LIF: A method to infer disease–gene relationships using literature data and impact factor

Jeongwoo Kim [a, †]
jwkim2014@naver.com

Jinyoung Lee [a, †]
wlsdyd09@naver.com

Heechul Kang [a]
quietus999@gmail.com

Chunghun Kim [a]
jhoney7374@gmail.com

Sanghyun Park [a, *]
sanghyun@yonsei.ac.kr

ABSTRACT

Biological relationships are important in discovering the causes of disease. Therefore, a number of studies have been conducted to extract information regarding the relationships between biological entities. However, given the large number of journals and amount of literature that is available, it is difficult to assess data regarding biological relationships. In this study, we present a method called LIF, which infers disease–gene relationships using literature data and impact factor. Since the impact factor is influenced by a large number of researchers, we considered that the impact factor can be used as a measure to evaluate relationships that are extracted from literature data. To implement the LIF method, we extracted genes from disease-specific literature data. We then calculated the weight of the genes based on the impact factor of the literature in which the genes were described. For validation, we investigated the top N inferred genes for lung cancer, using an answer set. The answer set comprised several databases that contained information on disease–gene relationships. We demonstrated that the LIF is a useful method to infer disease–gene relationships compared with existing methods.

CCS Concepts

• **Applied computing → Life and medical sciences → Bioinformatics**

Keywords

Text-mining; Disease; Gene; Impact Factor

1. INTRODUCTION

After the Human Genome Project (HGP), sequences of genes were determined. Genes of the human genome were identified and mapped to physical chromosome locations. Following this biomedical advance, large amounts of genetic information are stored in databases such as PubMed [29]. This database offers new opportunities to researchers who attempt to find unknown relationships between genes and diseases. Extracting these relationships is important because they can contribute to discovering the causes and novel treatments of diseases. There have been many approaches for discovering the relationships between genes and diseases. One current approach that has been an area of recent interest is text-mining.

Text-mining is a useful approach for extracting information from large amounts of literature data. As the amount of literature data increases, text-mining is becoming increasingly popular with researchers because of its several advantages. Text-mining is an efficient method for discovering new information. Additionally, it is possible to find unexpected but meaningful information by considering several sources of literature that are generated from biological studies. Given these advantages, text-mining has been widely used in biomedical research to identify relationships such as gene–gene interactions, protein–protein interactions, and disease–gene relationships [2, 10, 13].

In this study, impact factor was considered as one of the most useful measures to determine the reliability and significance of literature data. The impact factor is calculated and utilized as a quantitative measure for evaluating journals. The measure represents the frequency with which the average article in a journal has been cited in a given period of time. We therefore assumed that the experimental results of high impact factor papers have been verified by several researchers. Therefore, we considered that extracted information in a journal with high impact factor is more reliable and significant than that from papers published in low impact factor journals. Applying this concept to biological text-mining, we proposed an LIF method to extract more trustworthy disease–gene relationships.

In this paper, we propose a novel method using impact factor and literature data to infer disease–gene relationships. We extracted literature about lung cancer, using MeSH tags in PubMed. Then, we collected gene data from the HGNC (HUGO Gene Nomenclature Committee) database [12] and obtained the impact factor list from the omics online organization [26]. After preprocessing the literature, we extracted gene data from the literature. We then calculated a score for each gene based on impact factors. Finally, we inferred the top N genes with the highest score.

The rest of the paper is divided into four sections. In Section 2, we describe previous studies related to our current work. We describe the proposed method in Section 3 and present our results and a discussion in Section 4 and 5, respectively. We conclude the paper by discussing the implications of our findings in Section 6.

[a] Department of Computer Science, Yonsei University, Seoul, Korea.
Tel: +82-2-2123-7757

[*] Corresponding author

[†] These authors contributed equally to this work

SAC 2017, April 03-07, 2017, Marrakech, Morocco

© 2017 ACM. ISBN 978-1-4503-4486-9/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3019612.3019613

2. RELATED STUDIES

Several text-mining studies [4, 5, 8, 31, 32, 39] have been presented in the biological area. Relationship extraction, named entity recognition, and hypothesis generation are representative in biomedical text-mining. This study addresses relationship extraction to infer disease–gene relationships.

Hou et al. [13] presented a method to discover gene–disease associations from biomedical texts. The presented method has two main steps. In the first, they extracted correct sentences, which include information on human genetic diseases and genes. In the second step, they randomly selected incorrect sentences, which are not involved in human genetic diseases and genes. Using these sentences, they built rules to find pairs of human genetic diseases and genes within one sentence. Weeber et al. [40] introduced a discovery support tool to analyze the scientific literature in order to generate hypotheses. To develop the system, they used Swanson's ABC model [35, 36] structure. Using the system, they found evidence that thalidomide may be useful for treating acute pancreatitis, chronic hepatitis C, *Helicobactor pylori*-induced gastritis, and myasthenia gravis. Srinivasan et al. [31] proposed an open discovery algorithm to identify disease and substances that may have therapeutic potential. In their experiments, retinal diseases are extracted as top ranking entry. In further analysis, they found relationships between retinal diseases and curcumin.

