A Simple Approach for Representation of Gene Regulatory Networks (GRN)

Raza-ul-Haq, Javed Ferzund, Shahid Hussain
COMSATS University Islamabad
Sahiwal, Pakistan

***Abstract*—Gene expressions are controlled by a series of processes known as Gene Regulation, and their abstract mapping is represented by Gene Regulatory Network (GRN) which is a descriptive model of gene interactions. Reverse engineering GRNs can reveal the complexity of gene interactions whose comprehension can lead to several other details. RNA-seq data provides better measurement of gene expressions; however it is difficult to infer GRNs using it because of its discreteness. Multiple other methods have already been proposed to infer GRN using RNA-seq data, but these methodologies are difficult to grasp. In this paper, a simple model is presented to infer GRNs, using RNA-seq based coexpression map provided by GeneFriends database, and a graph-based database tool is used to create regulatory network. The obtained results show that it is convenient to use graph database tools to work with regulatory networks instead of developing a new model from scratch.**

Keywords*—Graph theory; graph database; gene regulatory networks; RNA-seq; Genes Co-Expression; Neo4j

I. INTRODUCTION

The required information to make proteins and other molecules is stored in DNA. The functional circuitry of all living organisms is formed by genes [1]and synergistic actions between inter related genes is the reason of all biological reactions inside a cell. Gene regulation is a mechanism of increasing or decreasing production of gene products. In this process, genes are regulated by regulators to produce proteins or RNA, which produces a complex network of regulatory relationships. To understand the cellular process, it is critical to understand the regulatory relationships between genes. These relationships are expressed by gene regulatory network and are used to understand functions of genes. A gene regulatory network consists of nodes which represent genes, and edges between nodes represent relationships between genes. GRNs play important role in cell transduction, metabolism, cell differentiation, cell cycle and every other biological mechanism. In-depth and comprehensive understanding of complex biological processes can be provided by gene regulatory networks. The study of gene regulatory networks not only unveils the dynamics of organisms but also reveals their behavior in different scenarios and shows how their fate is controlled. The interaction of genes can be understood by reconstructing gene regulatory networks. The representation of genetic data, reverse engineering of GRNs and performing analytics on regulatory networks to retrieve information is challenging tasks. It can not only help to diagnose diseases but can also shed light on those changes which became reason for a disease and what changes have occurred because of that disease. Different people can be on different stages of a diseases, have different medical history which can lead to different response to the treatment from others; GRNs can be helpful in individualized treatment [2]. Determining or knowing drug sensitivity on target is an important aspect of the process of individualized treatment and is a research area in which GRNs can be used to detect drug sensitivity easily. Using GRNs we can answer the question that if a gene is mutated can its function be restored or not? The main genes responsible for a differentiation of cell into an organ or responsible for a disease can be identified using GRNs. In the process of regulating wide range of activities which include cellular, physiological and behavioral, circadian rhythm plays fundamental role. The researchers know a small number of genes which play a key role in circadian rhythm, however by using these genes, the existence of other key genes and how they work can be unveiled through GRNs. There are two main types of data which are used to infer GRN: microarray and RNA-seq. Continuous probe intensities are measured by microarrays, but discrete digital sequencing read counts which are aligned to sequence and are quantified by RNA-seq. Genes differential expression is measured more accurately with transcriptome sequencing (RNA-seq) than with microarrays. Different splice variants and non-coding RNA (ncRNA) can play important role in regulation of gene expression. The measurement of levels of transcripts provided by RNA-seq is far more precise than other methods [3]. In a single experiment RNA-seq can identify novel isoforms, novel transcripts allele specific expression, alternative splice sites and rare transcripts beyond gene expression analysis. It provides abilities to perform such types of experiments which traditional microarray-based methods cannot provide. Huge size of RNA-seq data increases challenges to interpret results, due to which processing mechanism and interpretation of results from RNA-seq experiments was often impeded. There was no database available for bioscience community which could provide RNA-seq data in a form through which any useful information could be extracted without going through time consuming processes of analyzing raw RNA-seq data before GeneFriends [4]. It allows researchers to identify those genes which are poorly annotated and associated with genes under study. Genes that are responsible for lung cancer is used in this research work to infer regulatory network. In this research work, GeneFriends database is used to obtain the genes co-expression network and an existing graph database tool is used to infer GRN. Deployment method shows that it is much simpler than other existing methods.

