SSL: Inferring disease-related genes using Sentence Structure and Literature data

Jeongwoo Kim
Department of Computer Science, Yonsei University
Seoul, Korea
jwkim2014@naver.com

Jungrim Kim
Department of Computer Science, Yonsei University
Seoul, Korea
kimgogo02@yonsei.ac.kr

Won Gi Choi
Department of Computer Science, Yonsei University
Seoul, Korea
cwk1412@yonsei.ac.kr

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

Abstract— **Text mining is widely applied in biology to infer relationships between biological entities. In biology, disease–gene relationships are important to discover the cause of disease. Therefore, we propose a useful method called SSL, which infers disease-related genes, using sentence structure and literature data. Using sentence structure, the proposed method decreases the number of candidate disease-related genes and infers more meaningful disease-related genes than other comparable methods. Furthermore, our method extracts useful sentences that have information on the relationship between specific diseases and genes. By analyzing the structure of the sentences, we can obtain useful knowledge of disease-gene relationships. We applied our method to five diseases, including Alzheimer's disease, prostate cancer, gastric cancer, colorectal cancer, and lung cancer. For validation, we investigated the top 10 inferred genes for five diseases. Our method demonstrated up to 50% higher precision than existing methods, and showed 98% accuracy in inferring disease-related genes.**

Text mining; Disease-gene relationships; Alzheimer's disease; Prostate cancer; Gastric cancer; Colorectal cancer; Lung cancer

I. Introduction

Biomedical text data are generated from several biological experiments. These data include useful knowledge to describe complex biological relationships. We can obtain biomedical text easily from online databases such as PubMed [15], PMC [14], and OMIM [11]. Therefore, in biology, text analysis is widely performed to obtain biomedical knowledge from literature data. In particular, text mining is applied to infer biomedical relationships between biological entities such as disease-gene, disease-drug, and gene-drug, because the relationships are important to describe complex biological phenomenon. Furthermore, by analyzing various biomedical literature data, we can infer new relationships using existing relationships, which are included in biological experimental results.

The Swanson's ABC model [17, 18] is a representative biomedical text mining approach. This approach infers new relationships between biological entities, using existing relationships. Since text mining was shown to be a useful method for inferring biomedical relationships, a vast number of approaches have been presented [9, 12, 16].

However, many text-mining approaches infer too many disease-related genes to validate using biological experiments. This limitation is caused by a large amount of biomedical literature data. The other limitation is that previous studies infer only candidate relationships between biological entities from text-mining results. However, the biological literature includes information on relationships as well as relationships between biomedical entities. Therefore, extracting information for inferred relationships is important.

To consider these limitations, we propose a method to infer disease-related genes, using sentence structure and literature data. This study has two main goals: to decrease the number of inferred relationships, and to infer useful disease–gene relationships with information. To address these goals, we used sentence structure including auxiliary verb. Our assumptions are as follows:

- Biological experimental results cannot be described with 100 percent reliability.
- An auxiliary verb is widely used to describe biological experimental results.
- A sentence that has an auxiliary verb includes useful information

We considered that use of an auxiliary verb is key to describing biological experimental results in the literature. Therefore, we utilized auxiliary verbs to achieve our goals.

The main contributions of this work include:

- A decrease in the number of inferred relationships.
- Inferring of meaningful disease–gene relationships.
- Extraction of useful sentences to support relationships.

* Corresponding author.
Tel: +82 2 2123 5714; fax: +82 2 365 2579

978-1-5090-3015-6/17/$31.00 ©2017 IEEE

In this study, we propose a novel method to infer disease-related genes, using sentence structure and literature data. To implement our method, we defined sentence structures, which contain an auxiliary verb and gene symbol. By analyzing the part-of-speech of sentences, we identified sentences with defined sentence structure. Based on the sentences, we inferred meaningful disease-related genes and candidate genes. We validated our experimental results by applying an answer set and sentence validation. Furthermore, we presented supporting sentences that included information on the relationships between diseases and genes as well as disease-related genes.