Several studies have extracted relationships between genes and diseases [20, 33, 43]. Kim et al. [17] attempted to identify gastric cancer-related genes using microarray data. Based on analysis of gene expression for human gastric cancer cells, they identified 40 genes that are up-regulated or down-regulated in human gastric cancer cells. Among them, CDC20 and MT2A were identified as potential biomarkers of human gastric cancer. Browne et al. [2] attempted to investigate protein–protein interaction networks (PPINs) to identify an effect on disease network analysis. They used experimental protein–protein interaction data and curated protein–protein data for Alzheimer's disease. Using the protein–protein interaction data, they identified the Alzheimer's disease susceptible TRAF1 gene. Liu et al. [21] proposed a methodology to discover such genes involved in disease using gene expression data and computational protein localization prediction. Le et al. [18] presented ontology-based disease similarity network for disease gene prediction. Using the presented method, they predicted 19 relationships between Alzheimer's disease and genes among the 100 ranked candidate genes. Zickenrott et al. [47] proposed a method to predict disease–gene–drug relationships based on a differential network analysis.

PubMed provides a tremendous amount of publications in various formats, including xml format. Searches in PubMed are processed by inserting search keywords with MeSH tags. MeSH (Medical Subject Headings) is an advanced search engine provided by PubMed to produce better search results. It works based on MEDLINE subject indexing performed by the NLM (National Library of Medicine). Indexing is constructed systematically in three steps. First, researchers review all journal articles. Then, they determine their subject content and finally describe those contents using a controlled vocabulary. As a result, it is possible to search with specific biomedical keywords. For instance, when researchers search "Mucinous adenocarcinoma of the ovary," it is indexed as "Adenocarcinoma, Mucinous Ovarian Neoplasms." Therefore, it is possible to search data flexibly when using MeSH.

3. METHODS

In this section, we describe a proposed method in detail. First, we extracted journals related to lung cancer from PubMed in XML format. After obtaining literature data, we conduct preprocessing for literature data. We parsed the literature data to remove unnecessary information and to delete stop words. Then, we extracted genes from the literature data using the HGNC database. We obtained impact factor and journal names from the omics online organization.

We analyzed the abbreviations of journal names for matching impact factor and literature data, because the majority of journal names are represented as an abbreviation. Next, we calculated LIF scores for each extracted gene based on two values: frequency and impact factor. Then, we ranked genes based on the LIF score. Figure 1 shows a flow of how the LIF score was calculated.

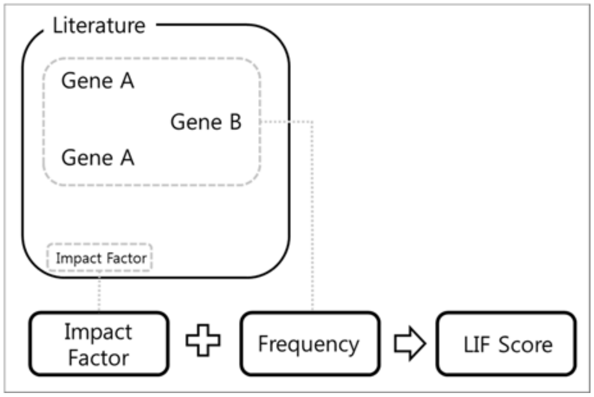
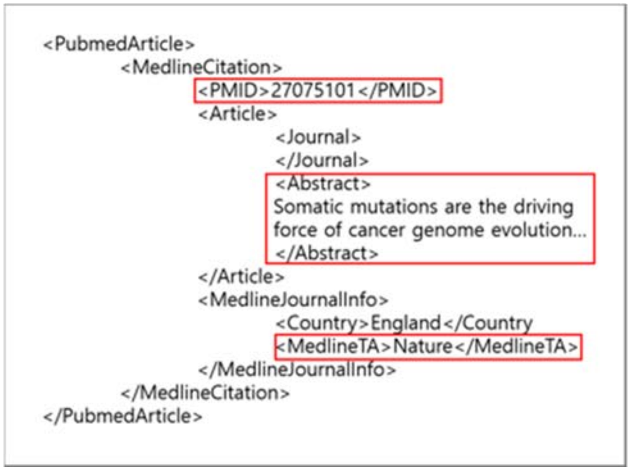


Figure 1. LIF score is calculated based on two values, frequency and impact factor.

Frequency indicates the number of appearances of a gene in the literature data. Impact factor indicates the impact factor of a journal referring to the gene.

