Designing Graphical Data Storage Model for Gene-Protein and Gene-Gene Interaction Networks

Hina Farooq
COMSATS Institute of Information Technology
Sahiwal, Pakistan 57000

Azka Mahmood
COMSATS Institute of Information Technology
Sahiwal, Pakistan 57000

Javed Ferzund
COMSATS Institute of Information Technology
Sahiwal, Pakistan 57000

Muhammad Atif Sarwar
COMSATS Institute of Information Technology
Sahiwal, Pakistan 57000

***Abstract*—Graph is an expressive way to represent dynamic and complex relationships in highly connected data. In today's highly connected world, general purpose graph databases are providing opportunities to experience benefits of semantically significant networks without investing on the graph infrastructure. Examples of prominent graph databases are: Neo4j, Titan and OrientDB etc. In biological OMICS landscape, Interactomics is one of the new disciplines that focuses mainly on the data modeling, data storage and retrieval of biological interaction data. Biological experiments generate prodigious amount of data in various formats(semi-structured or unstructured). The large volume of such data posses challenges for data acquisition, data integration, multiple data modalities (either data model of storage model, storage, processing and visualization. This paper aims at designing a well suited graphical data storage model for biological information which is collected from major heterogeneous biological data repositories, by using graph database.**

***Keywords*—Big Data; Graph Theory; Graph Database; Gene-Gene Interaction; Protein-Protein Interaction; Large Scale Biological Graphs; Storage Model; Neo4j**

I. Introduction

Big Data is defined as data that contains variety, volume, velocity, veracity, valance and value. Key term in Big data is data, not big. Data speed, frequency, volume and connectedness are being driven by the source of transmission of data. The data gathered from different sources are in different forms such as structured data, semi-structured data and unstructured data. Some major repositories of Biological Data include: Molecular Interaction Database (MINT)[1], Database of Interaction Protein (DIP)[2], Biomolecular Interaction Networks Database (BIND)[3] which is a component of Biomolecular Object Network Database, Reactcome[4], Search Tool for the Retrieval of Interacting Gene/Protein (STRING)[5], Unified Human Interactome (UniHI)[6], Online Mendelian Inheritance in Man (OMIM)[7], Kyoto Encyclopedia of Genes and Genomes (KEGG)[8], Human Protein Reference Databases (HPRD)[9], Biological General Repository for Interaction Datasets (BioGrid)[10], National Center for Biotechnology Information (NCBI)[11], and Universal Protein Resource Knowledgebase (UniprotKB)[12].

Graphs databases are trending in today's highly connected world where the flood of data is having dynamic and complex relationships. It is required in coming decades to get insight of vast graphs and highly connected data in order to achieve competitive advantages. Graphs formally consist of nodes (vertices) which represent entities and edges (relationships) which represent connections between nodes. From real world perspective, everything is connected and can be represented as graph.

With the emergence of recent tools and technologies, it is challenging to keep track of all of the storage, analytics and management frameworks. In this study, the scope of graph landscape is discussed in order to understand the presented graphical data storage model for Biological Interaction Data. There are two broader views of graph landscape: one perspective is the Graph Models and the other is Graph Processing.

Graph Model Perspective: The prominent graph models which are used by various other graph technologies are Property Labeled Graph Model[13], RDF (Resource Description Framework)[14] and HyperGraphs[15]. Property Graph model contains nodes which represent entities and edges which represent relationships. Both nodes and relationships can contain properties in the form of key-value pair. Relationships must have start and end node, and are directed and named. Hypergraph model is a generalized graph data model which allows any number of nodes connected with a relationship (called hyper-edge). It can be used to model many-to-many relationship scenarios. Hyperedges can be multi-dimentional. The concept of triple stores is originated from the movement of Semantic Web. Triple is the data model which contains subject-predicateobject structure. It is suitable to capture the semantically-rich information and logically connected data. Among aforementioned graph databases, OrientDB[16] provides Property Graph Model, Neo4j[17] provides Property Labeled Graph Model (Labels can be assigned to nodes) and HypergraphDB[18] provides Hypergraphs.

