Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2022, 11(2):62-72
https://doi.org/10.51847/VMnq3K5OmN

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Graphical Data Representation and Analytics to Link the Potential Interaction for Lung Cancer Genes

Bandar Hamad Aloufi[1]*, Ahmad Mohajja Alshammari[1]

[1]Department of Biology, College of Science, University of Hail, Kingdom of Saudi Arabia.

***Email:** bandaraloufi@yahoo.com*

ABSTRACT

Graph data representation is an efficient technique for highlighting the relationship among huge and highly linked biological data. With the advent of next-generation sequencing, a large volume of data is being generated. Currently, many graph tools exist to extract semantically associated data. Neo4j, Titan, and OrientDB are examples of well-known graphical representation tools. In this paper, a perfect graphical data storage and interaction retrieval model for lung cancer genes are presented which are collected from different types of databases such as Uniprot, COSMIC, and NCBI. The model contains interactions between genes, proteins, protein domains, their expression in various tissues, involvement in other diseases and their corresponding role, disorder type, mutation location, disease description, etc. By applying different types of queries, many types of unknown relationships have been uncovered which were not well studied earlier such as three proteins named KRAS, NRAS, and RIT1, which have a common domain. Similarly, groups of genes have shown no expression in any other organ except the lungs. Some groups of genes have some types of somatic disorders showing that they may have a related genetic basis. Through the deep analysis, we found different groups of genes which have the same disorder type, mutation location for different diseases, and different genes playing a crucial role in the development of various diseases including lung cancer. Such a type of analysis helps design drugs against the highlighted cause factors of diseases.

Key words: *Graph analytics, Graph database, Cancer, Protein-protein interaction, Protein domains, Mutations*

INTRODUCTION

Big data is one of the emerging fields across the world. There are five main components of big data namely Volume, Variety, Velocity, Veracity, and Value. Data sets are growing rapidly day by day and becoming big data [1]. We need new powerful techniques to manipulate such huge and varied data. Biological data is the most varied data as it can be structured, semi-structured, and unstructured. There are different types of famous databases such as the Molecular Interaction Database (MINT) [2], Database of Interaction Protein (DIP) [3], Biomolecular Interaction Networks Database (BIND) [4], Reactcome [5], STRING [6], Unified Human Interactome (UniHI) [7], Online Mendelian Inheritance in Man (OMIM) [8], Kyoto Encyclopedia of Genes and Genomes (KEGG) [9], Human Protein Reference Databases (HPRD) [10], Biological General Repository for Interaction Datasets (BioGrid) [11], National Center for Biotechnology Information (NCBI) [12], and Universal Protein Resource Knowledgebase (UniprotKB) [13] all of which could be used for storing and analyzing this data. The management and analysis of such huge and varied data has become a challenge for the scientific community.

Currently, graph databases have taken a central place in big data analytics, as there are a lot of important and complicated relationships among the data. Thus, the new era needs to analyze these complex graphs and extract

the most important entity relationships which play an important role in a process. Graphs contain nodes (vertices) which are used to show the entities and edges (relationships), which are used to demonstrate the connection between nodes. In our actual lives, each and everything is connected and can be shown in the form of a graph. With the advent of new technology, it has become very important that we should have complete knowledge of all the stored data, its management procedures, and possible analysis [13].

Cancer is famous for its other name "the genome variation disease". Sequence analysis leads to the recognition of mutational genes in cancer, especially those genes which play a vital role in cancer development like oncogenes [14]. Furthermore, the structural and functional analysis also helps us to check out the expression complexities of these mutated genes. Cancer is one of the leading causes of death. There are almost more than 7.4 million deaths worldwide per year, from which the lung cancer rate is over 1.4 million deaths per year [15, 16]. Due to the introduction of new DNA analysis technologies like next-generation sequencing the analysis of mutation has shifted from a single gene to multiple genes or groups of genes, as in cancer there is always more than one gene responsible for the disease development. Lung cancer is mainly divided into two categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). 85% of total lung cancers are NSCLC, which is further classified into two main subtypes: squamous cell carcinoma (SCC) and adenocarcinoma (AD). The recognition of total altered genes at an early stage can control lung cancer [17, 18].

Graphical representations of biological data are important in visualization and analytics. A graph database uses graph structure. The basic idea is to develop a graph structure that consists of nodes (biological entities) and edges (relationships among nodes) of stored data. The graph database can handle all types of flexibility in data easily. It can also manage all types of relationships in heavily linked data. Many scientists have moved to graph databases for the representation of biological data. For example, [19] used the graph databases for data storage and representation of the disease networks and launched a flexible structure to create new hypotheses.

Bio4j [20] launched a graph database (Neo4j) which can combine information from main repositories, for example, Gene Ontology (GO) [21] NCBI and Expasy Enzyme DB [22], RefSeq [23], Taxonomy [24]. Neo4j is a freely available open-source NoSQL graph database that can run at any type of database server and has its query language (like SQL) called Cypher query language (CQL), which is a declarative language.

In this paper, a graphical representation of biological data is designed for lung cancer. This model contains the mutated genes involved in lung cancer and these genes are connected with other corresponding information from these genes. Each gene has an encoding protein and this protein has one or more domains. Similarly, each gene has an expression in other organs and is also involved in other diseases. In other diseases protein has some specific role, e.g. playing a role in some signaling pathway, making a complex with other proteins, playing a big role due to mutation, or co-expressing with other genes and leading from the normal procedure to disease development. In this model, the role of protein in other diseases is also merged by literature searches. Similarly, other related information like mutation location, disorder type, and disease description are also merged. The model will give a short view that starts from the mutated gene and proceeds to its corresponding protein, domains, expression, etc. It ends with the role of this mutated protein in causing other different diseases.