3.1 Preprocessing

When obtaining lung cancer related literature data from PubMed, the journal data in xml format consists of diverse information including PMID, Medline TA, author, and abstract. As shown in Figure 2, we parsed the data to remove unnecessary data such as author and country. The PMID is used to check the meaning of gene symbols in the literature, because the gene symbol can be used as another meaning. We used Medline TA to assign impact factor values to the literature. Abstracts were used as literature data. After parsing the literature data, we removed useless stop words such as "a" and "the." The stop word lists were obtained from ranks nl [30].



```
<PubmedArticle>
        <MedlineCitation>
                <PMID>27075101</PMID>
                <Article>
                        <Journal>
                        </Journal>
                        <Abstract>
                        Somatic mutations are the driving
                        force of cancer genome evolution...
                        </Abstract>
                </Article>
                <MedlineJournalInfo>
                        <Country>England</Country
                        <MedlineTA>Nature</MedlineTA>
                </MedlineJournalInfo>
        </MedlineCitation>
</PubmedArticle>
```

Figure 2. Example for PubMed in xml format.

3.2 Extracting genes

To extract information on genes from the literature, we used approved gene symbols in the HGNC database (HUGO Gene Nomenclature Committee). Among the approved gene symbols, some short named symbols are used to denote other meanings in the literature. IV and PC are representative gene symbols that are frequently used to denote other meanings in the literature. For instance, “IV” is widely used as the roman character indicating a numerical value and “PC” is used to denote prostate cancer.

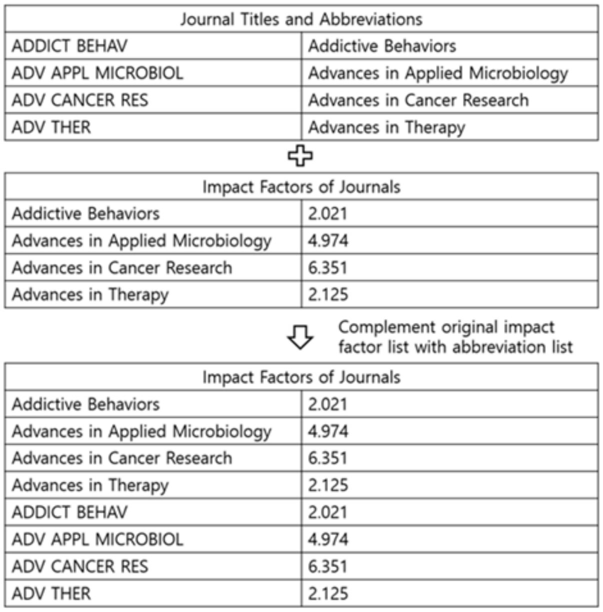
To address this problem, we used MeSH terms in PubMed data. We first extracted gene symbols with less than 3 letters from literature data. Then, we made a PMID list consisting of literatures that refer to the short gene symbols. Next, we searched all PubMed literature that contained a MeSH tag for the short gene symbol and extracted their PMID. Then, for each short gene symbol, we gathered two PMID lists. One of these contained PMIDs of literature that referred to gene symbols and the other contained PMIDs for literature that included MeSH tags for each short gene symbol. Using them, we calculated the SG score to identify short gene symbols in the literature. The SG score is calculated by the formula below.

$$\text{SG score (A)} = \frac{\textit{The number of journals that have MeSH A}}{\textit{The number of journals that refer to A}}$$

In the equation, SG score (A) indicates the SG score for gene symbol A. The MeSH A indicates the MeSH term for gene symbol A. The number of journals that have MeSH A are extracted from all PubMed literature. The number of journals referring to symbol A are extracted from lung cancer-related literature. After calculating the SG score for a few gene symbols with more than 2 letters, we compared the SG score of gene symbols between those that have less than 3 letters and those that have more than 2 letters, to determine a cutoff value. In this study, we determined 0.1 as a cutoff value, because the SG score of gene symbols with more than 2 letters is more than 0.1. For this reason, if the SG score is lower than 0.1, we determined that the symbol is more frequently used as another meaning in the literature data. Such gene symbols are excluded.

3.3 Impact factor matching

To utilize impact factor, we used Medline TA that indicates the name of the journal for the literature. After matching journal name and impact factor, we gathered the impact factors of 8411 journals from the omics online organization. Medline TA data uses many abbreviations for journal names. For this reason, we analyzed the abbreviations of journal names to use Medline TA data. We obtained abbreviations of journal names from Caltech library [3]. Then, we built pairs for journal name, abbreviations, and impact factor. Using the built pairs, we converted abbreviations into full name. For instance, impact factor of Advances in Cancer Research is 6.351 and its abbreviation is ADV CANCER RES. Then, ADV CANCER RES is converted as Advances in Cancer Research and is assigned an impact factor value of 6.351 for the journal. The process is presented in Figure 3.



Journal Titles and Abbreviations	
ADDICT BEHAV	Addictive Behaviors
ADV APPL MICROBIOL	Advances in Applied Microbiology
ADV CANCER RES	Advances in Cancer Research
ADV THER	Advances in Therapy

✚

Impact Factors of Journals	
Addictive Behaviors	2.021
Advances in Applied Microbiology	4.974
Advances in Cancer Research	6.351
Advances in Therapy	2.125

⇩ Complement original impact factor list with abbreviation list

Impact Factors of Journals	
Addictive Behaviors	2.021
Advances in Applied Microbiology	4.974
Advances in Cancer Research	6.351
Advances in Therapy	2.125
ADDICT BEHAV	2.021
ADV APPL MICROBIOL	4.974
ADV CANCER RES	6.351
ADV THER	2.125

Figure 3. Process of extracting impact factor.

