Journal of Biomedical Informatics 54 (2015) 270–282

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

LGscore: A method to identify disease-related genes using biological literature and Google data

Jeongwoo Kim [a], Hyunjin Kim [a], Youngmi Yoon [b], Sanghyun Park [a,*]

[a] Department of Computer Science, Yonsei University, 50 Yonsei-ro, Sinchon-dong, Seodamun-gu, Seoul 120-749, South Korea
[b] Department of Computer Engineering, Gachon University, 1342 Sengnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do, South Korea

ARTICLE INFO

Article history:
Received 3 June 2014
Accepted 5 January 2015
Available online 21 January 2015

Keywords:
Text-mining
Data mining
Gene
Disease
Google

ABSTRACT

Since the genome project in 1990s, a number of studies associated with genes have been conducted and researchers have confirmed that genes are involved in disease. For this reason, the identification of the relationships between diseases and genes is important in biology. We propose a method called LGscore, which identifies disease-related genes using Google data and literature data. To implement this method, first, we construct a disease-related gene network using text-mining results. We then extract gene–gene interactions based on co-occurrences in abstract data obtained from PubMed, and calculate the weights of edges in the gene network by means of *Z*-scoring. The weights contain two values: the frequency and the Google search results. The frequency value is extracted from literature data, and the Google search result is obtained using Google. We assign a score to each gene through a network analysis. We assume that genes with a large number of links and numerous Google search results and frequency values are more likely to be involved in disease. For validation, we investigated the top 20 inferred genes for five different diseases using answer sets. The answer sets comprised six databases that contain information on disease–gene relationships. We identified a significant number of disease-related genes as well as candidate genes for Alzheimer's disease, diabetes, colon cancer, lung cancer, and prostate cancer. Our method was up to 40% more accurate than existing methods.

1. Introduction

Since the human genome was sequenced, a large number of gene-based studies have been performed, and vast amounts of gene data have been generated. These data are stored in databases such as the Online Mendelian Inheritance in Man (OMIM) database [19]. Extracting hidden information from these databases offers new research opportunities and challenges. One of the best known tools with which to extract knowledge is text-mining.

In the biomedical area, text-mining has been used to identify biological entities such as protein and gene names in the literature. Furthermore, text-mining can reveal novel relationships among biological entities. Text-mining can provide opportunities to reduce the time and effort needed to extract relationships between biological entities from a large amount of publications. Interest in text-mining is increasing due to the increasing number of electronic publications stored in databases such as PubMed [26]. Furthermore, Swanson's ABC model [1,2] makes text-mining a feasible approach.

Network analysis also plays an important role in biological research. Gene networks, which describe gene–gene interactions, and protein networks, which describe protein–protein interactions, allow the visual relationships among biological entities in complex biological systems to be presented in a simple, clear manner. Network analysis also provides an opportunity to analyze which relationships are meaningful among various candidates. A network analysis provides several analysis measures as well, such as degree centrality, closeness centrality, and betweenness centrality to identify novel relationships among the large numbers of relationships in the network.

Several techniques have been developed to extract hidden information using text-mining and network analysis. Li et al. [16] tried to integrate both literature and microarray gene-expression data. They constructed a gene network using the co-occurrence-based text-mining method and then refined the network using microarray data. Their results showed that the network by Li et al. is more reliable than the co-occurrence-based network. Gonzalez et al. [10] presented a method which uses literature data and interactions. They extracted an initial set of genes and proteins from the literature and then integrated the set with interactions from the curated databases of BIND and DIP. They then constructed a network based on these data, ranking the genes and gene products using a combina-

* Corresponding author. Fax: +82 2 365 2579.
E-mail addresses: jwkim2013@cs.yonsei.ac.kr (J. Kim), chriskim@cs.yonsei.ac.kr (H. Kim), ymyoon0719@gmail.com (Y. Yoon), sanghyun@cs.yonsei.ac.kr (S. Park).

http://dx.doi.org/10.1016/j.jbi.2015.01.003
1532-0464/

tion of the two scores. One of the scores measures the strength of the relationship with the initial set of genes, and the other score measures the importance of each gene in maintaining the connectivity of the network. Their method showed high accuracy levels for atherosclerosis. Chen et al. [4] presented a method that constructs a gene-regulatory network using micro-array data and literature-based knowledge. They first extracted gene–gene relationships from the literature and then assigned random weights to the relationships. Through this process, they generated 2000 chromosomes. Subsequently, they used a genetic algorithm to optimize the strength of the interactions using a microarray and an artificial neural network fitness function. Their results demonstrated the advantage of combining gene interactions extracted from the literature with microarray analysis in generating contribution-weighted gene-regulatory networks. Ozgur et al. [20] determined the relationships between prostate cancer and genes using text-mining and network analysis. They constructed a disease-related gene network using the biomedical literature and seed genes, and extracted disease-related genes based on an analysis of the gene network using different scoring methods. A seed gene is a gene known to be involved in a disease. Although this approach by Ozgur et al. inferred prostate cancer-related genes successfully, it cannot be used to determine the relationships between genes and diseases for which there are no seed genes. Furthermore, the experimental results are influenced by the choice of the seed gene. The PRINCE algorithm [24,25] is another method that was developed to infer relationships among genes and diseases using network analysis based on disease–disease similarity and protein–protein interaction data. The PRINCE algorithm can be applied to all diseases; however, it is less accurate than the method by Ozgur et al.

In this paper, we propose a novel approach to infer disease-related genes based on Google data and literature data. We constructed a disease-related gene network by means of co-occurrence-based text-mining for specific disease-related studies in the literature. We then extracted the Google search result value for the every gene pair which has an edge in the gene network from Google. The Google search result value is then used to re-enforce the gene network. The disease-related gene network has two weights between gene pairs which are linked. One of the weights is a frequency value which is obtained from the literature data, and the other weight is Google search result value which is obtained from Google data. After constructing a disease-related gene network, we calculated the LGscore using the two weights in the network. Using the LGscore, we extracted disease-related genes from the disease-related gene network. Our method has three steps. First, we obtain genes and gene–gene relationships from the literature on a certain disease. We then construct a disease-specific gene network based on text-mining results. In the next step, we supplement the gene–gene relationships in the gene network using Google data. In the last step, we calculate the LGscores of the genes using the frequency and Google search result values to identify disease-related genes based on the LGscore.

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

2. Related works

2.1. ABC model

The ABC model refers to a method with which to determine a relationship between "A" and "C" using the A–B relationship and the B–C relationship. The ABC model can reveal novel relationships using two entities which are already known to be related. For instance, if a disease is related to a gene and the gene is linked to a drug, then a candidate relationship between a disease and a drug is inferred by the ABC model. In this way, the ABC model can infer indirect relationships using direct relationships which are known. The ABC model also provides an opportunity to identify new knowledge without special skills. For this reason, the ABC model is commonly used in bioinformatics. Swanson showed that with the ABC model, it is possible to use literature data to infer new relationships. Swanson inferred a relationship between Raynaud's disease and fish oil using the ABC model. A number of text-mining methods using the ABC model were subsequently introduced.

Srinivasan et al. [28] inferred relationships between curcuma and disease using the ABC model. They found papers on topic A in PubMed, and extracted the A–B relationships between topics A and B from the literature using MeSH terms. Likewise, they constructed B–C relationships from the biological literature. In their experiment, A denoted curcuma disease; B accounted for the genes, genomes, enzymes, amino acids, peptides, and proteins; and C consisted of body parts, organ components, diseases, syndromes, and the neoplastic process. They confirmed that curcumin plays a beneficial role in several diseases, such as retinal diseases, Crohn's disease, and disorders related to the spinal cord. The evidence is based on the relationships between curcumin and several genes. Lee et al. [17] inferred relationships between Alzheimer's disease and drugs using an advanced version of the ABC model. They incorporated context-term vectors into the previous ABC model to infer meaningful relationships. They extracted various relationships from the literature by means of text-mining and created a context-term vector based on biological entities which occur in conjunction with relationships in the literature. They calculated scores for relationships using context-term vectors and inferred more accurate relationships between Alzheimer's disease and drugs than the ABC model.

In their experiments, we confirmed that the ABC model is a useful method with which to infer more meaningful relationships. For this reason, we propose a method to identify meaningful disease–gene relationships using the ABC model enhanced by Google data.