The rest of the paper is organized as follows. Section 2 introduces related studies. The proposed method is described in Section 3. Section 4 describes the experimental results and discussion for this study. The conclusions and further studies are included in Section 5.

II. Related studies

Several text-mining approaches [1, 2, 8] have been developed in the biomedical field. Named entity recognition, text classification, terminology extraction, and relationship extraction are representative biomedical text mining approaches. Among them, this study addresses the relationship extraction field.

Jung et al. [6] presented a literature search tool for extraction of disease-associated genes. To implement this tool, they applied a rule-based text-mining algorithm with keyword matching to extract target diseases, genes, significant results, and the type of study described by the article. Pletscher-Frankild et al. [13] presented a system for extracting disease-gene associations from biomedical abstracts. To implement their system, they used a dictionary-based tagger from a named entity recognition and scoring scheme that takes into account co-occurrences. They also developed the DISEASES resource, which integrates the results from text mining with manually curated disease-gene associations, cancer mutation data, and genome-wide association studies from existing databases. Fang et al. [3] provided a database called MeInfoText. This database presents comprehensive association information about gene methylation and cancer based on association mining from literature data. The MeInfoText also presented a set of genes, which may contribute to the development of cancer by aberrant methylation. Tiffin et al. [19] attempted to extract candidate disease genes, using expression profiles. They used the eVOC anatomical ontology to integrate text mining of biomedical literature and data-mining of human gene expression data. Using the proposed approach, they successfully prioritized candidate genes according to their expression in disease-affected tissues.

Several studies based on identifying disease-gene relationships have also been presented. Le et al. [24] attempted to predict disease-related genes using an ontology-based disease similarity network. They constructed the disease similarity network by considering human phenotype ontology and semantic similarity measures. Using the disease similarity network, they inferred 100 Alzheimer's disease related genes. Among them, they found 19 candidate genes. Gottlieb et al. [4] presented the PRINCIPLE tool, which analyzes and visualizes disease specific gene networks based on the PRINCE [22] algorithm. The PRINCE algorithm was developed to infer disease–gene relationships by network analysis. To implement their algorithm, they used disease–disease similarity and protein–protein interaction data. Luo et al. [10] constructed a reliable heterogeneous network by fusing multiple networks including the PPI network, phenotype similarity network, and known associations between diseases and genes. After constructing the network, they analyzed it using RWRHN, which is devised based on a random walk-based algorithm. The proposed approach predicted novel causal genes for 16 diseases.

III. Methods

In this section, we describe a proposed method for inferring disease-related genes using sentence structure and literature data. Figure 1 outlines the proposed method.

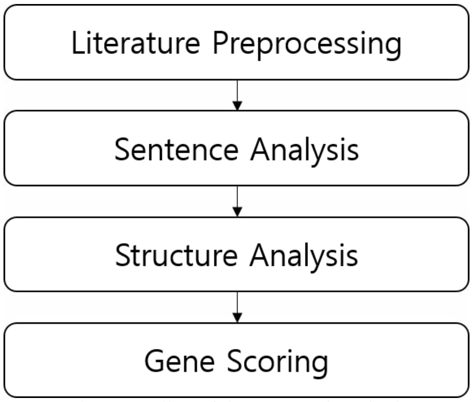


Fig. 1. Outline of the proposed method.

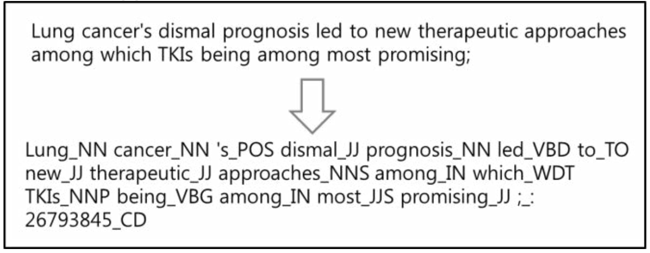
Our method has four steps. First, we obtained literature data, which are involved in specific diseases from PubMed using MeSH terms. After processing the literature data, we identified sentences that included auxiliary verbs and genes. In the next step, we analyzed sentence structure by considering the location of auxiliary verbs and gene symbols. Finally, we calculated the scores for each gene based on structure sentences extracted in the previous step.

