Study And Analysis Of Gene-Protein Interactions From Pubmed Articles Using Conditional Random Field- Named Entity Recognition Technique

G. Suganya, R. Porkodi

Abstract— Extraction of information from the large repositories offers new research opportunities and challenges and the best-known tool, which is used to extract knowledge, is text mining. Text mining is an innovation that can be applied to different tasks in biology and to solve the medical problems, biomedical literature mining is used more and more. The main aim of text mining is used to identify biological entities such as protein and gene names in the biomedical texts and it reduces the effort to extract relationships between biological entities abstracts. The number of studies associated with genes have been conducted to identify the genes involved in proteins. However, this research area remains the lot of scope and open challenges for biomedical text mining researchers. This paper developed a framework to identify gene and its related proteins from PubMed abstracts for Breast cancer. The framework includes two phases such as identify and extract the gene and its protein relations and constructs gene network from the identified relations using cytoscape network visualizer. It identifies the significant number of gene-protein relationships related to breast cancer disease. The identified relations are verified and validated with the benchmarking database and the network analysis results revealed that PFKP, EGFR, BMP9, RhoA, TRIM33, CDK5 and STAT3 are top genes that were found and the relations for the PFKP and EGFR genes are high when compared to the other top genes. Overall, the proposed framework produces 75% accuracy results.

Keywords: Text mining, Gene, Protein, Gene network, Pubmed abstracts

1 INTRODUCTION

One of the important area for text mining application is systems biology and systems biology involves the iterative interplay between the computational modelling, high-throughput and high content experimentation and technology development. The common functionalities of text mining are NER (Named Entity Recognition), Text classification, Synonym and abbreviation extraction, relationship extraction and hypothesis generation and in which NER is to identify the entities names like gene, protein, drug, mutants, chemical or disease from the biomedical text document [1]. In the biomedical area, text-mining has been used to identify biological entities such as protein and gene names in the literature and furthermore, text-mining can reveal novel relationships among biological entities. Text-mining can provide opportunities to reduce the time and effort needed to extract relationships between biological entities from a large amount of publications. Interest in text-mining is increasing due to the number of electronic publications stored in databases are increased such as PubMed [2]. The system generally involves annotating raw text with named entities and extracting relationships between these entities. Named entity recognition (NER) is the foundation of relationship extraction and the effect of entity recognition greatly affects relationship extraction results [3].

Biological Network deals with 3 network categories: Pathway, Similarity Network and Interaction Network. Interaction network, nodes represent biological entities edges represent some form of interaction/relationships. Based on the different levels of integration of cellular processes, the biological networks can be classified into 4 types.

i) Gene network, representing genome-wide interactions
ii) Protein network, representing proteome interactions
iii) Signal transduction network for interactions between genes, proteins and other cellular signaling molecules.
iv) Metabolic network for biochemical interactions between substrates and enzymes.

Network analysis also plays an important role in biological research. Gene networks, which describe gene-gene interactions, and protein networks, which describe protein-protein interactions, allow the visual relationships among biological entities in complex biological systems to be presented in a simple and clear manner. Network analysis also provides an opportunity to analyse which relationships are meaningful among various candidates. Several techniques have been developed to extract the hidden information using text mining and network analysis [4]. Several software tools

- *Suganya G is currently pursuing Ph.D in Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India. E-mail: suganyadheksha@gmail.com*
- *Porkodi R is currently working as Associate Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India. E-mail: porkodi_r76@buc.edu.in*

are available for network visualization and analysis including Osprey, VisANT, CellDesigner, GenMAPP, PIANA, Proviz, Biolayout, PATIKA and Cytoscape, Gene Science Visualizer [3]. The rest of the paper is organized as follows: Section 2 discusses some related work including existing approaches to identify the entity association. Section 3 illustrates the proposed framework for identifying gene-protein relationships and constructing the gene network. Section 4 describes the results and discussion. Finally, Section 5 summarizes the conclusion and future directions.

