Using Graph Databases for Portraying and Analysing Biological and Biomedical Networks

Blagoj Ristevski
Faculty of Information and Communication Technologies - Bitola University“St. Kliment Ohridski”-Bitola
Bitola, Republic of Macedonia
blagoj.ristevski@uklo.edu.mk
https://orcid.org/0000-0002-8356-1203

Snezana Savoska
Faculty of Information and Communication Technologies - Bitola University“St. Kliment Ohridski”-Bitola
Bitola, Republic of Macedonia
snezana.savoska@uklo.edu.mk
https://orcid.org/0000-0002-0539-1771

Zlatko Savoski
Borka Taleski General Hospital Internal department Prilep
Prilep, Republic of Macedonia
bisera89@yahoo.co.uk

***Abstract* - Nowadays, huge amounts of data are generated experimentally in systems biology as well in clinics and other healthcare and medical institutions. This has resulted in the emergence of new concepts: big data and NoSQL databases that are becoming more popular and promising especially for analyzing complex interactions that exist in biological networks. These heterogeneous and voluminous data, which are usually semi-structured or unstructured, highly connected and unpredictable, need to be integrated and stored properly. With the growth of data size and data complexity, NoSQL databases have outperformed traditional relational databases for analysis, access and querying. Particularly, to represent various complex relationships among entities in physics, in biological, social and computer networks, graph-based databases are very suitable. These databases are appropriate to represent, store and query heavily interconnected data, especially for large-scale network data that exist in biology. This paper describes the biological and other networks in biomedicine and surveys the most widely used graph databases and software tools and packages and their properties. Additionally, the portraying, analyzing and querying of biological networks are described. The analysis of the biological networks results in gaining very significant insights into the relevant information for the biological processes, such as diseases, interactions and regulatory mechanisms that occur in biological networks, as well as studying their properties.**

Keywords- bioinformatics, big data, biological networks, omics data, biomedical networks, graph databases.

I. INTRODUCTION

Nowadays, a huge amount of heterogeneous healthcare, medical and biological data are generated constantly. These data have big data properties and can be structured, semi-structured and unstructured, so the standard relational databases are not suitable to store and manage them [14]. To discover knowledge contained in various omics data, such as genomics, proteomics, metabolomics, transcriptomics, lipidomics, epigenomics, microbiomics, and immunomics data, the graph-based database can be used.

While standard relational databases are optimized to deal with aggregated structured data stored in tables with predefined relationships between them, graph databases are appropriate to deal with highly connected data [12]. A graph database has a flexible schema and it can store, manage, and update data and relationships. Differently from the tabular results of SQL queries, graph query languages like Cypher enable the results to be shown as graphs and/or tables.

There are various applications of graph theory in computer science such as cryptography, coding theory, human brain networks, image processing, image segmentation, information retrieval, social networks analysis, robotics, data science, pattern recognition [2], computer networks, complex networks, web pages ranking, algorithms, bioinformatics, computational biology etc.

Graph theory and methods of network analysis in the past few decades were shown to be effective when applied to large voluminous biological system data to discover meaningful information and knowledge [8].

A network (graph) is a collection of vertices (nodes) representing the particular subunits of the system and edges (arcs), representing the interactions among nodes. When values to the network’s edges are assigned, a weighted network is obtained. In a signed network these values can have either positive or negative weights. Directed networks have directed edges. Multidimensional networks cover various types of connections among nodes. Moreover, labels and categories can be added to the network’s nodes. The clustering coefficient of the network is defined as the ratio of the existing triangles (triples of nodes all connected pairwise with edges) and the number of potential triangles [1]. Regarding the network model, the network graphs are categorized as random, Erdős–Rényi, lattice, small-world and scale-free graphs. In biological and medical systems numerous high-order relationships cannot be covered by binary measures. These networks can be described by hypergraphs with hyperlinks among their nodes [1].