2.2. Networks in biology

A network can be used to present complex relationships between biological entities. In particular, a network is widely used to indicate gene–gene interactions such as activation and inhibition relationships. A gene-regulatory network (GRN) is a typical gene network. This type of network provides a variety of scoring measures for calculating node scores, such as degree centrality, closeness centrality, and betweenness centrality. Using these measures, we can determine more meaningful disease-related genes in a gene network for a specific disease.

Fig. 1 shows the gene network for prostate cancer constructed by the PRINCIPLE [25] tool. The PRINCIPLE tool describes gene networks based on the PRINCE algorithm. It uses node colors to represent degrees of gene–disease similarity. In Fig. 1, nodes indicate genes, while edges indicate gene–gene interactions. In their gene network, we confirmed that a network can be used to present useful knowledge between biological entities. Considering the network characteristics, we constructed a disease-related gene network. We used various node shapes to indicate various gene conditions, such as confirmed genes and candidate genes. In our research, degree centrality was used with Google data as a network analysis measure to calculate scores of nodes in a gene network.

2.3. Google data

Google data can be used to determine trends in diseases. Cook et al. [5] predicted influenza activity using Google Flu Trends

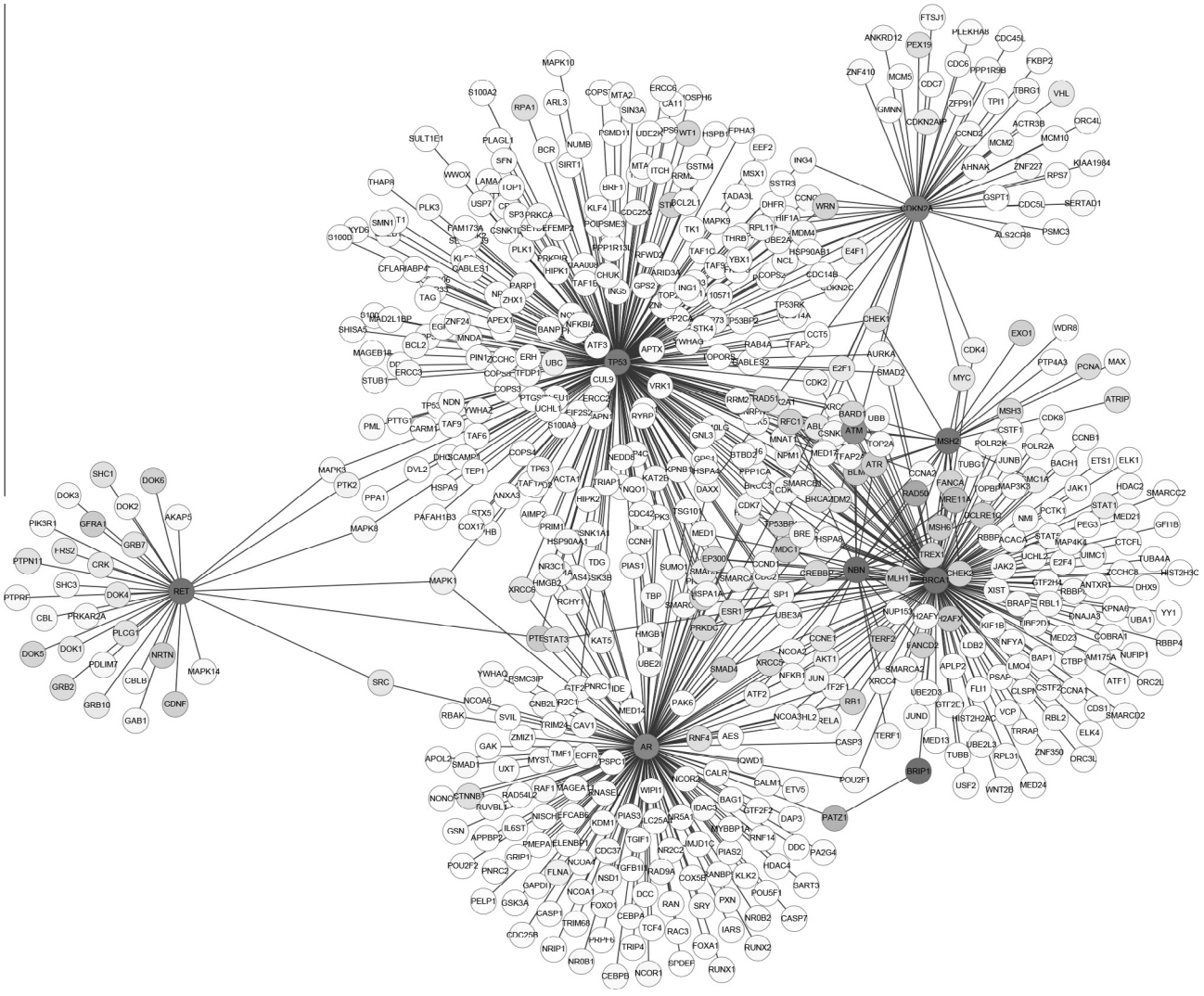

Fig. 1. Gene network for prostate cancer as described by the PRINCIPLE tool.

(GFT). GFT predicts influenza activity based on Internet search activity. Cook et al. confirmed that the influenza activity they inferred was closely related to official influenza surveillance data. Furthermore, they detected the 2009 influenza virus flow by analyzing changes in Google search terms such as "influenza complications" and "terms for influenza." In their experimental results, we confirmed that Google data can be used to infer biological information. For this reason, we used Google data to supplement gene–gene interactions in the gene network constructed by text-mining.

2.4. Text-mining

Current databases contain a vast number of biological publications. Making full use of these databases is difficult. One tool that has been used to extract hidden information from publication data is text-mining. Xie et al. [29] collected microRNA-related data using text-mining based on 75 rules. They extracted 878 relationships between 236 instances of microRNA data and 79 instances of cancer data in publications. Text-mining was thus shown to be a useful tool for extracting data from publications and identifying relationships among biological entities. In their research, we confirmed that text-mining is a useful traditional method that can also be used to extract relationships among biological entities.

In this study, we used co-occurrence-based text-mining to construct a gene network for diseases. After constructing the gene network, we used the ABC model to infer various disease-related genes. In our approach, the A–B relationship indicates the disease–gene interaction, and the B–C relationship indicates the gene–gene interactions. The disease–gene interactions are extracted from the text-mining results, and the gene–gene interactions are obtained from Google data. These two values are used as weights for the edges to analyze the network.

3. Methods

In this section, we propose a means of identifying disease-related genes using Google search data and literature data. Our method is illustrated in Fig. 2.

First, we mined the abstracts of publications in PubMed related to prostate cancer, colon cancer, lung cancer, diabetes, and Alzheimer's disease. PubMed provides biological literature data in an abstract format. In the PubMed database, abstract data is generated by search results for an input keyword. To obtain disease-specific abstract data, we used disease names as search keywords in PubMed. We obtained abstract data for each disease from PubMed using five disease names as search keywords. Alzheimer's disease

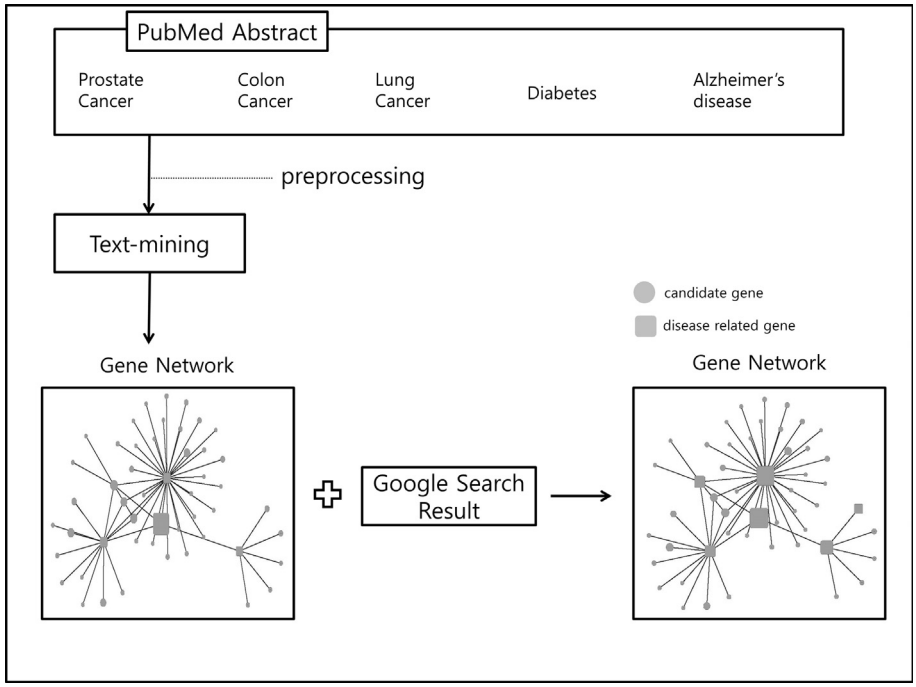

Fig. 2. Outline of the proposed method. In the gene network, all nodes indicate genes inferred by our method. The square nodes indicate disease-related genes from among the inferred genes, and circle nodes indicate genes which are not confirmed as disease-related genes from among the inferred genes.