MATERIALS AND METHODS

Dataset preparation

First, the mutated lung cancer genes in altered chromosomal regions were collected from the integrated genome database of non-small cell lung carcinoma (IGDB.NSCLC). These mutated genes are taken from mainly two types of regions: the amplified regions and the deleted regions from squamous cell carcinoma and adenocarcinoma, which are the two subcategories of non-small cell lung cancer (http://igdb.nsclc.ibms.sinica.edu.tw/mutation.php).

Other genes such as those reported in various works of literature are also retrieved. The genes' ID, expression, and involvement in other diseases were taken from Uniprot manually. Domains and Repeats are collected from Interpro. Similarly, the data regarding the role of lung cancer genes in causing other related diseases like Hereditary desmoid disease (HDD) and Colorectal cancer (CRC)was obtained by literature review, e.g. mutation in EPHA3 causes lung cancer but it is also involved in colorectal cancer CRC [25]. The type of disorder, mutation location, and disease description were taken from OMIM. The data was converted into CSV file format. The file with gene details of lung cancer was named "Lung_Cancer_Data" and the file with complete details of diseases and their role was named "Lung_Cancer_Data_Disease_role". These two files were placed into the import folder of Neo4j so that they can be directly accessed by CQL from the import folder.

Graphical model development

The CSV files which contain data related to lung cancer genes and their other details were uploaded at Neo4j for making the graphical representation. We got the complete graphical data model of the uploaded files in **Figure 1a**.

Different groups were generated in the graphical model which was hidden before. The genes which have similar expressions were grouped and linked together in graphical form. Similarly, the proteins which share the same domains, role, disorder type, and mutation location have formed relationships and come in the form of groups in the graphical model. Some protein-making hubs are also highlighted in the model, i.e. the proteins which are not involved in any other diseases except lung cancer are gathered in one place and their other detail like their domains and expression is also linked with them. Clusters were separated in the graph, e.g. the cluster of those proteins which have the same disorder type.

RESULTS AND DISCUSSION

Graph databases are very useful in explaining the relationships between genes and diseases. We illustrate different queries which are important in understanding the role of different genes and proteins and their domains in developing lung cancer, their expression in different organs other than the lungs in the aspect of spreading the disease, and the role of these lung cancer-causing genes in developing other diseases. We formulated different queries such that the query for extracting common domain among different proteins and a query to check the overlapping role of protein in developing a disease. A total of 657 nodes, 657 labels, and 630 relationships were created, and 391 properties were set.

Identification of different proteins with common domains

We describe a simple query that returns the different proteins, which have the same domains from the lung cancer genomics dataset **(Figure 1a)**. The query is given in Listing1.

```
$ MATCH p= ()-[r: hasDomainAndRepeats] -> () RETURN p limit 300 (1)
```

Listing 1. Cypher query to discover different proteins which have the same domains and repeats in lung cancer genes.

