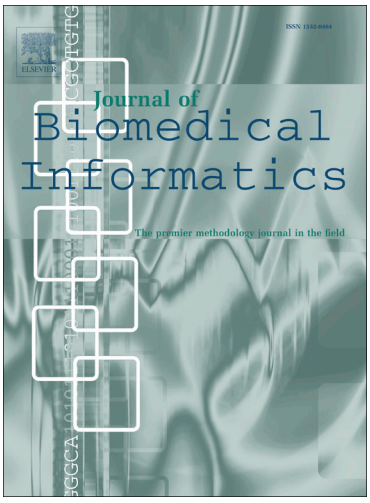

IMA: Identifying disease-related genes using MeSH terms and association rules

Jeongwoo Kim, Changbae Bang, Hyeonseo Hwang, Doyoung Kim, Chihyun Park, Sanghyun Park

PII: S1532-0464(17)30246-0
DOI: https://doi.org/10.1016/j.jbi.2017.11.009
Reference: YJBIN 2888

To appear in: *Journal of Biomedical Informatics*

Received Date: 15 June 2017
Revised Date: 29 October 2017
Accepted Date: 13 November 2017

Please cite this article as: Kim, J., Bang, C., Hwang, H., Kim, D., Park, C., Park, S., IMA: Identifying disease-related genes using MeSH terms and association rules, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.11.009

IMA: Identifying disease-related genes using MeSH terms and association rules

Jeongwoo Kim [a,*], Changbae Bang [a,*], Hyeonseo Hwang [a], Doyoung Kim [a], Chihyun Park [a], Sanghyun Park [a,†]

Affiliations

[a] Department of Computer Science, Yonsei University 50 Yonsei-ro, Sinchon-dong, Seodamun-gu, Seoul 120-749, South Korea

jwkim2013@yonsei.ac.kr, bcb225@gmail.com, hyeonseo0129@gmail.com, arbc139@gmail.com, chihyun.park@yonsei.ac.kr, sanghyun@yonsei.ac.kr

[*] These authors equally contributed to this paper.

[†] corresponding author Tel: +82 2 2123 5714; fax: +82 2 365 2579

Abstract Genes play an important role in several diseases. Hence, in biology, identifying relationships between diseases and genes is important for the analysis of diseases, because mutated or dysregulated genes play an important role in pathogenesis. Here, we propose a method to identify disease-related genes using MeSH terms and association rules. We identified genes by analyzing the MeSH terms and extracted information on gene-gene interactions based on association rules. By integrating the extracted interactions, we constructed gene-gene networks and identified disease-related genes. We applied the proposed method to study five cancers, including prostate, lung, breast, stomach, and colorectal cancer, and demonstrated that the proposed method is more useful for identifying disease-related and candidate disease-related genes than previously published methods. In this study, we identified 20 genes for each disease. Among them, we presented 34 important candidate genes with evidence that supports the relationship of the candidate genes with diseases.

Keywords: Association rules, data mining, gene, disease

1. Introduction

A gene is a locus of DNA that consists of nucleotides and has genotypes made up of different DNA sequences. Through various biological experiments, researchers confirmed that genotypes determine the resulting phenotypes. The phenotype describes various biological or physical traits, such as disease, eye color, and height. For this reason, identification of disease-gene relationships is important in biology. However, the size and number of human genes is too large to analyze for all disease-gene pairs. The biological experimental cost is

also prohibitively expensive. To solve this problem, several studies [6, 19, 31] have been performed to identify candidate disease-related genes. Among the various tools available to extract disease-gene relationships, biomedical text mining is well known.

A vast number of biological experiments have been presented as literature reports at conferences or in journals. These literature data include meaningful biological knowledge based on experimental results and are accumulated in online databases such as PubMed [34], PMC [33], and OMIM [30]. To obtain useful biological knowledge from the literature data, text mining is recommended. Text mining is widely used to extract information on biological entities or relationships between biological entities. However, using this approach, it is difficult to extract exact biological knowledge from the literature. To extract precise entity or relationship information, a named entity recognition (NER) step is required. NER is a method used to identify and classify words of interest (genes, drugs, and diseases in biology) from the stream of text. However, in biology, the accuracy of NER remains a challenge [11, 22, 27], because biological entities are reported as multiple synonyms, abbreviations, variations, etc. To avoid this problem and find exact biological information, we propose a method based on medical subject heading (MeSH) terms. MeSH is a medical vocabulary resource curated by the National Library of Medicine (NLM) [29]. The NLM provides biomedical keywords for all PubMed literature as MeSH terms. Professional researchers tag the MeSH terms for each article; therefore, the accuracy and correlation between MeSH terms and corresponding articles are high. In addition, representative MeSH terms are generated for synonyms by MeSH term generation rules. Due to this, MeSH terms can be used to address and process multiple synonyms.