2 Literature survey

Gonzalez et al [4] presented a method which uses literature data and interactions. The author extracted an initial set of genes and proteins from the literature and then integrated the set with interactions from the curated databases of BIND and DIP. Then constructed a network and rank the genes. Gene products using a combination of the two scores such as one scores measures the strength of the relationship with the initial set of genes, and the other score measures the importance of each gene in maintaining the connectivity of the network. The method shows high accuracy levels for atherosclerosis. Chen et al [5] presented a method that constructs a gene-regulatory network using microarray data and literature-based knowledge. Through this, first extracted gene-gene relationships from the literature and then assigned random weights to the relationships. Through this process, generated 2000 chromosomes. Subsequently, used a genetic algorithm to optimize the strength of the interactions using a microarray and an artificial neural network fitness function. The results demonstrated the advantage of combining gene interactions extracted from the literature with microarray analysis in generating contribution weighted gene-regulatory networks. Li et al [6] tried to integrate both literature and microarray gene-expression data. Authors constructed a gene network using the co-occurrence-based text-mining method and then refined the network using microarray data. The results showed that the network is more reliable than the co-occurrence-based network. Ozgur et al [7] determined the relationships between prostate cancer, genes using text-mining and network analysis. Constructed a disease-related gene network using the biomedical literature and seed genes, and extracted disease-related genes based on an analysis of the gene network using different scoring methods. A seed gene is nothing but a gene known to be involved in a disease. Although it inferred prostate cancer related genes successfully, it cannot be used to determine the relationships between genes and diseases for which there are no seed genes. Lee et al. [8] presented a method that inferred relationships between Alzheimer's disease and drugs using an advanced version of the ABC model. The method is incorporated context-term vectors into the ABC model to infer meaningful relationships. The author extracted various relationships from the literature by means of text-mining and created a context-term vector based on biological entities which occur in conjunction with relationships in the literature and calculated scores for relationships using context-term vectors and inferred more accurate relationships between Alzheimer's disease and drugs than the ABC model. Jie Zhou et al [9] proposed a method to integrate Mesh database, term weight and co-occurrence methods to predict gene-disease association based on cosine similarity between gene vectors and disease vectors and evaluated the performance of prediction results by comparing with HNEP method. The novelty is based on the combination of text mining and graphic model. The HNEP method integrates the graphic model and MLM to predict gene-disease linkages based on logistic regression analysis. The precision rate of the method was significantly higher than the HNEP method when the recall rate higher than 0.1. Komandur Elayavilli Ravikumar et al [10] developed a text mining system "MutD" which extracts protein, mutation-disease associations from Medline abstracts by incorporating discourse level analysis. The proposed algorithm achieves an F-measure of 64.3% for reconstructing protein mutation-disease associations. Kyubum Lee et al [11] proposed a computational method that utilizes all the PubMed articles as domain specific background knowledge in the extraction and curation of gene-mutation-drug relations from the literature. It includes two methods such as one uses the BEST scoring results as some of that are train the machine learning classifier, other one is focused on not only the BEST scoring results and it also uses the word vectors in Deep Convolutional Neural Network model that are constructed from and trained on numerous documents such as PubMed abstracts, Google news articles. It extracts mutation-gene and mutation-drug relations from the literature using Machine learning classifiers such as Random Forest and Deep Convolutional Neural Network. Using Deep learning, the classification accuracy and F-score of 0.96, 0.86 were obtained for the mutation-gene, mutation-drug relations respectively. Lada A. Adamic et al [12] presented a statistical method for identifying the set of genes knowns to be associated with the breast cancer disease. The gene symbols, alias names are collected from various databases such as Human Genome Organization, Online mendelian inheritance in man and Locuslink database. Performed an automated search of the abstract and title to find the PMID identifier and gene name or symbol. It does not search the full name of the gene and did not find the number of occurrences, only the gene symbol was identified. This algorithm also identifies most of the genes presented in breast cancer from the human edited database as well as identified many additional genes. Jeongkyun Kim et al [13] proposed a method to identify disease related genes that are involved in development of disease from medline abstracts. It identified the associations between 13,054 genes and 4,494 disease types which cover more disease related genes than the manually curated databases for all disease type. Digsee identifies more disease related genes than the OMIM, GWAS databases. For the comparison case study Alzheimer disease is used and hypertension. Sune Pletscher Frankild et al [14] presented a

- *Suganya G is currently pursuing Ph.D in Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India. E-mail: suganyadheksha@gmail.com*
- *Porkodi R is currently working as Associate Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India. E-mail: porkodi_r76@buc.edu.in*

system for extracting disease-gene associations from biomedical abstracts. The system consists of highly efficient dictionary-based tagger for named entity recognition for human genes and disease. This approach extract half of all manually curated associations with false positive rate as only 0.16%. For the above reason developed the "diseases" resource which integrates the results from text mining with manually curated disease-gene associations, cancer mutation data and genome wide association from existing databases. Chun Hw et al [15] proposed a system to extract the disease-gene relations from medline abstracts. It constructed a dictionary for disease and gene name from public database and extracted relation candidates by dictionary matching. The dictionary matching produces many false positive, so approach of machine learning based named entity recognition to filter out false recognitions of disease/ gene names. The performance of relation extraction was heavily dependent upon the performance of named entity recognition filtering. The filtering improved the precision of relation extraction by 26.7% at the cost of small reduction in recall. Jeongkyun Kim et al [16] constructed a disease related gene network using literature and google data. It identified disease related genes using analysis of gene network. The proposed method LGScore found more disease related genes than comparable methods. It contains three phases such as Construct a disease related gene network using text mining results and the second phase is to extract gene-gene interactions based on co-occurrences in abstracts data obtained from PubMed abstracts and the final phase is to calculate the weights of edges in the gen network by means of Z-Score for validation of top 20 genes for different disease using answer sets. It identified a significant number of disease related genes as well as candidate genes for Alzheimer, diabetes, colon cancer, lung cancer and prostate cancer. The proposed method achieved the 40% more accuracy than the existing method.