Understanding the fundamental diseases processes requires an integration of various heterogeneous data and then the study and representation of the complex interactions that exist between different biological entities [13].

The rest of this paper is structured as follows. Biological and biomedical networks and their properties are highlighted in Section II. The subsequent section describes the widely used graph databases and software tools for analyzing and querying biological networks. The last section provides discussion and directions for further work.

II. BIOLOGICAL AND BIOMEDICAL NETWORKS

There exist a wide variety of biological network data and numerous available public databases, that organize and provide data to the researchers, medical doctors, specialists, practitioners, and decision-makers in the healthcare institutions. As data amounts increase, the data and their relationships become more complicated, and using standard relational databases is not appropriate. On the other side, graph-based databases enable the causal relationship between biological and medical entities in the networks. For instance, when the causal relationships between particular diseases are depicted in graph-based databases, further deterioration of those diseases can be prevented in time [6].

Biological data represented by networks enable visualization and analysis of various phenomena such as drug-target interactions, diseases' spreading as well as evolutionary relationships between species [11]. Networks that modeled the association between different biological entities are usually portrayed by using bipartite networks. Integration of bipartite and general network portrayal is used to relate multiple types of networks into one complex, multi-relational and heterogeneous network. The most commonly used biological networks are protein-protein interaction (PPI) networks, genetic interaction networks, gene/microRNA/transcriptional regulatory networks, cell signaling networks, metabolic networks, disease-disease association networks, and heterogeneous biological networks as shown in Fig. 1.

To exploit medical and biological big data, network theory provides efficient tools [1]. Biological networks cover experimentally identified physical and functional interactions; computationally and statistically inferred relationships; evolutionary associations based on protein co-evolution, domain families and sequence homology; and functional associations and ontologies representing systematized knowledge [2].

To identify specific sub-networks or interactions involved in specific pathways or to retrieve neighbors of a specific network node, querying interfaces are used [2]. To study biological and biomedical networks, network properties like graph diameter, node degree distribution, centrality measures, network motifs and graphlets, link prediction, community detection, and clustering coefficient should be considered [13].

Moreover, specific network algorithms like community detection and network motif discovery algorithms are very appropriate to be applied to graph databases. A community is a subgroup of a network whose internal connections are much denser than external connections [9]. Identifying such communities can be essential for evaluating group behavior, discovering emergent phenomena, visualization, identifying network sub-clusters, and comprehending network structure.

Network medicine, as an approach to understanding human diseases from a network theory, has achieved significant results in the last decade. Also, the study of diseases as non-isolated elements and the understanding of how they relate to each other is crucial to insights into pathogenesis and etiology. The complete sequencing of the human genome at the beginning of the 21st century was a revolution in the study of the relationships between diseases. When the graphs provide a combination with the growing availability of transcriptomic, proteomic, and metabolomics data sources, it was clear that this will improve the classification of diseases [16]. As a part of the medicine network, the disease network succeeds to reveal hidden connections among apparently unconnected biomedical entities such as diseases, physiological processes, signaling pathways, and genes as an intuitive pathway for visual data understanding. Some researchers pointed out benefits from the identification of new opportunities for the use of old drugs, known as drug repurposing [16] which is important taking into consideration decreasing costs for regulatory purposes as well as the time for new drug registration.

Many researchers investigate some points connected with biological and medical networks, considering their ability to query and visualize data from a graph-based database. Some researchers proposed using directed disease networks to facilitate multiple-disease risk prediction [15], proposing a new framework that combines directed disease networks with system techniques to enhance the prediction risk assessment. The proposed model takes into account multiple-disease risk assessment scores for different diseases based on the patient's medical history as a recommendation system where input is the medical record of the patient and output is a list of ordered diseases with a fixed length, depending on the needs of medical experts, professionals or patients. They use ICD-10 (the 10th revision of the International Statistical Classification of Diseases and Related Health Problems) classification but aggregated in DALY (disability-adjusted life year), a comprehensive metric of disease burden designed by the World Health Organization [15].

