2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

ISN: Inferring disease-related genes using seed gene and network analysis

Jeongwoo Kim
Department of Computer Science
Yonsei University
Seoul, Korea
jwkim2013@yonsei.ac.kr

Sanghyun Park
Department of Computer Science
Yonsei University
Seoul, Korea
sanghyun@yonsei.ac.kr

***Abstract*— In biology, text-mining is widely used to extract relationships between biological entities. Gene prioritization is also important to analyze diseases, because mutated or dysregulated genes play an important role in pathogenesis. Here, we propose a method to identify disease-related genes using seed genes and network analysis. We constructed an integrating gene network for lung cancer by combining local gene networks for seed genes. Analyzing the integrating gene network, we inferred meaningful lung cancer-related genes and potential candidate genes. We also demonstrated that our method is more useful for extracting disease-gene relationships than previous methods. In this study, we extracted 21 lung cancer related genes and 11 candidate genes with supporting evidence of their association with lung cancer.**

Keywords—text-mining; gene; disease; network;

I. Introduction

The biomedical literature is generated based on the results and discussions of biological experiments. These data are stored in public online databases such as PubMed [26] and OMIM [24]. These databases allow researchers to search the biomedical literature for various data and information. It is therefore easy and convenient to access useful biomedical data. However, the size of the literature data is too large to be read thoroughly by researchers. To address this issue, text-mining is widely used to extract relevant biological knowledge from the literature.

Text-mining is a useful approach to extract interesting relationships such as gene-gene interactions, protein-protein interactions, and disease-gene relationships from a large amount of text data. Additionally, it is possible to infer unexpected information by considering several known relationships.

Biological relationships are important to describe biological phenomenon. Accordingly, several studies have attempted to extract useful biological relationships using text-mining. [5, 7, 10] Palakal et al. [25] presented a method for extracting meaningful relationships between biological entities. They considered several steps which include object identification, synonym discovery, and relationship extraction. Based on one thousand abstracts, they extracted 43 correct relationships among 53 extracting relationships. Sharma et al. [27] proposed another method to extract biological relationships using the main verb in a sentence. They showed that the main verb is meaningful for extracting biological relationships.

Gene prioritization is an interesting topic in biological text-mining, because genes play an important role in describing diseases. Several studies have demonstrated the effects of gene prioritization in text-mining. [2, 14, 29] Luo et al. [21] attempted to infer potential candidate genes based on the topological similarity of protein-protein interactions and phenotype data. They applied the method to several diseases including breast cancer, prostate cancer, diabetes mellitus type 2. Gottlieb et al. [8] designed a tool for associating genes with diseases using network propagation. Given a query disease, the tool prioritizes disease-related genes based on the protein-protein interaction network and similar diseases. Using the tool, various disease-gene relationships were extracted. Kim et al. [15] aimed to infer disease-related genes using literature and google data. They extracted gene-gene relationships from the literature, and extracted weights between genes from google data.

A number of previous studies [1, 16] have tried to establish biological relationships, and prioritize disease-related genes from disease-specific studies. However, these approaches confine the scope of literature data as a specific disease-related text. It is therefore possible to miss useful information.

To address this problem, we propose a novel method of inferring disease-related genes using seed gene and network analysis. First, we obtained lung cancer related seed genes, which are included in the OMIM database. The seed gene is a gene that is already known to be related to a particular disease. Second, we downloaded studies for each known gene from PubMed. In these articles, we extracted gene-gene interactions using the HGNC database [9]. Using the extracted interactions, we constructed local gene-gene interaction networks for each known gene. We then built an integrated gene-gene network by combining the local gene-gene networks. After analyzing the integrated gene-gene network, we prioritized lung cancer related genes.

The goal of this study was to prioritize lung cancer related genes using various literature data. We used seed genes for lung cancer to prioritize the literature data. By combining the results generated from several studies for each seed gene, we identified major lung cancer related genes.

The main contributions are described below:

- Secured various literature data using seed genes

- Constructed an integrated gene-gene network.

- Performed gene prioritization based on analysis of the integrated gene-gene network.

978-1-5386-1644-4/17/$31.00 ©2017 IEEE

In Section 2 of this paper, we describe the proposed method. The results and discussions for our experiments are described in Section 3, and we conclude our study with a discussion in section 4.

II. Methods

In this section, we describe a method for inferring disease-related genes using seed gene and network analysis. Fig. 1 outlines the proposed method.