A. Literature Data Preprocessing

We gathered data from abstracts describing five diseases, which include Alzheimer's disease, prostate cancer, gastric cancer, colorectal cancer, and lung cancer from PubMed. The abstracts included several sections such as author, data, and journal name. Among them, we used the abstract text and PMID in our analysis. The abstract section provides abstract text of the research study, and the PMID section provides the PubMed ID number to access the research article.

B. Sentence Analysis

After obtaining the literature data, we categorized sentences according to parts-of-speech tagging using a POS tagger [20, 21]. Fig. 2 shows a POS-tagging example using the POS tagger.



Lung cancer's dismal prognosis led to new therapeutic approaches among which TKIs being among most promising;

Lung_NN cancer_NN 's_POS dismal_JJ prognosis_NN led_VBD to_TO new_JJ therapeutic_JJ approaches_NNS among_IN which_WDT TKIs_NNP being_VBG among_IN most_JJS promising_JJ ;_: 26793845_CD

Fig. 2. Example for POS tagger.

The POS tagger analyzes a sentence word by word. By using tagging results, we can identify parts-of-speech for each word. The tagging results are used to extract sentences that have auxiliary verbs.

C. Structure analysis

First, we converted the paragraphs of the abstracts into sentence units. Using the tagging results, we extract sentences that include an auxiliary verb. Among the several auxiliary verbs, we used “may” and “might”, because they are the most commonly used verbs to present biological experiment results. In the next step, we filtered the extracting sentence by identifying gene symbols in the sentence. Sentences with gene symbols were selected. The gene symbol was obtained from the HGNC database. In this study, we used approved gene symbols. Among them, we excluded three gene symbols, specifically “T”, “PC”, and “GC”, because they are commonly used to denote other meanings than gene symbol in the literature data. The “T” is widely used to describe T cell, and “PC” and “GC” are used as abbreviations of prostate and gastric cancers, respectively.

After filtering sentences, we investigated sentence structure as shown in Fig. 3.

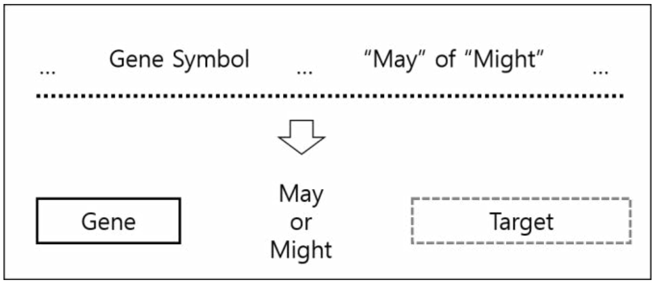


Fig. 3. sentence structure

Fig. 3 shows the sentence structure used in this method. We considered the location of the gene symbol and auxiliary verb in each sentence. We presumed that the sentence structure represents relationships between the gene and its target. Therefore, we used the sentence structure to infer disease-related genes.

D. Gene Scoring

To score genes, we used frequency as a measure. The frequency is a conventional approach to scoring in text mining. If an interesting term appears several times in sentences, the term is considered important. We calculate the frequency of genes that appear in structured sentences generated in the previous step. Using the frequency, we infer disease-related genes with ranking.

E. Validation

To validate our experimental results, we gathered an answer set from several databases such as OMIM, GHR, and KEGG disease. These databases provide known disease-related gene data. The answer set indicates genes that are already known to be related to disease. By using the answer set, we can calculate the precision of our experimental results. Precision is calculated as follows:

$$\text{Precision} = \frac{The\ number\ of\ known\ genes}{The\ number\ of\ inferred\ genes} \quad (1)$$

As shown in the equation, we calculate precision by considering the number of known genes among the inferred genes.

