HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1059.2021-0028
Natural Sciences 2021, Volume 66, Issue 2, pp. 57-65
This paper is available online at http://stdb.hnue.edu.vn

BUILDING AND MINING GRAPH DATABASES FROM BIOMEDICAL HETEROGENEOUS NETWORKS

Vu Duc Hung [1], Le Thi Huong [2] and Dang Xuan Tho[1]
[1]*Faculty of Information Technology, Hanoi National University of Education*
[2]*Faculty of Information Technology, University of Transport Technology*

Abstract. All things and phenomena in life, especially in the fields of life and biomedical medicine, are more or less related, interact with each other, forming a heterogeneous network. Therefore, when studying an object, we need to consider the relationships around it. However, current research often focuses on a specific object, not considering other subjects that are influencing it. Therefore, this paper proposes to use a graph database as an approach to dealing with heterogeneous networks solving the biomedical problem. Experimental results on two heterogeneous networks, miRNA-disease, and autism-miRNA-protein, has drawn the network of interactions, the relationships in a very intuitive way; shows the interaction between each specific object in the graph; and finally, statistics the interaction levels and shows the top 5 diseases, the top 5 miRNAs with the most interaction in the data. From there, it can be seen that the proposed method improves efficiency, increases accuracy, and reduces execution time compared to the traditional way of storing data before.

Keywords: biomedical network, heterogeneous biomedical data, data mining, graph database.

1. Introduction

Today, with the rapid development of Industrial Revolution 4.0, technology has been creeping into all areas of life such as biomedical, economics, finance, etc. In particular, the field of biomedical and biological development is very exciting, with a lot of techniques and experiments that have been invented and discovering knowledge, providing many kinds of useful information about biological components. genes, proteins, cells, tissues, diseases, etc., and their functions. These studies are mainly focusing on exploring individual biological components, although information about these individual biological constituents has its important implications in knowledge discovery [1].

Received May 7, 2021. Revised June 17, 2021. Accepted June 24, 2021.
Contact Dang Xuan Tho, e-mail address: thodx@hnue.edu.vn

However, biological characteristics are the result of many interactions between different biological factors, interacting with each other to produce specific reactions. Focusing only on individual components will be able to ignore key-factor relationships without fully understanding the nature of the biological problem [1]. Therefore, researching and understanding complex relationships between heterogeneous biological data components are very important to help us better understand the mechanisms and functions of the whole biological system. However, the problem of knowledge mining from heterogeneous biomedical networks is very difficult and complex [1-2].

To solve this difficult problem, many proposed methods have been developed to build a heterogeneous biomedical network, analyze and my knowledge from this network. However, for the majority of methods, the system currently uses traditional relational database management systems, such as SQL and Oracle [3]. Although, this traditional database management system can store biological network data with many different types of data such as genes, diseases, proteins, etc. But with these relational databases, it can face many limitations because of the way data is stored in the form of tables, records, properties. Meanwhile, the application of query statements, connecting tables by different relationships will become complicated in algorithm design, computationally expensive, making the algorithms real-time. current high, ineffective when the biological data is getting bigger and bigger [4].

In that context, a graph database was developed to easily integrate discrete components and objects to create a heterogeneous complex network to solve and explore the knowledge underlying these heterogeneous complex relationships. Graphical databases use the platform as a graph structure, with vertices being objects, entities, such as biological entities including genes, diseases, miRNAs, etc. and edges are the relationships between the top vertices, e.g. gene-disease relationship, miRNA-disease [5]. A graph database is essentially a graph structure built by vertices and edges so that connections and relationships between discrete data can be represented. From there, it can be seen those graph databases that represent relationships are significantly more efficient, simpler than previous traditional relational databases, and especially useful in fields. integration needs to represent many types of data integrated. Moreover, the graph database has inherited many graphing algorithms that have been developed with a long history. With algorithms on graphs, we can query as well as discover a lot of knowledge that on traditional databases is very complicated, expensive, and almost impossible to do. For example, what diseases are associated with the X gene? Or which miRNA is most likely to have a potential relationship with liver cancer? In recent years, the graph database has been widely applied in many fields including biomedical, social network analysis, computer science, data mining, and has achieved many successes.