As shown in Figure 3, we used two lists that included information on each journal. Using the two lists, we assigned impact factor values for abbreviations.

3.4 Scoring

We calculated the LIF score for scoring inferred genes. The LIF score consists of two values consisting of frequency and impact factor. In the case of frequency, the value indicates the number of appearances in the literature. We considered that a gene has a high frequency value in the literature then the gene is important and less important genes have a lower frequency in the literature. The LIF score used impact factor value as well as the frequency. We considered that the impact factor value can be used as a measure to score literature data.

The frequency values of genes are calculated by counting the number of appearances in the literature data. The impact factor value of a gene is assigned by the journal mentioning the gene.

LIF score is calculated as follows:

$$\text{LIF score(A)} = \sum_{i=1}^{n}\sum_{j=1}^{m}(1 + IF(i))$$

Here, *LIF score (A)* denotes the LIF score of gene A. n denotes the number of journals containing gene A, whereas m denotes the frequency of gene A in the i^{th} journal. *IF(i)* denotes the impact factor of the i^{th} journal that includes gene A. We added 1 as the default value to assign scores for journals with an impact factor of 0.

4. RESULTS

In this section, we describe experimental results for the LIF method. To implement our method, we used lung cancer-related literature as experimental data. We also compared this to other methods that infer disease–gene relationships.

4.1 Experimental data properties

The lung cancer-related literature data were obtained from PubMed. By using the biological journal lists that are obtained from the omics online organization, we built a list containing 8,411 journals and impact factors. Among the literature data (123,346), we used those that were matched by the journal list (70,059). To extract gene symbols from the literature, we used approved gene symbol lists (39816), which were obtained from the HGNC. Among them, we excluded 18 genes that had an SG score lower than 0.1 to infer more meaningful results. In our experimental results, we used answer sets to validate inferred genes. The answer set is constructed by integrating several databases, which includes KEGG disease [16], OMIM [27], GHR [9], LuGend [22], and IGDB.NSCLC [15]. These databases have information on disease–gene relationships.

Table 1. Answer Set

	Answer Set				
Database	KEGG disease	OMIM	GHR	LuGend	IGDB.NSCLC
The number of genes	18	38	13	73	16
Total	120				

Table 1 shows the number of genes that are involved in lung cancer for each database. In the table, the "Total" indicates the number of total genes that exclude overlapped genes.

4.2 Top N gene inferred by LIF

To validate our method, we confirmed the top 10 genes that are inferred by the LIF method, using an answer set. Table 2 shows the inferred top 10 genes.

Table 2. Top 10 gene inferred by LIF

Rank	Gene	Evidence
1	EGFR	OMIM, GHR, LuGend
2	ALK	KEGG, OMIM, GHR
3	KRAS	KEGG, OMIM, GHR, LuGend
4	EGF	Literature [19]
5	MET	OMIM
6	ERCC1	LuGend
7	TP53	KEGG, OMIM, LuGend
8	GSTM1	LuGend
9	CYP1A1	LuGend
10	PTEN	KEGG, GHR

In the Table 2, the "Evidence" indicates databases that include information on disease–gene relationships for each gene. As shown in Table 2, our method found 9 lung cancer-related genes among the top 10 genes. Furthermore, we confirmed that the "EGF" gene is involved in lung cancer by literature validation. As a result, our method found 9 lung cancer-related genes and 1 meaningful candidate gene among the top 10 inferred genes.

We compared our method to random values to confirm that the impact factor can be used as a scoring measure.

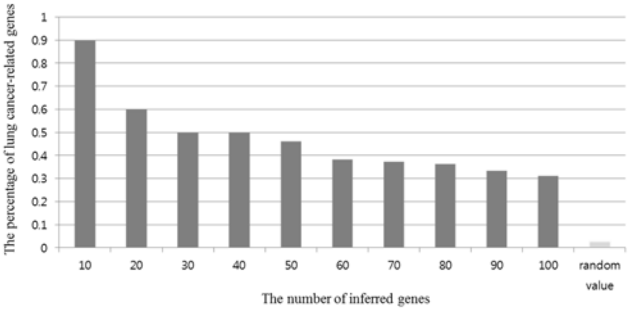


Figure 4. Comparison results with a random value for top n genes.

Figure 4 shows a comparison of results with random values. Figure 4 indicates the percentage of lung cancer-related genes among the inferred top N genes. The random values mean probability for randomly selected genes have a relationship with lung cancer. The random values are calculated by the formula below:

$$\text{Random value} = \frac{The\ number\ of\ lung\ cancer - related\ genes}{The\ number\ of\ all\ genes\ inferred\ by\ the\ LIG\ method}$$

In this experiment, we used answer sets to confirm the lung cancer-related genes. As shown in Figure 4, our results show that the LIF method has a higher percentage of lung-cancer related genes than a random value. In addition, the result means that the LIF method successfully ranked lung cancer-related gene, using impact factor.