In this study, we present a method (which we call IMA) to identify disease-related genes using MeSH terms and association rules. Our goals in this study are the extraction of exact biological knowledge from the literature and identification of meaningful disease-gene relationships. To achieve these goals, we used MeSH terms to extract biological knowledge and association rules to extract meaningful relationships. Association rule learning is a concept used to identify relationships based on the frequency of sets of items. For example, if two items coappear in the several data sets, then association rules consider that they have a relationship. Many applications (such as web mining, scientific data analysis, and marketing) have utilized association rules to reveal hidden relationships within large data sets. For this reason, we utilized the concept to find hidden relationships between genes. Our assumptions are as follows:

- Within the same literature references, the MeSH terms have relationships.

- By using association rules, we can extract gene-gene relationships from the MeSH terms.

Our method consists of three main steps, including MeSH processing, association rule mining, and network analysis. First, we obtain literature data for each disease from PubMed. After preprocessing, we extract MeSH terms from each literature report. Among the various MeSH types, we extract human gene symbols. Based on these human gene symbols and association rules, we generate gene-gene interactions, which we use to construct a gene-gene interaction network for each disease. After analyzing the network, we identify disease-related genes based on centrality scores for indicators such as degree, betweenness, and eigenvector.

Our primary contributions are below:

- We propose a method to identify disease-related genes using MeSH terms and association rules.

- We construct gene-gene interaction networks for each disease.

- We identify disease-related genes and meaningful disease-related candidate genes.

The rest of this paper is organized as follows. Section 2 discusses related work on methods for extracting disease-related genes. Section 3 describes our proposed method for using MeSH terms and association rules to extract disease-related genes. Section 4 presents the experimental results for the proposed method and compares the results with those of previous studies. Section 5 contains a discussion of the experimental results. Section 6 presents the conclusions of the study that highlight the implications of our findings.

2. Related works

2.1 Association rules and apriori algorithms

Association rules are useful for discovering hidden relationships within large data sets. This technique is used in many application domains, including bioinformatics, web mining, and scientific data analysis. Wright et al. [38] used association rules to elucidate the association between medications and medical problems. Using patient data, they conducted association mining to find related pairs between medications and medical problems. From the results of association mining, they inferred a large number of relationships between medications and patients' problems. Harpaz el al. [12] attempted to identify adverse drug events using an association rule mining technique. With association rule mining, items are organized in the form A to B, where A and B are disjoint item sets, and they designed a set of drugs as A and adverse symptoms as B. Using an apriori algorithm, an association rule mining technique that is useful for large dataset analysis, they obtained known adverse drug events and unknown potential events. Nahar et al. [28] attempted to find factors affecting heart disease, one of the most common fatal diseases. They also considered differential factors between genders. To find affecting and differential factors, they utilized three association rule mining techniques, including apriori, predictive apriori, and tertius.

Apriori [1] is an algorithm used for frequent item set mining and association rules. The algorithm finds the most frequent items and extends them to larger and larger item sets, as long as the number of appearances is greater than the minimum support value (threshold).

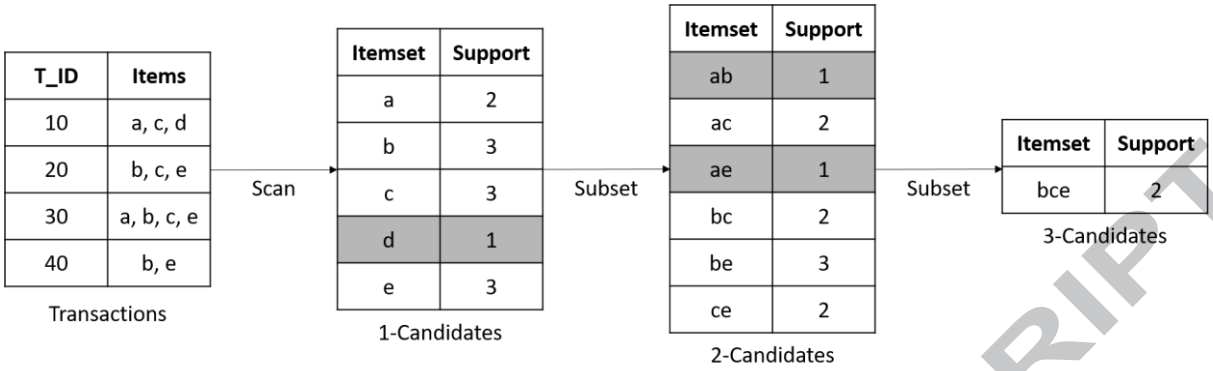

Fig. 1. Example of an apriori algorithm

Figure 1 shows an example of an apriori algorithm. First, a single item set is selected from the transactions with support. The support value represents the number of appearances of each transaction. If the support value of an item set is greater than the threshold value, subsets of the item set are created with support values. These results can be used to make a relationship with several values. In case of the relationship (A=>B), we can calculate support, confidence, and lift based on the support values for A and AB.