and diabetes have been widely studied, as have cancers. In particular, a large amount of abstract data is available for prostate cancer and lung cancer, while less data is available for colon cancer. After preprocessing the abstract data, we connected genes that appeared in the same sentence to construct disease-related gene networks. Next, we rebuilt the gene networks using Google search results. Afterward, we analyzed the gene networks and then extracted disease-related genes. Our method processes were applied for each disease.

3.1. Preprocessing and gene network construction

We removed unnecessary data, such as the author, institute, date, and journal name from the abstract data. We categorized sentences according to parts-of-speech tagging using POS tagger [22,23]. Fig. 3 shows how a sentence is analyzed using a POS tagger. Rectangles in the figure indicate nouns.

As shown in Fig. 3, identified parts of speech are separated using the '_' character. In our experiment, we selected nouns. Noun symbols consisted of NN, NNP, NNPS, and NNS. We compared extracted words with human gene symbol lists to identify gene names in the sentences. The human gene symbol list was obtained from the HUGO Gene Nomenclature Committee (HGNC) [12], [13]. Nodes and edges of the gene network were constructed based on co-occurrences. We linked genes that appeared in the same sentence, and assigned weights to each edge between two genes using the frequency. The frequency of an edge between two genes indicates the number of sentences that refer to both genes.

3.2. Google search results

We obtained Google search result values by entering two genes at a time in the Google search box. Google search results indicate the number of documents that have the search term – in our case the names of the two genes. We used the Google search results to enhance the weights of edges in the gene network. We did not use gene symbol names but the full names of the genes in the search box to obtain accurate results. An example of a Google search result is presented in Fig. 4.

In Fig. 4, the two genes searched for were estrogen receptor 1 and the epidermal growth factor receptor. The circle in the figure indicates the Google search result value.

3.3. Scoring

We calculated a score for each gene using LGscore. LGscore consisted of two values which include the frequency and the Google search result. The frequency was obtained from literature data,

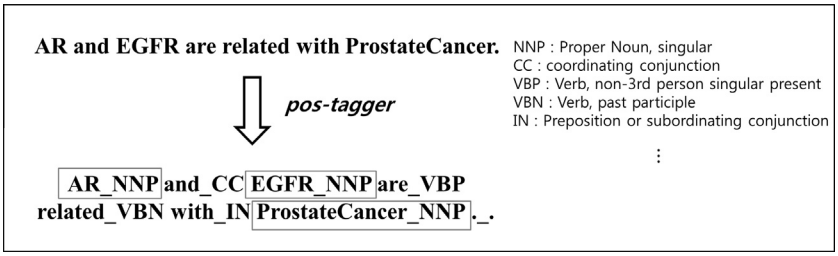

Fig. 3. Application of POS-tagger to sentences. The rectangles indicate nouns selected by POS-tagger in an example sentence.

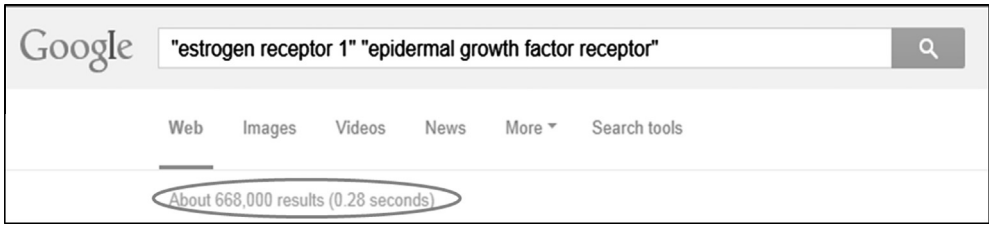

Fig. 4. Google search results.

and it is used to infer similarity with a disease. If a gene appears numerous times with other genes in the literature related to a specific disease, then the gene is considered to be a disease-related gene for that disease. The Google search result value obtained from Google was used to represent the degree of similarity in the relationships between candidate disease genes. If a gene has a high Google search result value in relation to other candidate disease genes, the gene is considered to be closely related to the candidate disease genes. In this case, we extract the gene as a disease-related gene. In this process, we assumed that all of genes in the gene network are candidate disease-related genes because they appeared in the disease-related literature. We used the Google search results to represent the weight of the gene–gene interactions to extract various disease-related genes which cannot be found by methods which use only the frequency. For instance, certain genes are cited in fewer papers, even when they are meaningfully linked to a disease, as they were recently confirmed to be disease-related genes. In this case, the frequency-based method cannot extract these genes, as the measure considers only the number of papers which contains the gene. Hence, our approach uses one more measure, i.e., the Google search result, to consider relationships with other candidate disease genes. Consequently, our approach considered indirect relationships between candidate disease genes as well as direct relationship with disease.

Fig. 5 shows a flow chart of the process used to calculate the LGscore. In the literature, we extracted gene pairs using text-mining based on co-occurrences. The frequency value between the genes pairs is calculated by the number of sentences which contain two genes in the literature, and the Google search result value is obtained from the Google search. The two values are normalized using the z-score measure. After scaling, the frequency and Google search result are used to calculate the LGscore.

LGscore was calculated as follows:

$$\mathrm{LGscore}(A) = \mathrm{Zscore}(\mathrm{Fre}(A)) + \mathrm{Zscore}(\mathrm{GSR}(A))$$

Here, Fre(A) denotes the score calculated based on the frequency values, while GSR(A) denotes the score calculated from the Google search result. Zscore(x) denotes the Z-scoring value of the number x. LGscore(A) is the summation of two values to consider direct and indirect relationships. The two values are the z-scored results of the frequency value and Google search result. The Google search result values were much larger than the frequencies before Z-scoring. We used the Z-score as a scaling factor to make the frequency and Google search result values comparable. The frequency indicates the number of sentences in the biological literature that contained both genes. The formulae we used to calculate Fre(A) and GSR(A) are shown below.

$$\mathrm{Fre}(A) = \sum_{n=1}^{N(A)} \mathrm{Frequency}(A, A_n^+)$$

$$\mathrm{GSR}(A) = \sum_{n=1}^{N(A)} \text{Google Search Result}(A, A_n^+)$$

In these equations, A_n^+ denotes the n-th neighbor node linked by node A, and $N(A)$ is the number of neighbor nodes linked by node A. The Frequency(A, B) is the number of appearances of node A and node B in the same sentence, and Google Search Result(A, B) is the Google search result value between node A and node B. Fre(A) is the score of node A calculated based on the frequency value, and GSR(A) is the score of node A calculated based on the Google search result value.

As shown in Fig. 6, LGscore is classified into four cases: "high Fre(A) + high GSR(A)," "high Fre(A) + low GSR(A)," "low Fre(A) + high GSR(A)," and "low Fre(A) + low GSR(A)." Case 1 indicates that gene A is closely related to the disease and closely related to the candidate disease genes. In this case, gene A has the highest LGscore, and gene A is extracted as a disease-related gene. In case 2, the score of gene A is affected by the direct relationship with the disease as opposed to indirect relationships with the candidate disease genes. If gene A has a high frequency value, gene A is extracted by case 2. In contrast to this, case 3 indicates that the score of gene A is affected by indirect relationships with the candidate disease genes more than it is by direct relationship with the disease. In case 3, LGscore can extract disease-related genes which are cited in fewer papers. Case 4 means that gene A is not significantly linked to the disease and candidate disease genes. When using the LGscore method, GSR is used to offset the weakness of a frequency value when used only with the number of cited papers. Using these two weights, the LGscore can extract disease-related genes which are cited in fewer papers as well as disease-related genes which have high frequency values.