Human disease and health networks can be used to quantify the contagion effects, sometimes denoted as spillover or influence effects [17]. The estimation and identification of the contagion effects can be important taking into account understanding the spread of human disease and health behavior as well as various implications for creating effective public health interventions. Disease and health networks also can be used together with traditional statistical analysis and several state-of-the-art statistical methods [17] to quantify some advancements and remaining challenges of cognition effects together with some diseases.

To explore medical data focused on the analysis of drugs for different diseases with complex network analysis, graph-based data are used in [18]. They used the drug as a node, the relationship as an edge connecting the two nodes, and the co-occurrence frequency of the drug, as the weight of the edge to establish a network graph. The clustering algorithm of the weighted network graph was used as the center of a diffusion method combined with the network topology and the edge weights. The purpose is to divide the network graph into communities, where Clustering Algorithm on Weighted Networks (SCW) is used to study the medical prescriptions. The SCW community classification algorithm is used also to analyze the association relationship between treatment and medicine in medical case data. Due to research of case study data, the commonly used medicine combination rules are used, conducive to assisting medical diagnosis [18].

Graph databases are also used to represent and create models of biological networks that provide a framework for comprehension of disease by depicting the relationships between the mechanisms engaged in the biological processes' regulation. An example of such a network consisting of a set of 46 biological networks relevant to lung and chronic obstructive pulmonary disease is described in [19] [22]. The network is built using Biological Expression Language (BEL) with detailed information for each node and edge, including supporting verification from the literature. The mentioned network scoring contains public transcriptomic data for some subset of mechanisms and networks that correspond to the measured outcomes. The obtained results are used to identify novel mechanisms that activate disease, compare different treatments and time points, and permit the assessment of low signal data. Some toxicology and drug discovery applications already use the network [19] [22].

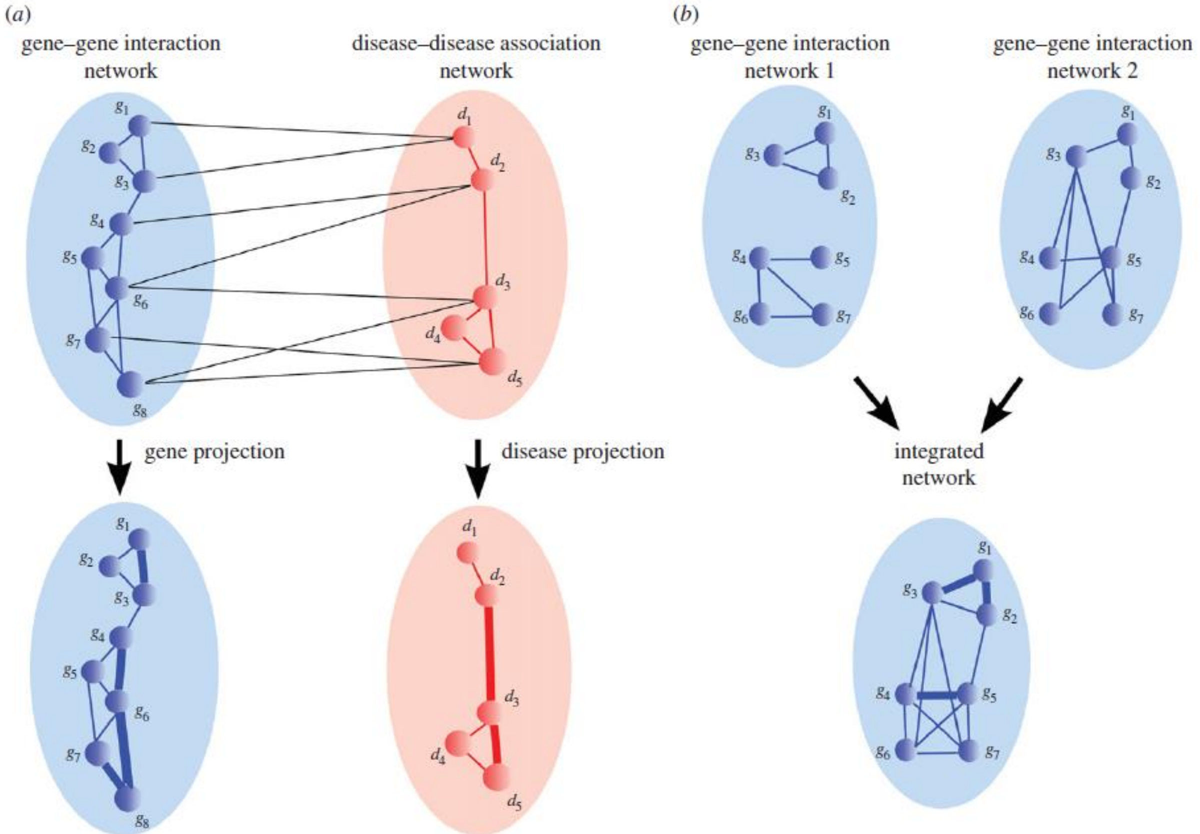
(a)
gene–gene interaction network
disease–disease association network
g1
g2
g3
g4
g5
g6
g7
g8
d1
d2
d3
d4
d5
gene projection
disease projection
(b)
gene–gene interaction network 1
gene–gene interaction network 2
integrated network