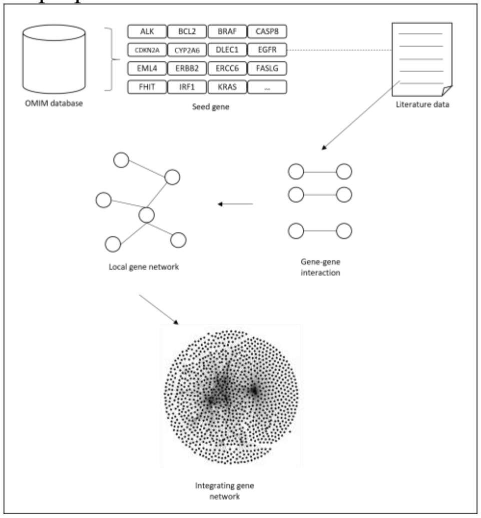

Fig. 1. Example of the proposed method for each seed gene

First, we obtained seed genes from the OMIM database [24]. Then, we downloaded abstract text for each seed gene from PubMed [26] with the MeSH terms. MeSH terms indicate keywords in the literature generated by PubMed. After collecting the literature data, we extracted gene-gene interactions based on co-occurrence. The co-occurrence indicates that if interesting terms appear in the same sentence, then these terms are considered as being related. Using the extracted gene-gene interactions, we constructed local gene-gene network for each seed gene. By combining these local gene-gene networks, we constructed an integrated gene-gene network. Analyzing the integrated gene network based on network analysis measures, we identified lung cancer-related genes.

A. Seed gene and literature data

We obtained 32 seed genes for lung cancer from the OMIM database, which includes several biological data. For each seed gene, we searched and downloaded abstract text from PubMed. The PubMed database also provides various biological data. In this step, the seed gene is used as a search keyword to obtain literature data. Therefore, the obtained literature data is related to the seed gene, and not lung cancer. After obtaining abstract texts for 32 seed genes, we extracted gene-gene interactions from the literature.

B. Co-occurrence based relationship extraction

Using the literature data generated from the previous step, we extracted gene-gene interactions. If two genes appeared in the same sentence, we assumed that there was a relationship between the two genes. This concept is referred to as co-occurrence based text-mining, and is widely used to extract relationships between interesting terms in the text-mining field. Using this concept, we extracted gene-gene interactions from seed gene-related literature data. In case of gene identification, we used an approved gene symbol list of human genes obtained from the HGNC database.

C. Construction of a local gene network

Using the interactions extracted from the relationship extraction step, we constructed a local gene network. The local gene network refers to the gene-gene interaction network for each seed gene. Therefore, we obtained 32 local gene networks. In this step, we used a frequency value as a weight for interactions. The frequency describes the number of sentences that include related genes. The frequency is widely used to determine the weight of relationships. For example, if gene A and B appear in three sentences, the weight of the relationship between gene A and B is three.

D. Integration of local gene network

We constructed an integrated gene network by combining 32 local gene networks. Fig. 2 shows the cases for constructing the integrated gene network.

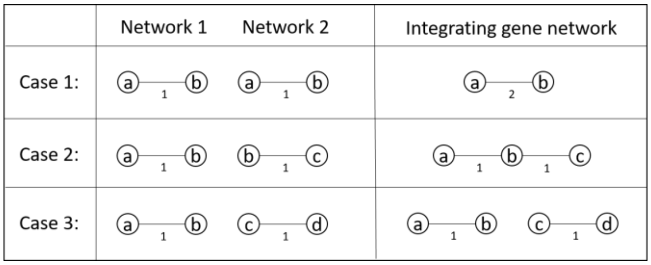

Fig. 2. Example of the construction of the integrated gene network

As shown in Fig. 2, three cases exist in integrated network construction because the type of network is undirected. Case 1 shows that two networks have the same interaction. In this case, we added weights for each interaction. Case 2 indicates that two interactions share one node. In this case, we linked the two interactions through the shared node. If two interactions were independent, the interactions remained unchanged. Using these cases, we combined all of interactions, which are included in 32 local gene networks.

E. Network Analysis and gene prioritization

After constructing the integrated gene network, we analyzed the network using several network analysis measures including degree, weighted degree, eigenvector, and betweenness centrality. The degree centrality indicates the number of neighbor nodes that are linked. The weighted degree considers a weight in degree centrality. The eigenvector is calculated based on the influence of high-scoring and low-scoring nodes. The betweenness indicates the number of times a node is included in the shortest path between two other nodes. For each measure, we extracted the top 20 high scoring genes.