IV. Results and Discussion

In this section, we describe experimental results and discussions for our study. We also present comparison experimental results by comparing previous studies that infer disease-related genes.

A. Expermental Data

In this experiment, we obtained literature data for five diseases from PubMed. These data are presented in Table 1.

Table 1. Literature data properties

	Literature	**Sentence**
Alzheimer’s disease	16,639	85,072
Prostate cancer	218,99	104,575
Gastric cancer	18,034	81,546
Colorectal cancer	47,541	222,931
Lung cancer	36,180	176,118

In Table 1, “Literature” indicates the number of literature data for each disease, and “Sentence” indicates the number of sentences in the literature for each disease.

To validate the experimental results, we used answer sets. The answer set is shown in Table 2.

Table 2. Answer set

	OMIM	GHR	KEGG	Total
Alzheimer’s disease	9	26	4	29
Prostate cancer	18	38	13	59
Gastric cancer	8	7	16	24
Colorectal cancer	26	28	14	50
Lung cancer	16	31	17	43

Table 2 shows the number of known genes included in the databases. The “Total” indicates the number of answer sets for each disease, and the value of “Total” is calculated by excluding common genes included in several databases.

B. Auxiliary verb analysis

We analyzed the distribution of auxiliary verbs and the number of sentences with the proposed sentence structure.

Table 3. Distribution of auxiliary verbs

	may/ might	will/ would	can/ could	should/ must
Alzheimer’s disease	6,033 (55.11%)	1,188 (10.85%)	3,286 (30.01%)	441 (4.03%)
Prostate cancer	6,078 (50.95%)	1,176 (9.86%)	4,199 (35.20%)	476 (3.99%)
Gastric cancer	4,962 (58.62%)	461 (5.45%)	2,743 (32.41%)	298 (3.52%)
Colorectal cancer	12,145 (49.00%)	2,278 (9.19%)	8,723 (35.20%)	1,638 (6.61%)
Lung cancer	9,676 (49.84%)	1,759 (9.06%)	7,140 (36.78%)	840 (4.33%)

Table 3 indicates the number of sentences that include an auxiliary verb. As shown in Table 3, we confirmed that “may” and “might” are the most common auxiliary verbs used to describe biological research.

C. Comparison of the number of inferred genes

One of our goals was to decrease the number of inferred disease-related genes. Table 4 and Fig. 4 demonstrate the number of structure sentences used in this experiment and the number of inferred genes, respectively.

Table 4. The number of sentences

	All sentence	All gene	SSL sentence	SSL gene
Alzheimer’s disease	85,072	1,269	6,033	313
Prostate cancer	104,575	2,399	6,078	577
Gastric cancer	81,546	2,249	4,962	653
Colorectal cancer	222,931	3,587	12,145	981
Lung cancer	176,118	3,328	9,676	883

Table 4 shows the number of “all sentences” and SSL sentences. The “All sentence” indicates the number of original sentences included in the literature. The “SSL sentence” indicates the number of sentences that are filtered by the proposed SSL method. The “ALL gene” and “SSL gene” indicate the number of inferred genes for each sentence. As shown in Table 4, we confirmed that a large amount of sentences are filtered by the SSL method. Therefore, the number of inferred genes also decreased. Fig. 4 presents the number of inferred genes.

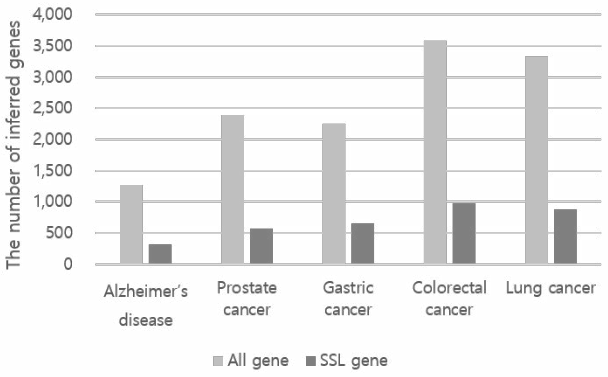


Fig. 4. The number of inferred genes

The results shown in Fig. 4 were generated by extracting the genes for each sentence, including all and SSL sentences. In the case of all sentences, we confirmed that a lot of genes are inferred by text-mining results. However, the SSL method decreases the number of inferred genes by considering sentence structure for five diseases (Fig. 4). The validation for inferred genes is presented in the following sections.