2.2 Gene prioritization

To prove relationships between genes and disease, biological experiments are expensive and time-consuming. To address these issues, various gene prioritization studies have been published. Gene prioritization is the identification of causal genes for disease based on various biological resources, such as gene expression, sequence, and published literature. Since the necessity and significance of gene prioritization was identified, a large volume of studies has been published. Taub et al. [36] used computational image analysis of microarray results to obtain the genotype of total cDNA probes. They hybridized two cDNA probes that were in pre-stimulus and post-stimulus states, respectively. Then, they used computational methods to analyze the cDNA probes. The study demonstrated that shifts in abundance between the states can be compared without purification by using computational methods. Maher et al. [25] suggested a robust pipeline to find novel genes using RNA-seq data. They analyzed the transcriptome of cancer cells to detect novel gene fusions, and identified experimentally proven novel gene fusions in tumors from RNA-seq data. Shim et al. [35] used genome-wide association study (GWAS) data to identify new candidate genes related to diseases. To overcome statistical limitations, they boosted weak association signals through a gene network, rather than increasing the number of samples. By combining GWAS P-values of single nucleotide polymorphisms and the gene network, they reprioritized candidate genes. This study showed that the gene network structure is useful for finding related genes.

2.3 Gene Networks

The network is a useful structure to describe biological relationships. The structure can present relationships between nodes and illustrate the weight of nodes and edges using size, color, and shape. Networks also permit various network analysis measures, such as degree, closeness, betweenness, and eigenvector centrality. Based on these network analysis measures, we can identify crucial nodes in the network. These advantages of networks have led many researches to utilize the network structure to describe and analyze relationships between biological entities.

Hoffmann and Valencia [14] attempted to build a network of genes and proteins that extends through the scientific literature, touching on phenotypes, pathologies, and gene function. By using genes and proteins as hyperlinks between sentences and abstracts, they converted the information in PubMed into one navigable resource, bringing all the advantages of the internet to scientific literature investigation. The network, called Information Hyperlinked over Proteins (iHOP), shows that distant medical and biological concepts can be connected by a few intermediate genes. Montojo et al. [26] designed the GeneMANIA application which provides the possibility to construct a gene-gene functional interaction network from a gene list. The constructed network includes genes related to the input gene list and functional annotations from gene ontology. Kim et al. [18] attempted to construct a disease-specific gene network to find novel disease-related genes. They extracted gene-gene interactions from the biomedical literature and calculated weights for genes using Google search data. By analyzing the constructed gene network, they extracted disease-related genes for five diseases.

3. Methods

In this section, we describe our proposed method to identify disease-related genes using MeSH terms and association rules. To extract meaningful biological data, we identify MeSH terms in the literature. We also extract gene-gene interactions based on association rule results for MeSH terms. An outline of the proposed method is described in Figure 2.