III. Results and Discussions

In this section, we present experimental results and discuss our findings. To validate the inferred disease-related genes, we used an answer set in which known genes are extracted from several databases. Furthermore, we present comparison results by comparing other gene prioritization methods. We also visualize the integrated gene network.

A. Data properties and Answer Set

We used 32 seed genes, which were extracted from the OMIM database. These seed genes are described in Table 1.

Table 1. Seed genes for lung cancer

Gene symbol	# Literature	Gene symbol	# Literature
ALK	822	MET	1,529
BCL2	2,772	MYC	2,092
BRAF	3,368	PIK3CA	910
CASP8	986	PPP2R1B	1,874
CDKN2A	3,719	PARK2	85
CYP2A6	485	PTEN	2,803
DLEC1	46	RARB	383
EGFR	8,082	RASSF1	773
EML4	55	RB1	1,328
ERBB2	4,431	RET	1,315
ERCC6	235	ROS1	173
FASLG	1,157	SLC22A18	45
FHIT	575	SLC34A2	65
IRF1	389	STK11	545
KRAS	3,781	TFG	69
MAP3K8	226	TP53	12,451

Table 1 shows 32 seed genes and the number of studies in the literature for each seed gene. The literature data include unstructured and structured abstract texts. The answer set is described in Table 2.

Table 2. Answer Set

Database	# Gene	Total
GHR	14	104
KEGG	16	
LuGend	72	
IGDB.NSCLC	15	

The answer set consists of four databases which include GHR [6], KEGG [13], LuGend [20], and IGDB.NSCLC [11]. These databases provide lung cancer-related genes. We require the answer set in order to validate genes that are excluded from seed genes among the inferred genes.

B. Inferred Top 20 genes

By analyzing the integrated gene network, we extracted the top 20 genes for each network analysis measure. These genes are described in the Table 3. Table 4 shows the validation results for inferred genes.

Table 3. Inferred top 20 genes for each measure

Rank	Degree	Weighted degree	Betweenness	Eigenvector
1	EGFR	EGFR	EGFR	EGFR
2	BRAF	KRAS	BRAF	KRAS
3	EGF	BRAF	KRAS	EGF
4	KRAS	PIK3CA	EGF	BRAF
5	PIK3CA	EGF	PIK3CA	PIK3CA
6	ERBB2	PTEN	PTEN	TP53
7	PTEN	NRAS	ERBB2	PTEN
8	TP53	TP53	IRF1	ERBB2
9	CASP8	ERBB2	PARK2	CDKN2A
10	NARS	MLH1	CASP8	AKT1
11	PARK2	ALK	PINK1	NRAS
12	CDKN2A	MET	TP53	MET
13	APC	AKT1	FAS	APC
14	AKT1	APC	FADD	CCND1
15	IRF1	MGMT	TNF	CTNNB1
16	CCND1	CASP8	APC	MGMT
17	STAT3	CDKN2A	ATM	STAT3
18	FAS	HRAS	NRAS	BRCA1
19	MET	CTNNB1	RASSF1	MLH1
20	MGMT	STAT3	CYP2A6	IGF1R

Table 3 shows the top 20 genes for each network analysis measure. The yellow color indicates seed genes and the blue color shows genes that are validated by the answer set. We also present other genes with potential relevance to lung cancer.

Table 4 describes the inferred top 20 genes. Table 5 defines common terms and descriptions used in Table 4. Among the measures, the degree centrality identified more lung cancer-related genes than other measures, and the eigenvector inferred more candidate genes for lung cancer.

Answer set validation is limited in validating candidate genes. To address this problem, we conducted literature validation. Literature validation refers to finding sentences or studies describing disease-gene relationships. The literature validation for candidate genes is described in Section C.

C. Validation of Candidate genes

To verify candidate genes not validated by the answer set, we conducted literature validation. Literature validation involves finding studies that include experimental results for disease-gene relationships. The literature validation findings are shown in Table 6. Table 6 shows the candidate gene symbols and key sentences. These gene symbols are extracted from the inferred Top 20 genes in Table 3. The key sentence provides evidence that describes the disease-gene relationship. In the case of the APC gene, for example, we found a study that describes the gene as a potential diagnostic marker for lung cancer diagnosis. As shown in Table 6, we identified key sentences for all candidate genes. As a result, our method presented 11 meaningful lung cancer-related candidate genes with evidence.