In this paper, in order to explore the underlying knowledge in heterogeneous network data, we propose an approach to the Neo4J graph database and some graph algorithms such as PageRank, Community detection, and Similarity algorithm. Performing experiments on some heterogeneous data, we have obtained significant results, we can see that the proposed method improves efficiency, increases accuracy, and reduces execution time compared to the traditional way of storing data.

2. Content

2.1. Related works

In recent years, a number of researchers have begun to analyze biological networks of many complex components in order to uncover the hidden knowledge between the components that were previously often studied in fragments. The analysis of this heterogeneous network helps researchers better understand the nature and biological mechanisms, and the application of a graph database helps researchers to apply graph algorithms to analyze and provide quick and effective queries.

According to Lysenko *et al.* [6], today biomedical experiments have been generating enormous amounts of data with a variety of phenotypes, formats, and modes. These complex, heterogeneous diverse data with diverse semantics create a challenge for their integration. In this study, the authors pointed out that the graphical database is very suitable for representing biomedical information networks, because information of this type often has a large association, and is very difficult to make. predictable. The authors use the database of graph Neo4j to build a database of graphs and experiment several queries to extract and explore some potential relationships between genes related to asthma. The experimental results of the authors have provided a flexible and effective solution for integrating many types of mixed biomedical data, thereby creating a foundation for exploring hidden knowledge in these links.

Henkel *et al.* [7] has shown that the current biomedical data archive is frequently accessed to extract computational models for the exploration of biomedical systems. However, the current biomedical database archives are mostly relational, so there are many limitations in querying, some databases are still very difficult to access. and especially it is almost impossible to apply mathematical methods, data mining to extract models. In this study, the authors present a method of storage based on graph databases to solve these challenges by improving the structure of the model, combining semantic annotations of objects. and functional relationships between components that connect different types of data. This graph data structure helps to connect heterogeneous biomedical data and components, thereby enabling efficient search and query of biological data, the discovery of new functions and relationships such as gene-disease, miRNA-disease. The experimental results of the authors have shown that a graph database is an effective approach to store, represent, and query heterogeneous biomedical data.

Mullen *et al.* [8] approached the problem of drug repositioning by applying a database of graphs with a method of ranking the relationships between genes and diseases. The authors applied the Bayesian statistical method to rank 309,885 genetic-disease associations, and other components were also ranked and integrated with other biological data to create a heterogeneous lattice for extraction. destroy information related to drugs, particularly central nervous system (CNS) medicines. The data set is built and installed on the graphical database management system Neo4j. The authors then applied an association rating algorithm, identified and ranked for 275,934 drug-disease associations. Experimental results have shown that the proposed algorithm applied on the Neo4j graph database is highly efficient and reliable in practice with showing some potential relationships.

Balaur *et al.* [9] Access to colorectal cancer (CRC), the third most common cancer, and the StatEpigen database currently dedicated to CRC research. The authors point out that the

StatEpigen database system is managed and annotated very manually, although a graphical user interface is provided. However, if we integrate the components of this data, we can obtain more advanced queries, capable of exploring more hidden knowledge and understanding more about colorectal cancer. Since then, the authors have proposed to develop a database of graphs derived from original StatEpigen data. Experimental results have shown many advanced functions of this new graphical database, in particular detecting genetic and epigenetic interdependencies in intestinal adenomas; or prognosis of links between genes and CRC, etc.

2.2. Method overview

We have proposed a framework for building and mining latent relationships in graph data based on heterogeneous networks consisting of three steps. First, we construct a heterogeneous network-based graph database from three available biological databases. Second, we apply graph algorithms to the graph data we just built. Finally, mine the latent knowledge from the graph data. An illustration of the framework is shown in Figure 1.