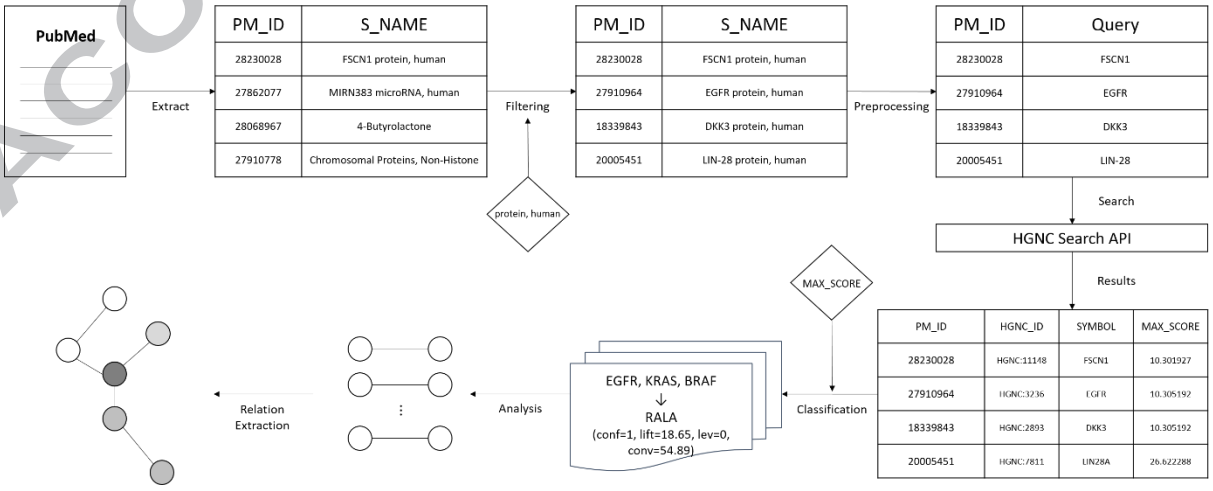

Fig. 2. Outline of the IMA method

First, we downloaded data on disease-specific literature from PubMed. Among the information available in the literature, we extracted MeSH terms and PMIDs. We filtered MeSH terms based on the "* protein, human" form to extract human gene symbols. Then, we conducted a preprocessing step to delete meaningless words in order to build queries. Queries were sent to the HUGO Gene Nomenclature Committee (HGNC) [10, 13] Search API to determine an approved gene symbol by considering similarity score. After obtaining gene symbols, we extracted relationships between genes using an apriori algorithm. In this step, the item set consisted of MeSH terms for each literature sample. Extracted relationships have a weight that is generated by lift value. Using these relationships, we constructed a gene-gene interaction network for a specific disease.

3.1 Literature and MeSH term preprocessing

To obtain disease-specific literature, we used the disease name as a search keyword in PubMed. PubMed provides literature in XML format. Among the available XML headers, we extracted “name of substance” and “PMID”. The “name of substance” is used to extract genes that are assigned for each literature reference, and “PMID” is used to make an item set for each literature reference. The PMID indicates the reference number that is generated by PubMed. To extract gene names from the MeSH terms, we filtered the MeSH terms based on their structure. The MeSH term presents substances using comma-delimited text. To the left of the comma is the name of the substance and to the right is the explanation of the substance. For example, the MeSH term for BRAF genes is “BRAF protein, Human.” By analyzing the structure, we found terms including “* protein, Human” to extract human gene symbols. However, all genes cannot be identified through structure analysis because a gene can have various synonyms. To consider this issue, we utilized the HGNC REST API service. This step is described in section 3.2.

3.2 Gene symbol extraction

To extract approved gene symbols from the filtered MeSH gene terms, we used the HGNC REST API service. The service provides a gene list for a sending query based on similarity. By using the return results and score, we can identify the approved gene symbol among synonyms.

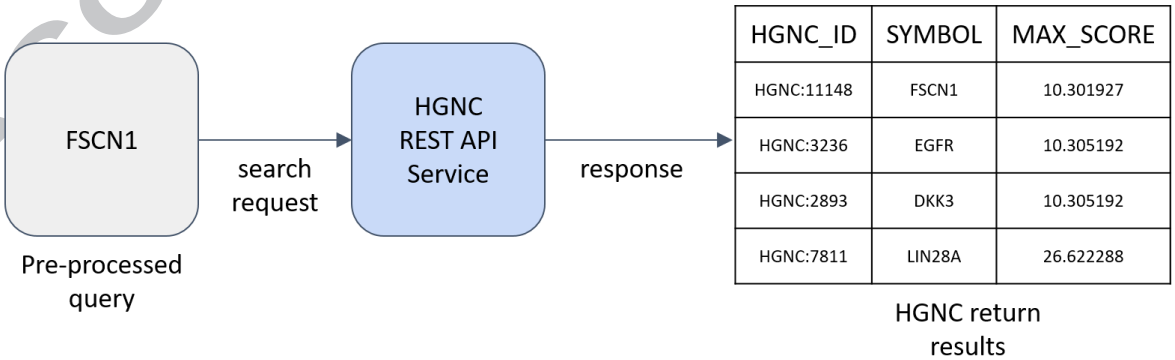

Fig. 3. Example of the HGNC REST API service

Figure 3 shows example of the HGNC REST API service. The HGNC_ID indicates the reference number for the gene generated by the HGNC. First, we extracted the substance part by removing the description part; the

substance is used as the query sent to the HGNC service. As shown in Figure 3, we can obtain approved gene symbols and similarity scores calculated by HGNC. Among the various results for each query, we used the gene symbol with the maximum score.