Graph Processing Perspective: The technologies that are exploiting the concept similar to the OLTP (Online Transactional Processing)[19] of traditional relational space are termed as Graph Databases. Graph Databases offers online transactional processing and provides access in real time either from a user or an application. From another perspective, the technologies that are exploiting concepts similar to OLAP (Online Analytical Processing)[20] or Data Mining are cat-

egorized as Graph Processing Engines (GPE)[21][22]. These are typically designed to perform analytics on bulk of data in batch steps.

Graph Databases (Graph Database Management Systems)[23] are online transactional systems that expose graph data model by exploiting CURD (Create, Update, Read, Delete)[24] approach, and are designed for better transactional performance, integrity and availability. The distinguished properties of graph databases include graph storage and graph processing. Some Graph databases offer their native graph storage while others store graph data serially into general purpose database such as relational database[25], object-oriented database[26] and NoSQL store[27] (other than graph store). The approach used by graph database in which adjacent nodes directly point to each other is termed as index free adjacency. In other words: a graph database qualifies as a graph database when it behaves like real graphs from the user's perspective. Some graph databases use native graph processing means that they provide index free adjacency[28].

Relational Databases are used to store data in tabular and structured form and they are doing it exceedingly well. But today's technologies are facing challenges to store data which is highly connected and semi-structured, which should be well modeled and suitable for ad-hoc queries. Almost everything is connected in this world and it is needed to understand the influence of connections in order to thrive and progress. In Biological Domain, data is more connected and have complex relationships. This research is aimed at designing storage model for connected data which is collected from major biological data repositories, by using Graph Database (Neo4j). Neo4j[17] provides Native Graph Storage and Native Graph processing. Other prominent Graph Databases are discussed in table I.

TABLE I. EXISTING GRAPH DATABASES WHICH ARE PROVIDING NATIVE/NON-NATIVE STORAGE AND PROCESSING

Graph Database	Graph Storage	Graph Processing
Neo4j	Native	Native
OrientDB	Native	Native
Affinity	Native	Native
Dex	Native	Native
HypergraphDB	Native	Non-Native
Allegrograph	Native	Non-Native
FlockDB	Non-Native	Non-Native
Titan	Non-Native	Native
Trinity	Non-Native	Native
InfiniteGraph	Non-Native	Native

Biological interaction networks are typically dense, semi-structured, unpredictable and highly connected. For example, in protein-protein interaction network[29], a gene may be interacted with other proteins, or may it be participated in biological pathways[30], or may it be involved in disease relevant network. This type of connected biological information leads to highly connected networks. Therefore, traditional database storge models are not suitable to handle such datasets. Because classical database storage models are naturally design to handle the datasets which are less-connected (few number of relationships among data entities) with the entities represent limited data types and querying the data need joins that make it computationally expensive. Graph storage models provide an easy way of modeling, understanding and visualizing data of a domain. In Biological domain, the problem is to get data from heterogeneous biological data sources, integration of collected datasets, designing storage model based on the information-rich graph model which helps to understand the connectedness of data with several other aspects. With the less-familiarity of graph databases, biologists (people from other domains) face difficulty to design graph storage models.

The objectives of this research include:

- Biological data acquisition from heterogeneous data sources like NCBI[11], RefSeq[31], EntrezGene [32], BioGrid[10], OMIM[7], HGNC[33], HPRD[9] and STRING[5] etc. (Selection of datasets of Gene-Gene and Gene-Protein Interactions)
- Transformation, Cleaning and Integration of datasets
- Data modeling of Gene-Gene and Gene-Protein Interaction data using Labeled Property Graph Model
- Designing data storage model for Graph Database (using Neo4j)
- Evaluation of implemented storage model

The outline followed in this paper is as: In section 2, Graphical Data Storage Model is presented for Interaction Networks by using Graph Database. In section 3 it is discussed, how a data model(Labeled Property Graph Model) can be represented as a graph storage model specifically for biological interaction graphs. Further in section 4, evaluation of storage model is discussed by using Cypher Query Language in Neo4j[17]. Related work is presented in section 5, followed by the conclusion in section 6.

