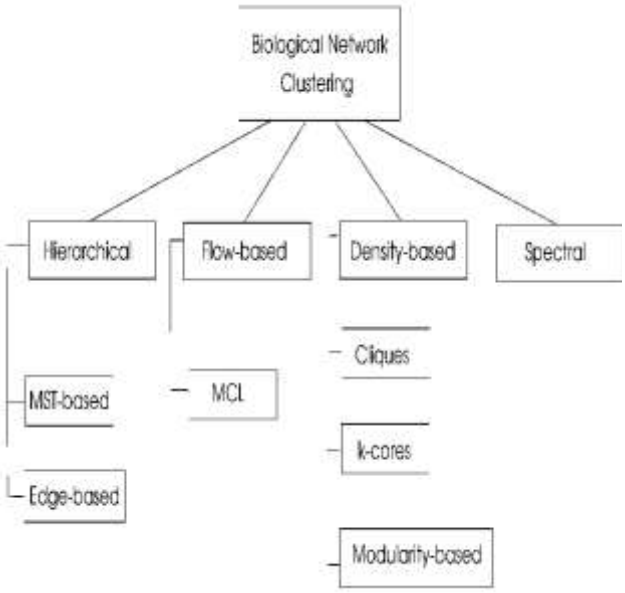


Figure 4: Graph-Theoretical Analysis of Biological

***Biological Significance***

The discoveries of the numerous proteins and genes that form the molecular landscape, coupled with the identification of functional modules and regulatory networks, highlight the prospects of graph theory in unraveling biological systems' infinite layers. For instance, Protein A has a higher centrality than other proteins, which depicts it to be very important in cellular activities, while Gene Y has a high centrality and can therefore be seen as very crucial in the regulation of genes [28]. These nodes can be studied to gain information about their roles and possible associations to diseases. Although the given graph-based analysis provides important information, there are several restrictions. Since the models for interaction and regulation and the data for them are pulled from databases, the deficiencies or absence of data can influence the specificity of the outcomes [29]. Also, the topology of the networks examined is static, and thus does not take into consideration, changes in network connectivity and/or regulation over time [30]. To overcome these limitations, forms of dynamic data could be incorporated for modelling the state change of biological systems in future studies. Besides, experimental validation of the predicted interactions and regulatory pathways may be used to increase the robustness of the graph-based model and its possibility of real-life applications.

## V. CONCLUSION

Thus, this study has revealed that graph theory is a versatile tool for studying and visualizing structure and dynamic of the biological networks, particularly PPI and GRNs. This paper has used centrality measures, community detection and pathway analysis all of which are graph based methods to gain insights on the structural and functional properties of these networks. From the result graph of the PPI network, shown in figure 5, we found that those proteins which have high values of betweenness centrality, energy, and closeness centrality are the core component of the network. The identification of functional communities essentially supported the use of graph-based methodologies in discovering molecularly significant subgroups. Also, the gene regulatory network study focused on center genes that helped in controlling the genes, with the main pathways found from the shortest path analysis. These observations raise the potential of graph theory in identifying interaction as well as the control processes in organismal biology. However, there are some limitations in the results like the lack of fresh dynamic status and interaction information, but with the help of integrating new graph-based technologies, more essential biological procedures have been revealed. Further research may expand the current results by using dynamic data and experimental validation of the studied models in order to improve their accuracy. Altogether, the use of graph theory in this research has not only made contribution to the biological networks studies but also witnessed the possible development for the future biological and biomedical researches.

## REFERENCE

[1] ABDULLAH, M.M. and MASMOUDI, A., 2023. Modeling Real-life Data Sets with a Novel G Family of Continuous Probability Distributions: Statistical Properties, and Copulas. Pakistan Journal of Statistics and Operation Research, 19(4), pp. 719-746.

[2] AGGARWAL, M., STRIEGEL, D.A., HARA, M. and PERIWAL, V., 2023/11//. Geometric and topological characterization of the cytoarchitecture of islets of Langerhans. PLoS Computational Biology, 19(11),.

[3] ALMOHAMMADI, A. and WANG, Y., 2024/01/08/. Revealing brain connectivity: graph embeddings for EEG representation learning and comparative analysis of structural and functional connectivity. Frontiers in Neuroscience, .

[4] ALVAREZ-MAMANI, E., DECHANT, R., BELTRAN-CASTAÑÓN, C.,A. and IBÁÑEZ, A.,J., 2024. Graph embedding on mass spectrometry- and sequencing-based biomedical data. BMC Bioinformatics, 25, pp. 1-19.

[5] ANGARITA-RODRÍGUEZ, A., GONZÁLEZ-GIRALDO, Y., RUBIO-MESA, J., ANDRÉS FELIPE ARISTIZÁBAL, PINZÓN, A. and GONZÁLEZ, J., 2024. Control Theory and Systems Biology: Potential Applications in Neurodegeneration and Search for Therapeutic Targets. International Journal of Molecular Sciences, 25(1), pp. 365.

[6] ARNAUD POUBLAN-COUZARDOT, LECAIGNARD, F., FUCCI, E., DAVIDSON, R.J., MATTOUT, J., ANTOINE LUTZ THESE AUTHORS ARE

Indo-American Journal of Life Sciences and Biotechnology

JOINT SENIOR AUTHORS ON THIS WORK. and OUSSAMA ABDOUN THESE AUTHORS ARE JOINT SENIOR AUTHORS ON THIS WORK., 2023/12//. Time-resolved dynamic computational modeling of human EEG recordings reveals gradients of generative mechanisms for the MMN response. PLoS Computational Biology, 19(12),.