D. Inferred Top 10 genes

To verify our experimental results, we investigated the top 10 inferred genes. We also used an answer set to validate relationships between inferred genes and disease. Table 5 shows the top 10 genes inferred by the proposed method.

Table 5. Inferred Top 10 genes

Rank	Alzheimer's disease	Prostate cancer	Gastric cancer	Colorectal cancer	Lung cancer
1	APP	AR	CDH1	APC	EGFR
2	APOE	ERG	GCA	KRAS	KRAS
3	BDNF	PTEN	RUNX3	BRAF	ALK
4	BACE1	TMPRSS2	APC	EGFR	XRCC1
5	IDE	BRCA1	XRCC1	FAP	ERCC1
6	PSEN1	EGFR	TFF1	MLH1	FHIT
7	SORL1	VDR	EGFR	MTHFR	GSTM1
8	ACE	GSTP1	MTHFR	PTEN	MET
9	A2M	SRD5A2	GSTM1	DCC	CYP1A1
10	GAB2	BRCA2	ERCC1	PIK3CA	XPC

In Table 5, the gray gene symbol indicates known genes validated by the answer set. In the case of prostate cancer, we inferred eight known genes and two candidate genes. However, several inferred genes were not validated by the answer set (Table 5). The size of the answer set was too small to cover the inferred genes. Therefore, we conducted sentence validation for inferred candidate genes. The sentence validation is described in section F.

E. Comparison of the experimental results

We compared our method to other methods that infer disease-gene relationships. One of the methods is the PRINCE algorithm [22], and the other is RWRHN [10]. For the PRINCE algorithm, we implemented the method using the PRINCIPLE tool [4]. For "RWRHN", we extracted the top 10 genes inferred by RWRHN, from results in the paper. To validate the genes inferred by each method, we used the answer set.

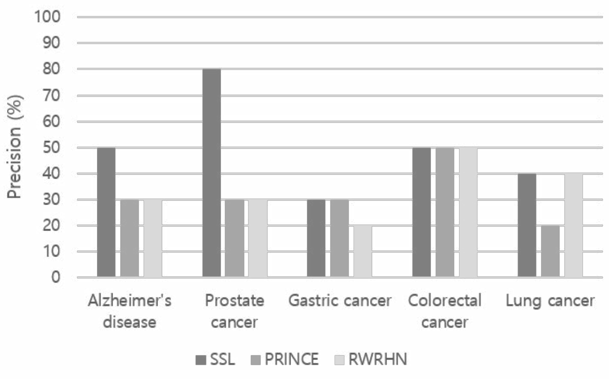


Fig. 5. comparison results

Figure 5 shows the precision of the inferred top 10 genes for five diseases. The y-axis indicates the precision for the inferred top 10 genes, and the x-axis indicates the diseases. Our method identified more known genes than comparable methods for two diseases (Fig. 5). For other diseases, the SSL demonstrated the highest precision. In the case of prostate cancer, the proposed method demonstrated up to 50% higher precision than existing methods. This demonstrates that the SSL is a useful method for inferring disease-related genes.

F. Extracted Sentence validation

To extract information for inferred candidate genes, we investigated structure sentences. The structure sentences are sentences that include an auxiliary verb and gene symbol. They are extracted from the structure analysis step in our method. Using the structure sentences, we can obtain specific knowledge of candidate genes. Table 6 shows candidate genes and structure sentences.