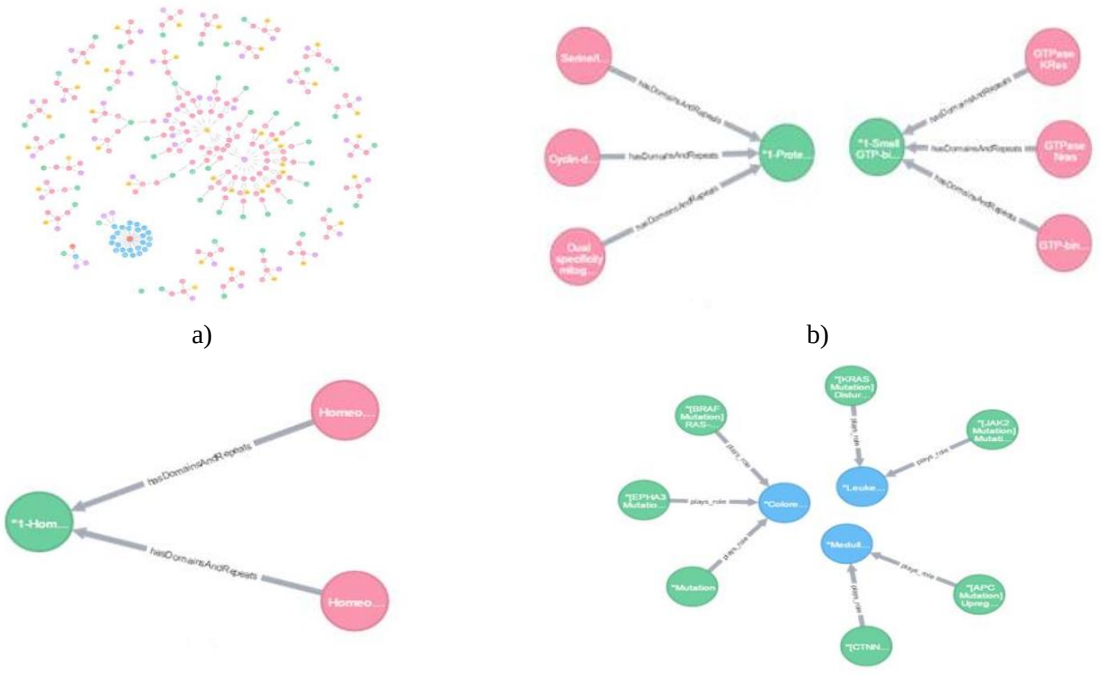
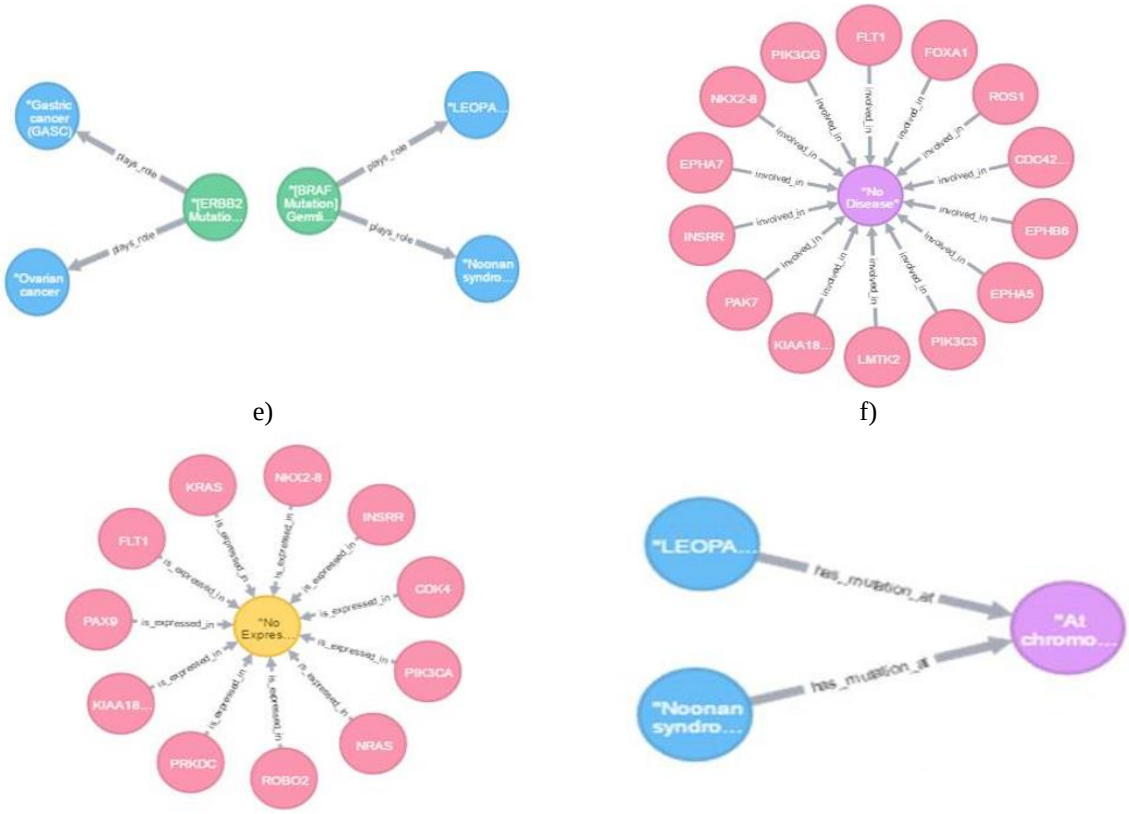


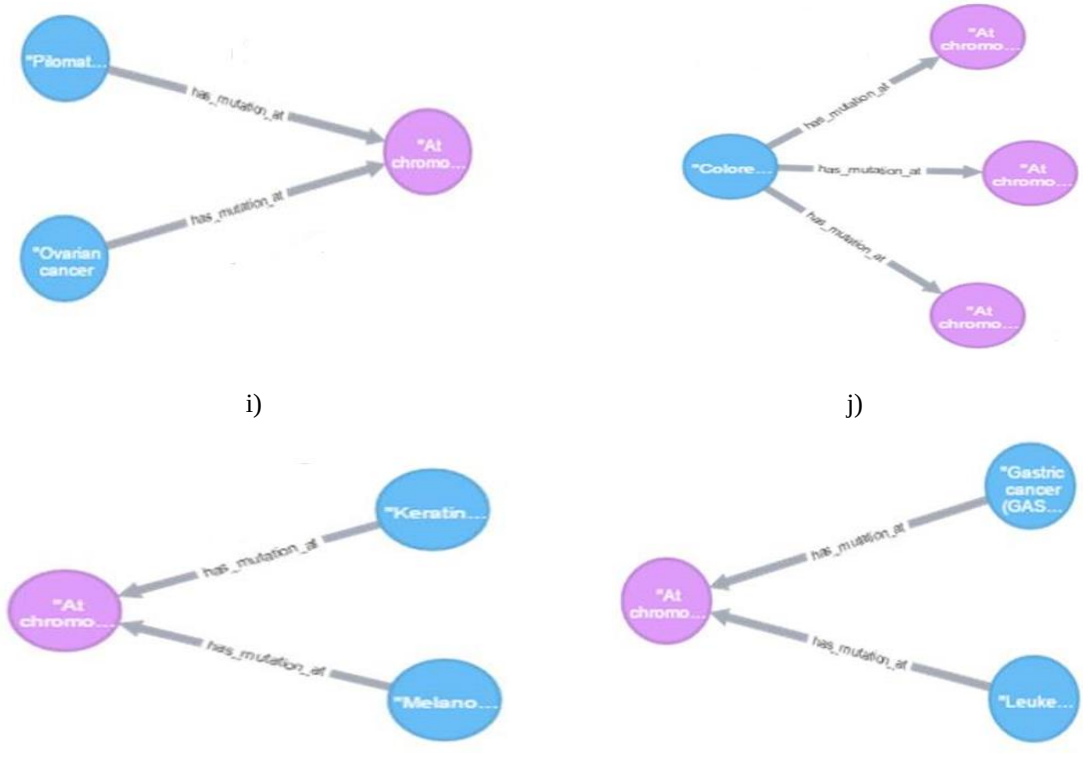
Figure 1. a) Graphical Data model of lung cancer genes and its other linked details. b) Two groups of proteins having common domains. c) A group of two proteins that share a common domain. d) various groups of lung cancer proteins that play a special role in developing other diseases.