4.3 Comparison with frequency-based method

To verify the effect of impact factor, we presented comparison results with frequency-based method. The method is considered to assess only the frequency of genes without impact factor in the literature.

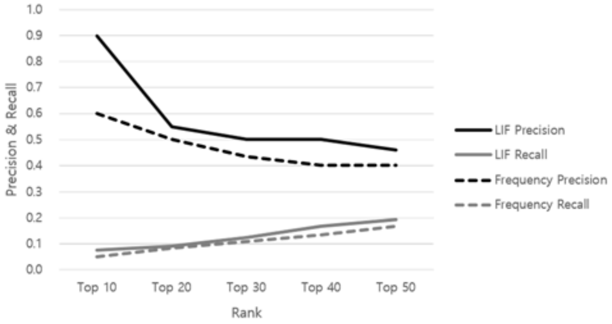


Figure 5. Comparison results with a frequency based method for top 50 genes

Figure 5 shows precision and recall for the LIF and frequency-based methods. As shown in Figure 5, the LIF method has higher precision and recall value than frequency-based methods for all sections. These results showed that impact factor can be used as a measure to extract meaningful relationships in the literature.

4.4 Comparison with existing methods

We compared our method to existing methods that infer disease–gene relationships. One of the methods is the PRINCE algorithm [37], and the other is the RWRHN [23] method. The PRINCE algorithm infers disease–gene relationships based on network analysis. To implement the algorithm, they used disease–disease similarity and protein–protein interaction data. The RWRHN method also infers disease–gene relationships by fusing multiple networks, which include PPI network, a phenotype similarity network, and known associations between disease and genes. In the case of the PRINCE algorithm, they offer the PRINCIPLE [10] tool to implement the PRINCE algorithm. For this reason, we used this tool to infer disease–gene relationships for the PRINCE algorithm. In the case of the RWRHN, we extracted experimental results for lung cancer from the literature.

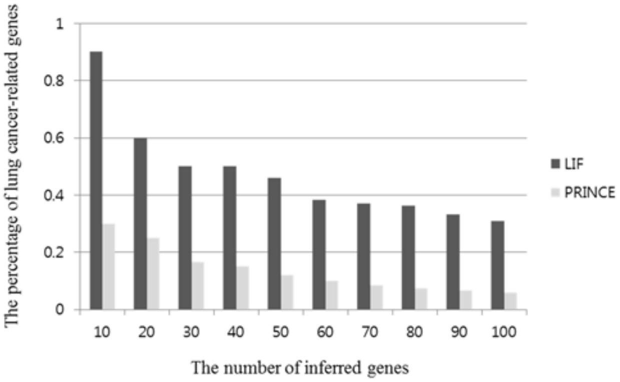


Figure 6. Comparison results with a PRINCE algorithm for top n genes.

To validate inferred genes, we confirmed the number of lung cancer-related genes among the inferred genes using the answer set. As shown in Figure 6, the LIF method inferred more lung cancer-related genes than the PRINCE algorithm for all cases.

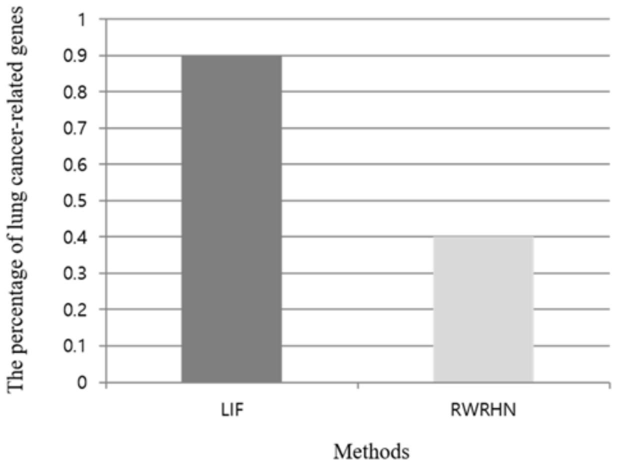


Figure 7. Comparison results with a RWRHN method for top 10 genes.

The experimental results for the RWRHN [23] are extracted from the literature for lung cancer. As shown in Figure 7, the LIF method inferred more lung cancer-related genes than the RWRHN method for the top 10 genes.

In summary, the LIF method found more lung cancer-related genes than random value, frequency-based method and two conventional methods. These results indicate that the proposed LIF is a useful method to infer disease–gene relationships and impact factor can be used as a measure to score literature data.

5. DISCUSSION

In this section, we describe genes that are not included in answer sets but have a high LIF score. Additionally, we investigate literature that has a high impact factor to confirm the role of impact factor in literature analysis.