3 Methodology framework

The framework describes the methodology for classification of gene and its related protein names and constructed the gene network. The methodology comprised of following pipelining architecture is illustrated in figure 1.

i) Identifying the entities
ii) Entities relations identification
iii) Validating the identified relations which includes the entities
iv) Constructing the network using the network visualizer

The research work collected the 500 abstracts from NCBI: PubMed (http://www.ncbi.nlm.nih.gov/PubMed) related to Breast cancer disease. PubMed provides biological literature data in an abstract format. In the PubMed database, abstract data is generated by search results for an input keyword. To obtain disease specific abstract data, use disease names as search keywords in PubMed and a large amount of abstract data is available for breast cancer. Table 1 shows the number of abstracts and size of the dataset.

Table 1 Dataset description

Name	Breast Cancer
Number of Abstracts	500
Size	3634KB

3.1 Pre-processing

Two pre-processing techniques are applied for experimental study such as Stop word removal and Stemming. In stop word removal, it removes the unnecessary data such as author, institute, date and journal name from the abstract data. It also categorizes the sentences according to the POS tagging using MedPOSt Tagger. To perform the stemming, Porter Stemmer algorithm is used.

3.2 Entity Identification and Tagging

Noun phrases can be identified from the input text document and Noun symbol consists of NN, NNP, NNPS and NNs. The genes are extracted from the abstracts using Conditional Random Field (CRF) based Named Entity Recognition (NER) techniques. The extracted entities based on comparison with Unified Medical Language System (UML). UML classifier identified the entities by semantic type; entities are extracted only if semantic type was gene or genome. It also used the UMLs concept for unique identifiers to identify a preferred term for each extracted entity. The filtered terms are related to genes such as gene or genome but not gene itself. The system extracts a simple statement from the sentence by finding the right subject (noun) and object term (verb).

3.3 Identifying the gene-protein relations

Named entity recognition (NER) task is used for extracting gene associations. Genes are usually represented by symbols and names in literature. The names usually are the long terms of their symbols and describe the functions of the genes operations. Once the entities are identified the next step is to identify the relations in the input document. After identifying the relations, must identify what the relationships is. The gene-protein interactions are extracted from the input data using text-mining [2]. The target is to analyse the text to find out the interaction between genes which are usually expressed by frequently seen verbs in the biomedical domain such as "activate"," interact", "bind", etc.

The gene relationship was computed for all connected gene in the network. To identify the associations between one entity to other entities could not validate predictions for all combinations of available genes. Gene associated with other protein is mapped in gene network. The figure 2 represents the entities identified from the two abstracts and its relations.

3.4 Validation

Identified entities are validated with the benchmarking HUGO gene Nomenclature committee database. Nodes and edges of the gene network were constructed based on the co-occurrences of the entities. It identifies the groups of nodes in a network which are more similar to each other and also optimizes the detection the community structure in network. Each component identifies the gene/proteins and representative genes belonging to the specific component and the network shows broadly spread genes association.