This demonstrates that the proposed method is useful for extracting disease-gene relationships and inferring candidate genes.

D. Integration gene network visualization

In this section, we illustrate an integrated gene network for lung cancer.

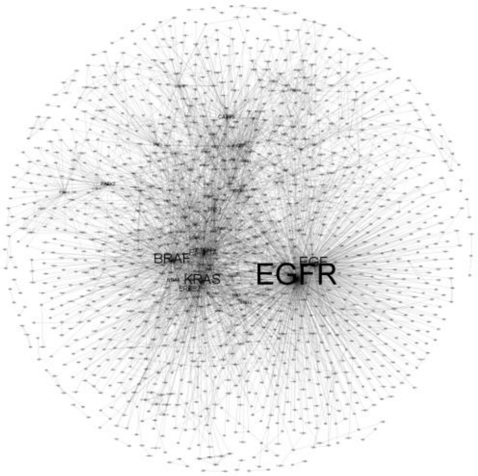

Fig. 3. Integrating gene network

Fig. 3 shows the integrated gene network for lung cancer. The network has 1,339 nodes and 4,092 edges. In the gene network, the node label is proportion to the degree centrality. Therefore, *EGFR*, *BRAF*, *KRAS* and other seed genes are larger than other genes. However, the network is too complicated to confirm genes, because potential genes have low degree centrality values. To address this issue, we re-scaled the network using the edge weight.

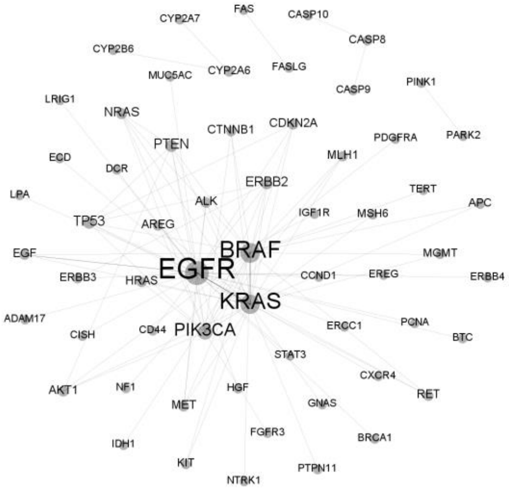

Fig. 4. Gene network for 100 edges with high weight

The gene network has 60 nodes and 100 edges. As shown in Fig. 4, seed genes are in the center of the network. By considering several attributes that include weight, centrality, relationships with seed genes, and the number of edges, we can infer various candidate genes. Accordingly, if a gene network is meaningful, we can extract a lot of information by analyzing the network. Therefore, the proposed method can infer several disease-related genes using the integrating gene network.

E. Comparison results

To validate the proposed method, we present comparison results by comparing previous methods which infer disease-related genes. We extracted the inferred top 10 genes for lung cancer from comparable methods including RWRHN [21] and SSL [17]. These are also methods for inferring disease-related genes using biological data. The RWRHN utilized protein-protein interaction and phenotype data, and the SSL used literature data and auxiliary verbs.

Using the method proposed in this study, we present the top 10 genes based on degree centrality. To validate the inferred genes, we conducted an answer set validation. The comparison results are described in Fig. 5. As shown in Fig. 5, the proposed method identified more lung cancer-related genes than the SSL and RWRHN methods. Among the top 10 genes, our method found 9 lung cancer-related genes. RWRHN and SSL, identified 8 and 4 genes, respectively. This shows that the proposed method is useful for extracting disease-related genes.

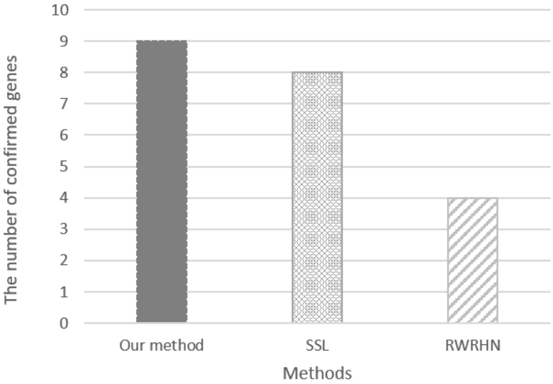

Fig. 5. Descriptions for comparison results

IV. CONCLUSION

We proposed a method for inferring disease-related genes using seed genes and network analysis. We applied our method to lung cancer and demonstrated that it is useful for extracting disease-related genes and potential candidate genes. We also validated our algorithm by comparing two previous studies. Furthermore, by analyzing the integrated gene network based on several network features, we identified various candidate genes.