e) f)

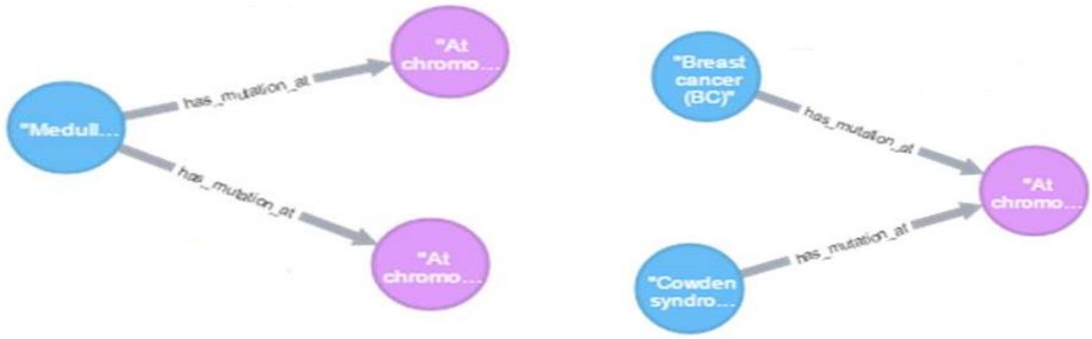
g) h)

Figure 2. e) Two proteins that are involved in different diseases. f) 15 proteins that have no role in causing any other disease except lung cancer. g) Proteins that have no expression in any other organs except the lungs. h) Two different diseases are caused by the same lung cancer gene mutation.



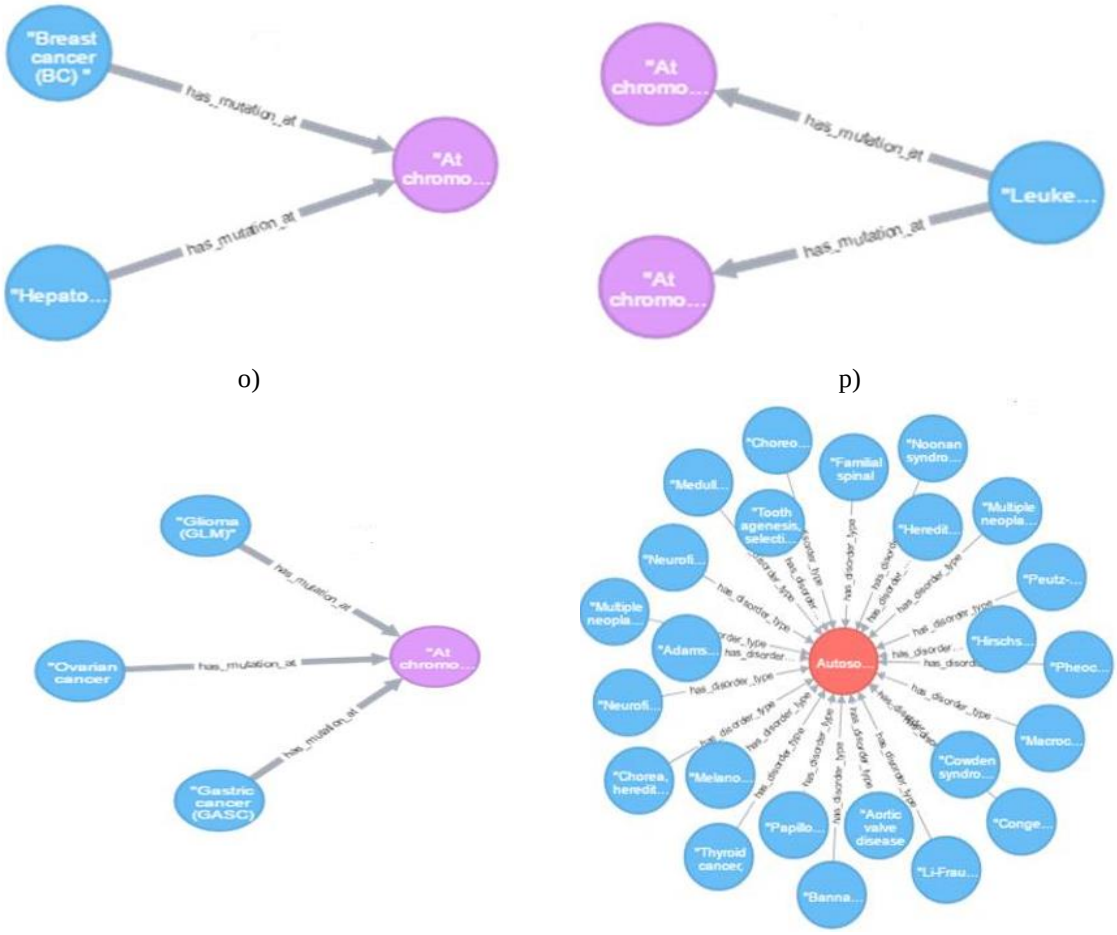
i) j)

k) l)