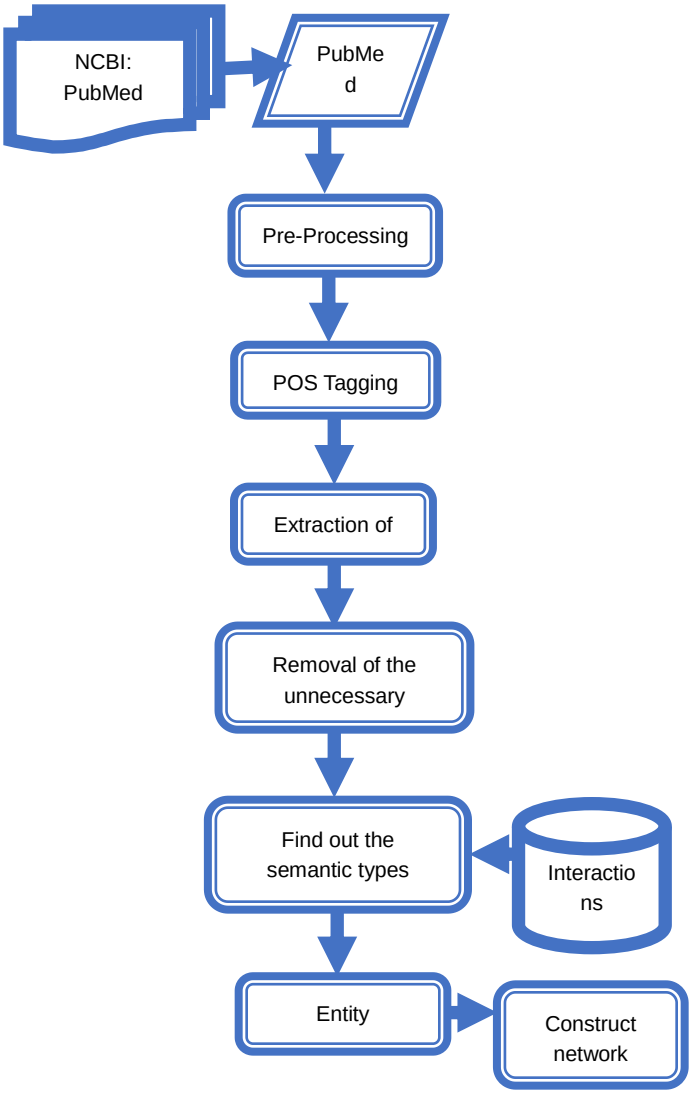


Fig 1. Methodology framework

3.5 Cytoscape-Network Visualizer

There are three open source tools are mostly used for network visualization such as Cytoscape, Social network Visualizer and Gephi. This paper used Cytoscape as the network visualizer to visualize the relations found by the framework. It is freely available and distributed which allows the activities of software including feature extension by programming. Here, nodes representing biological entities such as protein or genes relate to edges representing pairwise interactions such as experimentally determined gene-protein interactions. Nodes and edges can have associated data attributes describing properties of the protein or interactions. A key feature is its ability to self-visual aspects of nodes and edges such as shape, color and size based in attribute values. It allows users to extend its functionality by creating or downloading additional software modules known as "plugin" [4].

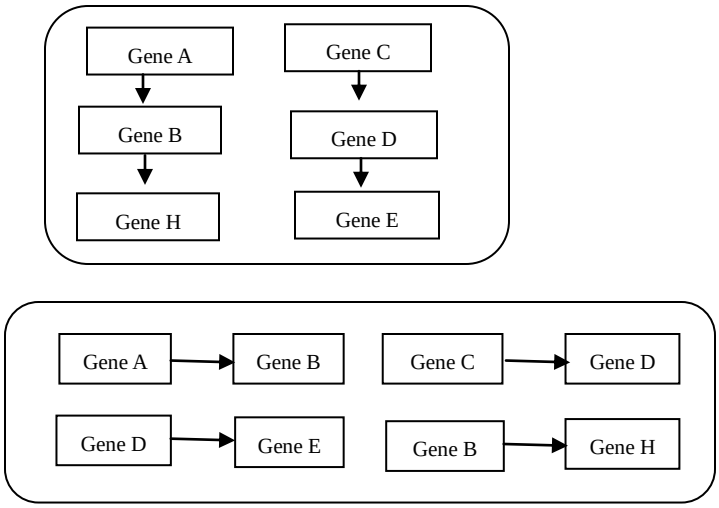


Fig. 2 (a) Entity finding from two abstracts (b) Relationships are identified by the identified Entities

4 Results and Discussion

4.1 Extraction of Entities

The named entities refer to gene and protein names within the abstracts were retrieved. The Named Entity Recognition (NER) is used for identifying the entities present in the input text document. The identified entities along with the semantic type and corresponding POS tag are shown in Table 2. The snippet of entity tagging is shown in example 1 and 2.

Example 1:

Original Sentence: *Aberrant activation of receptor tyrosine kinases (RTKs), including platelet-derived growth factor receptor (PDGFR), are frequently observed in glioma.*

Tagged sentence: *Aberrant activation of* ***<entity> receptor tyrosine kinases (RTKs) <entity>****, including* ***<entity> platelet-derived growth factor receptor (PDGFR) <entity>****, are frequently observed in glioma.*

Example 2:

Original Sentence: *Importantly, the occupancy of L3mbtl1 on the Ctnnb1 gene was regulated by neuronal activity.*

Tagged Sentence: *Importantly, the occupancy of* ***<entity> L3mbtl1 <entity>*** *on the* ***<entity>Ctnnb1<entity>*** *gene was regulated by neuronal activity.*

4.2 Identification of gene-protein interactions

The identified entities are tagged in the biomedical text document and the relations are identified by the interaction keywords. There are 40 interaction keywords are considered and some of the important and frequently used keywords are "interacts", "binding", "activates", "includes", "associations", "inhibits", "regulates", etc. Table 3 depicts the gene-protein relations identified by the framework with following information such as entity name, association keyword and the complete sentence where these two features are present.

The snippet of entities associations is shown in Example 3 and 4.

Example 3:

Original Sentence*: CHD8 activates expression of BRG1-associated SWI/SNF complexes that in turn activate CHD7, thus initiating a*

successive chromatin remodeling cascade that orchestrates oligodendrocyte lineage progression.

***Relationships Identified**: **<entity> CHD8<entity> activates** expression of **<entity> BRG1-associated SWI/SNF <entity>** complexes that in turn activate **<entity> CHD7 <entity>**, thus initiating a successive chromatin remodeling cascade that **<entity> orchestrates <entity><entity> oligodendrocyte <entity>**lineage progression.*

Example 4:

***Original Sentence:** Western blot results of autophagy-associated protein (LC3 II, Beclin-1) and apoptosis-associated proteins (caspase-3, Bcl-2, Bax) revealed that AG-1031 could activate apoptotic signal pathway via inhibiting autophagy process in cancer cells.*

***Relationships Identified:** Western blot results of autophagy-associated protein (**<entity> LC3 II <entity>, <entity> Beclin-1 <entity>**) and apoptosis **associated** proteins (**<entity> caspase-3 <entity> , <entity> Bcl-2 <entity> ,<entity> Bax <entity>**) revealed that AG-1031 could activate apoptotic signal pathway via inhibiting **<entity> autophagy <entity>** process in cancer cells.*

Based on the identified genes and its relations first construct the gene network using network visualizer. The relations are identified by the co-occurrence of the entities appeared in the same article as well as sentences. The focus is on identifying a relationship between genes, the directionality of citation provided no additional understanding as it might not reflect gene-to-protein directionality. Based on these analyse, gene-protein network is formed.