II. GRAPHICAL DATA STORAGE MODEL

This paper aims at offering a unifying, gene-centric view over the data made available by the heterogeneous data sources and designing graphical data storage model for integrated data. In order to achieve this objective, available typologies of biological information are formulated as:

- **Gene**, i.e., Identification of a gene of a dataset through data source identifier. For example: a Gene, symbolically represented as RXRA is identified by its data source identifier. In this data model, diverse datasets are integrated from heterogeneous data sources including HGNC [33], HPRD[9], UniProt[12], Ensembl[34], EntrezGene[32]. BioGrid[35], NCBI[11], STRING[5] and RefSeq[31]. Properties of gene include Gene-Family Identifier, Gene-Symbol, Gene-Aliases, Gene-Description, Genomic-Coordinates and Cytogenetic-Location.
- **Protein**, i.e., Identification of a protein of a dataset through data source identifier. In this data storage model, diverse datasets are integrated from heterogeneous data sources including HGNC[33], HPRD[9], UniProt[12], Ensembl[34], EntrezGene[32]. BioGrid[35], NCBI[11], STRING[5] and RefSeq[31]. Properties of protein include Protein-Identifier, Protein-Symbol and Protein-Aliases.
- **Locus**, i.e., Information about Locus Type and Locus Family.

- **External Links**, i.e., Identification of a gene or a protein of a dataset through data source identifier. For example: a Gene, symbolically represented as RXRA is identified in HGNC as 10477, in UniProt as Q6P3U7, its Ensembl identifier is ENSG00000168824, HPRD identifier is 1577 and so on. In this data model, diverse datasets are integrated from heterogeneous data sources including HGNC[33], HPRD[9], UniProt[12], Ensembl[34], EntrezGene[32]. BioGrid[35], NCBI[11], STRING[5] and RefSeq[31]
- **Molecular Information**, i.e., Molecular Weight (unit: Dalton) of a Gene, information about Molecular Class from which a Gene belongs and Information about Molecular Function a gene may be performed.
- **Disease**, i.e., Information about participation of a Gene in Disease-Association[36] Networks for example a gene can can be associated to a certain kind of Tumor or other kind of disease.
- **Publication**, i.e., Reference of existing biological literature[37] for Gene that includes information about Author, Publication Year and Publication Identifier.
- **Sequences**, i.e., biological sequences include DNA Sequence and Protein Sequence.
- **Specie**, i.e., NCBI [11] Taxonomy Information about Organisms and Species (For example: HomoSapien taxonomy identifier is 9606).
- **Pathways**, i.e., Information about participation of a Gene in biological processes for example a gene can take part in cell communication or in signal transduction etc.
- **Gene-Gene Interaction Information**, i.e., Interaction of Gene with other Genes carries information about the experiment method through which the G-G interaction is detected and recorded (by the data sources). Examples of Interaction Experiment Methods are: Two-Hybrid[38], Affinity Chromatography[39] and Mass Spectrometry[40].
- **Gene-Protein Interaction Information**, i.e., Interaction of Gene with other Proteins carries information about the Interaction Detection Method through which the G-P interaction is recorded (by the data sources). Examples of Interaction Detection Methods are: Direct Interaction, Physical Association and Co-Localization.