Table 1. Example of MeSH term processing

MeSH Term	Query	HGNC Query Response (Max score)		
		HGNC ID	Symbol	Score
FSCN1 protein, human	FSCN1	HGNC: 11148	FSCN1	10.30512
EGFR protein, human	EGFR	HGNC: 3236	EGFR	10.305192
DKK3 protein, human	DKK3	HGNC: 2893	DKK3	10.305192
LIN-28 protein, human	LIN-28	HGNC: 7811	LIN28A	26.622288

The official gene symbol is important to validate or analyze genes, because several web services and tools for gene analysis (such as GO [2] and DAVID [15, 16]) require official gene symbols as inputs. Also, the answer sets consist of approved gene symbols. However, several MeSH terms indicating genes are actually gene synonyms or unofficial gene symbols, and these can be a problem for analysis. To address this issue, an official gene symbol conversion process is needed.

Table 1 shows the overall process for extracting gene symbols from the MeSH terms. In this example, the genes *FSCN1*, *EGFR*, and *DKK3* are listed by their official gene symbols. However, *LIN28A* is described using a different symbol. Using the HGNC service, we could extract the official gene symbols for *LIN28A* and all other MeSH terms.

3.3 Relation extraction

After the gene extraction step, we built gene-literature pairs based on the PMID. These pairs were used as input for association rule mining. In this study, we used an apriori algorithm, one of several association rule techniques that is widely used to extract relationships from large data sets. By using association rule learning, we can obtain several statistic measures such as support, confidence, and lift. These measures can be utilized statistically to extract meaningful relationships and weights for relationships. After applying the apriori algorithm to the genes, we obtained directed N:M relationships between genes with lift values. In reality, the directed relationships are meaningless because we did not consider direction when we built the gene-literature pairs. We also have to convert N:M relationships to 1:1 relationships to extract gene-gene interaction pairs. For this reason, we converted directed N:M relationships to undirected 1:1 relationships by dividing the

relationships by the lift value. These processes are described in Figure 4.

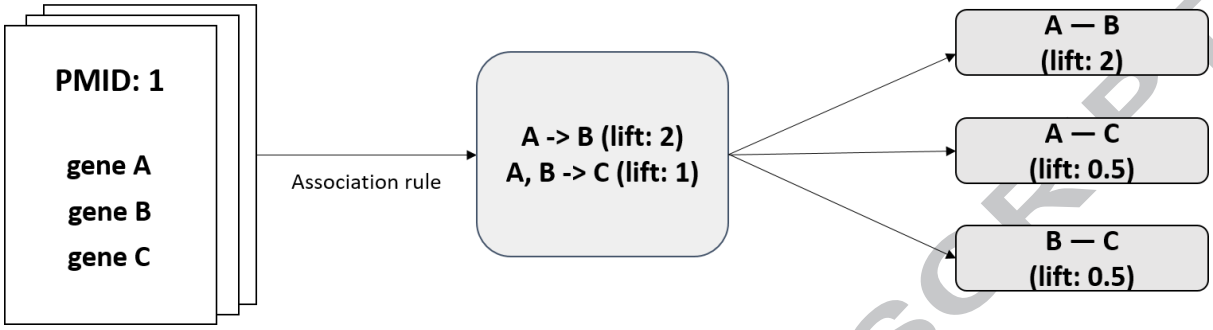

Fig. 4. Example of Relation extraction

Figure 4 shows an example of the conversion of N:M directed relationships to 1:1 undirected relationships. We used lift value as a weight of edge between genes. The lift value in association rules is described in Eq. (1):

$$lift(X \Rightarrow Y) = lift(Y \Rightarrow X) = \frac{conf(X \Rightarrow Y)}{supp(Y)} = \frac{conf(Y \Rightarrow X)}{supp(X)} = \frac{P(X \cap Y)}{P(X)P(Y)} \qquad (1)$$

The lift measures how many more times X and Y occur together than would be expected if they were statistically independent. A lift value of 1 indicates independence between X and Y. The lift is robust with regards to the rare item problem, because rare item sets with low frequency have large lift values. In other words, if a relationship is meaningful in terms of statistical analysis, then the relationship is extracted with high weight, although the relationship is not frequently observed. The weights of the relationships between genes are calculated as below:

Let U denote the set of rules by association rule mining:

$$\mathrm{U} = \{r_1, r_2, r_3 \ldots r_n\}$$

Where r_i is a rule from the association rule mining results.