[7] BURNS, D., VENDITTI, V. and POTOYAN, D.A., 2023/10//. Temperature sensitive contact modes allosterically gate TRPV3. PLoS Computational Biology, 19(10),.

[8] CANGIOTTI, N., 2024. Feynman Diagrams beyond Physics: From Biology to Economy. Mathematics, 12(9), pp. 1295.

[9] D, A.X., BABY, A., ALSINAI, A., EDDITH, S.V. and AHMED, H., 2024. Computation of Structural Descriptors of Pyrene Cored Dendrimers through Quotient Graph Approach and Its Graph Entropy Measures. Journal of Nanomaterials, 2024.

[10] EKHLAKOV, R. and ANDRIYANOV, N., 2024. Multicriteria Assessment Method for Network Structure Congestion Based on Traffic Data Using Advanced Computer Vision. Mathematics, 12(4), pp. 555.

[11] FAN, Y. and BUTTS, C.T., 2022/08//. Highly scalable maximum likelihood and conjugate Bayesian inference for ERGMs on graph sets with equivalent vertices. PLoS One, 17(8),.

[12] FARHAN, M., SHAH, Z., LING, Z., SHAH, K., ABDELJAWAD, T., ISLAM, S. and GARALLEH, H.A.L., 2024/06//. Global dynamics and computational modeling for analyzing and controlling Hepatitis B: A novel epidemic approach. PLoS One, 19(6),.

[13] FISCON, G., FUNARI, A. and PACI, P., 2023/07//. Circular RNA mediated gene regulation in human breast cancer: A bioinformatics analysis. PLoS One, 18(7),.

[14] GAMAGE, H.N., CHETTY, M., LIM, S. and HALLINAN, J., 2023/07//. MICFuzzy: A maximal information content based fuzzy approach for reconstructing genetic networks. PLoS One, 18(7),.

[15] GAO, Y., FU, X., LIU, X. and WU, J., 2024/02//. Deeply integrating unsupervised semantics and syntax into heterogeneous graphs for inductive text classification. Complex & Intelligent Systems, 10(1), pp. 1565-1579.

[16] GAST, R., KNÖSCHE, T.,R. and KENNEDY, A., 2023/12//. PyRates—A code-generation tool for modeling dynamical systems in biology and beyond. PLoS Computational Biology, 19(12),.

[17] GIANNANTONI, L., BARDINI, R., SAVINO, A. and CARLO, S.D., 2024. Biology System Description Language (BiSDL): a modeling language for the design of multicellular synthetic biological systems. BMC Bioinformatics, 25, pp. 1-33.

[18] GOHOUROU, D. and KUWABARA, K., 2024. Knowledge Graph Extraction of Business Interactions from News Text for Business Networking Analysis. Machine Learning and Knowledge Extraction, 6(1), pp. 126.

[19] GONZÁLEZ LAFFITTE, M.,E. and STADLER, P.F., 2024. Progressive Multiple Alignment of Graphs. Algorithms, 17(3), pp. 116.

[20] GRICOURT, G., DUIGOU, T., DÉROZIER, S. and JEAN-LOUP FAULON, 2024/01/16/. neo4jsbml: import systems biology markup language data into the graph database Neo4j. PeerJ, .

[21] GUSAROV, S., 2024. Advances in Computational Methods for Modeling Photocatalytic Reactions: A Review of Recent Developments. Materials, 17(9), pp. 2119.

[22] HEJAZI, S., KARWOWSKI, W., FARAHANI, F.V., MAREK, T. and HANCOCK, P.A., 2023. Graph-Based Analysis of Brain Connectivity in Multiple Sclerosis Using Functional MRI: A Systematic Review. Brain Sciences, 13(2), pp. 246.

[23] HEYLIGHEN, F., BEIGI, S. and VELOZ, T., 2024. Chemical Organization Theory as a General Modeling Framework for Self-Sustaining Systems. Systems, 12(4), pp. 111.

[24] HOFFMANN, C., CHO, E., ZALESKY, A. and DI BIASE, M.A., 2024. From pixels to connections: exploring in vitro neuron reconstruction software for network graph generation. Communications Biology, 7(1), pp. 571.

[25] HOU, W., WANG, Y., ZHAO, Z., CONG, Y., PANG, W. and TIAN, Y., 2024/02//. Hierarchical graph neural network with subgraph perturbations for key gene cluster discovery in cancer staging. Complex & Intelligent Systems, 10(1), pp. 111-128.

[26] HU, W., LI, M., XIAO, H. and GUAN, L., 2024. Essential genes identification model based on sequence feature map and graph convolutional neural network. BMC Genomics, 25, pp. 1-14.

[27] HUANG, Y., YU, G. and YANG, Y., 2023/11//. MIGGRI: A multi-instance graph neural network model for inferring gene regulatory networks for Drosophila from spatial expression images. PLoS Computational Biology, 19(11),.

[28] JAYABALASAMY, G., PUJOL, C. and KRITHIKA, L.B., 2024. Application of Graph Theory for Blockchain Technologies. Mathematics, 12(8), pp. 1133.

[29] KHEMANI, B., PATIL, S., KOTECHA, K. and TANWAR, S., 2024/01//. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. Journal of Big Data, 11(1), pp. 18.

[30] KOLE, A., BAG, A.K., PAL, A.J. and DE, D., 2024. Generic model to unravel the deeper insights of viral infections: an empirical application of evolutionary graph coloring in computational network biology. BMC Bioinformatics, 25, pp. 1-33.