TABLE II. TYPES OF NODES AND RELATIONSHIPS INCLUDED IN GRAPHICAL STORAGE MODEL

Node	Relationship	Node
Gene	GGI-INTERACTS-WITH	Gene
Gene	GPI-INTERACTS-WITH	Protein
Gene	LOCUS-INFORMATION-IS	Locus
Gene	ASSOCIATES-TO	Disease
Gene	OF-ORGANISM	Specie
Gene	PARTICIPATES-IN	Pathway
Gene	HAVE-SEQUENCE	Sequences
Gene	IN-LITERATURE	Publication
Gene	REPRESENTED-IN	External Links
Protein	PPI-INTERACTS-WITH	Protein

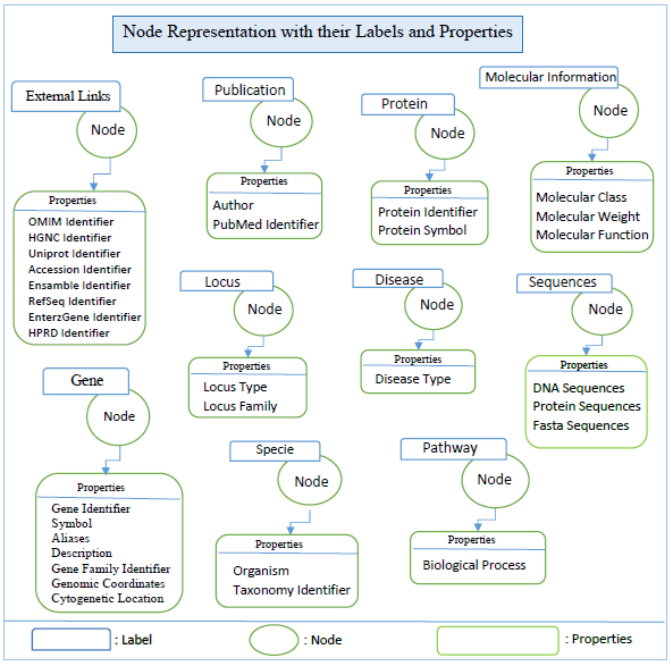

Fig. 1. Graphical Data Storage Model

In Table II, entities are represented as nodes and edges are represented as relationships between biological entities. Nodes have properties and can have one or more labels. Relationships are directed and can have properties as well. In figure 1, Graphical Data Storage Model is presented that is based on Labeled-Property Graph Model. Nodes are representing aforementioned entities of biological domain along with the label and properties of each node.

III. PHYSICAL DATA STORAGE IN GRAPH DATABASE

The way in which graphs are stored in graph database is one of the key aspects of the designing graph database. Neo4j is one of the prominent graph databases which provides index-free adjacency, native storage, native processing and native query language(Cypher). Storage model is designed in section 2, for graph databases. This section aims at illustrating that how biological interactions(binary) are physically stored in a graph database(Neo4j). Neo4j is designed to store graph data in different store files, i.e., Nodes, Relationships, Properties and Labels have different physical stores on disk. There is structural dissimilarity between the actual graphical view of a graph and the actual view of stored records on disk.

Protein-Protein interaction networks are usually very diverse and have various properties. The reason is the generation of data from heterogeneous sources both experimentally and computationally.Mostly, Protein interaction networks follow the characteristics of scale-free networks. In such networks, higher degree of protein connectivity shows the higher biological significance of that protein. Fig 2 presents, how a Protein-Protein interaction is physically stored in Neo4j.

Gene-Gene interaction networks are usually sparse and highly connected networks, also known as Gene-Regulatory Networks. In fig 3, it is presented that how Gene-Gene interactions are physically stored in Neo4j.