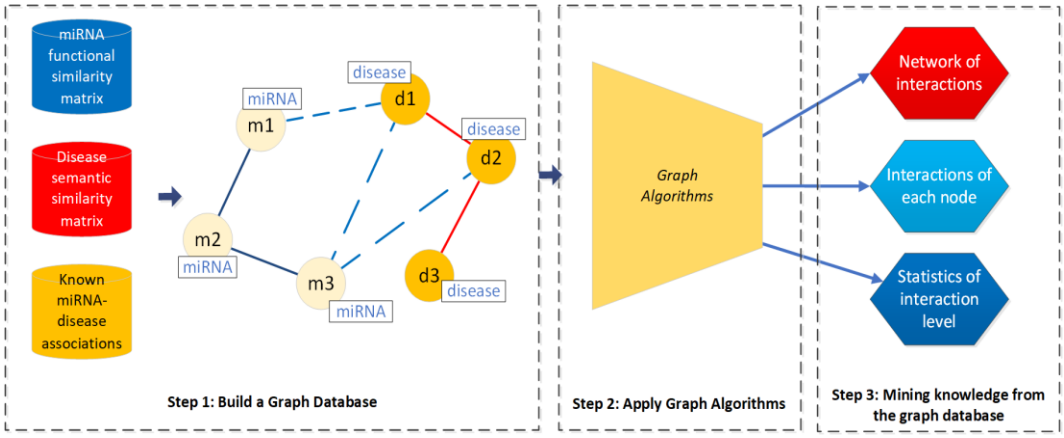

Figure 1. General workflow containing three main steps

2.2.1. Experimental biomedical heterogeneous networks

In this study, we use biomedical datasets to construct heterogeneous networks of miRNA-disease and autism-miRNA-protein.

The first data, miRNA-disease, we downloaded experimentally verified miRNA-disease links from HMDD v2.0, this dataset included 383 diseases, 495 miRNA and 5430 miRNA-disease pairs, 258 disease-disease pairs, 5386 miRNA-miRNA pairs have been identified experimentally before [10]. MiRNA functional similarity (miRNA-miRNA pairs) is built on the assumption that functionally miRNAs are more likely to be involved in similar and vice versa diseases. This method used the semantic similarity of the disease and the known associations between miRNAs and disease to the structure of the functional homology matrix of miRNAs. Disease-disease pairs were constructed to describe the relationship between diseases using the Directed Acyclic Graph (DAG) built against the MeSH database.

The second data we constructed to predict autism-related genes, this dataset consisted of 3 types of nodes (autism, 197 miRNAs, and 5175 proteins), and 3 types of edges (330 pairs of protein-autism, 3148 miRNA-protein pairs, 19988 protein-protein pairs) [5].

Details on how the training dataset was developed according to the method are presented in [5]. For each integration network built on the protein-protein interaction network (HINT, HPRD, BIOGRID) with 4 L_{max} values from 4 to 7. The data sets are named after protein-protein interaction network names and values of L_{max} (eg: HINT04, HPRD05, etc.).

The two data sets are depicted graphically as shown in Figure 2 and Figure 3.