Let $S(v_1,v_2)$ denote subsets of U, as follows:

$S(v_1, v_2) =$ All subsets of U that have v_1 on the left hand side and v_2 on the right hand side.

$$\mathrm{numL}(r_i) = \text{ The number of elements on the left hand side of the rule } r_i.$$

$$\mathrm{numR}(r_i) = \text{The number of elements on the right hand side of the rule } r_i.$$

Denote LiftSum as follows:

$$\mathrm{LiftSum}(v_1, v_2) = \sum_{r_i \in S(v_1,v_2)} \frac{Lift\ value\ of\ the\ rule\ r_i}{numL(r_i) * numR(r_i)}$$

Then we can obtain the weight of the edge that connects node v_1 and v_2 as follows:

$$W(v_1, v_2) = LiftSum(v_1, v_2) + LiftSum(v_2, v_1)$$

3.4 Network construction and analysis

By integrating extracted relationships, we can construct a disease-specific gene-gene network. In the case of integration, we combined edges based on shared nodes. After constructing the gene network, we analyzed it using various network analysis measures, such as degree, betweenness, and eigenvector. These measures provide a score for each node. The score is used to prioritize disease-related genes. In the case of validation for extracted genes, we conducted an answer set validation. The answer set indicates genes that are already known to be related with disease. These data were collected from several databases.

4. Results

In this section, we describe the experimental results for the IMA method of identifying disease-related genes. To demonstrate our method, we conducted a known-gene based validation. We also present comparative results with previous methods, including PRINCE [9, 37], RWRHN [24], and DISEASES [32]. These are also methods of identifying disease-related genes. In this study, we applied the IMA method to five cancers, including prostate, lung, breast, stomach, and colorectal cancer.

4.1. Data properties and parameter setting

In this study, the literature data was obtained from PubMed. Among the available information in PubMed literature, the section of substance was used to obtain MeSH terms. By analyzing the substance section, we extracted genes that are described as MeSH terms. These data are summarized in Table 2.

Table 2. Experimental data properties

Disease	# Literature	# MeSH genes	# Answer set	Reference
Prostate cancer	13717	1219	167	GHR, KEGG, OMIM, PGDB
Lung cancer	22777	351	96	GHR, KEGG, OMIM, LuGend
Breast cancer	39857	584	78	GHR, KEGG, OMIM
Stomach cancer	10096	1001	24	GHR, KEGG, OMIM,
Colorectal cancer	29086	416	50	GHR, KEGG, OMIM,

Table 2 shows the number of literature references and genes that are described as MeSH terms in PubMed literature. In the case of prostate cancer, we obtained 13,717 literature results, from which 1,219 genes were extracted by analysis. In Table 2, “Answer set” indicates the number of known disease-related genes extracted

from various databases (Genetics Home Reference (GHR) [8], Kyoto Encyclopedia of Gene and Genomes (KEGG) [17], Online Mendelian Inheritance in Man (OMIM) [30], the Lung Cancer Gene Database (LuGenD) [23], and the human Prostate Gene Database (PGDB) [20]). The answer set was used to verify the inferred genes by the proposed method and those of previous studies. GHR provides biological information by analyzing online scientific databases. From the GHR, we extracted several disease-related genes by searching by disease name. KEGG is a database that serves to understand the functions of biological systems. Among the several KEGG databases, KEGG DISEASE includes knowledge regarding disease pathways, and from this, we extracted disease-related genes for specific diseases. OMIM is a database regarding the relationships between genes and diseases. In particular, OMIM focuses on the relationships between phenotype and genotype. From these databases (GHR, KEGG, and OMIM), we confirmed disease-related genes using disease names as queries. By integrating the extracted genes, we constructed the answer set. LuGenD and PGDB are disease-specific databases. LuGenD provides lung cancer-related genes by analyzing and integrating web-based databases related to human lung cancer genes and their genomic loci, while PGDB provides human prostate cancer-related genes. We extracted prostate cancer-related genes by filtering categories.

In this experiment, we used an apriori algorithm as an association rule technique. To implement this algorithm, two parameters (confidence and minimum support value) are required. We assigned that the confidence was 0.01 and the minimum support value was 0.0001. The confidence value indicates the reliability of extracted relationships, and the minimum support value is used as a threshold for support.