5.1 Literature validation

In this study, we used answer sets to validate inferred genes. However, the answer set has limitations to validate all genes inferred by the LIF method. To consider this issue, we validated for top 11–20 genes, which are not included in the answer set using literature. Table 3 shows the top 11–20 genes inferred by the LIF method. Among them, 3 genes are validated by the answer set. For the other 6 genes, we found evidence that they are involved in lung cancer. However, we cannot find evidence for the CAD gene by answer sets and literature validation.

Table 3. Top 11-20 genes inferred by the LIF method

Rank	Gene	Evidence
11	STAT3	Literature [42]
12	TNF	LuGend
13	CD44	Literature [45]
14	PCNA	Literature [41]
15	HGF	Literature [11]
16	CXCR4	Literature [44]
17	CAD	None
18	FHIT	KEGG, LuGend
19	GRP	Literature [28]
20	MDM2	LuGend

We describe key sentences that are extracted from the literature for each gene below.

Yu et al. [42] described that “pSTAT3 (phosphorylated STAT3) overexpression is an important factor related to prognosis of NSCLC patients and indicates new anticancer strategies.”

Zhao et al. [45] described that “CD44v6 (CD44 variant exon 6) expression can be used as a novel prognostic marker in NSCLC cases.”

Weng et al. [41] described that “CD44v6 and PCNA play important roles in invasion and metastasis of NSCLC.”

Han et al. [11] described that “High HGF levels are significantly associated with resistance to gefitinib and can be used as a predictive marker for the differential outcome of gefitinib treatment in NSCLC irrespective of EGFR mutation status.”

Zhang et al. [44] described that “The present meta-analysis indicated that CXCR4 protein expression is associated with an increased risk and worse survival in NSCLC patients.”

Oremek et al. [28] described that “These results show that Pro-GRP (Pro-gastrin-releasing peptide) may be a potential tumor marker for small cell lung carcinoma.”

Based on these sentences, we confirmed that 6 genes are involved in lung cancer. As a result, we inferred 3 lung cancer-related genes and 6 lung cancer-related candidate genes among the 10 genes (in Top 11–20). These results showed that the LIF method is useful to infer candidate genes as well as lung cancer-related genes.

5.2 Analysis for literature with high impact factor

To investigate the role of impact factor, we selected the top 10 percent of literature as data by considering impact factor. The selected literature data have an impact factor more than 4.57. Using the top 10 percent of literature, we inferred the top 50 genes by the LIF method. Among them, we confirmed 11 genes that are not included in the Top 50 genes that are inferred by using all literature data.

Table 4. The genes (inferred using the top 10 percent of literature) which are not included in Top 50 genes (inferred using the all literature)

Gene	Before Rank	After Rank	Evidence
E2F1	57	24	Literature [14]
AR	51	32	Literature [25]
MYC	61	35	LuGend
MIF	87	39	Literature [6]
BRMS1	85	40	Literature [1]
ERBB2	76	42	OMIM, GHR, LuGend
IGF1R	60	44	Literature [46]
SOX2	54	46	Literature [38]
BAP1	59	48	Literature [7]
CXCR2	147	49	Literature [24]
CXCL12	74	50	Literature [34]

Table 4 shows 11 genes among the top 50 genes that are inferred from the top 10 percent of literature. They are not included in the top 50 genes that are inferred by using all literature. In Table 4, the "Before Rank" indicates the rank when using all the literature data. The "After Rank" indicates the rank when using the top 10 percent of literature data. Among the 11 genes, we can validate 2 genes using answer set. For the other 9 genes, we found literature that outlines evidence that they are involved in lung cancer. We described key sentences for the literature below.

Haung et al. [14] described that "During the progression of NSCLC (non-small cell lung cancer), E2F1 overexpression could produce more aggressive tumors with a high proliferation rate and chemo-resistance."

Mikkone et al. [25] described that "Our results include that adult lung is an AR target tissue and suggest that AR plays a role in lung cancer biology."

Fallica et al. [6] described that "highlighting a critical role of MIF in EC (endothelial cell) homeostasis with in the lung."

Balgkouranidou et al. [1] described that "Methylation of BRMS1 promoter in cfDNA (cell-free DNA) isolated from plasma of NSCLC patients provides important prognostic information and merits to be further evaluated as a circulating tumor biomarker."

Zhao et al. [46] described that "Our results suggested IGF1R positive expression as an unfavorable factor for DFS (disease free survival) in NSCLC patients, and IGF1R expression was associated with smoking status and tumor size."

Velcheti et al. [38] described that "SOX2 is an independent positive prognostic marker in NSCLC."

Fan et al. [7] described that "BAP1 may be a useful prognostic factor of NSCLC patients and potential target for anticancer drugs."

Massarelli et al. [24] described that "We concluded that the CXCR2 axis may be an important target in smoking-related lung adenocarcinoma."

Suzuki et al. [34] described that "Analysis for CXCL12 may provide novel opportunities for prognosis and therapy of resected NSCLCs."