Using co-occurrence frequency, identified frequently co-occurring genes in bioinformatics. Assuming the frequently occurring, genes represent the core genes in bioinformatics, then use network analysis to gain insight into how these genes are interact with each other. Table 4 shows some of the gene-protein associations for constructing the network.

4.3 Visualizing the gene network

Visual analysis of the gene-protein network shown in fig 4. The visualizer identifies and create a group of nodes in a network and each group identifies the association between the two entities and representative genes belonging to the specific cluster. The identified relations are depicted in fig.4. Gene to protein network shows the broadly spread genes associated with the diseases. The visualizer used the node colours to represent degrees of gene-protein similarity. The nodes indicate genes, while edges indicate gene-protein interactions. In their gene network, it confirmed that a network can be used to present useful knowledge between biological entities. The gene-protein network shows that a set of genes with similar properties tend to form a fragmented cluster. To identify the characteristic of the gene to protein network based on the extracted gene names from abstract in the field of bioinformatics. The gene-proteins network consists of 260 nodes and 763 edges. Analysing all pairs of the network reveal 789 pairs are shown in gene-protein network which represents about number of pairs identifies by the system. This shows the top gene pairs commonly appearing in the networks are significant in bioinformatics. By using the same process, analysed gene pairs appearing in the gene-protein networks based on the top 5, 10, 50 and 100 genes and are shown in figure. For the breast cancer, there are 10 components are formed in which component 1,2,3, and 4 are small number of components that have common protein names associated with most of the genes. Cluster 10 has genes with more numbers of genes such as 879. The network analysis results revealed that PFKP, EGFR, BMP9, RhoA, TRIM33, CDK5 and STAT3 are top genes that were found and the relations for the PFKP and EGFR genes are high when compared to the other top genes. In figure 5 represents the subnet network for the identified relations using Cytoscape visualizer.

5 Conclusion

The study has focused on the co-occurring genes in Breast cancer literature and discussed the complete list of genes tagged by the Named Entity Recognition (NER) and demonstrated the feasibility of network assisted identification of entities. The paper presents the construction of gene-protein relations network that significantly improve the predictive power of probabilistic functional gene network of the Breast cancer. The framework contains the four major aspects. First is Pre-processing the abstracts, Second identify the entities using NER, Third is identify the semantic types of relationships by using interaction words and finally the semantic relationships are visualized using Cytoscape. The paper discussed the framework of identifying the relations and construction of gene-protein network for functional prediction and prediction of essential genes relations. It demonstrated that the gene relationships based on citation relation extends the assumption of gene interaction being limited to the same articles and opens a new opportunity to analyse gene interaction from the wider spectrum of datasets. It identifies the significant number of gene-protein relationships related to breast cancer disease. The identified relations are verified and validated with the benchmarking HUGO database and the network analysis results revealed that PFKP, EGFR, BMP9, RhoA, TRIM33, CDK5 and STAT3 are top genes that were found and the relations for the PFKP and EGFR genes are high when compared to the other top genes. Overall, the proposed framework produces 75% accuracy results. The method has following limitations i) the semantic relation analysis is important and necessary to reduce the errors. ii) Complex relations are not extracted. iii) It suitable for abstract level and used simple co-occurrence based on text mining algorithms to extract the relations. In future with the use of full-text articles and advanced techniques can be used to overcome these limitations.

References

[1] Balu Bhasuran, Devika Subramanian, Jeyakumar Natarajan, “Text mining and network analysis to find functional associations of genes in high altitude disease”, Computional biology and chemistry 75, pp 101-110, 2018.

[2] Jeongwoo Kim, Hyunjin Kim, Youngmi Yoon, Sanghyun Park, “LGscore: A method to identify disease-related genes using biological literature and Google data”, Journal of Biomedical informatics, Vol 54, April, pp 27-282, 2015.

[3] PubMed: MEDLINE Retrieval on the World Wide Web

[4] Luo Y , "Bridging semantics and syntax with graph algorithms state-of-the-art of extracting biomedical relations", Brief. Bioinformatics , 18, pp 160–178, 2017.

[5] G. Gonzalez, J.C. Uribe, L. Tari, C. Brophy, C. Baral, "Mining gene–disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures", Pacific Symp Biocomput, 12, pp. 18-39, 2007.

[6] G.Chen, M.J. Cairelli, H. Kilicoglu, D. Shin, T.C. Rindflesc h, "Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference", PLoS Comput Biol, Vol 10, Issue 6, 2014.