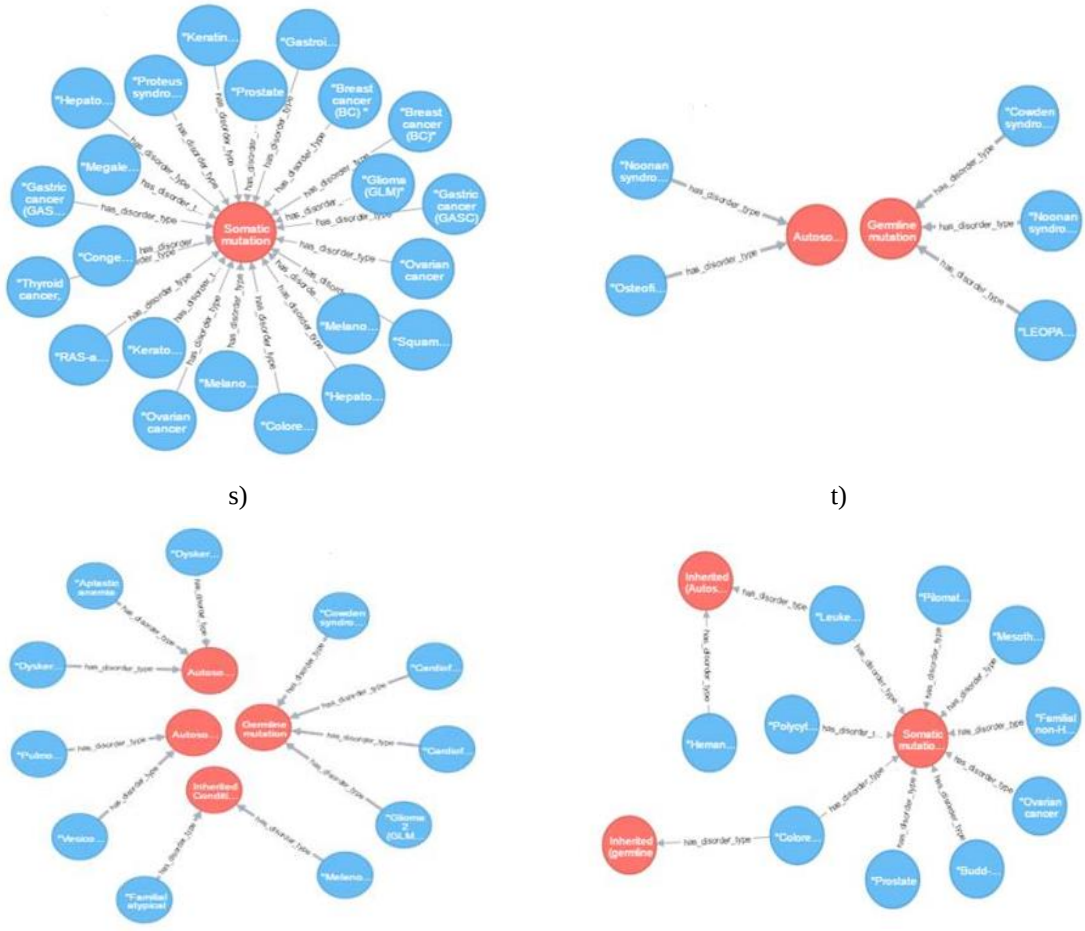
m) n)

Figure 3. i) Two diseases that have the same mutation in a lung cancer gene. j) Group of three different mutations which can cause colorectal cancer (CRC). k) Two diseases having the same mutation. l) Two different diseases having the same mutation. m) Two genes having the same mutation in a disease. n) Group of two diseases having the same mutation.



q) r)

Figure 4. o) Group of two diseases having the same mutation. p) Mutation in two genes involved in the same disease. q) A group of diseases caused due to the same mutation in ERBB2. r) Cluster of diseases that have an autosomal dominant mutation.



s) t)

u) v)

Figure 5. s) Cluster of diseases that have a somatic mutation. t) One group of diseases has germline mutation and the other group has an autosomal dominant mutation. u) Three groups of diseases that have overlapping disorder types. v) Group of diseases which share the same disorder type.

This query gives the proteins along with their domains which we identify as two groups of proteins that have common domains. Three different proteins named Serine/threonine-protein kinase STK11, Cyclin-dependent kinase 4, and Dual specificity mitogen-activated protein kinase 1 have a common domain called Protein kinase domain (IPR000719). Similarly, another group is identified which consists of three proteins known as GTPase KRas, GTP-binding protein Rit1, and GTPase Nras having the same domain known as Small GTP-binding protein domain (IPR005225) **Figure 1b**.

Similarly, the last group of genes which we have identified consists of two different proteins which have the same domain. This group of proteins includes two proteins named Homeobox protein Nkx-2.1 and Homeobox protein Nkx-2.8 which have the same domain called Homeobox domain (IPR001356) and Homeobox domain, metazoa (IPR020479) **Figures 1c and 1d**.

Identification of different proteins' roles in the same or different diseases

Similarly, the query for identifying the different groups of proteins that have an important role in developing the same disease as well as different diseases is given in Listing 2.

```
$ MATCH p= ()-[r: plays_role] -> () RETURN p LIMIT 300 (2)
```

Listing 2. Cypher query to discover a group of proteins that have an important role in developing the same disease as well as different diseases.