In summary, we investigated 10 genes (in Top 11–20) based on answer set and literature validation. The results showed that the proposed method can infer various candidate genes as well as disease-related genes. Furthermore, we presented that our method can extract various results using same literature data and different impact factor. Through our experimental results, we showed that inferred genes that are extracted by high impact factor literature also are involved in lung cancer. These results demonstrated that the LIF method can infer various and useful knowledge in the literature.

6. CONCLUSION

In the present study, we attempted to infer disease–gene relationships using literature data and impact factor. To utilize impact factor, we built journal lists. Using the journal lists and literature data, we inferred meaningful disease–gene relationships and candidate genes, which could potentially be disease-related genes. Additionally, four distinct experiments demonstrated that the proposed method is useful to infer disease–gene relationships.

Furthermore, we investigated the role of impact factor in literature analysis. Our experimental results showed that we can infer various candidate genes by selecting literature based on impact factor. Furthermore, new information that has a few citations can be utilized by considering impact factor.

In this study, we only used lung cancer-related literature as experimental data. In further work, we will implement our method to several genetic diseases such as other types of cancer, Alzheimer's disease, and diabetes. Furthermore, we will present various approaches based on impact factor to extract useful knowledge from the literature.

7. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2015R1A2A1A05001845).

8. REFERENCES

[1] Balgkouranidou, I., Chimonidou, M., Milaki, G., Tsarouxa, E. G., Kakolyris, S., Welch, D. R., ... & Lianidou, E. S. (2014). Breast cancer metastasis suppressor-1 promoter methylation in cell-free DNA provides prognostic information in non-small cell lung cancer. British journal of cancer, 110(8), 2054-2062.

[2] Browne, F., Wang, H., & Zheng, H. (2016). Investigating the impact human protein–protein interaction networks have on disease-gene analysis.International Journal of Machine Learning and Cybernetics, 1-10.

[3] Caltech Library https://library.caltech.edu

[4] Chiang, J. H., & Yu, H. C. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics,19(11), 1417-1422.

[5] Corney, D. P., Buxton, B. F., Langdon, W. B., & Jones, D. T. (2004). BioRAT: extracting biological information from full-length papers. Bioinformatics, 20(17), 3206-3213.

[6] Fallica, J., Boyer, L., Kim, B., Serebreni, L., Varela, L., Hamdan, O., ... & Bucala, R. (2014). Macrophage Migration Inhibitory Factor Is a Novel Determinant of Cigarette Smoke–Induced Lung Damage. American journal of respiratory cell and molecular biology, 51(1), 94-103.

[7] Fan, L. H., Tang, L. N., Yue, L., Yang, Y., Gao, Z. L., & Shen, Z. (2012). BAP1 is a good prognostic factor in advanced non-small cell lung cancer. Clinical & Investigative Medicine, 35(4), 182-189.

[8] Gaizauskas, R., Demetriou, G., Artymiuk, P. J., & Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system.Bioinformatics, 19(1), 135-143.

[9] Genetics Home Reference <https://ghr.nlm.nih.gov/gene/GHR>

[10] Gottlieb, A., Magger, O., Berman, I., Ruppin, E., & Sharan, R. (2011). PRINCIPLE: a tool for associating genes with diseases via network propagation. Bioinformatics, 27(23), 3325-3326.

[11] Han, J. Y., Kim, J. Y., Lee, S. H., Yoo, N. J., & Choi, B. G. (2011). Association between plasma hepatocyte growth factor and gefitinib resistance in patients with advanced non-small cell lung cancer. Lung Cancer, 74(2), 293-299.

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

[13] Hou, W. J., & Kuo, B. Y. (2016). Discovery of Gene-disease Associations from Biomedical Texts. Computer Science and Information Technology, 4(1), 1-8.

[14] Huang, C. L., Liu, D., Nakano, J., Yokomise, H., Ueno, M., Kadota, K., & Wada, H. (2007). E2F1 Overexpression Correlates with Thymidylate Synthase and Survivin Gene Expressions and Tumor Proliferation in Non–Small-Cell Lung Cancer. Clinical Cancer Research, 13(23), 6938-6946.

[15] Integrated Genomic Database of Non-Small Cell Lung Carcinoma <igdb.nsclc.ibms.sinica.edu.tw>

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

[17] Kim, J. M., Sohn, H. Y., Yoon, S. Y., Oh, J. H., Yang, J. O., Kim, J. H., ... & Kim, J. G. (2005). Identification of gastric cancer–related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. Clinical Cancer Research, 11(2), 473-482.

[18] Le, D. H., & Dang, V. T. (2016). Ontology-based disease similarity network for disease gene prediction. Vietnam Journal of Computer Science, 1-9.

[19] Lemos-Gonzalez, Y., Rodriguez-Berrocal, F. J., Cordero, O. J., Gomez, C., & de la Cadena, M. P. (2007). Alteration of the serum levels of the epidermal growth factor receptor and its ligands in patients with non-small cell lung cancer and head and neck carcinoma. British journal of cancer, 96(10), 1569-1578.

[20] Li, D., Wang, L., Xue, Z., & Wong, S. T. (2016, February). When discriminative K-means meets Grassmann manifold: Disease gene identification via a general multi-view clustering method. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 364-367). IEEE.