II. Methodology

A. Data Set Selection

Genes which are involved in commonly identified genomic/genetic alterations of lungs including chromosomal fusion rearrangements, nonsense or missense mutations, alternative splicing and small deletions or insertions are *EGFR, KRAS, MET, LKB1, BRAF, PIK3CA, ALK, RET,* and *ROS1* or in other words these genes are the most relevant to lung cancer [5]. These genes are selected in this research work to infer regulatory network.

B. Data Extraction and Transformation

Genomic databases are produced by a project named Ensembl for vertebrates. It is used to extract Ensembl transcript IDs of selected lung cancer genes Table I.

TABLE I. Lung Cancer Genes and Their Corresponding Ensemble Id

Gene	Ensemble ID
EGFR	ENSG00000146648
KRAS	ENSG00000133703
MET	ENSG00000105976
LKB1	ENSG00000118046
BRAF	ENSG00000157764
PIK3CA	ENSG00000121879
ALK	ENSG00000171094
RET	ENSG00000165731
ROS1	ENSG00000047936

These IDs are used in GeneFriends database to generate co-expression network of all the genes. Connection strength of direct partners within the seed list is 1, the genes having 0.75 connection strength are strongly co-expressed, 0.5 is connection strength of other direct partner and indirect partners have strength of 0.25. More than 1700 gene entries were generated as co-expression network in CSV format.

C. Tool

In our research scenario, graph databases are appropriate tools to infer network. These are database engines which model nodes and edges as first-class entities. Complex interactions between nodes can be represented in natural form. There are multiple graph databases available to work with and ArangoDB, Neo4j, Oracle Spatial and Graph, IBM System G Native Store and OrientDB are a few of them. Neo4j [6] graph database is used in this research work because it protects data integrity while providing fast reads and writes, and it is easy to learn.

D. GRN Construction

Cypher query language is used to query dataset. The network data available in CSV form was loaded into network. This data file contains genes and their connection strengths. The first gene was treated as source node, second one was treated as destination node and connection strength between two genes was used to create an edge between them, Figure 1. While creating nodes MERGE function was used instead of CREATE as values in data file are repeating multiple times and if a value is repeating, CREATE function always creates a new node having same label and property value which is not required in this research work. We could not afford repetition because multiple nodes with same properties will add extra work to get information about a single gene and it will also need more computing resources. On the other hand, only one node is created by MERGE against all the repeating values. But it does not happen in every case as if unbounded elements of an existing graph are being used to MERGE with a pattern it will start repeating nodes so always use bounded elements if an existing graph is being used. As in this research work there was no existing graph available, this issue was not considered. A value from data file was required to create a relationship, the default mechanism provided by neo4j could not be used as it does not provide this functionality. There is another function APOC.CREATE.RELATIONSHIP provided by apoc plugin, which was used for this purpose.

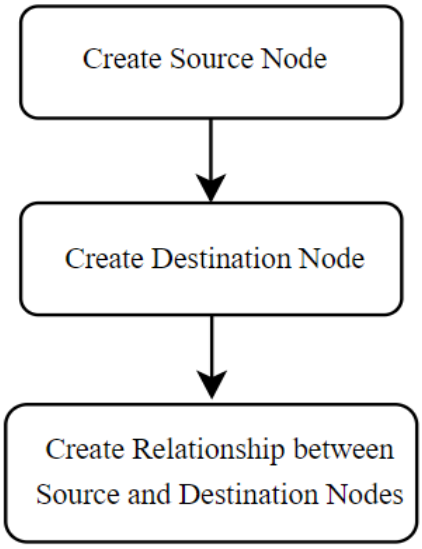

Fig. 1. Gene Regulatory Network Inference Model.