We noted three proteins, EPHA3, BRAF, and CTNNB1, which are involved in the same disease called colorectal cancer (CRC). Similarly, proteins KRAS and JAK2 are also involved in developing the disease called Leukemia, acute myelogenous (AML). In the same way, APC and CTNNB1 proteins play important role in

causing the Medulloblastoma (MDB) **Figure 2e**.

ERBB2 protein plays an important role in the development of gastric cancer (GASC) and Ovarian Cancer (OC). Similarly, BRAF is playing a key role in causing two different diseases called LEOPARD syndrome 3 (LPRD3) and Noonan syndrome 7 (NS7) **Figure 2f**.

By applying another query given in Listing 3 we got 15 proteins, NKX2-8, EPHB6, EPHA5, LMTK2, INSRR, PIK3C3, PAK7, ROS1, KIAA18, PIK3CG, FOXA1, CDC42 and FLT1, which are purely involved in lung cancer and are not involved in any other disease **Figure 2g**.

```
$ MATCH p= ()-[r: involved_in] -> () RETURN p LIMIT 300 (3)
```

Listing 3. Cypher query for discovering those proteins which are not involved in any other disease except lung cancer.

Similarly, by applying another Cypher query. we were able to extract those proteins which have no expression in any other tissue or organ except in the lungs **Figure 2h**. The query is given in Listing 4. The proteins NKX2-8, CDK4, KRAS, PIK3CA, NRAS, ROBO2, INSRR, PRKDC, KIAA18, FLT1and PAX9 have expression only in the lungs and show no expression in other organs.

```
$ MATCH p= ()-[r: is_expressed_in] -> () RETURN p LIMIT 300 (4)
```

Listing 4. Cypher query to find those proteins which have no expression in other organs except the lungs. Similarly, we can also uncover diseases that have the same mutation location of different genes. The query is given in Listing 5.

```
$ MATCH p= ()-[r: has_mutation_at] -> () RETURN p LIMIT 300 (5)
```

Listing 5. Cypher query for discovering those diseases which have the same mutation location of different proteins.

We found a total of 10 clusters, some of which include different diseases which have the same mutation location of a protein. In the same way, we found a group of different protein mutations involved in the same diseases. LEOPARD syndrome 3 (LPRD3) and Noonan syndrome 7 (NS7) have the same mutation at chromosome 7q34 in exons 6,11,12,13,15,16,17 of BRAF protein **Figure 3i**. Similarly, Pilomatrixoma (PTR) and ovarian cancer (OC) have the same mutation at chromosome 3p22.1 in exon 3 of CTNNB1 protein **Figure 3j**. Mutation in BRAF, EPHA3, and CTNNB1 can also cause colorectal cancer (CRC) **Figure 3k**.

Keratinocytic non-epidermolytic nevus (KNEN) and Melanocytic nevus syndrome, congenital (CMNS) have the same mutation in NRAS at chromosome 1p13.2 in codon 61 **Figure 3l**. Leukemia, juvenile myelomonocytic (JMML) and Gastric cancer (GASC) have the same mutation in KRAS at chromosome 12p12.1 in codon 12 **Figure 3m**. Similarly, we can detect those genes which have the same mutation location for a disease. The same mutations in CTNNB1 and APC cause Medulloblastoma (MDB) **Figure 3n**. Similarly, Breast cancer (BC) and Cowden syndrome 6 (CWS6) have the same mutation at chromosome 14q32.33 in E17K and AKT1 genes **Figure 4o**.

Breast cancer (BC) and Hepatocellular carcinoma (HCC) have the same mutation in PIK3CA at chromosome 3q26.32 in exons 9 and 20 **Figure 4p**. A mutation in KRAS and JAK2 causes Leukemia, acute myelogenous (AML) **Figure 4q**. Glioma (GLM), Ovarian cancer (OC), and Gastric cancer (GASC) have the same mutation at chromosome 17q12 in the kinase domain in the ERBB2 gene **Figure 4r**.

Similarly, we can group those diseases which have the same disorder type as if the nature of the disease is the same then there is a possibility that they will also share their genetic factors **Figures 5s-5v**. The query to extract these diseases is given in Listing 6.

```
$ MATCH p= ()-[r: has_disorder_type] -> () RETURN p LIMIT 300 (6)
```

Listing 6. Cypher query to find the cluster of diseases which share the same disorder type.

When the data model of lung cancer genetics came into a graphical representation, a bulk of information came

into existence that was neglected in the table form like overlapping domains of different proteins, overlapping expression of different genes, overlapping mutation type and disorder type of different genes and overlapping disease of different genes.

In the results, we have seen that listing 2 returns the proteins with the same domains or share some domains. In this regard, the same domains of different proteins lead us to the possibility that they will surely share their function also, as domains are the functional part of the proteins. We have uncovered three proteins, Serine/threonine-protein kinase STK11, Cyclin-dependent kinase 4, and Dual specificity mitogen-activated protein kinase 1, which have an overlapping domain called Protein kinase domain (IPR000719). Protein kinase has a key role in several cellular processes which include division, proliferation, apoptosis, and differentiation [26]. As domains are the functional part of the protein, if some proteins have similar or same domains then they will perform the same function no matter in which organism they exist [27]. Similarly, we retrieved different proteins which have a role in developing a disease other than lung cancer. We found three proteins EPHA3, BRAF, and CTNNB1, which are involved in the disease called colorectal cancer (CRC). This means that these cancerous genes are not only involved in lung cancer but are also responsible for causing other types of diseases. Likewise, the overlapping role of the disease gave us the idea that there exist some pathways or processes which are important with respect to the development of different types of diseases. This means that if these shared pathways of processes are controlled then we can overcome different types of diseases [28]. As in **Figure 2e**, the proteins which are not involved in any other disease except lung cancer led us to the concept that these proteins are playing a critical role in causing lung cancer [29, 30]. By targeting these genes, we can also design some drugs to cure lung cancer properly [31]. Similarly, **Figure 2f** shows the proteins which have no expression in any other organ and tissue except the lungs, which is why they are also the leading cause of lung cancer as they are expressed in the lungs only, which means that they produce all their proteins in lungs [32].