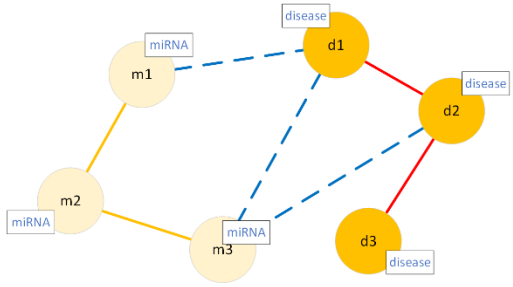

Figure 2. miRNA-disease heterogeneous network model

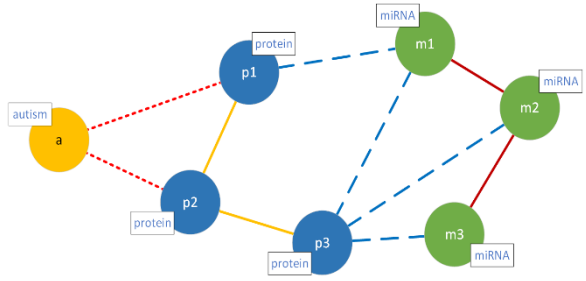

Figure 3. autism-miRNA-protein heterogeneous network model

2.2.2. Graph algorithms

Graph algorithms on Neo4j are developed to compute and explore latent patterns and patterns of nodes and relationships between nodes with algorithms such as behavior middleware, community discovery, link finder, etc. [11-13]. Many graph algorithms use computational algorithms like using random step patterns, depth traversing or width traversing, or pattern matching, etc. Based on these algorithms, we can learn a lot of important and valuable information such as center, ranking, community discovery, zoning, graph grouping, etc. Some of the graph algorithms used in this research to demonstrate the power of graph analysis in a graph database include Page Rank Algorithm, Overlapping Similarity Algorithm, Cosine Similar Algorithm [14-15].

The PageRank algorithm was proposed by the co-founder of Google, Larry Page, for ranking websites in Google search results. This algorithm will measure the influence of the links or the connectivity of the nodes by counting the number and quality of the links to a node to determine and estimate the importance of the node. there. The PageRank formula in the original Google article is defined by the formula as follows:

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}) \quad (1)$$

where suppose that page A has pages T_1 through T_n linking to it; d is the extinction factor valid in the range [0...1] and is usually set to 0.85; $C(A)$ is the number of outbound links from page A.

The Overlapping Similarity algorithm will calculate the overlap or similarity between two data sets based on the mathematical concept that the size of the intersection of two sets divided by the size of the smaller set of the two sets. The formula for calculating Overlap similarity is as follows:

$$O(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \quad (2)$$

Cosine similarity is defined as the cosine of the angle between two n-dimensional vectors in n-dimensional space, by calculating the dot product of two vectors divided by the product of the length (or magnitude) of the two vectors. The formula for calculating Cosine similarity is as follows:

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

2.3. Results and discussion

2.3.1. Draw a visual interactive network on the graph

In order to visualize the network of interactions between heterogeneous network elements, we proceed to build visual interactive network visualization. This functionality is extended by the active support of the Neo4j library.

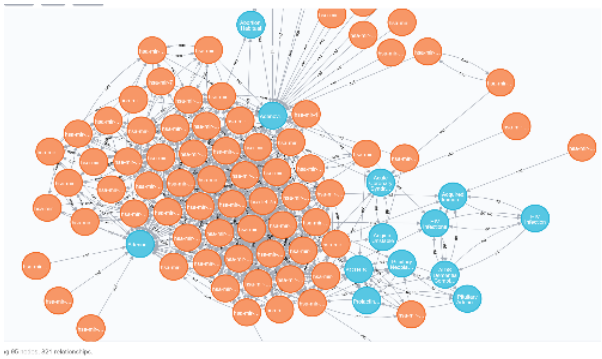

Figure 4. miRNA-disease heterogeneous network

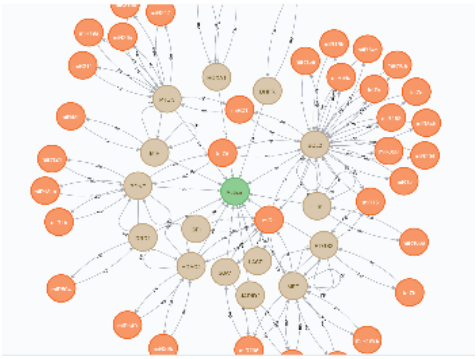

Figure 5. miRNA-protein-autism heterogeneous network

From the data of PPI Network, MiRNA Target Genes, Disease Genes, etc. initially put in, the program will conduct pre-processing of biological data to remove redundant data, then classify biological data learners need to display, and eventually cluster biological data. Experimental results we can reconstruct the network of interactions between genes, between genes and diseases, thereby giving us a better visualization and understanding of the relationship between genes and diseases.

2.3.2. Displays interactions of each specific object

The next function helps biomedical specialists analyze relationships and interactions of specific objects such as PPI Network, MiRNA Target Genes, Disease Genes, etc. Display functions can include such as displaying the interaction of each object, displaying the interaction of a group of objects, displaying objects by interaction type. In Figure 6 we can see the association of a miRNA named "has-mir-25" in the overall context with all diseases and associated miRNAs. Similarly, in Figure 7, we can see that the protein named "RBM19" is related to two miRNAs named "LNAlet7b" and "let7b". With this function, biomedical specialists can easily display the relationships of an object they are interested in.