III. Results and Discussion

Using neo4j, a network of more than 1700 genes was created in which nodes indicate genes; relationships among nodes indicate connection strength between genes, and Ensemble IDs were assigned as property values of nodes. In this network, first value is treated as source node and is indicated by blue nodes, second value is treated as relationship type, third value is treated as destination node and is indicated by green nodes. The whole network is shown in Figure 2. There are 322 source nodes and 852 destination nodes having 1791 relationships, out of which three source nodes are connected to three destination nodes with connection strength 1. There are 65 source nodes connected to only 8 destination nodes with connection strength of 0.75. On the other hand, 9 source nodes are connected to 313 destination nodes with 0.5 connection strength and 0.25 is the connection strength of 1599 relationships among 1068 nodes out of which 322 are source nodes and 746 are destination nodes.

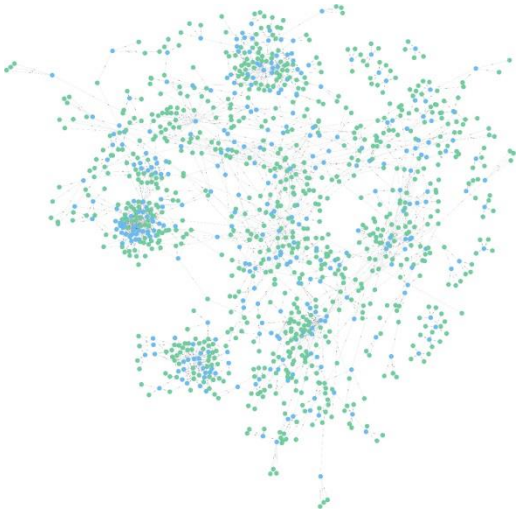

Fig. 2. Lung Cancer Regulatory Network Having 1791 Relationships, 322 Source Nodes and 852 Destination Nodes.

There is a node in the entire network which has the most relationships with other nodes among all. It means that, the gene named ALK associated with this node ID is the most important gene in the entire network.

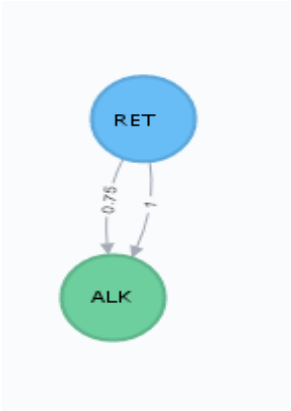

Fig. 3. *ALK* and *RET* Gene are connected with 0.75 and 1 Connection Strengths.

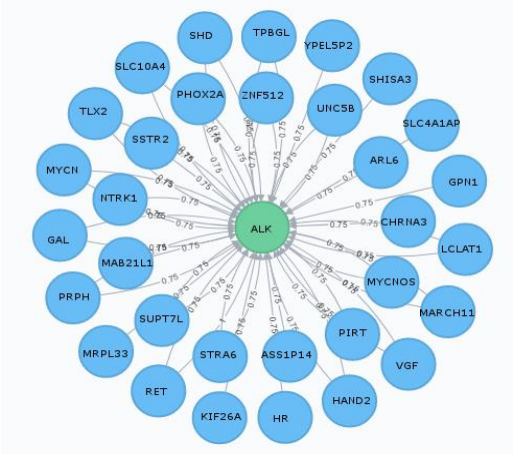

Fig. 4. *ALK* Gene is Connected to 32 Genes as Destination Node.

A gene named RET is a direct partner to ALK within seed list Figure 3. The production of protein involved in signaling within cells receives instructions from RET gene. Several kinds of nerve cells are developed by this protein. Mutation in this gene is the reason for Hirschsprung disease, pheochromocytoma, and most importantly lung cancer.

There are 32 genes associated with ALK with 0.75 connection strength, Figure 4. These 32 genes act as sources for ALK, which means it is co-expressed with these 32 genes. On the other hand, 39 genes are connected as destination nodes with ALK having 0.5 connection strength, Figure 5, which means it is direct partner with these 39 genes. It has 8 relationships of strength 0.25 as source node with 4 destination nodes, Figure 6.