In this study, we applied our method to lung cancer only. Future studies will involve applying our method to several genetic diseases such as other cancers, Alzheimer's disease, and Parkinson's disease. In addition, we also plan to extract various relationships between biological entities such as disease-drug, drug-gene, and disease-disease.

Table 4. Descriptions for the inferred top 20 genes

	Number of seed genes	Number of inferred genes	Percentage of inferred genes	Number of confirmed inferred genes	Percentage of confirmed inferred genes	Percentage of confirmed genes
Degree	12	8	0.40	4	0.50	0.80
Weighted degree	11	9	0.45	3	0.33	0.70
Eigenvector	9	11	0.55	4	0.36	0.65
Betweenness	12	8	0.40	3	0.37	0.75

Table 5. Term definition

Term	Definition
Number of seed genes	The number of seed genes among the inferred genes
Number of inferred genes	The number of inferred genes which are not seed genes
Percentage of inferred genes	(The number of inferred genes)/20
Number of confirmed inferred genes	The number of genes included in the answer set among the inferred genes
Percentage of confirmed inferred genes	(The number of confirmed inferred genes)/ (The number of inferred genes)
Percentage of confirmed genes	(The number of seed genes and confirmed inferred genes)/20

Table 6. Literature validation for candidate genes

Gene symbol	Key sentence
APC	Hypermethylation of the APC gene promoter in plasma is a potential diagnostic marker for lung cancer diagnosis [3]
ATM	Our study indicates that ATM may serve as a potential molecular target for MDR formation in lung cancer chemotherapy [12]
CCND1	Increased nuclear CCND1 is a potential unfavorable prognostic factor for lung adenocarcinoma patients, especially those with clinical early stage (stage I+II) [30]
CTNNB1	Moreover, the data confirm a crucial role of CTNNB1 mutations in the pathogenesis of PB, and indicate that CTNNB1 gene sequencing may be a useful in distinguishing PB from other types of lung cancer [22]
EGF	The present study revealed that the EGF A61G genotype may be a novel independent prognostic marker to identify patients at higher risk of occurrence and an unfavourable clinical outcome [23]
FADD	The release of FADD by human NSCLC could be a new marker of poor prognosis as it correlates positively with both tumor progression and aggressiveness [4]
HRAS	The present study indicated that there was P21(ras) in human lung cancer and normal control and the expression level of ras gene in lung cancer was related to the differentiation of cancer [18]
IGF1R	Targeting IGF1R and EMT may be a potential therapeutic strategy for advanced NSCLC with acquired EGFR-TKIs resistance [32]
MLH1	This study also suggests that MLH1 -93A>G polymorphisms and ETS exposure have a role in the tumorigenesis of lung adenocarcinoma among never smokers [19]
PINK1	Together, our findings indicate that PINK1 plays a significant role in NSCLC progression and chemoresistance, and highlights its potential role as a target in future anticancer therapies [31]
STAT3	Therefore, STAT3 represents a potential therapeutic target in the treatment of lung-to-brain metastases [28]

ACKNOWLEDGEMENT

Sanghyun Park is the corresponding author. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2017M2A2A7A02020213).

REFERENCES

[1] J. Cha, J. Kim, and S. Park, "GRiD: Gathering rich data from PubMed using one-class SVM," Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on. IEEE, 2016.

[2] J. Chen, et al. "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization." Nucleic acids research, vol 37.suppl 2 pp. W305-W311, July 2009.

[3] Chen, JinBBu, et al. "Methylation quantification of adenomatous polyposis coli (APC) gene promoter in plasma of lung cancer patients." Chinese Journal of Cancer, vol 28 pp. 384-389, April 2009.

[4] Y. Cimino, et al. "FADD protein release mirrors the development and aggressiveness of human non-small cell lung cancer." British Journal of Cancer, vol 106 pp. 1989-1996, June 2012.

[5] Y. Garten, and R.B. Altman. "Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text." BMC Bioinformatics, vol 10 suppl 2 pp S6, February 2009.

[6] Genetics Home Reference https://ghr.nlm.nih.gov/gene/GHR

[7] C.B. Giles and J.D. Wren. "Large-scale directional relationship extraction and resolution." BMC Bioinformatics, vol 9 suppl 9 pp. S11, August 2008.