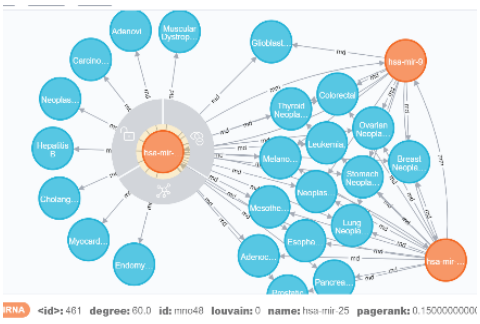

Figure 6. The has-mir-25 network of interactions

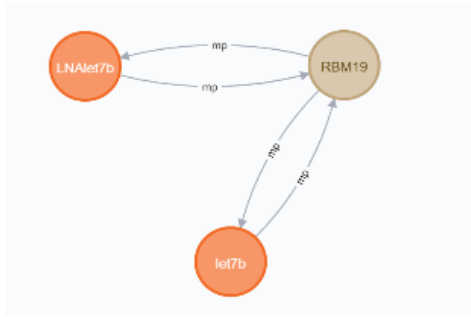

Figure 7. The RBM19 network of interactions

2.3.3. Statistics of interaction level information between the nodes

The next function helps biomedical specialists to statistic and display information about the interaction between PPI Networks, MiRNA Target Genes, Disease Genes. Based on these statistics, the system will discover nodes that have many relationships or many interactions with other nodes on this heterogeneous graph. From there, experts can analyze these subjects in more detail because they can play a very important role in life, as well as play important roles in biological functions, in healing, and assist in finding new drugs or repositioning drugs (Table 1). Statistics of the top 5 diseases with the highest degree of interaction in the graph data are the 5 diseases that are most likely to play an important role.

Based on the results of graph data mining combined with medical citations. We have some information about the outcomes of several highly interactive diseases in this data, as follows: Sézary syndrome is an aggressive form of a type of blood cancer called cutaneous T-cell lymphoma, and it is the second most common form of cutaneous T-cell cancer after fungicides. Lymphoma is a group of malignancies in the blood that develop from lymphocytes (a type of white blood cell) and is the most common form of hematologic malignancy, or "blood cancer" in developed countries. Osteoma is the 8th most common childhood cancer, accounting for 2.4% of all pediatric malignancies and about 20% of all primary bone cancers

Table 1. Top 5 diseases with the highest degree of interaction

No.	Protein	Score
1	Sezary Syndrome	9
2	Lymphoma	7
3	Lymphoma, Large-Cell, Anaplastic	6
4	Osteosarcoma	5
5	Lymphoma, Mantle-Cell	5

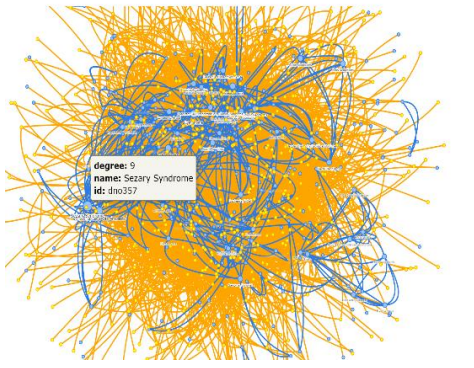

Figure 8. Visualize highly interactive diseases

Likewise, Table 2 shows the top 5 most correlated miRNAs in the heterogeneous network, if we can control these miRNAs, the more likely we are to control the diseases involved.