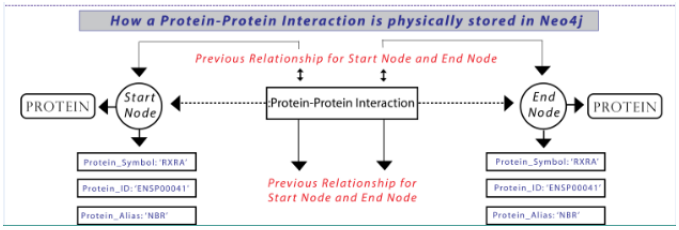

Fig. 2. Graphical Data Storage for Protein-Protein Interaction

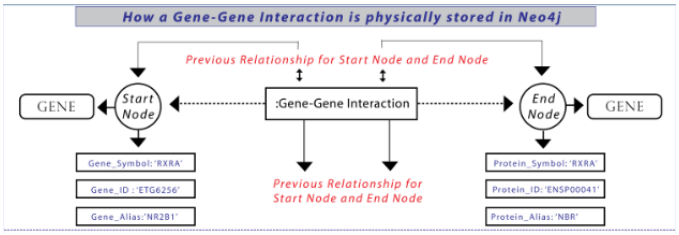

Fig. 3. Graphical Data Storage for Protein-Protein Interaction

Gene-protein interaction networks is presented in fig 4, i.e., how Gene-Protein interactions are physically stored in Neo4j.

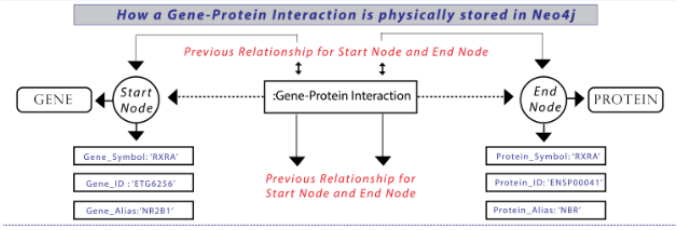

Fig. 4. Graphical Data Storage for Gene-Protein Interaction

IV. Evaluation of Biological Interaction Graphs using Neo4j

The Biological Networks are naturally more complex, and the complexity increases with the accumulation of data. The variability of biological information is one of the major cause of data inaccuracy. As for this research, data is integrated from different major data repositories, and storage model is presented for querying and visualization on Neo4j. The results are evaluated by the verification of queried information with the major sources of biological information.

In order to demonstrate, how the biological data can be accommodated in neo4j, some queries results are presented. The diverse data sets are polled in Neo4j, particularly for Biological Domain and Gene-Gene and Gene-Protein Interaction scenario and are queried by using Cypher Query Language. Query results are evaluated on the basis of designed storage model and its potential to capture all the information, a biological network have, about its entities and relationships. Additionally, query results are verified from the heterogeneous data sources from where the data had been collected.In fig 5, the way is depicted which is used in Neo4j for the representation of G-P and G-G interaction networks.

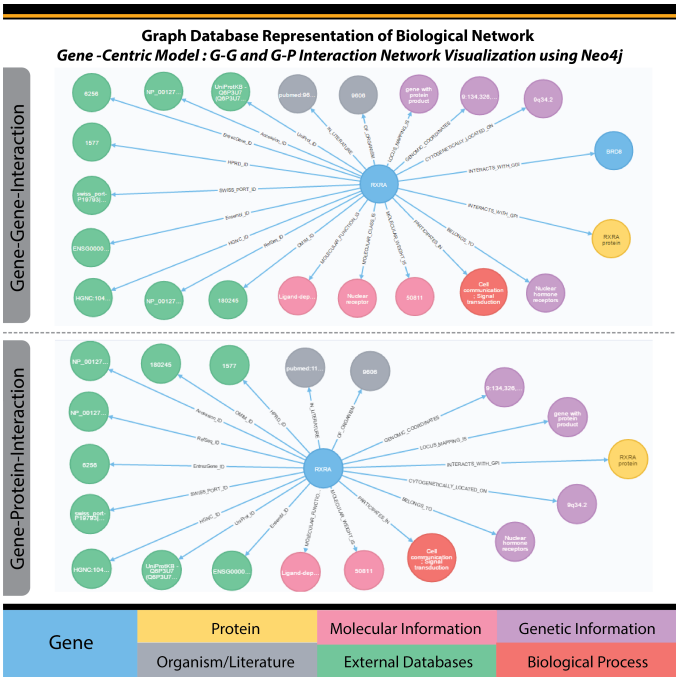

Fig. 5. Neo4j Results based on presented Data Storage Model

V. Related Work

Study of protein-protein, protein-gene and gene-gene interactions are becoming increasingly important to understand human diseases on a system-wide level. These proteinprotein interactions provide significant information for new perceptions in different ways that can impact biomedical research. Protein functionality often modulate with other interactors which can either be proteins, or genes or other molecules. Biochemical Interaction Detection Methods are used to detect interactions among biological entities, such methods include protein affinity chromatography, affinity blotting, co-immunoprecipitation, and cross-linking etc. Other prominent experimental methods for interaction detection in molecular biology are protein probing and two-hybrid system. Examples of genetic interaction detection methods include suppressors [41], synthetic mutants [42], and non-complementing mutants [43] etc.