4.2. Inferred Top 10-20 genes

After constructing each disease-specific gene network, we analyzed them using several centralities, including degree, betweenness, and eigenvector. Centrality is a factor that conveys how important a certain node is based on its network position, and there are several ways to evaluate the network centrality of each node. Degree centrality is a simple and straightforward concept. It is calculated by analyzing the number of nodes connected to a given node. If the degree of a node is high, it is more likely be a hub within the network. However, degree centrality only considers the local area of the node, not the entire network. In the case of betweenness centrality, this measure considers not only neighbors, but also the entire network structure, to overcome degree centrality's local distortion. This centrality is calculated by assessing the number of times a node is within the shortest path of other nodes. Nodes with high betweenness centrality are essential for the network's connectivity. Eigenvector centrality measures the influence of a node. It does not only consider the degree of each node, but also the importance of connected nodes. If a node is connected with a lot of important nodes, the node has a high eigenvector value. Based on these centralities, we inferred the "top 10" genes for each disease. These results are shown in Figure 5.

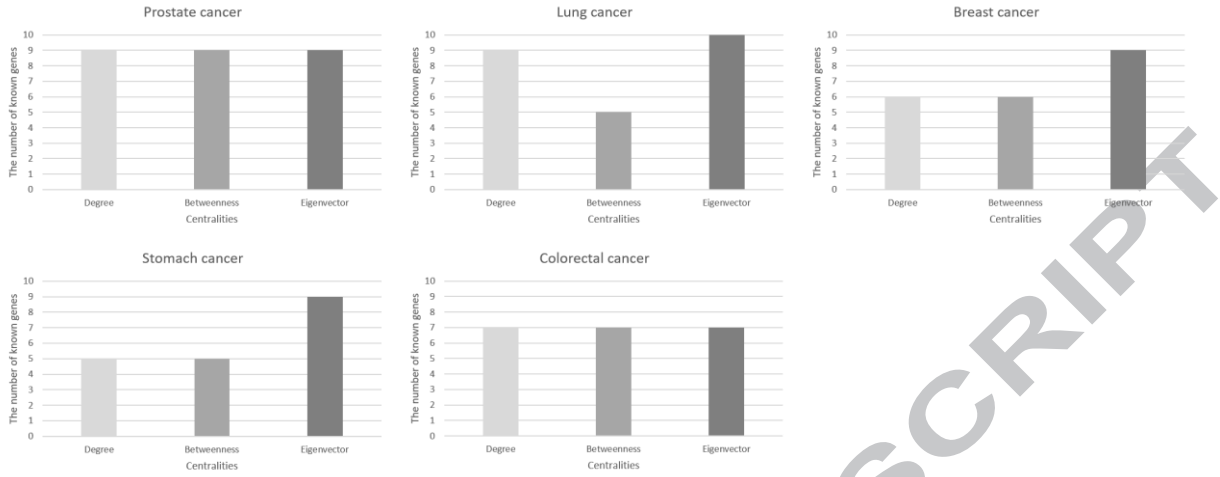

Fig. 5. Inferred "top 10" genes for each centrality

In Figure 5, the x-axis indicates type of centrality and the y-axis indicates the number of known genes among the inferred top 10 genes. Eigenvector centrality found more disease-related genes than the others for all diseases. For this reason, we used eigenvector centrality as a network analysis measure and to compare our result with those of previous studies.

To demonstrate that the proposed method is useful to infer disease-related genes, we conducted comparative experiments with three other methods: RWRHN, PRINCE, and DISEASES. RWRHN is a method of inferring disease-related genes by fusing multiple networks, including a protein-protein interaction (PPI) network, a phenotype similarity network, and known disease-gene associations. PRINCE identifies disease-related genes based on disease-disease interactions and PPI networks. In case of DISEASES, literature data is used as a resource to infer disease-related genes. The method extracts disease-gene relationships based on a co-occurrence text-mining approach.

In our comparative experiments, we compared the number of known genes (included in the answer set) among the inferred top 10-20 genes for each method. We inferred the top N genes based on the tools provided by PRINCE and DISEASES for five diseases. However, in the case of RWRHN, we extracted inferred genes from the literature. RWRHN presented the top 10 genes for lung cancer, stomach cancer, and colorectal cancer; however, the literature presented the top 20 genes for prostate cancer and breast cancer. For this reason, we have illustrated the comparative results based on the top 10 and 20 inferred genes. The results are shown in Figure 6.