In the same way, **Figure 2g to 5t** shows different proteins which have the same mutation location which means that these mutations have a very special link that generates a group of diseases collectively and form some type of syndrome by causing these mutations [33]. Such analysis can generate multidimensional results which can cover a large amount of linked information, due to which we can easily discover the main cause of a disease by checking every possible cause and factor. In the future, by detecting the cause of disease we can easily design specialized drugs against specific diseases. This will open the scope for new methods of treatment of a disease.

A large volume of data is being generated by different biological sources which are high in semantics; it provides a challenge for the scientific community to integrate such type of data. Graph databases beat this challenge very efficiently due to their powerful support and storage system etc. In medical science, it has become very important to study the molecular basis of diseases to identify the genes, proteins, and corresponding pathways. For this purpose, the integrated analysis of data is useful to understand the complete process of the working of disease pathology. For example, Neo4j is widely used for forming a network of different biological processes and disease protein interaction networks e.g. the asthma-related genes network, etc. [19]. A different type of biological data is generated daily, and based on this data different biological networks are also made. These networks are very useful for understanding biological processes. However, analyzing these highly connected and interlinked networks and understanding them deeply is also a very hectic task. Different types of techniques are developed to understand these networks to extract the important relationship between different entities. Most of the research is being conducted to develop different methods for different types of biological networks or to solve different biological problems. However, there are different benefits and drawbacks of these methods. Different integration techniques need specific analysis methods accordingly [34].

Cancer is a result of many mutations in genes leading to disturbance in many complex biological pathways and networks and influencing the many functions of cells. Therefore, to cure cancer scientists need to analyze and understand many networks and signaling pathways computationally. The Atlas of cancer signaling networks (ACSN) provides a map of different interaction networks specialized for cancer [35]. It provides a lot of tools for the visualization and analysis of cancer-related data in detail. It is very helpful in understanding different mechanisms, e.g. cancer spreading, apoptosis, cell survival, etc. [36, 37].

Consensus PathDB is widely used to collect the human molecular data which are gathered from 32 various repositories. It is also a web interface that provides the features to visualize the interaction between this data and offers different methods to analyze this data more deeply. Biomolecules are used to form a network of biological processes [38, 39]. The tool is mainly used to form the interaction networks between biological molecules or the interaction pathways. These networks and pathways are then analyzed by applying different

statistical methodologies. The results are very efficient and easy to analyze. These networks are used to recognize different functions of the cell, identify clusters, and develop topology for generating different other models [40-44].

CONCLUSION

A large amount of biological data has been generated from different resources. Recently, a great interest has developed to analyze this huge and super complex data and its understanding can open the door to the solutions to many different biological problems such as finding the root cause of a disease, searching for the interacting cluster of proteins which plays important role in producing different diseases and drug designing against these diseases by the combined analysis of these genes. By making a graphical model of such data we can increase understanding, which can help us to reveal the important relationships between different entities or groups of entities.

In this study, we found that there are overlapping domains of different lung cancer proteins which are purely involved in lung cancer, which means that these genes work in the form of groups or clusters to develop the disease. These findings provide us with the cause factor of disease and by designing some drugs against these clusters to stop their domain functionality; we can control lung cancer or its spread. Similarly, those genes which are involved in some pathways are also involved in more than one disease as those pathways are responsible for performing different functions. Any disturbance in those pathways disrupts all those functions which were performed with the help of those pathways; that's the main reason that one gene was playing a crucial role in more than one disease. This way we can identify those pathways and try to fix those genes which were causing a disturbance in those pathways, which will be very helpful in treating those groups of diseases at the same time.

ACKNOWLEDGMENTS : None

CONFLICT OF INTEREST : None

FINANCIAL SUPPORT : None

ETHICS STATEMENT : None

REFERENCES

1. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inform. 2015;19(4):1209-15.
2. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40(Database issue):D857-61.
3. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004;32(Database issue):D449-51.
4. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. 2005;33(Database issue):D418-24.
5. Haw R, Stein L. Using the reactome database. Curr Protoc Bioinformatics. 2012;Chapter 8:Unit8.7.
6. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362-D8.
7. Kalathur RK, Pinto JP, Hernández-Prieto MA, Machado RS, Almeida D, Chaurasia G, et al. UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. Nucleic Acids Res. 2014;42(Database issue):D408-14.
8. Feramisco JD, Sadreyev RI, Murray ML, Grishin NV, Tsao H. Phenotypic and genotypic analyses of genetic skin disease through the Online Mendelian Inheritance in Man (OMIM) database. J Invest Dermatol. 2009;129(11):2628-36.
9. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109-14.

10. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. A. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009;37(Database issue):D767-72.
11. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):D529-D41.
12. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013;41(Database issue):D43-7.
13. Debrouvier A, Parodi E, Perazzo M, Soliani V, Vaisman A. A model and query language for temporal graph databases. VLDB J. 2021;30(5):825-58.
14. McGuire S. World Cancer Report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO Press, 2015. Adv Nutr. 2016;7(2):418-9.
15. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015;520(7547):353-7.
16. Arora A, Gera S, Maheshwari T, Raghav D, Alam MJ, Singh RK, et al. The dynamics of stress p53-Mdm2 network regulated by p300 and HDAC1. PLoS One. 2013;8(2):e52736.
17. Nikiforov YE, Carty SE, Chiosea SI, Coyne C, Duvvuri U, Ferris RL, et al. Impact of the Multi-Gene ThyroSeq Next-Generation Sequencing Assay on Cancer Diagnosis in Thyroid Nodules with Atypia of Undetermined Significance/Follicular Lesion of Undetermined Significance Cytology. Thyroid. 2015;25(11):1217-23.
18. Fehringer G, Brenner DR, Zhang ZF, Lee YA, Matsuo K, Ito H, et al. Alcohol and lung cancer risk among never smokers: A pooled analysis from the international lung cancer consortium and the SYNERGY study. Int J Cancer. 2017;140(9):1976-84.
19. Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Min. 2016;9:23.
20. Pareja-Tobes P, Tobes R, Manrique M, Pareja E, Pareja-Tobes E. Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv. 2015:016758.
21. Gene OC. Gene ontology consortium: going forward. Nucleic Acids Res. 2015;43(Database issue):D1049-56.
22. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 2000;28(1):304-5. doi:10.1093/nar/28.1.304
23. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. 2014;42(Database issue):D553-9.
24. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40(Database issue):D136-43.
25. Nasri B, Inokuchi M, Ishikawa T, Uetake H, Takagi Y, Otsuki S, et al. High expression of EphA3 (erythropoietin-producing hepatocellular A3) in gastric cancer is associated with metastasis and poor survival. BMC Clin Pathol. 2017;17(1):8.
26. Manning G, Plowman GD, Hunter T, Sudarsanam S. Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci. 2002;27(10):514-20.
27. Azevedo RF, Gonçalves-Vidigal MC, Oblessuc PR, Melotto M. The common bean COK-4 and the Arabidopsis FER kinase domain share similar functions in plant growth and defence. Mol Plant Pathol. 2018;19(7):1765-78.
28. Sever R, Brugge JS. Signal transduction in cancer. Cold Spring Harb Perspect Med. 2015;5(4):a006098.
29. Wei J, van der Wekken AJ, Saber A, Terpstra MM, Schuuring E, Timens W, et al. Mutations in EMT-Related Genes in ALK Positive Crizotinib Resistant Non-Small Cell Lung Cancers. Cancers (Basel). 2018;10(1):10.
30. Su C, Zhang J, Yarden Y, Fu L. The key roles of cancer stem cell-derived extracellular vesicles. Signal Transduct Target Ther. 2021;6(1):1-5.
31. Nicolaou KC, Erande RD, Yin J, Vourloumis D, Aujay M, Sandoval J, et al. Improved Total Synthesis of Tubulysins and Design, Synthesis, and Biological Evaluation of New Tubulysins with Highly Potent Cytotoxicities against Cancer Cells as Potential Payloads for Antibody-Drug Conjugates. J Am Chem Soc. 2018;140(10):3690-711.
32. Kim JH, Sherman ME, Curriero FC, Guengerich FP, Strickland PT, Sutter TR. Expression of cytochromes P450 1A1 and 1B1 in human lung from smokers, non-smokers, and ex-smokers. Toxicol Appl Pharmacol. 2004;199(3):210-9.

33. Iacovazzo D, Flanagan SE, Walker E, Quezado R, de Sousa Barros FA, Caswell R, et al. MAFA missense mutation causes familial insulinomatosis and diabetes mellitus. Proc Natl Acad Sci U S A. 2018;115(5):1027-32.
34. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface. 2015;12(112):20150571.
35. Nobari H, Saedmocheshi S, Johnson K, Suzuki K, Maynar-Mariño M. A Brief Overview of the Effects of Exercise and Red Beets on the Immune System in Patients with Prostate Cancer. Sustainability. 2022;14(11):6492.
36. Kuperstein I, Bonnet E, Nguyen HA, Cohen D, Viara E, Grieco L, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis. 2015;4(7):e160.
37. Lonjou C, Eon-Marchais S, Truong T, Dondon MG, Karimi M, Jiao Y, et al. Gene-and pathway-level analyses of iCOGS variants highlight novel signaling pathways underlying familial breast cancer susceptibility. Int J Cancer. 2021;148(8):1895-909.
38. Gupta OP, Deshmukh R, Kumar A, Singh SK, Sharma P, Ram S, et al. From gene to biomolecular networks: a review of evidences for understanding complex biological function in plants. Curr Opin Biotechnol. 2022;74:66-74.
39. Zhang Y, Shen L, Zhong QZ, Li J. Metal-phenolic network coatings for engineering bioactive interfaces. Colloids Surf B Biointerfaces. 2021;205:111851.
40. Alam MJ, Kumar S, Singh V, Singh RK. Bifurcation in Cell Cycle Dynamics Regulated by p53. PLoS One. 2015;10(6):e0129620.
41. Devi GR, Alam MJ, Singh RK. Synchronization in stress p53 network. Math Med Biol. 2015;32(4):437-56.
42. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc. 2016;11(10):1889-907.
43. Malik MZ, Alam MJ, Ishrat R, Agarwal SM, Singh RK. Control of apoptosis by SMAR1. Mol Biosyst. 2017;13(2):350-62.
44. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011;39(Database issue):D712-7.