Table 6. Sentence validation for inferred candidate genes

<table>
<tr><th>Gene</th><th>Disease</th><th>PMID</th></tr>
<tr><th colspan="3">Structure sentence</th></tr>
<tr><td>BDNF</td><td>Alzheimer's disease</td><td>25364831</td></tr>
<tr><td colspan="3">Our work suggested that peripheral BDNF promoter methylation might be a diagnostic marker of AD risk, although its underlying function remains to be elaborated in the future.</td></tr>
<tr><td>BACE1</td><td>Alzheimer's disease</td><td>22709416</td></tr>
<tr><td colspan="3">Dysregulation of the intracellular trafficking of BACE1 may affect Aβ generation, contributing to AD pathology.</td></tr>
<tr><td>IDE</td><td>Alzheimer's disease</td><td>21873424</td></tr>
<tr><td colspan="3">Targeting the regulation of IDE may be a promising therapeutic approach to sporadic AD.</td></tr>
<tr><td>ACE</td><td>Alzheimer's disease</td><td>17401152</td></tr>
</table>

Several studies have shown that a common insertion (I)/deletion (D) polymorphism of angiotensin-converting enzyme (ACE) gene may confer an increased risk of late-onset Alzheimer disease (LOAD).		
GAB2	Alzheimer's disease	24161894
The current meta-analysis further supports previous findings that the GAB2 gene may be associated with SAD risk.		
EGFR	Prostate cancer	20736346
Therefore, inhibition of EGFR may effectively suppress prostate cancer growth and progression.		
VDR	Prostate cancer	19255064
Results from the most comprehensive evaluation of serum vitamin D and its related genes to date suggest that tag SNPS in the 3' UTR of VDR may be associated with risk of prostate cancer in men with low vitamin D status.		
GSTT1	Gastric cancer	19960261
This meta-analysis suggests that GSTT1 gene polymorphism may be not associated with increased gastric cancer risk among Europeans, Americans, and East Asians.		
RUNX3	Gastric cancer	16367921
The detection of hypermethylation at multiple regions within the RUNX3 CpG island may be useful in the diagnosis and risk assessment of gastric cancer.		
XRCC1	Gastric cancer	11058877
These findings support the hypothesis that these 2 XRCC1 variants may contribute to the risk of developing gastric cancer, particularly gastric cardia cancer.		
TFF1	Gastric cancer	23329884
Reduced expression of TFF1 and increased expression of TFF3 may play a role in the carcinogenesis of gastric cancer.		
MTHFR	Gastric cancer	15643524
These findings suggest that the MTHFR common variants and their haplotypes may play a role in the etiology of gastric cancer, particularly gastric cardia adenocarcinoma.		
GSTM1	Gastric cancer	10667466
The homozygous deletions or null genotypes of GSTT1 (theta class) and GSTM1 (mu class) genes may be associated with an increased risk of cancer.		
ERCC1	Gastric cancer	24793015
In conclusion, we found that ERCC1 rs11615 and XPF rs2276465 may substantially contribute to the future design of individualized cancer treatment in gastric cancer patients.		
BRAF	Colorectal cancer	21742054
BRAF mutations also may play a role in treatment decisions.		
EGFR	Colorectal cancer	21559018
EGFR promoter hypermethylation, after confirmation in larger data set, may represent a valuable asset in further studies investigating EGFR as a therapeutic target in colorectal cancer.		
FAP	Colorectal cancer	None
MTHFR	Colorectal cancer	20726304
The MTHFR gene polymorphism may influence the risk of developing sporadic CRC.		
PTEN	Colorectal cancer	19724853
PTEN expression may be a good marker for the prognosis of colorectal carcinoma.		
XRCC1	Lung cancer	26097609
In conclusion, we found that XRCC1 Arg194Trp polymorphism may be associated with NSCLC risk, especially in smokers.		
ERCC1	Lung cancer	25375151
Our analysis suggested ERCC1 expression may be a prognostic factor in SCLC patients receiving platinum-based chemotherapy, especially for LS-SCLC		
GSTM1	Lung cancer	19669596
Our results suggest that GSTM1 and GSTT1 polymorphisms may play a role in the development of lung cancer for some histological subtypes and modifies the risk of smoking-related lung cancer.		
MET	Lung cancer	25416047
MET has been suggested to have an intimate relationship with small cell lung cancer (SCLC) and might be a promising therapeutic target.		
CYP1A1	Lung cancer	24964616
CYP1A1 Ile462Val polymorphisms may contribute to		

the decreased susceptibility of small cell lung cancer.		
XPC	Lung cancer	22166526
Polymorphisms of the XPC gene, Lys939Gln, may be a predictive marker of treatment response for advanced NSCLC patients in stage III.		