[8] A. Gottlieb, et al. "PRINCIPLE: a tool for associating genes with diseases via network propagation." Bioinformatics, vol 27 pp. 3325-3326, October 2011.

[9] HGNC Database, HUGO Gene Nomenclature Committee (HGNC). EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB 10 1SD; UK

[10] W.J. Hou and B.Y. Kuo. "Discovery of gene-disease associations from biomedical texts." Computer Science and Information Technology, vol 4 pp. 1-8, 2016.

[11] S. Kao et al. "IGDB. NSCLC: integrated genomic database of non-small cell lung cancer." Nucleic Acids Research, vol 40 pp. D92-D977, January 2012.

[12] S.Z. Ke et al. "Camptothecin and cisplatin upregulate ABCG2 and MRP2 expression by activating the ATM/NF-κB pathway in lung cancer cells." International Journal of Oncology, vol 42 pp. 1289-1296, April 2013.

[13] KEGG: Kyoto Encyclopedia of Genes and Genomes www.genome.jp/kegg/.

[14] J.M. Kim et al. "Identification of gastric cancer–related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells." Clinical Cancer Research, vol 11 pp. 473-482, January 2005.

[15] J. Kim, H. Kim, Y. Yoon, and S. Park. "LGscore: A method to identify disease-related genes using biological literature and Google data." Journal of Biomedical Informatics, vol 54 pp. 270-282, April 2015.

[16] J. Kim, Y. Yoon, and S. Park. "IDO: inferring describable disease-gene relationships using opinion sentences." Proceedings of the 31st Annual ACM Symposium on Applied Computing. ACM, 2016.

[17] J. Kim et al. "SSL: Inferring disease-related genes using sentence structure and literature data." Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on. IEEE, 2017.

[18] X. Lin, X. Shu, X. Sun, X. Zhao, and X. Sun. "A study on P21 (ras) in human lung cancer and body fluid of cancer patients." Hua xi yi ke da xue xue bao= Journal of West China University of Medical Sciences= Huaxi yike daxue xuebao, vol 24 pp. 233-236, September 1993.

[19] Y.L. Lo et al. "Polymorphisms of MLH1 and MSH2 genes and the risk of lung cancer among never smokers." Lung Cancer, vol 72 pp. 280-286, June 2011.

[20] LuGend: Lung cancer gene database www.bioinformatics.org/lugend/

[21] J. Luo and S. Liang. "Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data." Journal of Biomedical Informatics, vol 53 pp. 229-236, February 2015.

[22] S. Macher-Goeppinger et al. "Expression and mutation analysis of EGFR, c-KIT, and β-catenin in pulmonary blastoma." Journal of Clinical Pathology, vol 64 pp. 349-353, April 2011.

[23] M. Masroor et al. "Clinical implication of EGF A61G polymorphism in the risk of non small cell lung adenocarcinoma patients: a case control study." Asian Pac J Cancer Prev, vol 16 pp. 7529-34, 2015.

[24] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetics Medicine, Johns Hopkins University (Baltimore, MD).

[25] M. Palakal et al. "A multi-level text mining method to extract biological relationships." Bioinformatics Conference, 2002. Proceedings. IEEE Computer Society. IEEE, 2002.

[26] PubMed: MEDLINE Retrieval on the World Wide Web https://www.ncbi.nlm.nih.gov/pubmed

[27] A. Sharma, R. Swaminathan, and H. Yang. "A verb-centric approach for relationship extraction in biomedical text." Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on. IEEE, 2010.

[28] Singh, Mohini, et al. "STAT3 pathway regulates lung-derived brain metastasis initiating cell capacity through miR-21 activation." Oncotarget 6.29 (2015): 27461.

[29] Sun, Hui. "Identification of key genes associated with gastric cancer based on DNA microarray data." Oncology letters 11.1 (2016): 525-530.

[30] Xu, Ping, et al. "Elevated nuclear CCND1 expression confers an unfavorable prognosis for early stage lung adenocarcinoma patients." International journal of clinical and experimental pathology 8.12 (2015): 15887.

[31] Zhang, Rui, et al. "High expression of PINK1 promotes proliferation and chemoresistance of NSCLC." Oncology Reports 37.4 (1899): 2137-2146.

[32] Zhou, Juan, et al. "Implication of epithelial-mesenchymal transition in IGF1R-induced resistance to EGFR-TKIs in advanced non-small cell lung cancer." Oncotarget 6.42 (2015): 44332.