[7] S. Li, L. Wu, Z. Zhang, "Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach", Bioinformatics, Vol 17, Issue 22, pp 2143-2150, 2015.

[8] Ozgur A, Vu T, Erkan G, Radev DR,"Identifying gene–disease associations using centrality on a literature mined gene–interaction network", 24:i277–85, 2008.

[9] S.J. Lee, J. Choi, K. Park, M. Song, D. Lee, "Discovering context-specific relationships from biological literature by using multi-level context terms", BMC Med Inform Dec Mak, Vol 12, Issue 1, 2012.

[10] Jie zhou, Bo-quan Fu, "The research on gene-disease association based on text mining of Pubmed" , BMC Bioinformatics, Feb, Vol 19, Issue 37, 2018.

[11] Komandur Elayavilli Ravikumar, Kavishwar B. Wagholikar, Dingcheng Li, Jean Pierve Kocher, "Text mining facilitates database curation-extraction of mutation-disease associations from biomedical literature", BMC Bioinformatics, April, Vol 13, Issue 17, 2016.

[12] Kyubum Lee, Byounggun Kim, Yonghwa choi, Sunkyu Kin, "Deep Learning of mutation-gene-drug relations from the literature", 2018.

[13] Lada A. Adamic, Dennis Wilkinson, Bernardo Huberman, Eytan Adar, "A literature-based method for identifying gene-disease connections", IEEE Computer Society Bioinformatics Conference, 2012 Feb.

[14] Jeongkyun Kim, Jung-jae Kim and Hyunju Lee, "An analysis of disease-gene relationships from medline abstracts by Digsee", Scientific reports, 2017 Jan.

[15] Sune Pletscher-Frankild, Albert Palleja, Kalliopi Tsafou, Janos X.Binder, Lars Juhl Jensen, "Diseases: The text mining and data integration of disease-gene associations", Methods, Mar, Vol 74, pp 83-89, 2014.

[16] Chun Hw, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J, "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning", Pac Symposium on Biocomputing, 2006.

[17] Jeongwoo Kim, Hyunjin Kim, Youngmi Yoon, Sanghyun Park, "LGscore: A method to identify disease-related genes using biological literature and Google data", Journal of Biomedical Informatics, Jan, Vol 54, pp 270-282, 2015.

[18] Arzucan Ozgur, Thuy Vu, Gunes erkan, Dragomir R Radev, "Identifying gene-disease associations using centrality on a literature mined gene interaction network", 2018.

Table 2 (a) Named entity tagging

[1] Word	[2] Semantic Type	[3] POS tag
[4] f1a	[5] ['gene']	[6] NN
[7] Rbpj	[8] ['protein']	[9] NN
[10] AEP	[11] ['Protein']	[12] NN
[13] PP2A	[14] ['Gene']	[15] NN
[16] Synthetase	[17] ['Protein']	[18] NN
Superoxide	['Protein']	NN
Lipid	['Protein']	NN
Hydroperoxide	['Protein']	NN
LPH	['Protein']	NN
4-hydroxynonenal	'Protein']	NN
3-nitrotyrosine	['Protein']	NN
Glutathione	['Protein']	NN
GPx	['Gene']	NN

Table 2 (b) Named entity tagging

Word	Semantic Type	POS tag
ATF4	['gene']	NN
DDIT3	['protein']	NN
TRIB3	['protein']	NN
pseudokinase	['Gene']	NN
Boston	['Protein']	NN
Kinase	['Protein']	NN
MSK1	['protein']	NN
proliferation	['Protein']	NN
EGFR	['Gene']	NN
Exon	['Protein']	NN
L858R	['Protein']	NN

Entities	Action	Sentence
"C/EBP : ['Protein']", "DR5 : ['Protein']", "ATF4 : ['gene']"	Induced	Specifically, BIX-01294 induced C/EBP homologous protein (CHOP)-mediated DR5 gene transcriptional activation and DR5 promoter activation was induced by upregulation of the protein kinase R-like endoplasmic reticulum kinase-mediated activating transcription factor 4 (ATF4).
"GSK3 : ['Gene']", "proliferation : ['Protein']"	activation	GSK3 is constitutively active in T cells and is transiently inactivated during T cell activation resulting in rapid T cell proliferation.
"EGFR : ['Gene']", "erbitux : ['Protein']", "tyrosine : ['Protein']", "LTC : ['Protein']"	inhibitor	Irrespective of ligand or receptor expression, neither an EGFR antibody, erbitux, nor an EGFR tyrosine kinase inhibitor (TKI), gefitinib, were particularly active against LTC or GIC at clinically relevant concentrations.
"Smad6 : ['Gene']", "STAT3 : ['PROTEIN']", "PIAS3 : ['gene']"	regulates	In this study, we report that Smad6 is overexpressed in nuclei of glioma cells, which correlates with poor patient survival and regulates STAT3 activity via negatively regulating the Protein Inhibitors of Activated STAT3 (PIAS3).
"OGD : ['gene']", "Akt : ['Protein']"	Promotes	These findings suggest that PTEN inhibition promotes post-ischemic angiogenesis in HUVECs after exposure to OGD and this enhancing effect might be achieved through activation of the Akt signal cascade.
"cannabinoid : ['Protein']", "CB1 : ['Gene']", "CB1R : ['Gene']"	Activation	Co-administration with selective cannabinoid receptor subtype blockers revealed that PrNMI's anti-allodynic effects are mediated by CB1 receptor (CB1R) activation.
"HGF : ['protein']", "HGFAC : ['gene']", "protease : ['Protein']"	Activator	On the other hand, all glioblastoma lines expressed mRNA for HGF activator (HGFAC), a target protease of HAI-2/SPINT2.