In Table 6, the "PMID" is the PubMed identification number. By using the PMID, we can access literature data for the sentence. "Structure sentences" indicate sentences that are extracted using the structure analysis step in our method. "Disease" indicates the disease used in our experiments. As shown in Table 6, our method provides information for inferred genes. By analyzing the structure sentences, we can obtain meaningful biological knowledge. We described 25 candidate genes, which were not validated by the answer set in Table 5. Through sentence validation, we found supporting sentences, which demonstrate that they are involved in the disease for the inferred candidate genes. In the case of "FAP", we cannot find any evidence for the relationship between FAP and colorectal cancer. By investigating structure sentences for colorectal cancer, we confirmed that FAP is used as familial adenomatous polyposis, which is not a gene symbol. However, we found 24 supporting sentences to confirm the disease-gene relationships among the 25 candidate genes.

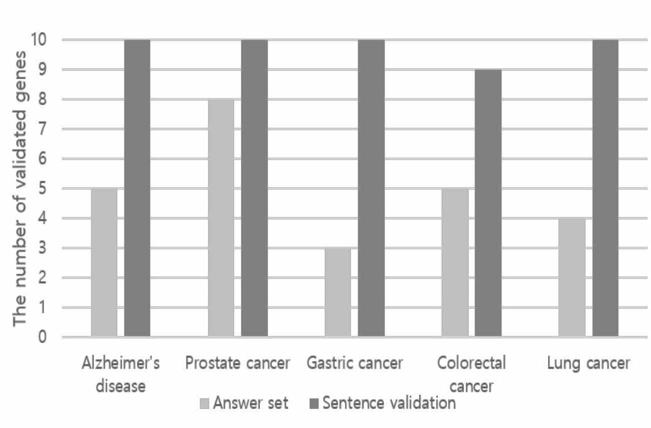


Fig. 6. The number of validated genes

Figure 6 indicates the number of validated genes among the inferred top 10 genes. As shown in Fig. 6, we validated inferred candidate genes, which were not validated by the answer set, as well as known genes. The proposed method identified 49 genes that are involved in disease among the inferred 50 genes. Therefore, our results showed 98% accuracy in inferring disease-related genes.

Using structure sentences, we can find specific information for relationships between inferred genes and diseases as well as disease-related genes. Therefore, our method can provide information for inferred candidate genes that are not validated by the answer set. Furthermore, our method is better at inferring disease-related genes than existing methods. These results demonstrate that the SSL is a useful method to infer candidate genes with specific information.

V. Conclusions

In the present study, we attempted to infer disease-related genes, using auxiliary verb and literature data. Among the several auxiliary verbs, we predominantly used "may" and "might" in our analysis. We also used sentence structure by considering the location of auxiliary verbs and genes. We applied our method to five genetic diseases, namely Alzheimer's disease, prostate cancer, gastric cancer, colorectal cancer, and lung cancer. We validated the proposed method by presenting the top 10 inferred genes. We also presented comparison results by comparing existing methods, which infer disease-related genes.

Our experimental results showed that the proposed method is more useful than comparable methods. Furthermore, our method can extract useful structure sentences, which provide further information about the relationships between the disease and gene. Using the sentences, the proposed method can find specific knowledge for inferred candidate genes not validated by the answer set.

Of the 50 inferred genes, we identified 25 known genes and 24 meaningful candidate genes. We also provide supporting sentences by the structure sentence method. Our method demonstrated up to 50% higher precision than existing methods, and showed 98% accuracy in inferring disease-related genes. We also decreased the number of inferred genes by considering sentence structure.