Fig. 1. a) Heterogeneous networks consisted of a gene-gene interaction network (blue), a disease-disease association network (red) and a gene-disease association network (black edges); b) homogeneous gene-gene interaction network [23].

Many examples show the increased usage of drug-target and disease networks to increase the comprehension of complex diseases and the concept of a new “multi-target, multi-drug” model that aims at systemically modulating multiple targets [20]. Identification of the interaction between drugs and target proteins plays an essential role in the process of genomic drug discovery and development. The aim is to discover novel drugs or novel targets for existing drugs due to the demanding and costly experimental process of drug-target interaction prediction and provide beneficial information for supporting experimental interaction data. In postgenomic drug discovery, the large-scale integration of genomic, proteomic, signaling and metabolomics data can provide helpful information to construct complex networks of the cells that are capable to increase the understanding of physiological or pathophysiological states on a molecular basis. The authors in [20] have constructed an emerging model of poly-pharmacology in the postgenomic era that states that drug, target and disease spaces can be correlated to study the effect of drugs on different spaces and their interrelationships can be utilized for designing drugs or cocktails which can functionally target one or more diseases. They also state that it is possible to create a computational platform that will integrate genome-scale metabolic pathways, protein-protein interaction networks, and gene transcriptional analysis, to build a thorough network for multi-target multi-drug discovery [20].

III. Software tools for analyzing biological networks

When data sets become very large and complex, current software tools for graph-based analysis and querying are appropriate to handle the numerous nodes and edges in the networks as well as to represent them in a suitable layout and format. Relational databases use tables to store and manage data, as well as to interact with the data by using structured queries. Relational databases store data in a predefined static schema with a low ability to adapt the database to the new data types and formats. Besides their low horizontally scalability, they show deficient performances when working with large datasets and big data. Differently, graph databases are very suitable to store highly variable data in graph format and querying on multiple levels the relationships between instances [7]. Numerous heterogeneous biological and healthcare clinical data, especially those data are structured as networks and/or ontologies, are very suitable to be stored in the graph databases such as Neo4j, AllegroGraph, Datastax, Apache Giraph, AnzoGraph DB, Hypergraph DB, RedisGraph, TigerGraph, etc.