a) Mechanistically, BMX bypasses the suppressor of cytokine signaling 3 (SOCS3)-mediated inhibition of Janus kinase 2 (JAK2), whereas NPCs dampen the JAK2-mediated STAT3 activation via the negative regulation by SOCS3, providing a molecular basis for targeting BMX by ibrutinib to specifically eliminate GSCs while preserving NPCs.

b) Mechanistically, **BMX** bypasses the suppressor of **cytokine** signaling 3 (**SOCS3**)-mediated inhibition of Janus kinase 2 (**JAK2**), whereas NPCs dampen the **JAK2-mediated STAT3** activation via the negative regulation by SOCS3, providing a molecular basis for targeting BMX by ibrutinib to specifically eliminate GSCs while preserving NPCs.

c) Mechanistically, **BMX** bypasses the suppressor of **cytokine** signaling 3 (**SOCS3**)-mediated **inhibition** of Janus kinase 2 (**JAK2**), whereas NPCs dampen the **JAK2-mediated STAT3 activation** via the negative regulation by **SOCS3**, providing a molecular basis for targeting **BMX** by ibrutinib to specifically eliminate GSCs while preserving NPCs.

d)
BMX-**activation**-STAT3
BMX-**activation**-GSC
JAK2-mediated-**activation**-STAT3
SOC3-**inhibition**-JAK2

e)
BMX-**activation**-STAT3
BMX-**activation**-GSC
JAK2-mediated-**activation**-STAT3

Fig 3 Example of a scheme for relationships identification. a) Original sentence b) After identifying the entities (Gene and Protein) c) Identify the semantic types (i.e keywords with Entities) d) Predicted result e) Database result

Table 4 Identified gene-protein association

Gene	Action	Protein
NLRP3	activation	IL-18
NLRP3	activation	caspase-1
BMP9	activation	SMADs

BMP9	activation	GBM
BMP9	activation	proliferation
BMP9	activation	PI3K/AKT/MAPK
BMP9	activation	RhoA/Cofilin
AT2R	activation	redox
AT2R	activation	TRPA1
DDR	activation	DNA
C/EBP	induced	DR5
C/EBP	induced	ATF4
C/EBP	activation	DR5
C/EBP	activation	ATF4
LTC	activation	ERBB

Gene	Action	Protein
Ptf1a	activation	Rbpj
p53	activation	chr10
SIRT1	activator	GBM
SIRT1	induces	GBM
HCC	inhibition	PROX1
HCC	inhibition	HBx
HCC	activation	PROX1
HCC	activation	HBx
NLRP3	activation	inflammasome