Future studies will evaluate other auxiliary verbs, other than "may" or "might", by analyzing features for auxiliary verbs. Future studies will also consider other parts-of-speech as opposed to auxiliary verbs to discover more meaningful sentences. Furthermore, we will present several validations for experimental results, which include the top 10 inferred genes.

Acknowledgment

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

References

[1] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining." Brief. Bioinform, vol. 6, pp. 57-71, March 2005.

[2] K. B. Cohen and L Hunter, "Getting Started in Text Mining," PLoS. Comput. Biol, vol. 4, pp. e20, Jan 2008.

[3] Y. C. Fang, H. C. Haung, and H. F Juan, "MeInfoText: associated gene methylation and cancer information from text mining," BMC Bioinformatics, vol. 9, Jan 2008.

[4] A. Gottlieb, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. "PRINCIPLE: a tool for associating genes with diseases via network propagation," Bioinformatics, vol. 27, pp. 3325-3326, December 2011.

[5] 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>.

[6] J. Y. Jung, T. F. Deluca, T. H. Nelson, and D. P. Wall, “A literature search tool for intelligent extraction of disease-associated genes,” J. Am. Med. Inform. Assoc, vol. 21, pp. 399-405, May-June 2014.

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

[8] M. Krallinger, R. A. A. Erhardt, and A Valencia, “Text-mining approaches in molecular biology and biomedicine,” Drug Discov. Today, vol. 10, pp. 439-445, March 2005.

[9] S. J. Lee, J. Choi, K. Park, M. Song, and D. Lee, “Discovering context-specific relationships from biological literature by using multi-level context terms,” BMC Med. Inform. Dec. Mak, vol. 12, Suppl. 1, April 2012.

[10] J. Luo and S. Liang, “Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data,” J. Biomed. Inform, vol. 53, pp. 229-236, February 2015.

[11] Online Mendelian Inheritance in Man http://www.omim.org/[April 2016]

[12] I. Petric, T. Urbancic, B. Cestnik, and M Macedoni-Luksic, “Literature mining method RaJoLink for uncovering relations between biomedical concepts,” J. Biomed. Inform, vol. 42, pp. 219-227, April 2009.

[13] S. Pletcher-Frankild, A. Palleja, L. Tsafou, J. X. Binder, and L. J. Jensen, “DISEASES: Text mining and data integration of disease-gene associations,” Methods, vol. 74, pp. 83-89, March 2015.

[14] PubMed Central http://www.ncbi.nlm.nih.gov/pmc/

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

[16] C. Senger, B. A. Gruning, A. Erxleben, K. Doring, H. Patel, S. Flemming, I. Merfort, and S Gunther, “Mining and evaluation of molecular relationships in literature,” Bioinformatics, vol. 28, pp. 709-714, March 2012.

[17] D. R. Swanson, “Undiscovered public knowledge,” Libr. Quart, vol. 56, pp. 103-118, April 1986.

[18] D. R. Swanson, “Medical literature as a potential source of new knowledge,” Bull. Med. Libr. Assoc, vol. 78, pp. 29-37, January 1990.

[19] N. Tiffin, J. F. Kelsom, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide, “Integration of text- and data-mining using ontologies successfully,” Nucleic Acids Res, vol. 33, pp. 1544-1552, March 2005.

[20] K. Toutanova and C. Manning, “Enriching the knowledge sources used in a maximum entropy part-of-speech tagger,” in: Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora (EMNLP/VLC), 2000, pp. 63–70.

[21] K. Toutanova, D. Klein, C. Manning, Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in: Proceedings of the HLT-NAACL, 2003, pp. 252–259.

[22] O. Vannu, O. Magger, E. Ruppin, and R. Shlomi, “Associating genes and protein complexes with disease via network propagation,” PLoS Comput. Biol, vol. 6, pp. e1000641, January 2010.

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

[24] L. Duc-Hau and D. Vu-Tung, "Ontology-based disease similarity network for disease gene prediction." Vietnam Journal of Computer Science, vol. 3, pp. 197-205, August 2016.