More enlightened queries require integration of multiple networks and employing either suitable developed algorithms or software for analysis and visualization, like Cytoscape and its plugins, or using network libraries or packages like SNAP and igraph [2], as well as Graphviz, Payek, NodeXL, Gephi, Graphia and BioLayout [11]. To explore biological networks and to provide use of multiple database sources, several platforms such as UniProt and NDeX are developed.

A web portal with an integrated database, that contains genomic and transcriptomic data and analysis tools, named OmicsDB::Pathogens was developed [4]. This database explores functional networks, searches genomic data and inspects and compares biological networks across multiple species. OmicsDB stores the high-throughput and other biological networks data mostly in the graph-based database Neo4j. Neo4j is based on a property graph model, where key/value properties can be associated with the vertices and edges. For querying the network/graph, Neo4j uses its query language Cypher [4]. Cypher enables to query the graph paths.

Angles *et al.* have proposed a conceptual model to represent protein-ligand interactions and developed a bioinformatics web tool entitled GSP4PDB [5]. This tool provides visualization, searching and exploring protein-ligand structural patterns within the Protein Data Bank. GSP4PDB employs a PostgreSQL database to store and manage protein data.

By running network queries, users can discover novel knowledge and new insights on clusters, associations and relations that exist in the datasets. The main disadvantage of Neo4j is the difficulty in creating relationships between node labels when working with a graph consisting of more than 10 thousand nodes [7]. Neo4j graph algorithms and its Cypher query language are suitable to find the shortest network route, determining the centrality and community detection.

Neo4j is used for protein-protein interface identification by transforming information about protein complexes from Protein Data Bank (PDB) into Neo4j graph data in [10]. The obtained knowledge about protein-protein interactions can be used to identify inhibitors, that might prevent the formation of interaction and thus restrict the protein complex functionality, and this knowledge is also relevant in drug design.

Many researchers used the Neo4j graph database for building and querying prototype networks to provide beneficial biological understandings. The researchers used the Neo4j in the context of asthma-related genes [21]. Specifically, systems biology experiments produce large amounts of data of multiple modalities. These data present a challenge for integration due to a combination of complexity together with rich semantics. The graph databases provide an important framework for storage, querying and visualizing biological data on the example of asthma-related genes in [21]. Graph databases enable an adaptable solution for the combining of multiple types of biological data and speed up exploratory data mining to support hypothesis generation [21]. Neo4j performs quite well when applied to small query graphs. However, Neo4j has shown that the queries for subgraph matching, for relatively large databases with an increasing number of edges, are not very well optimized [10].

IV. Discussion and Further work

Recent advances in high-throughput technologies, as well as ICT, enable a huge volume of biological, medical, healthcare and biomedical data to be created. To store, analyze and query these heterogeneous datasets, the graph-based database model is very promising. As further work, more software applications, packages and tools based on graph databases should be developed. Graph databases can improve the network analysis, particularly the comprehension and reveal of the interactions between different biological, biomedical and other entities, usually using multiple networks. Graph databases are very suitable for representing biological and biomedical data that are usually highly connected, semi-structured and unpredictable. Considering the growing volume and complexity of the data, the development of effective computation methods and algorithms for data integration, analyzing and querying, particularly for large-scale networks and heterogeneous networks, as well as introducing standardization assessment, are still challenging tasks.

References

[1] V. Vasiliauskaite and F. E. Rosas, "Understanding complexity via network theory: a gentle introduction," arXiv Prepr. arXiv2004.14845, 2020.

[2] T. Cowman, M. Coşkun, A. Grama, and M. Koyutürk, "Integrated querying and version control of context-specific biological networks," Database, vol. 2020, 2020.

[3] A. Majeed and I. Rauf, "Graph theory: A comprehensive survey about graph theory applications in computer science and social networks," Inventions, vol. 5, no. 1, p. 10, 2020.

[4] B. O. Hansen and S. Olsson, "OmicsDB:: Pathogens-A database for exploring functional networks of plant pathogens," bioRxiv, 2020.