The score of each node is proportional to the number of neighboring nodes and the weight of the edges with the neighboring nodes. This equation is similar to the degree centrality measure in terms of how it uses the edges of a node to calculate the score of the node. However, our scoring function uses the frequency and Google search result for the weight of the edges to consider useful knowledge with neighboring nodes.

4. Results

In this section, we describe our experimental results and present comparisons of our method with comparable methods. We applied our approach to five diseases: Alzheimer's disease, diabetes, prostate cancer, colon cancer, and lung cancer. After extracting the 20 genes with the highest LGscores for each disease, we compared the genes with the answer set to verify the feasibility of our method.

4.1. Data and properties of the gene networks

We downloaded abstracts from publications related to the five diseases from PubMed. We obtained 41,257 human gene symbols from HGNC. Table 1 summarizes the data used in our experiment. Table 1 shows the number of abstracts, nodes, and edges in the gene network for each disease.

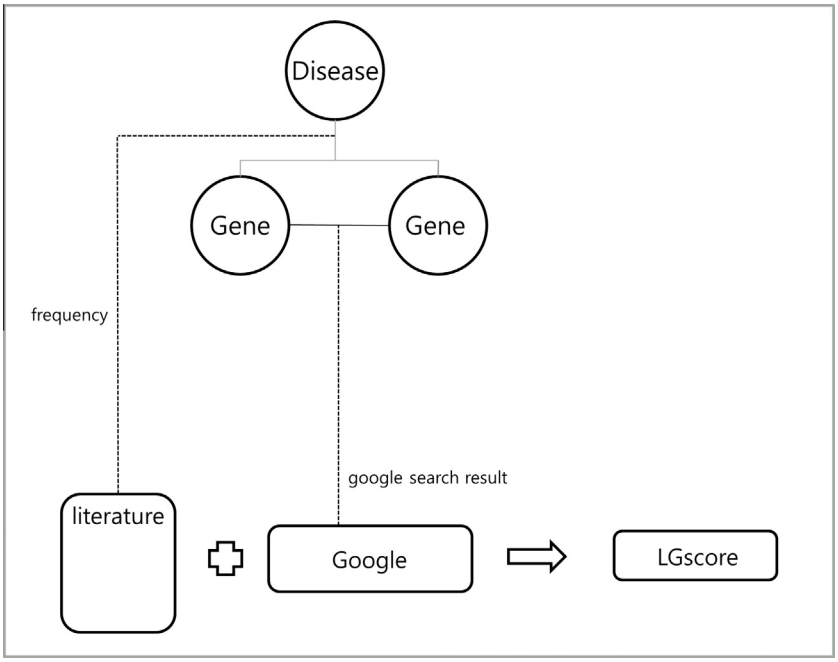

Fig. 5. Scoring process for LGscore. The frequency indicates the number of papers which include two genes, and the Google search result indicates the value obtained by a Google search using the two genes as keywords.

case 1: high Fre(A) + high GSR(A)
case 2: high Fre(A) + low GSR(A)
case 3: low Fre(A) + high GSR(A)
case 4: low Fre(A) + low GSR(A)

Fig. 6. Four cases in LGscore. Here, Fre(A) denotes the degree of similarity with a certain disease, and GSR(A) is the similarity with candidate disease genes.

Table 2 shows the answer set databases that we used to validate our results. The CTD [6,7], NCI [18], Sanger [27], KEGG [15], PGDB [21], and DDPC [8] databases contain information about gene–disease relationships. Additionally, we described the number of answer sets used in our experiments. CTD data was used to validate our results for Alzheimer's disease and diabetes. The results for the cancers were validated using KEGG and Sanger data. The NCI, PGDB, and DDPC databases were also used as answer sets for colon cancer and prostate cancer. We extracted disease-related genes for each disease from the databases listed in Table 2 and compared these genes to the top 20 genes inferred using our method and three comparable methods.

4.2. Comparison with existing methods

Terms used to describe our experimental results are defined in Table 3. "Genes by LGscore method only" denotes disease-related genes contained in the top 20 genes as inferred by LGscore but not contained in the top 20 genes inferred by other methods (see Table 4).

A Venn diagram to help understand what "genes by LGscore method only" refers to is presented in Fig. 7. The values in the Venn diagram are genes confirmed by the answer set to be among the top 20 inferred genes. The dark area in the Venn diagram corresponds to "genes inferred by the LGscore method only."

Fig. 8 shows a comparison of the results with random values for three cancers. The random values are calculated by the formulae below.

$$\text{Random value} = \frac{\text{The number of answer set}}{\text{The number of genes in the gene network}}$$

As shown in Fig. 8, our results show that LGscore has a higher percentage of confirmed genes than a random value. Furthermore, the results indicate that our method successfully ranked disease genes using LGscore.

A comparison of a results obtained using LGscore and the frequency-based model is presented in Table 4 and Fig. 9. The frequency-based model is a method which uses only the frequency as an edge weight in the gene network. All of the processes of the frequency-based model are identical to the LGscore except for the scoring function. The frequency-based model constructs a disease-related gene network based on co-occurrences in the literature. This approach links genes that appear in the same sentence and assigns weights to each edge between two genes using the frequency. The frequency indicates the number of sentences that mentioned both genes. The model infers disease-related genes using the frequency value with regard to other neighboring nodes. The scoring process is identical to the equation for Fre(A). For this reason, the nodes and edges in the frequency-based gene network

Table 1
Data and properties of gene networks.

	Alzheimer's disease	Diabetes	Colon cancer	Lung cancer	Prostate cancer
Number of abstracts	70,649	430,553	28,538	43,850	65,196
Number of nodes in the gene network	1051	3008	1431	2209	2058
Number of edges in the gene network	4017	13,800	4192	7616	8771

Table 2
Answer dataset.

	Alzheimer's disease	Diabetes	Colon cancer	Lung cancer	Prostate cancer
Data	CTD	CTD	NCI, Sanger, KEGG	Sanger, KEGG	PGDB, DDPC, KEGG, Sanger
Number of answer sets	>500	>500	33	16	177

Table 3
Definitions of terms.

Term	Definition
Confirmed genes	Genes inferred by each method known to be related to a disease
Genes by the LGscore method only	Confirmed genes in the top 20 genes as inferred by the LGscore but not contained in the top 20 genes inferred by other methods
Percentage of confirmed genes	(Number of confirmed genes/20) $*$ 100
Percentage of genes inferred using the LGscore method only	(Number of genes determined using the LGscore method only/number of confirmed genes) $*$ 100

Table 4
Comparison of LGscore and the frequency-based model.

	Alzheimer's disease		Diabetes		Colon cancer		Lung cancer		Prostate cancer	
	Frequency	LGscore	Frequency	LGscore	Frequency	LGscore	Frequency	LGscore	Frequency	LGscore
Number of confirmed genes	20	20	19	19	9	9	4	5	14	16
Number of genes identified by the LGscore method only	–	8	–	9	–	2	–	1	–	4
Percentage of confirmed genes	100.00	100.00	95.00	95.00	45.00	45.00	20.00	25.00	70.00	80.00
Percentage of genes identified by the LG score method only	–	40.00	–	47.37	–	22.22	–	20.00	–	25.00

and the nodes and the edges in the LGscore-based gene network are identical. However, the disease-related genes extracted by each method differ because the weights of edges are different in the scoring function. The *x*-axis indicates the disease and the *y*-axis indicates the number of inferred genes known to be related to the disease. For prostate cancer, we found that of the top 20 genes inferred based on their LG score, 16 were related to a disease, whereas the frequency-based model only found 14 confirmed genes. LGscore and the frequency-based model found the same number of disease-related genes for Alzheimer's disease, diabetes, and colon cancer. For lung cancer and prostate cancer, LGscore found more disease-related genes than the frequency-based model. Thus, LGscore was able to identify the same or a higher percentage of confirmed genes than the frequency-based model for the five diseases. Furthermore, LGscore was able to identify genes not identified by the frequency-based model.

For Alzheimer's disease and diabetes, both the LGscore method and the frequency-based model returned a high percentage of confirmed genes. For colon cancer and lung cancer, the percentage of confirmed genes was low for both approaches because the size of the answer sets was relatively small. For diabetes, 19 of the top 20 inferred genes were confirmed to be related to disease, and nine of these 19 genes were inferred by LGscore only. For all diseases, the LGscore method inferred a set of genes that the frequency-based model could not identify.