Table 2. Top 5 miRNAs with the highest degree of interaction

No.	miRNA	Score
1	hsa-mir-25	60
2	hsa-mir-419	59
3	hsa-let-7d	57
4	has-let-7f	57
5	has-let7a	55

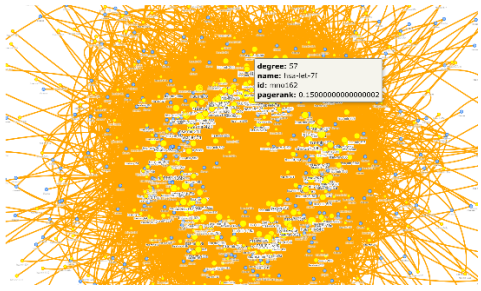

Figure 9. Visualize highly interactive miRNAs

3. Conclusions

In this scientific research, we have presented an overview of graph databases as a new approach to improve efficiency in exploring hidden knowledge in biomedical heterogeneous networks. By applying algorithms on a graph database, we have drawn the network of interactions and relationships between miRNA-disease, miRNA-protein-autism in a very intuitive way; show the interaction between each specific object in the graph; and finally, statistics the interaction levels and shows the top 5 diseases, the top 5 miRNAs with the most interaction in the data. The test results also show that the proposed method improves efficiency, increases accuracy, and reduces execution time compared to the traditional way of storing data before.

Based on the studies and results achieved, we find that there are still many issues that need further research. Specifically, we will continue to research and develop more algorithms on the heterogeneous network to better solve the problems of prediction, ranking, and clustering.

Acknowledgments. This research was supported by the Vietnam Ministry of Education and Training, project B2020-SPH-11.

REFERENCES

[1] Yoon, B. H., Kim, S. K., & Kim, S. Y., 2017. Use of graph database for the integration of heterogeneous biological data. *Genomics & informatics*, 15(1), 19.

[2] Chandrababu, S., & Bastola, D., 2019. Graph Model for the Identification of Multi-target Drug Information for Culinary Herbs. *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, Cham., p. 498-512.

[3] Celko, J., 2013. *Joe Celko's complete guide to NoSQL: What every SQL professional needs to know about non-relational databases*. Newnes.

[4] Istephan, S., & Siadat, M. R., 2016. Unstructured medical image query using big data-an epilepsy case study. *Journal of Biomedical Informatics*, 59, pp. 218-226.

[5] D. X. Tho, G. T. Trung, T. D. Hung, 2018. Autism-associated gene prognosis by machine learning model combined with data balance method. *HNUE Journal of Science*, Vol. 63, Issue 11A, pp. 124-133 (in Vietnamese).

[6] Lysenko, A., Roznovăţ, I. A., Saqi, M., Mazein, A., Rawlings, C. J., & Auffray, C., 2016. Representing and querying disease networks using graph databases. *BioData Mining*, 9(1), 1-19.

[7] Henkel, R., Wolkenhauer, O., & Waltemath, D., 2015. Combining computational models, semantic annotations and simulation experiments in a graph database. *Database*.

[8] Mullen, J., Cockell, S. J., Woollard, P., & Wipat, A., 2016. An integrated data driven approach to drug repositioning using gene-disease associations. *PloS one*, 11(5), e0155811.

[9] Balaur, I., Saqi, M., Barat, A., Lysenko, A., Mazein, A., Rawlings, C. J., & Auffray, C., 2017. EpiGeNet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. *Journal of Computational Biology*, 24(10), pp. 969-980.

[10] Van Thai, T., Bui, D. H., Dang, X. T., Nguyen, T. P., & Tran, D. H., 2021. A New Computational Method Based on Heterogeneous Network for Predicting MicroRNA-Disease Associations. *Soft Computing for Biomedical Applications and Related Topics*. Springer, Cham., pp. 205-219.

[11] Robinson, I., Webber, J., & Eifrem, E., 2015. *Graph databases: new opportunities for connected data*. O'Reilly Media, Inc..

[12] Van Bruggen, R., 2014. *Learning Neo4j*. Packt Publishing Ltd.

[13] Kemper, C., 2015. *Beginning Neo4j*. Apress.

[14] Vukotic, A., Watt, N., Abedrabbo, T., Fox, D., & Partner, J., 2015. *Neo4j in action* (Vol. 22). Shelter Island: Manning.

[15] Needham, M., & Hodler, A. E., 2019. *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly Media.