[5] R. Angles, M. Arenas-Salinas, R. García, J. A. Reyes-Suarez, and E. Pohl, "GSP4PDB: a web tool to visualize, search and explore protein-ligand structural patterns," BMC Bioinformatics, vol. 21, no. 2, pp. 1–15, 2020.

[6] J. Zhao, Z. Hong, and M. Shi, "Analysis of disease data based on Neo4J graph database," in 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), 2019, pp. 381–384.

[7] J. A. M. Stothers and A. Nguyen, "Can Neo4j replace PostgreSQL in healthcare?," AMIA Summits Transl. Sci. Proc., vol. 2020, p. 646, 2020.

[8] G. W. Huff and K. Cooper, "Correlation networks: Biologically driven relationships from gene expression data," in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 1712–1715.

[9] T. T. Aung and T. T. S. Nyunt, "Community detection in scientific co-authorship networks using neo4j," in 2020 IEEE Conference on Computer Applications (ICCA), 2020, pp. 1–6.

[10] D. Hoksza and J. Jelíinek, "Using Neo4j for mining protein graphs: a case study," in 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), 2015, pp. 230–234.

[11] T. C. Freeman et al., "Graphia: A platform for the graph-based visualisation and analysis of complex data," bioRxiv, 2020.

[12] B. Ristevski, "Using Graph Databases for Querying and Network Analysing," in IX International Conference on Applied Internet and Information Technologies (AIIT 2019), Zrenjanin, Serbia, 2019, pp. 28–32.

[13] B. Ristevski and S. Savoska, "Analysis and Simulation of Biological Complex Networks," in IX International Conference on Applied Internet and Information Technologies (AIIT 2019), Zrenjanin, Serbia, 2019, pp. 33–36.

[14] B. Ristevski and S. Savoska, "Healthcare and medical Big Data analytics," in Applications of Big Data in Healthcare, Elsevier, 2021, pp. 85–112.

[15] T. Wang, R. G. Qiu, M. Yu, and R. Zhang, "Directed disease networks to facilitate multiple-disease risk assessment modeling," Decis. Support Syst., vol. 129, p. 113171, 2020.

[16] E. P. G. Del Valle, G. L. García, L. P. Santamaría, M. Zanin, E. M. Ruiz, and A. Rodriguez-Gonzalez, "Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources," J. Biomed. Inform., vol. 94, p. 103206, 2019.

[17] R. Xu, "Statistical methods for the estimation of contagion effects in human disease and health networks," Comput. Struct. Biotechnol. J., vol. 18, pp. 1754–1760, 2020.

[18] H. Wu, Z. Fu, and Y. Wang, "A medical network clustering method with weighted graph structure," Meas. Control, vol. 53, no. 9–10, pp. 1751–1759, 2020.

[19] A. A. Namasivayam et al., "Community-reviewed biological network models for toxicology and drug discovery applications," Gene Regul. Syst. Bio., vol. 10, p. GRSB--S39076, 2016.

[20] A. Masoudi-Nejad, Z. Mousavian, and J. H. Bozorgmehr, "Drug-target and disease networks: polypharmacology in the post-genomic era," silico Pharmacol., vol. 1, no. 1, pp. 1–4, 2013.

[21] A. Lysenko, I. A. Roznovăţ, M. Saqi, A. Mazein, C. J. Rawlings, and C. Auffray, "Representing and querying disease networks using graph databases," BioData Min., vol. 9, no. 1, pp. 1–19, 2016.

[22] Q. Huang, "A Novel Important Node Discovery Algorithm Based on Local Community Aggregation and Recognition in Complex Networks," Int. J. Wirel. Inf. Networks, vol. 27, no. 2, pp. 253–260, 2020.

[23] V. Gligorijević and N. Pržulj, "Methods for biological data integration: perspectives and challenges," J. R. Soc. Interface, vol. 12, no. 112, p. 20150571, 2015.