In [44], a practical analysis guidance of interactions in genetic, biochemical and molecular biological methods is presented. In [45], protein interaction fundamentals, publicly available protein interaction databases with their useful data significant information which facilitate genome or genetic studies, are briefly discussed. A systematic prediction method of protein-protein interaction type is proposed in [29], based on solely techniques used to detect interactions. Lactose effect investigation on structural variation of aging induced by changing lactose content is presented in [46].

In biological literature[37], systematic views of human genome are presented from antiquity evolution to precision medicine against diseases . For research purpose, biological databases are increasing their importance with rapid growth of data. In [47], a review of biological databases is presented followed by the challenges such as data volume, processing, data exchange and curation from big data perspective.

Human(Homo-sapiens) databases are categorized by the information provided by database such as DNA [34], RNA [48], protein [2] [12], Expression [49], Pathway [4], disease [50], and literature [37]. Ancestral networks mechanisms of human and mouse genomes that are characterized by the new gene integration, and gene evolutionary significance are discussed in [51]. Exploration of their generation frequencies and patterns of new gene-driven evolution of Gene Gene Interaction networks is also discussed.

In [52], interaction pattern discovery with characterization of different types of interactions is discussed along with their use in protein-protein interaction. Graph databases enable efficient storage and processing of the encoded biological relationships. Systems biology graphical notation (SBGN) [53] represent STON [54] (SBGN TO Neo4j), a framework that exploits the Neo4j graph database to store biological pathways. In [30], a novel algorithm for the identification of spurious curves is presented where curves are used for different unfolding pathways. An evaluation of different resulting graphs generated from statistical analysis is presented in [55]. [56] shows detailed description of protein domains, functional sites, and families as well as associated patterns and their profiles identification methods. A brief description of major biological interaction databases such as BIND [3], DIP [2], HPRD [9], IntAct [57], MINT [1], MIPS [58], PDZBase [59] and Reactome [4] is represented in [60]. BioGrid[10] database is an open access database that houses protein interactions and genetic curated data from the primary biomedical literature for all major model organism/species[35]. Currently, BioGRID [35] contains 749912 interactions as drawn from 43149 publications that represent 30 model organisms.

VI. Conclusion

We are living in the age of Big Data and graphs are the most suitable choice for representing large scale multimodel biological data as they can effectively represent the relationships of data that is being collected by heterogeneous data sources. Large scale biological graphs have been used for analysis of complex data sets from biological domain like Interaction Networks, Bioinformatics, Health Informatics, Molecular Networks, Gene-Disease and Gene-Phenotypes Association Networks and applications that produce large amount of biological data. To fully utilize the information represented by graphs, efficient storage model and graph database are required. In this paper, a storage model has been presented for diverse data sets, collected from major biological data repositories by using one of the prominent Graph databases, Neo4j. Storage Model is described according to various types of biological information. Moreover, potential Graph Theory in Biology and tools and techniques used in biological research activities has been presented. This article will be helpful for the researchers to get firsthand knowledge of existing Graph Databases and techniques to plan for future research.

References

[1] P. D. Licata L, Briganti L, " Molecular INTeraction database," http://mint.bio.uniroma2.it/, 2012.

[2] DIP, "Database of Interacting Proteins," http://dip.doe-mbi.ucla.edu/dip/Main.cgi, 2014.

[3] BIND, " Biomolecular Interaction Networks Database ," https://www.bindingdb.org/bind/index.jsp, 2016.