Since it is present in every type of connection strength, it the most important node in the entire Network. There are strong evidences available to prove that it is the driving force of different types of cancers, including Non-Small Cell Lung Cancer and neuroblastoma [7] and inferred network in this research confirms this as well.

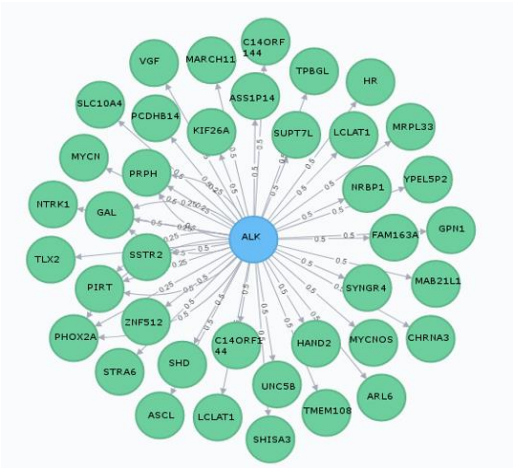

Fig. 5. *ALK* Gene is Source Node for 39 Genes Having 0.5 Connection Strength.

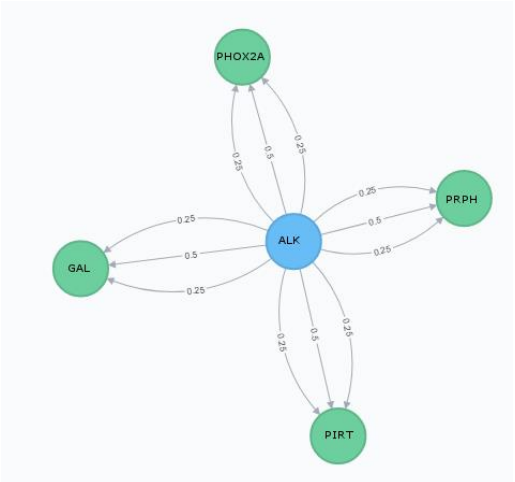

Fig. 6. *ALK* Has 12 Relationships of 0.25 Connection Strength with Four Other Genes.

TABLE II. GENES WHICH ARE COMMON IN BOTH 0.75 AND 0.5 CONNECTION STRENGTHS

Sr. No.	0.75	0.5
1	*TPBGL*	*TPBGL*
2	*HR*	*HR*
3	*MRPL33*	*MRPL33*
4	*YPEL5P2*	*YPEL5P2*
5	*GPN1*	*GPN1*
6	*MAB21L1*	*MAB21L1*
7	*CHRNA3*	*CHRNA3*
8	*ARL6*	*ARL6*
9	*SHISA3*	*SHISA3*
10	*LCLAT1*	*LCLAT1*
11	*STRA6*	*STRA6*
12	*PHOX2A*	*PHOX2A*
13	*PIRT*	*PIRT*
14	*TLX2*	*TLX2*
15	*SSTR2*	*SSTR2*
16	*GAL*	*GAL*
17	*ZNF512*	*ZNF512*
18	*SHD*	*SHD*
19	*UNC5B*	*UNC5B*
20	*HAND2*	*HAND2*
21	*MYCNOS*	*MYCNOS*
22	*NTRK1*	*NTRK1*
23	*MYCN*	*MYCN*
24	*SLC10A4*	*SLC10A4*
25	*VGF*	*VGF*
26	*PRPH*	*PRPH*
27	*KIF26A*	*KIF26A*
28	*ASS1P14*	*ASS1P14*
29	*SUPT7L*	*SUPT7L*
30	*MARCH11*	*MARCH11*
31	*LCLAT1*	*LCLAT1*