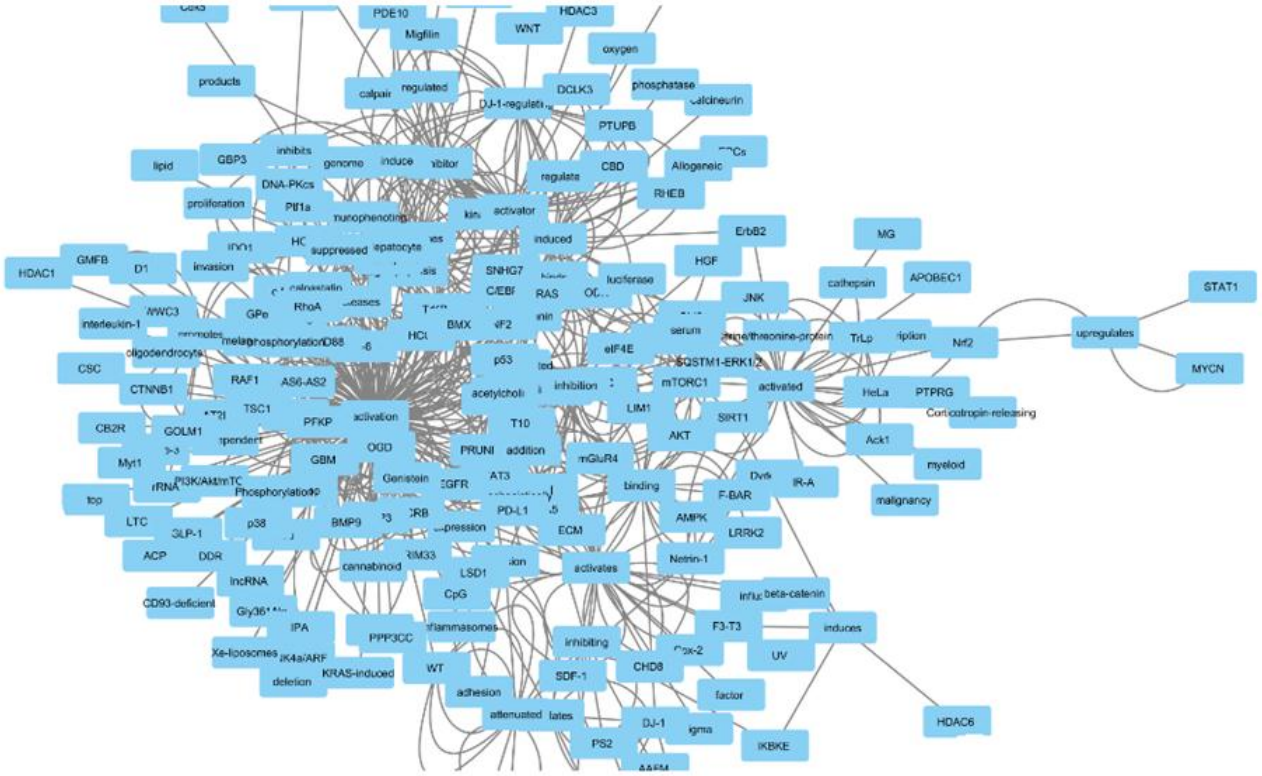
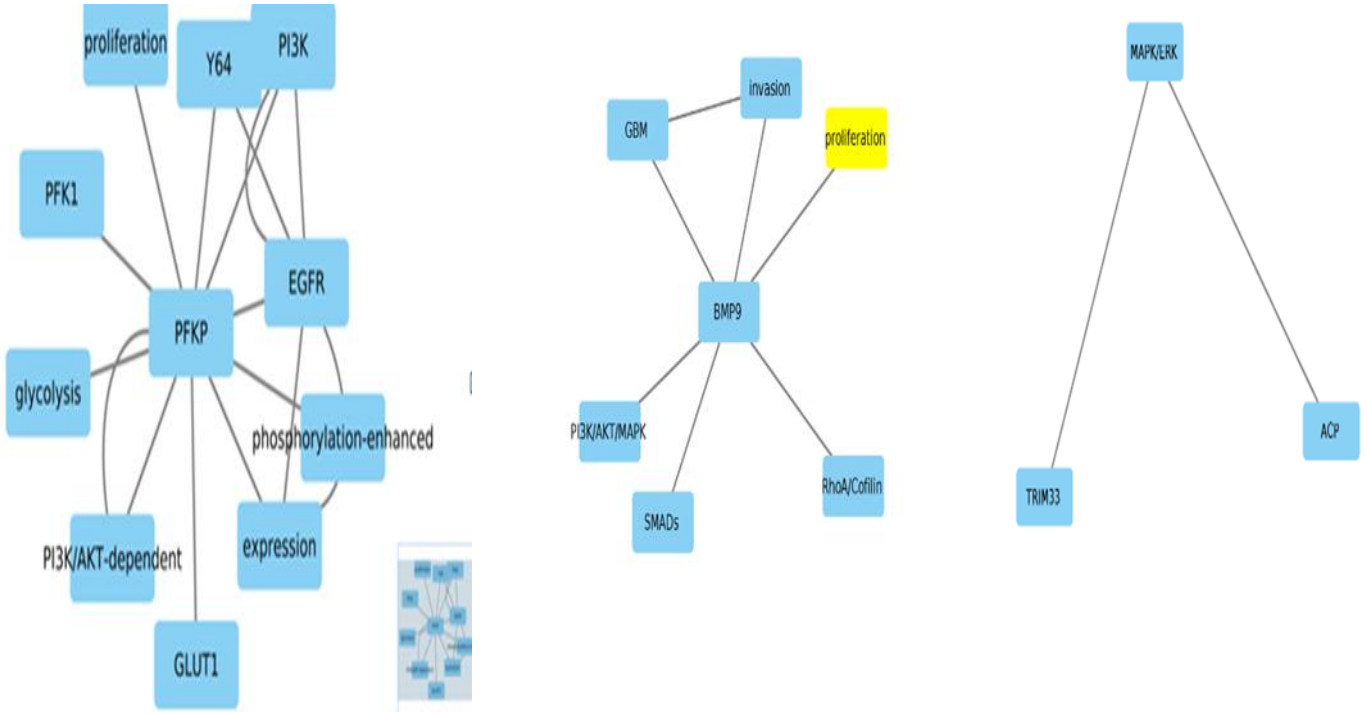


Fig.4 Visualization analysis of network



(a) PFKP (b) BMP9 (c) MAPK

Fig.5 Subnet of gene-protein network a) Graph showing the relationships between genes (PFKP to others) b) Genes associated with BMP9 – other entities c) Gene relationships with other protein (MAPK/LRK- TRIM33 and ACP)