[4] Reactcome, " REACTCOME Curated Pathway Database," http://www.reactome.org/, 2016.

[5] L. J. J. Peer Bork, " Search Tool for the Retrieval of Interacting Gene/Protein," http://string-db.org/, 2016.

[6] UniHi, "Unified Human Interactome ," http://www.unihi.org/, 2014.

[7] OMIM, " Online Mendelian Inheritance in Man," https://www.omim.org/, 2017.

[8] KEGG, " KEGG Pathway Databses," http://www.genome.jp/kegg/pathway.html, 2017.

[9] HPRD, "Human Protein Reference Database," http://www.hprd.org/, 2009.

[10] L. Boucher, "biogrid," https://thebiogrid.org/, 2017.

[11] NCBI, " National Center for Biotechnology Information," https://www.ncbi.nlm.nih.gov/, 2017.

[12] UniProtKB, "Universal Protein Resource Knowledgebas," http://www.uniprot.org/help/uniprotkb, 2014.

[13] J. W. Robinson, Ian and E. Eifrem, *Graph databases: new opportunities for connected data*, 2015, ch. Property Labeled Graph Model.

[14] e. a. Campinas, Stephane, "Introducing rdf graph summary with application to assisted sparql formulation." *IEEE 23rd International Workshop on Database and Expert Systems Applications (DEXA)*, 2012.

[15] e. a. Tan, Shulong, "Using rich social media information for music recommendation via hypergraph model." *Social media modeling and computing,Springer, London*, 2011.

[16] C. Tesoriero, *Getting Started with OrientDB*, 2013, ch. OrientDB.

[17] H. Huang and Z. Dong, "Research on architecture and query performance based on distributed graph database neo4j," in *IEEE 3rd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2013.

[18] B. Iordanov, "Hypergraphdb: a generalized graph database," in *International Conference on Web-Age Information Management. Springer*, 2010.

[19] C. C. Pavlo, Andrew and S. Zdonik., "Skew-aware automatic database partitioning in shared-nothing, parallel oltp systems." in *ACM SIGMOD International Conference on Management of Data*, 2012.

[20] e. a. Zhao, Peixiang, "Graph cube: on warehousing and olap multidimensional networks." in *ACM SIGMOD International Conference on Management of data*, 2011.

[21] I. M. Roy, Amitabha and W. Zwaenepoel., "X-stream: edge-centric graph processing using streaming partitions." in *ACM, 24th Symposium on Operating Systems Principles.*, 2013.

[22] e. a. Malewicz, Grzegorz, "Pregel: a system for large-scale graph processing." in *ACM SIGMOD International Conference on Management of data.*, 2010.

[23] S. G.-V. Martinez-Bazan, Norbert and F. Escale-Claveras, "Dex: A high-performance graph database management system." in *IEEE 27th International Conference on Data Engineering Workshops (ICDEW)*, 2011.

[24] L. Zhang, "Research and design of geospatial metadata deployment prototype system based on php framework," *International Journal of Interdisciplinary Telecommunications and Networking (IJITN)*, 2014.

[25] e. a. Jiang, Haifeng, "Xparent: An efficient rdbms-based xml database system," in *IEEE 18th International Conference on Data Engineering*, 2002.

[26] e. a. Bertino, Elisa, "Object-oriented databases." *Springer, Indexing Techniques for Advanced Database Systems.*, 1997.

[27] e. a. Han, Jing, "Survey on nosql database," in *IEEE 6th international conference on Pervasive computing and applications (ICPCA)*, 2011.

[28] A. P. Nayak, Ameya and D. Poojary, "Type of nosql databases and its comparison with relational databases," *International Journal of Applied Information Systems*, 2013.

[29] M. K. Silberberg, Yael and R. Sharan., "A method for predicting protein-protein interaction types." *PloS one:accelerating the publication of peer-reviewed science*, 2014.

[30] e. a. Marsico, Annalisa, "A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy," *International Secutriy for Computational Bioinformatics*, 2007.