A comparison of LGscore and the PRINCE algorithm is provided in Fig. 10.

The PRINCE algorithm identifies disease-related genes using disease–disease similarity data and protein–protein interaction data. The PRINCE algorithm initially selects the target disease, and the disease shows phenotypic similarity to other diseases based on the disease–disease similarity data. These similar diseases have known causal genes which are used as prior information. The PRINCE algorithm constructs a protein network using known causal genes and other proteins which are connected to the known causal genes in a protein–protein interaction network. The network is computed using an iterative network propagation method. After the amounts of the flow are determined, the proteins have a score which is used as a standard to extract candidate genes for the target disease. To compare with the PRINCE algorithm, we used the PRINCIPLE tool, which provides a user interface for the PRINCE algorithm. The tool has three parameters to execute: *a*, *k* and *t*. Here, a is a weighting parameter, *k* denotes the number of top-ranked genes to return, and the *t* is the number of iterations performed by the algorithm. We changed only the *k* parameter to obtain the top 20 genes and used the default values of the other parameters ($a = 0.9$, $t = 10$). As shown in Fig. 10, LGscore and the PRINCE algorithm found the same number of disease-related genes for diabetes. For the other diseases, the LGscore method found more disease-related genes than the PRINCE algorithm. Our approach was therefore able to identify the same or even a higher percentage of confirmed genes (up to 40%) than the PRINCE algorithm for five diseases. Furthermore, the proposed LGscore method was able to infer a set of genes that the PRINCE algorithm was not able to identify. As mentioned previously, the low percentages of confirmed genes for colon cancer and lung cancer are due to the small size of the answer sets. The results of the comparison between the LGscore and the PRINCE algorithm are presented in Table 5.

LGscore showed better performance for four of the five diseases. For prostate cancer, the percentage of confirmed genes inferred using our approach was twice as high as the percentage inferred using the PRINCE algorithm. For all diseases, the proposed LGscore method inferred a set of genes that the PRINCE algorithm was not able to identify. These findings indicate that LGscore can be used to find disease-related genes not found using conventional methods.

Finally, we compared our approach to that of Ozgur et al. to identify disease-related genes. In their experiment, they started

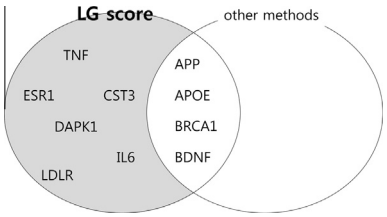

Fig. 7. Genes inferred using the LGscore method only.

with 15 seed genes already known to be related to the disease of interest. Tables 6 and 7 were extracted from the paper of Ozgur et al., and we incorporated our results. Terms are defined in Table 6, and we provide a comparison of our results to those of Ozgur et al. in Table 7.

Ozgur et al. used five scoring methods to analyze gene networks, namely Degree, Eigenvector, Betweenness, Closeness, and Baseline. They validated their results using prostate cancer data from literature data as well as the PGDB and KEGG pathways. Degree and Eigenvector identified 19 confirmed genes among the top 20 genes. LGscore identified 18 confirmed genes among the top 20 genes. Although these results could be interpreted to indicate that Degree and Eigenvector perform better than LGscore, the confirmed genes in Ozgur et al. include the seed genes. When these were excluded, Degree extracted 14 confirmed genes. Fig. 11 shows both cases.

As shown in Fig. 11, LGscore found more disease-related genes than the method in Ozgur et al. after excluding seed genes.

$$\text{Percentage of confirmed genes using LGscore} = \frac{\text{number of confirmed genes in top } (n-k)}{n-k} * 100$$

$$\text{Percentage of confirmed genes using Ozgur's method with } k \text{ seeds} = \frac{\text{number of confirmed genes in top } (n) - k}{n-k} * 100$$

Table 8 shows the percentage of confirmed genes among the genes inferred using each method. We validated the results using PGDB only as an answer set because, in the paper by Ozgur et al.,

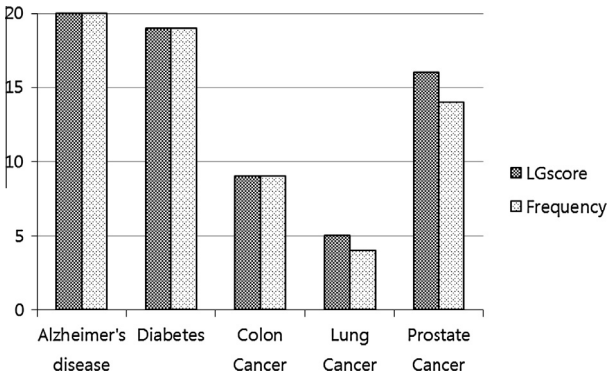

Fig. 9. The number of genes confirmed by LGscore and the frequency-based model. The *x*-axis indicates diseases, and the *y*-axis indicates the number of confirmed genes among the top 20 genes inferred by each method. LGscore indicates the proposed method, and Frequency indicates the frequency-based model.

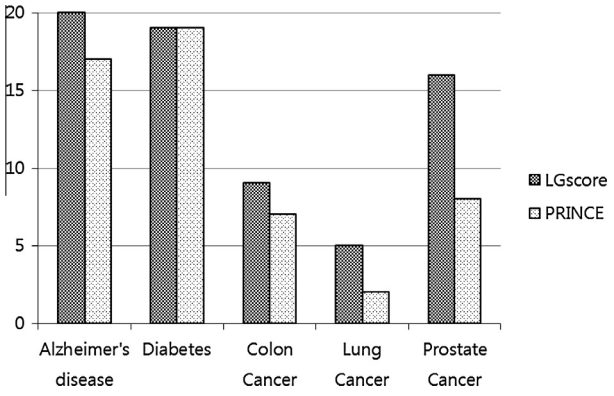

Fig. 10. The number of confirmed genes based on LGscore or the PRINCE algorithm. The *x*-axis indicates diseases, and the *y*-axis indicates the number of confirmed genes among the top 20 genes inferred by each method. The LGscore indicates the proposed method, and the PRINCE indicates the PRINCE algorithm.

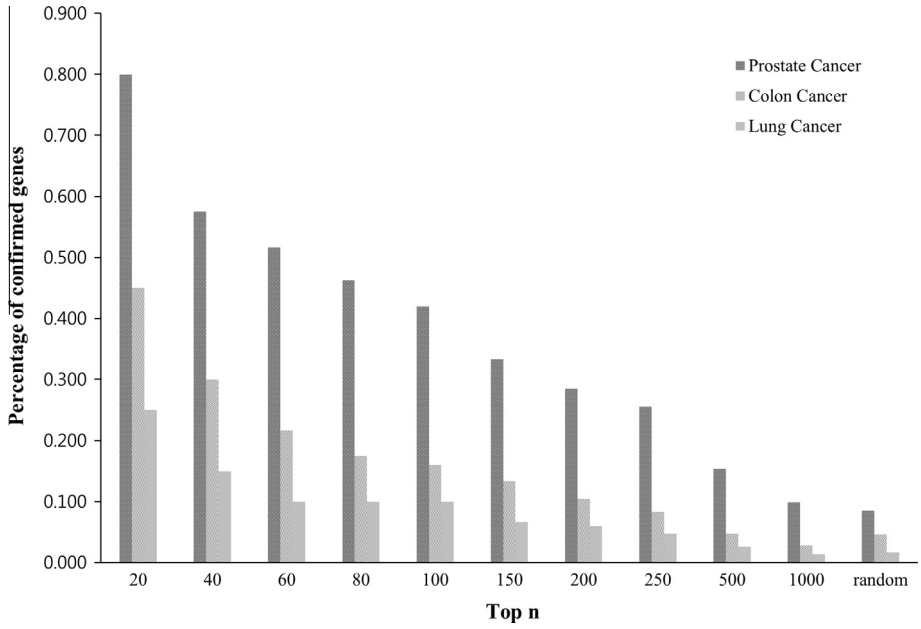

Fig. 8. Comparison of results with a random value for the top n genes inferred by LGscore: the *x*-axis indicates the number of genes inferred by LGscore, and 'random' indicates the probability that a gene which is selected randomly will be related to the disease. *Y*-axis indicates the percentage of confirmed genes for the three cancers.