A list of 31 genes is given in Table II, these are common genes connected to ALK gene with both connection strengths 0.75 as well as 0.5. But when these genes have 0.75 strength they are being source genes of ALK and when they have 0.5 strength they are being destination genes of ALK. It means that if ALK is being directly co-expressed with these gene then it is also their direct partner at the same time. These genes are not only connected with mentioned genes but there are multiple other genes interacting with them at the same time in the entire network. We have discussed multiple techniques already available to infer regulatory network. There is one thing common in above mentioned methods and even most of the other methods proposed in last two decades to infer GRN, that is, they need a highly expert person in graph theory or mathematics in general to implement these methods and infer GRNs; however, the method and already existing tool presented in this research work does not need a researcher to be highly expert in such areas. Even though graph theory is working at backend of the tool, but researchers don't need to know how it is working and how it is inferring networks. So instead of spending time on comprehending those methods and then implementing them, researchers can easily infer GRN and spend their most of the time on analysis on inferred regulatory network which is the actual process of unveiling underlying mechanisms of genes interactions.

IV. RELATED WORK

For the integration of graph database with the analytical process of transcriptome data, a platform is presented in [8] through which data coming from Affymetrix platforms on rhesus, rat, mice and humans can be analyzed. An algorithm bLARS which is based on regression is used to construct GRNs using steady state gene expression data [4], which allows different genes to have different regulatory mechanisms. It uses bootstrapping for scoring purpose. Based on FA, PSO, BA-PSO which are swarm intelligence techniques, RNN formalism is used to investigate reverse engineering of GRNs from time series microarray datasets [1]. For refinement of classical network thresholding, a GRN post processing tool is represented in [9], linking nodes that belong to the same cluster with nodes that have higher weight, get favor by this method. It uses random walker to compute an optimal gene and select an optimal edge jointly. GRN inference is improved when clustering process is introduced in edge selection process. A novel technique for discovery of gene regulatory network is proposed [10] in which discovery process is integrated into heuristic information. To construct large scale gene regulatory networks a dynamic multi-agent genetic algorithm is represented [11], which is based on FCM. The method proposed in [12] uses low rank property to construct a common GRN structure from other inferred GRNs, drug effect is also inferred and estimated by this method. Through anti-diabetic drug, Metformin, the benefits to target tumor cell metabolism are investigated using simulations [13]. The use of S-System modelling formulation is proposed in [14], by combining standard system identification procedures with this modelling formalism, the type of regulation between each gene is established and then a model which is suitable for designing a synthetic genetic feedback controller is derived. To predict transcription factors and gene interactions, a method which uses iterative SVM and clustering is implemented in [15]. To discover relationships between genes this method [16] combines PCA-CMI and GA algorithms. For a target gene to obtain the best predictor, GA was performed and PCA-CMI method was used to create initial population to reduce search space. For encoding dynamics of multi-valued network, use of an extension of ASP named FASP as the language in continuous domain is proposed [17]. A methodology [18] in which observed count data is modelled as being negative binomially distributed is proposed to infer gene regulatory networks using RNA-seq time series data. To identify multivariate gene interaction in RNA-seq data, an application of BEE and OBC is demonstrated [19] to differentiate biological phenotypes. To infer gene regulatory network Legendre neural network (LNN) is proposed [20] and to

optimize the parameters of Legendre neural network, Firefly algorithm is used.

V. Conclusion

Complex biological mechanisms are widely elucidated by gene expression information, which is expressed by interacting along with one or multiple other genes and this interaction with other genes formulates a regulatory network. Reverse engineering of this regulatory network is an important task to get the insight of biological mechanisms. To study cell trascriptome at system level, RNA sequencing is a revolutionary technique. In this research work, a graph database neo4j is used as a tool to construct GRN and RNA-seq data of lung cancer genes and is used as dataset provided by GeneFriends database. Many techniques are available to infer a GRN but most of them implement complex mathematical models during the process. In this research work, we have used an already existing tool, even though graph theory is working behind the scene to infer network in this tool as well, but researchers don't have to pay any attention on background process instead they can focus on network analysis part so that, the complex underlying mechanism can be understood.

References

[1] Abhinandan Khan, et al. "A swarm intelligence based scheme for reduction of false positives in inferred gene regulatory networks," IEEE Congress on Evolutionary Computation (CEC), pp. 40-47, Jul 2016.