[31] NCBI, "Reference Sequence," https://www.ncbi.nlm.nih.gov/refseq/, 2017.

[32] NCBI, "Enter Gene," https://www.ncbi.nlm.nih.gov/gene, 2017.

[33] HGNC, "HGNC gene nomenclature," http://www.genenames.org/, 2016.

[34] Ensembl, " Ensamble genome database," http://asia.ensembl.org/index.html, 2017.

[35] e. a. Chatr-Aryamontri, Andrew, "The BioGRID interaction database: 2015 update," Tech. Rep., 2015.

[36] e. a. Zou, Dong, "Biological databases for human research." *Genomics, proteomics and bioinformatics*, 2015.

[37] NCBI, " Pubublication Medical," https://www.ncbi.nlm.nih.gov/pubmed, 2016.

[38] S. K. Suter, Bernhard and I. Stagljar., "Two-hybrid technologies in proteomics research," *Current Opinion in Biotechnology*, 2008.

[39] BIORAD, "Affinity Chromatography," http://www.bio-rad.com/en-mu/applications-technologies/introduction-affinity-chromatography, 2016.

[40] S. L. Berggrd, Tord and P. James., "Methods for the detection and analysis of proteinprotein interactions," *Proteomics*, 2007.

[41] B. Lewin, " Suppressor analysis method to identify interacting genes," http://bioscience.jbpub.com/cells/GNTC2721.aspx, 2014.

[42] e. a. Babu, Mohan, "Array-based synthetic genetic screens to map bacterial pathways and functional networks in escherichia coli," *Strain Engineering: Methods and Protocols*, 2011.

[43] P. Dyson, *Streptomyces: Molecular Biology and Biotechnology*, 2011, ch. gene non-complementing mutants detection method.

[44] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiological reviews*, 1995.

[45] K. A. Pattin and J. H. Moore., "Role for proteinprotein interaction databases in human genetics," *Expert review of proteomics*, 2009.

[46] e. a. Norwood, Eve-Anne, "Crucial role of remaining lactose in whey protein isolate powders during storage." *Journal of Food Engineering*, 2017.

[47] e. a. Zou, Dong, "Biological databases for human research," *Genomics, proteomics and bioinformatics*, 2015.

[48] A. Kiran, " Identification of RNA editing in the human exome and development of DARNED DatabaseSF," http://darned.ucc.ie, 2014.

[49] H. P. Atlas, "The Human Protein Atlas," http://www.proteinatlas.org/, 2017.

[50] miR2Disease, " MiR2Disease Base," http://www.mir2disease.org/, 2017.

[51] e. a. Zhang, Wenyu, "New genes drive the evolution of gene interaction networks in the human and mouse genomes," *Genome biology*, 2015.

[52] e. a. Park, Sung Hee, "Prediction of protein-protein interaction types using association rule based classification," *BMC bioinformatics*, 2009.

[53] e. a. Le Novere, Nicolas, "The systems biology graphical notation," *Nature biotechnology*, 2009.

[54] e. a. Tour, Vasundra, "Ston: exploring biological pathways using the sbgn standard and graph databases," *BMC BioMedCentral, bioinformatics*, 2006.

[55] S. Grunert and D. Labudde., "Graph representation of high-dimensional alpha-helical membrane protein data," *BioData mining*, 2013.

[56] e. a. Sigrist, Christian JA, "New and continuing developments at prosite," *Nucleic acids research*, 2012.

[57] IntAct, "Molecular Interaction Database," http://www.ebi.ac.uk/intact/, 2017.

[58] MIPS, " MIPS Mammalian Protein-Protein Database," http://mips.helmholtz-muenchen.de/proj/ppi/, 2014.

[59] PDZbase, " PDZbase database," http://abc.med.cornell.edu/pdzbase, 2010.

[60] e. a. Mathivanan, Suresh, "An evaluation of human protein-protein interaction data in the public domain," *BMC BioMedCentral, Bioinformatics*, 2006.