Table 5
Comparison of LGscore with the PRINCE algorithm.

	Alzheimer's disease		Diabetes		Colon cancer		Lung cancer		Prostate cancer	
	PRINCE	LGscore	PRINCE	LGscore	PRINCE	LGscore	PRINCE	LGscore	PRINCE	LGscore
Number of confirmed genes	17	20	19	19	7	9	2	5	8	16
Number of genes inferred by the LGscore method only	–	16	–	19	–	5	–	4	–	12
Percentage of confirmed genes	85.00	100.00	95.00	95.00	35.00	45.00	10.00	25.00	40.00	80.00
Percentage of genes inferred by the LG score method only	–	80.00	–	100.00	–	55.56	–	80.00	–	75.00

Table 6
Definitions of terms.

Term	Definition
Seed gene	A prostate cancer gene retrieved from the OMIM Morbid Map
Inferred gene	A non-seed gene
Percentage of inferred genes	(Number of inferred genes/20) * 100
Confirmed inferred gene	An inferred gene found to be related to prostate cancer based on the PGDB and KEGG pathways for prostate cancer and published articles
Percentage of confirmed inferred genes	(Number of confirmed inferred genes/Number of inferred genes) * 100
Percentage of confirmed genes	((Number of confirmed inferred genes + Number of seed genes)/20) * 100

Table 7
Comparison of LGscore with the method of Ozgur et al.

	LGscore	Ozgur et al. scoring metrics				
		Degree	Eigenvector	Betweenness	Closeness	Baseline
Number of seed genes	0	5	6	7	2	3
Number of inferred genes	20	15	14	13	18	17
Percentage of inferred genes	100	75	70	65	90	85
Number of confirmed inferred genes	18	14	13	8	13	10
Percentage of confirmed inferred genes	90.00	93.33	92.86	61.54	72.22	58.85
Percentage of confirmed genes	90	95	95	75	75	65

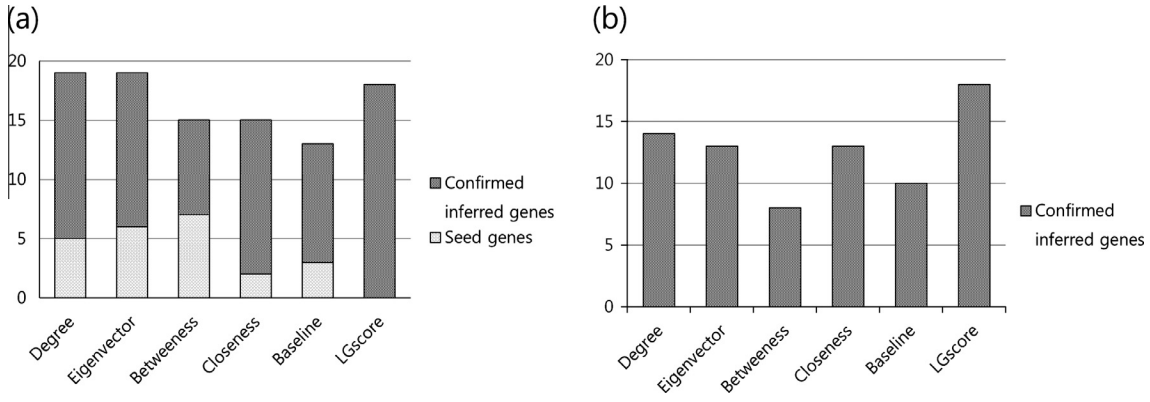

Fig. 11. Comparison of LGscore with the method of Ozgur et al.: (a) the *x*-axis indicates methods which include five measures in the method in Ozgur et al. and LGscore. The *y*-axis indicates the number of confirmed genes among the genes inferred by each method for prostate cancer. (b) The *x*-axis indicates methods which include five measures in the method in Ozgur et al. and LGscore. The *y*-axis indicates the number of confirmed genes among the genes inferred by each method for prostate cancer; (a) indicates the results from Ozgur et al. with the seed genes included, while (b) indicates the results with the seed genes excluded.

they used only PGDB as an answer set for validation of the top 226 prostate cancer-related genes. PGDB indicates whether a gene is related to prostate cancer or not. LGscore showed better performance than the approach in Ozgur et al. when the number of top genes in the set ranged from 75 to 226, and poorer performance when the number of top genes ranged from 10 to 50. However, these results changed when seed genes were considered, as shown in Table 9. We recalculated the percentages of confirmed genes again using the equation shown above to consider seed genes. The variable k indicates the number of seed genes among the inferred top 20 genes. The number of seed genes differed for each disease. For example, variable k for Degree was 5, as Degree used five seed genes among the top 20 inferred genes.

Fig. 12 shows results of the comparison of LGscore and the method in Ozgur et al. for prostate cancer. Eigenvector has largest value for the percentage of confirmed genes among the methods

Table 8
Comparison of "percentage of confirmed genes" for the top 226 prostate cancer-related genes inferred using LGscore and the method in Ozgur et al.

Top *n*	LGscore	Ozgur et al. scoring metrics				
		Degree	Eigenvector	Betweenness	Closeness	Baseline
10	60.00	80.00	80.00	90.00	70.00	50.00
20	70.00	75.00	80.00	70.00	55.00	45.00
30	56.67	60.00	63.33	63.33	56.67	43.33
40	50.00	55.00	57.50	52.50	47.50	32.50
50	48.00	46.00	50.00	48.00	42.00	28.00
75	41.33	33.33	36.00	34.67	33.33	34.67
100	38.00	26.00	28.00	26.00	27.00	27.00
125	30.40	23.20	25.60	23.20	23.20	22.40
150	27.33	20.67	22.00	20.00	20.00	18.67
175	24.57	18.29	20.57	18.29	18.29	17.14
200	23.50	17.50	19.00	18.50	17.00	15.00
226	21.68	17.70	17.70	17.70	17.70	13.27

used in Ozgur et al. As shown in Fig. 12, LGscore showed a higher percentage of confirmed genes than the method used in Ozgur et al. except for the top (20-*k*) genes.

Table 10 shows the comparison of the results for precision and recall. As shown in Table 10, LGscore showed higher precision and recall values in the entire interval, except for the top (20-*k*) cases.

In summary, LGscore recovered a higher percentage of confirmed genes than three other conventional approaches. The proposed LGscore method found more confirmed genes for lung cancer and prostate cancer than the frequency-based model. LGscore also returned a higher percentage of confirmed genes than the PRINCE algorithm and the method in Ozgur et al. when seed genes were excluded. Importantly, it is possible to use LGscore to identify disease-related genes without using seed genes. Together, these results indicate that LGscore is a more useful method than existing methods to identify relationships between diseases and genes.

5. Discussion

5.1. Analysis for the top 20 genes inferred by LGscore

In this section, we describe a sub-network of genes related to prostate cancer for which we scored the weights of edges using our LGscore method.

Fig. 13 shows part of a gene for prostate cancer. We included the top 20 genes identified by LGscore and other genes with a frequency weight greater than 5. The size of nodes is proportional to the LGscore of the node. TNF and IGFBP3 in the right part of Fig. 13 only have a few edges. It is difficult to identify these genes using methods that use frequency-based scores only. Incorporation of the Google search score helps to identify genes with fewer edges, such as these genes.

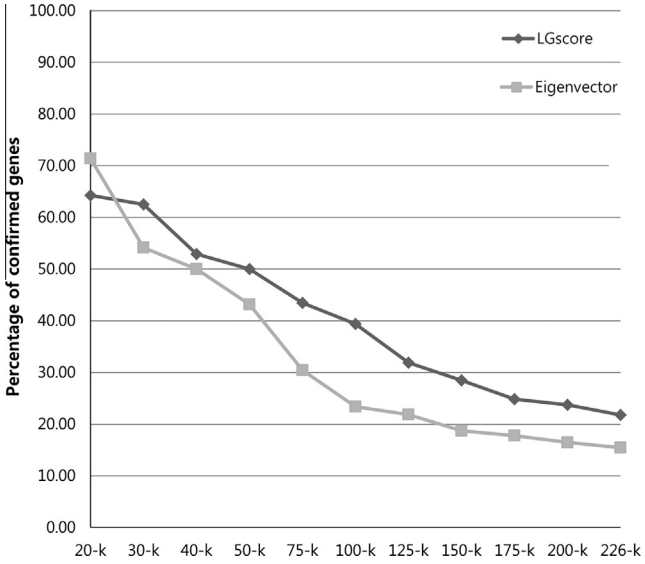

Fig. 12. Comparison of LGscore with the method in Ozgur et al. for prostate cancer. The *x*-axis indicates the number of inferred genes for prostate cancer. The *y*-axis indicates percentage of confirmed genes for LGscore and Eigenvector. Eigenvector shows best performance among the methods used in Ozgur et al.