[2] Baiyi, An, and Shen Wei. "A Novel Gene Regulatory Network Construction Method Based on Singular Value Decomposition." 2016 IEEE International Conference on Big Data Analysis (ICBDA), 2016, doi:10.1109/icbda.2016.7509844

[3] Lehrach, Hans. "Faculty of 1000 Evaluation for RNA-Seq: a Revolutionary Tool for Transcriptomics." F1000 - Post-Publication Peer Review of the Biomedical Literature, Sept. 2013, doi:10.3410/f.1128830.793481779.

[4] Nitin Singh, et al. "bLARS: An Algorithm to Infer Gene Regulatory Networks," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, pp. 301-314, Jun 2015.

[5] El-Telbany, A., and P. C. Ma. "Cancer Genes in Lung Cancer: Racial Disparities: Are There Any?" Genes & Cancer, vol. 3, no. 7-8, Jan. 2012, pp. 467–480., doi:10.1177/1947601912465177.

[6] López, Félix Melchor Santos, and Eulogio Guillermo Santos De La Cruz. "Literature review about Neo4j graph database as a feasible alternative for replacing RDBMS." Industrial Data, vol. 18, no. 2, 2015, p. 135., doi:10.15381/idata.v18i2.12106.

[7] Ardini, E., et al. "Anaplastic Lymphoma Kinase: Role in specific tumours, and development of small molecule inhibitors for cancer therapy." Cancer Letters, vol. 299, no. 2, 2010, pp. 81–94., doi:10.1016/j.canlet.2010.09.001.

[8] Costa, Raquel L., et al. "GeNNet: An Integrated Platform for Unifying Scientific Workflow Management and Graph Databases for Transcriptome Data Analysis." Dec 2016, doi:10.1101/095257.

[9] Aurelie Pirayre, et al, "BRANE Clust: Cluster-Assisted Gene Regulatory Network Inference Refinement," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, Mar 2017.

[10] Armita Zarnegar, et al, "A heuristic gene regulatory networks model for cardiac function and pathology," Computing in Cardiology Conference (CinC), pp. 353-355, Sept 2016.

[11] ing Liu, et al, "A Dynamic Multiagent Genetic Algorithm for Gene Regulatory Network Reconstruction Based on Fuzzy Cognitive Maps," IEEE Transactions on Fuzzy Systems, vol.24, pp. 419-431, Jul 2015.

[12] Young Hwan Chang, et al. "Retrieving common dynamics of gene regulatory networks under various perturbations," 54th IEEE Conference on Decision and Control (CDC), pp. 2531-2536, Dec 2015.

[13] Osama A. Arshad, et al. "Using Boolean Logic Modeling of Gene Regulatory Networks to Exploit the Links Between Cancer and Metabolism for Therapeutic Purposes," IEEE Journal of Biomedical and Health Informatics, vol.20 pp. 399-407, Nov. 2014.

[14] Mathias Foo, et al. "Modelling and Control of Gene Regulatory Networks for Perturbation Mitigation," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, Jan 2018.

[15] Jisha Augustine, et al. "Gene regulatory network inference: A semi-supervised approach," Electronics, Communication and Aerospace Technology (ICECA), Dec. 2017.

[16] Sima Iranmanesh, et al. "Inferring gene regulatory network using path consistency algorithm based on conditional mutual information and genetic algorithm," Computer and Knowledge Engineering (ICCKE), Dec. 2017.

[17] Mushthofa Mushthofa, et al. "Computing attractors of multi-valued Gene Regulatory Networks using Fuzzy Answer Set Programming," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Nov. 2016.

[18] Thomas Thorne, "Approximate inference of gene regulatory network models from RNA-Seq time series data," Cold Spring Harbor Laboratory, bioRxiv journal, Jan. 2017.

[19] Jason Knight, et al. "Detecting Multivariate Gene Interactions in RNA-Seq Data Using Optimal Bayesian Classification," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, Oct. 2015.

[20] Bin Yang, et al. "Inference of Gene Regulatory Network based on Legendre Neural Network," International Conference on Information Technology in Medicine and Education, pp. 192-194, July 2017