[21] Liu, Z., & Hu, J. (2016). Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction. Methods,93, 119-127.

[22] Lung cancer gene database <www.bioinformatics.org/lugend/>

[23] Luo, J., & Liang, S. (2015). Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. Journal of biomedical informatics, 53, 229-236.

[24] Massarelli, E. (2013). CXCR2 EXPRESSION IN TUMOR CELLS IS A POOR PROGNOSTIC FACTOR AND PROMOTES INVASION AND METASTASIS IN LUNG ADENOCARCINOMA.

[25] Mikkonen, L., Pihlajamaa, P., Sahu, B., Zhang, F. P., & Jänne, O. A. (2010). Androgen receptor and androgen-dependent gene expression in lung. Molecular and cellular endocrinology, 317(1), 14-24.

[26] OMICS International <www.omicsonline.org>

[27] Online Mendelian Inheritance in Man <http://www.omim.org/>

[28] Oremek GM, Sapoutzis N. Pro-gastrin-releasing peptide (Pro-GRP), a tumor marker for small cell lung cancer. Anticancer Res. 2003 Mar-Apr;23(2A) 895-898. PMID: 12820319.

[29] PubMed: MEDLINE Retrieval on the World Wide Web www.ncbi.nlm.nih.gov/pubmed [April 2016].

[30] Ranks NL Webmaster Tools <www.ranks.nl>

[31] Srinivasan, P., Libbus, B., & Sehgal, A. K. (2004, May). Mining medline: Postulating a beneficial role for curcumin longa in retinal diseases. InWorkshop BioLINK, linking biological literature, ontologies and databases at HLT NAACL (pp. 33-40).

[32] Srinivasan, P. (2004). Text mining: generating hypotheses from MEDLINE.Journal of the American Society for Information Science and Technology,55(5), 396-413.

[33] Sun, H. (2016). Identification of key genes associated with gastric cancer based on DNA microarray data. Oncology Letters, 11(1), 525-530.

[34] Suzuki, M., Mohamed, S., Nakajima, T., Kubo, R., Tian, L., Fujiwara, T., ... & Yasufuku, K. (2008). Aberrant methylation of CXCL12 in non-small cell lung cancer is associated with an unfavorable prognosis. Corrigendum in/ijo/47/2/791. International journal of oncology, 33(1), 113-119.

[35] Swanson, D. R. (1986). Undiscovered public knowledge. The Library Quarterly, 103-118.

[36] Swanson, D. R. (1990). Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association, 78(1), 29.

[37] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., & Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol, 6(1), e1000641.

[38] Velcheti, V., Schalper, K., Yao, X., Cheng, H., Kocoglu, M., Dhodapkar, K., ... & Rimm, D. L. (2013). High SOX2 levels predict better outcome in non-small cell lung carcinomas. PLoS One, 8(4), e61427.

[39] Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., De Jong-van Den Berg, L. T., & Vos, R. (2000). Text-based discovery in biomedicine: the architecture of the DAD-

system. In Proceedings of the AMIA Symposium (p. 903). American Medical Informatics Association.

[40] Weeber, M., Vos, R., Klein, H., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide.Journal of the American Medical Informatics Association, 10(3), 252-259.

[41] Weng MX, Wu CH, Yang XP. [Expression and significance of E-cadherin, CD44v6, and proliferating cell nuclear antigen in non-small cell lung cancer]. Ai Zheng. 2008 Feb;27(2) 191-195. PMID: 18279620.

[42] Yu, Y., Zhao, Q., Wang, Z., & Liu, X. Y. (2015). Activated STAT3 correlates with prognosis of non-small cell lung cancer and indicates new anticancer strategies. Cancer chemotherapy and pharmacology, 75(5), 917-922.

[43] ZENG, X., LIAO, Y., & Zou, Q. (2016). Prediction and validation of disease genes using HeteSim Scores.

[44] Zhang, C., Li, J., Han, Y., & Jiang, J. (2015). A meta-analysis for CXCR4 as a prognostic marker and potential drug target in non-small cell lung cancer. Drug design, development and therapy, 9, 3267.

[45] Zhao, S., He, J. L., Qiu, Z. X., Chen, N. Y., Luo, Z., Chen, B. J., & Li, W. M. (2013). Prognostic value of CD44 variant exon 6 expression in non-small cell lung cancer: a meta-analysis. Asian Pacific journal of cancer prevention: APJCP, 15(16), 6761-6766.

[46] Zhao, S., Qiu, Z., He, J., Li, L., & Li, W. (2014). Insulin-like growth factor receptor 1 (IGF1R) expression and survival in non-small cell lung cancer patients: A meta-analysis. International journal of clinical and experimental pathology, 7(10), 6694.

[47] Zickenrott, S., Angarica, V. E., Upadhyaya, B. B., & Del Sol, A. (2016). Prediction of disease–gene–drug relationships following a differential network analysis. Cell Death & Disease, 7(1), e2040.