Table 10
Comparison of Precision and Recall for the top 226 prostate cancer-related genes inferred using LGscore and the method in Ozgur et al. with seed genes excluded.

Top *n*	LGscore		Eigenvector	
	Precision	Recall	Precision	Recall
20-k	64.29	7.38	71.43	8.20
30-k	62.50	12.30	54.17	10.66
40-k	52.94	14.75	50.00	13.93
50-k	50.00	18.03	43.18	15.57
75-k	43.48	24.59	30.43	17.21
100-k	39.36	30.33	23.40	18.03
125-k	31.93	31.15	21.85	21.31
150-k	28.47	33.61	18.75	22.13
175-k	24.85	34.43	17.75	24.59
200-k	23.71	37.70	16.49	26.23
226-k	21.82	39.34	15.45	27.87

Table 11 describes the top 20 genes identified by LGscore. Eighteen out of these 20 genes were found to be involved in prostate cancer based on other lines of evidence. Fourteen genes were validated with PGDB, and two genes were validated with Sanger. Tumor necrosis factor (TNF) and CYP1A1 are both reportedly related to prostate cancer. Berhane et al. [3] demonstrated a signif-

Table 9
Comparison of "percentage of confirmed genes" for the top 226 prostate cancer-related genes inferred using LGscore and the method in Ozgur et al. with seed genes excluded.

Top *n*	LGscore *k* = 5	Degree	LGscore *k* = 6	Eigenvector	LGscore *k* = 7	Betweenness	LGscore *k* = 2	Closeness	LGscore *k* = 3	Baseline
20-k	66.67	66.67	64.29	71.43	61.54	53.85	66.67	50.00	64.71	35.29
30-k	60.00	52.00	62.50	54.17	65.22	52.17	57.14	53.57	59.26	37.04
40-k	54.29	48.57	52.94	50.00	54.55	42.42	52.63	44.74	51.35	27.03
50-k	48.89	40.00	50.00	43.18	48.84	39.53	47.92	39.58	46.81	23.40
75-k	42.86	28.57	43.48	30.41	42.65	27.94	42.47	31.51	43.06	31.94
100-k	38.95	22.11	39.36	23.40	39.78	20.43	37.76	25.51	38.14	24.74
125-k	31.67	20.00	31.93	21.85	32.20	18.64	30.89	21.95	31.15	20.49
150-k	28.28	17.93	28.47	18.75	27.97	16.08	27.70	18.92	27.89	17.01
175-k	24.71	15.88	24.85	17.75	25.00	14.88	24.86	17.34	25.00	15.70
200-k	24.10	15.38	23.71	16.49	23.32	15.54	23.74	16.16	23.86	13.71
226-k	21.72	15.84	21.82	15.45	21.92	15.07	21.88	16.96	21.52	12.11

k: number of seed genes in the top 20 genes inferred using the scoring methods in Ozgur et al.

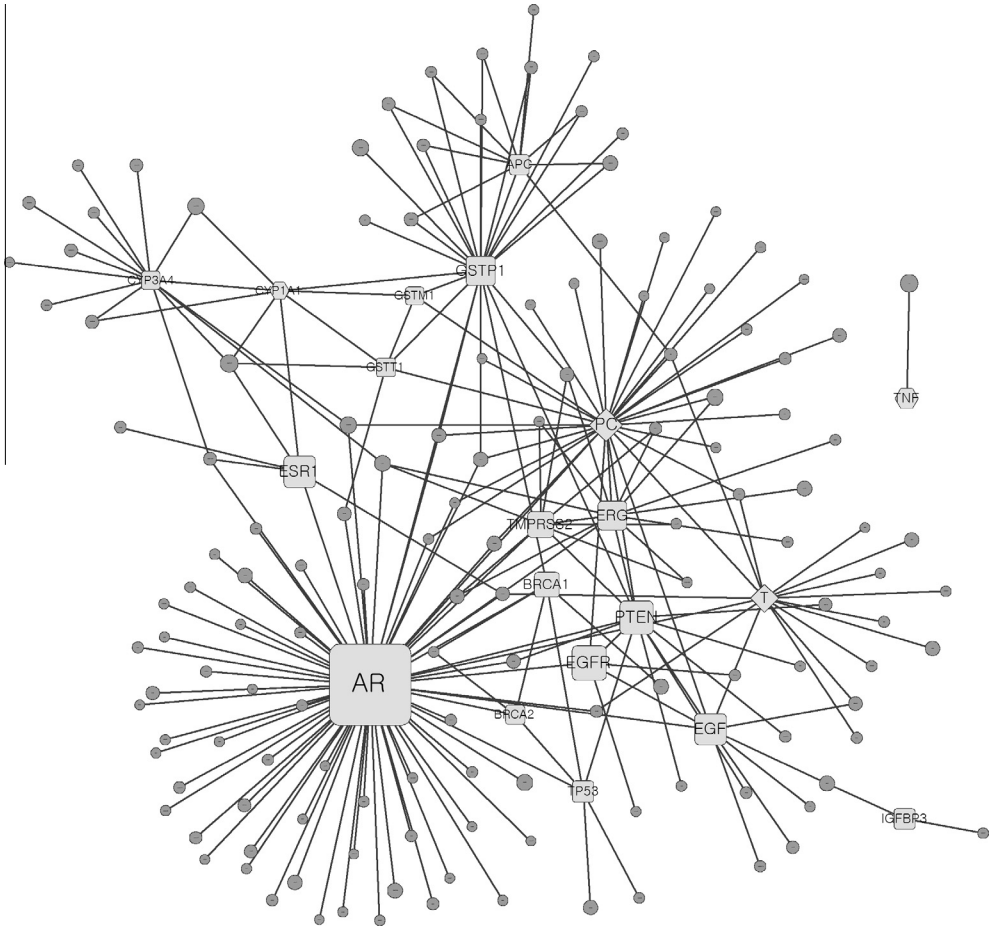

Fig. 13. Part of the gene network associated with prostate cancer. The nodes which include gene symbols indicate the top 20 genes extracted by LGscore, while the square nodes indicate confirmed genes validated by the PGDB answer set. The hexagon nodes denote confirmed genes validated based on the literature, while the diamond nodes indicate candidate genes which may be related to prostate cancer.

Table 11
Top 20 genes involved in prostate cancer as inferred by LGscore.

Rank	Gene	Evidence
1	AR	PGDB
2	EGFR	PGDB
3	PTEN	PGDB
4	PC	None
5	ESR1	PGDB
6	EGF	Sanger
7	GSTP1	PGDB
8	ERG	Sanger
9	T	None
10	TMPRSS2	PGDB
11	BRCA1	PGDB
12	TNF	Literature
13	TP53	PGDB
14	IGFBP3	PGDB
15	APC	PGDB
16	CYP1A1	Literature
17	BRCA2	PGDB
18	GSTT1	PGDB
19	CYP3A4	PGDB
20	GSTM1	PGDB

icantly increased risk of prostate cancer among individuals who carried the TNF a-308 allele. Ding [9] confirmed a significant association between the CYP1A1 MsqI polymorphism and the risk of prostate cancer among Asians. Brachyury (T) and pyruvate carboxylase (PC) are candidate genes that may be associated with prostate cancer. Huang [11] demonstrated that an in vivo treatment of tumor xenografts with chemotherapy resulted in the selective growth of resistant tumors characterized by high levels of Brachyury expression. Ihnatko [14] reported that the levels of pyruvate carboxylase were altered in tumor-bearing mice with anorexia.

To summarize, we showed that 18 genes among the inferred top 20 genes were meaningfully linked to prostate cancer, while the remaining two genes are reportedly involved in other cancers. These results indicate that the LGscore method can extract disease–gene relationships successfully.

5.2. Google data and literature data

In this section, we describe the Google data and literature data used to test that which components of the two LGscores play a more important role in identifying disease-related genes. LGscore uses the frequency value obtained from the literature and the Google search result value which is obtained from Google. We use the frequency value to denote similarity to the disease and Google search result value to indicate similarity to other candidate disease genes to consider special cases. A special case arises when disease-related genes cannot be extracted y using frequency weights because they are cited in fewer studies.

Fig. 14 indicates the number of confirmed genes among the top 20 inferred genes according to each weight for the five diseases. LGscore indicates the proposed method, and Google indicates the

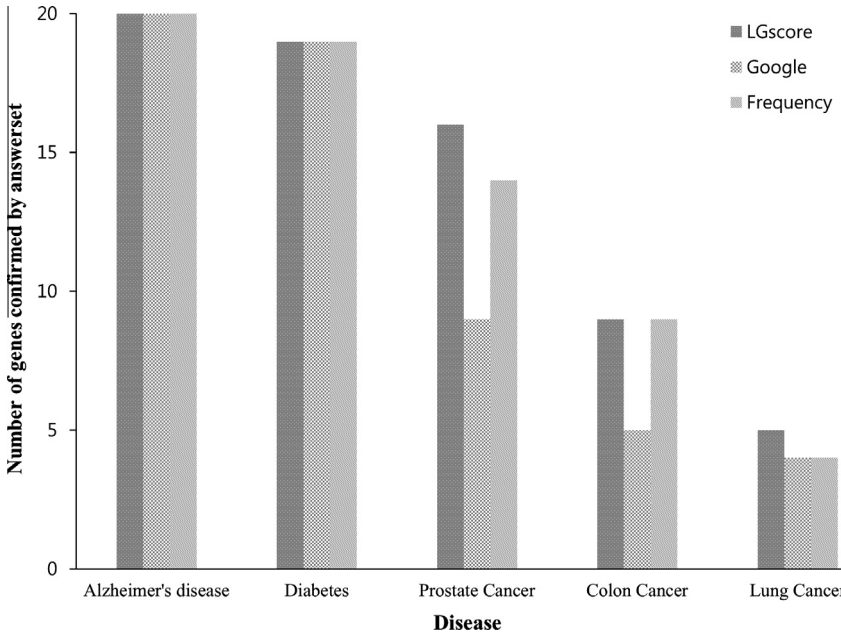

Fig. 14. Comparison of the three weights. The *x*-axis indicates diseases, and the *y*-axis indicates the number of confirmed genes among inferred 20 genes for each weight.

method which uses only the Google search result value as the weight of the edges. Likewise, frequency denotes the method which uses only the frequency value as the weight of the edges. As shown in Fig. 14, LGscore showed better performance than other methods for three diseases. In contrast, Google showed poorer performance than the other methods for three diseases. The Google search results, however, are useful information with which to supplement the frequency weight in LGscore. For this reason, LGscore has a higher value for the number of genes confirmed by an answer set as compared to the frequency. Furthermore, the top 20 inferred genes by LGscore included several genes which differ from the genes inferred by the frequency. These results show that Google search results contain useful knowledge for identifying relationships between genes, playing an important role in identifying disease-related genes as indirect information.

6. Conclusion

LGscore is a method that identifies disease-related genes using the literature and Google search results to increase the accuracy of extracted relationships. We applied our method to five diseases (Alzheimer's disease, diabetes, colon cancer, lung cancer, and prostate cancer) and demonstrated that LGscore extracted a higher percentage of genes known to be related to diseases than three other, comparable methods. LGscore is therefore an effective method with which to identify disease-related genes. In this paper, we used only nouns among the parts of speech. For further work, we will use other word classes such as verbs, adjectives, and adverbs to improve LGscore. Furthermore, we will use other biological information including protein, drugs, and miRNA data.

Acknowledgments

This research was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean Government (MSIP) (2012R1A2A1A01010775). Sanghyun Park is the corresponding author of this paper.

References

[1] Swanson DR. Undiscovered public knowledge. Libr Quart 1986;56(2):103–18.
[2] Swanson DR. Medical literature as a potential source of new knowledge. Bull Med Libr Assoc 1990;78(1):29–37.
[3] Berhane N, Sobti RC, Melesse S, Mahdi SA, Kassu A. Significance of Tumor necrosis factor a-308 (G/A) gene polymorphism in the development of prostate cancer. Mol Biol Rep 2012;39(12):11125–30.
[4] Chen G, Cairelli MJ, Kilicoglu H, Shin D, Rindflesch TC. Augmenting microarray data with literature-based knowledge to enhance gene regulatory network inference. PLoS Comput Biol 2014;10(6):e1003666.
[5] Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 Influenza Virus A(H1N1) pandemic. PLoS ONE 2011;6(8).
[6] Davis AP et al. The comparative toxicogenomics database: update 2013. Nucl Acids Res 2013;41:D1104–14.
[7] Curated[gene-disease] data were retrieved from the Comparative Toxicogenomics Database (CTD). North Carolina State University, Raleigh, NC and Mount Desert Island Biological Laboratory, Salisbury Cove, Maine. World Wide Web <http://ctdbase.org> [December 2013].
[8] Maqungo M, Kaur M, Kwofie S, Radovanovic A, Schaefer U, Schmeier S, et al. DDPC: dragon database of genes associated with prostate cancer. Nucl Acids Res 2010. http://dx.doi.org/10.1093/nar/gkq849.
[9] Ding G, Xu W, Liu H, Zhang M, Huang Q, Liao Z. CYP1A1 MspI polymorphism is associated with prostate cancer susceptibility: evidence from a meta-analysis. Mol Biol Rep 2013;40(5):3483–91.
[10] Gonzalez G, Uribe JC, Tari L, Brophy C, Baral C. Mining gene–disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. Pacific Symp Biocomput 2007;12:18–39.
[11] Huang B, Cohen JR, Fernando RI, Hamilton DH, Litzinger MT, Hodge JW, et al. The embryonic transcription factor Brachyury blocks cell cycle progression and mediates tumor resistance to conventional antitumor therapies. Cell Death Dis 2013;4.
[12] Gray KA, Daugherty LC, Gordon SM, Seal RL, Weight MW, Bruford EA. genenames.org: the HGNC resources in 2013. Nucl Acids Res 2013;41:D545–52.
[13] HGNC Database, HUGO Gene Nomenclature Committee (HGNC). EMBL Outstation – Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD; UK <www.genenames.org>.
[14] Ihnatko R, Post C, Blomqvist A. Proteomic profiling of the hypothalamus in a mouse model of cancer-induced anorexia-cachexia. Brit J Can 2013;109:1867–75.
[15] KEGG: KYoto Encyclopedia of Genes and Genomes <www.genome.jp/kegg/> [December 2013].
[16] Li S, Wu L, Zhang Z. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics 2006;17(22):p2143–50.

[17] Lee SJ, Choi J, Park K, Song M, Lee D. Discovering context-specific relationships from biological literature by using multi-level context terms. BMC Med Inform Dec Mak 2012;12(Suppl. 1):S1.
[18] National Cancer Institute: Comprehensive Cancer Information <www.cancer.gov/> [December 2013].
[19] Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetics Medicine, Johns Hopkins University (Baltimore, MD) <http://omim.org/> [December 2013].
[20] Ozgur A, Vu T, Erkan G, Radev DR, Identifying gene–disease associations using centrality on a literature mined gene–interaction network 2008;24:i277–85.
[21] Li LC, Zhao H, Shiina H, Kane CJ, Dahiya R. PGDB: a curated and integrated database of genes related to the prostate. Nucl Acids Res 2003;31(1):291–3.
[22] Toutanova K, Manning C. 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. p. 63–70.
[23] Toutanova K, Klein D, Manning C, Singer Y, Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the HLT-NAACL; 2003. p. 252–9.
[24] Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 2010;6(1).
[25] Gottlieb A, Magger O, Berman I, Ruppin E, Sharan R. PRINCIPLE: a tool for associating genes with diseases via network propagation. Bioinformatics 2011;27(23):3325–6.
[26] PubMed: MEDLINE Retrieval on the World Wide Web <www.ncbi.nlm.nih.gov/pubmed> [December 2013].
[27] Wellcome Trust Sanger Institute <http://www.sanger.ac.uk> [December 2013].
[28] Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substance and disease. Bioinformatics 2004;20:i290–6.
[29] Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics 2013;29(5):638–44.