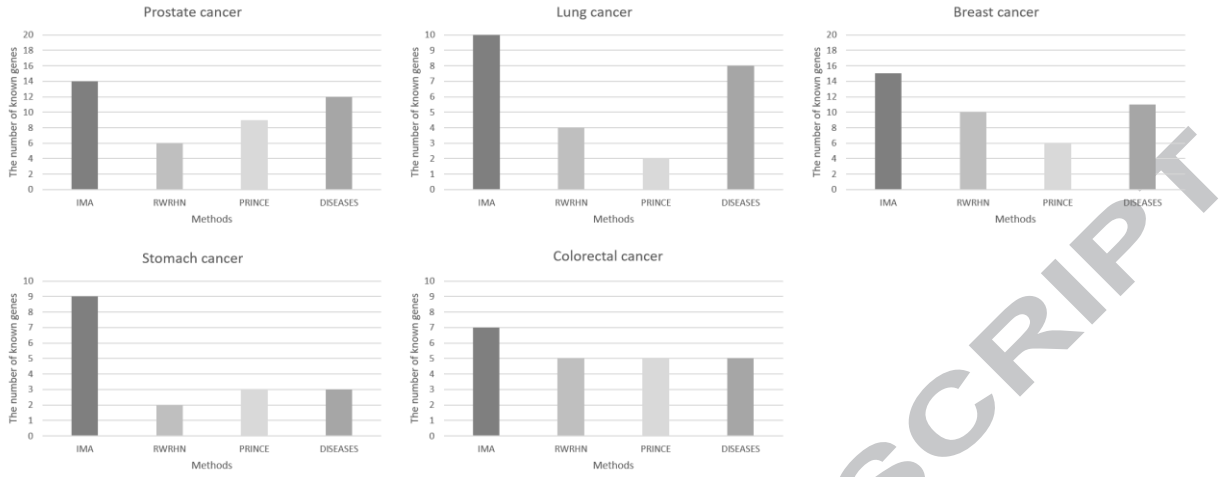

Fig. 6. Inferred top 10-20 genes for each method

Figure 6 shows the top 10-20 inferred genes for five diseases. The x-axis indicates method and the y-axis indicates the number of known genes among the inferred genes. Our proposed IMA method found more known disease-related genes than the other methods for all diseases. In particular, the IMA method found 10 known disease-related genes among the 10 inferred genes for prostate cancer. Furthermore, in the case of stomach cancer, the precision for inferring disease-related genes was 3 times that of PRINCE or DISEASES. The DISEASES method showed the highest precision for identifying disease-related genes by disease. These results showed that the proposed method found more known disease-related genes than the other methods, and IMA is robust for various diseases.

4.3. Top N precision and recall

To compare the top N inferred genes when the value of N was high, we calculated the precision and recall for each method. Precision was calculated by dividing the number of disease-related known genes by the number of inferred genes. The recall was calculated by dividing the number of disease-related known genes among the inferred Top N genes by the number in the answer set. We compared our results to those of the PRINCE and DISEASES algorithms for five diseases. The results are shown in Figure 7.

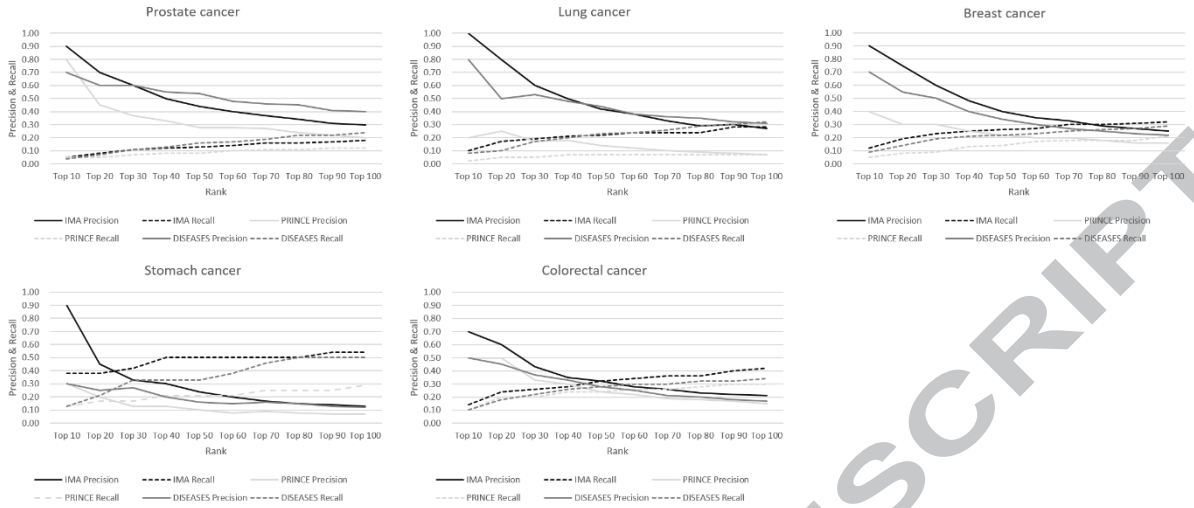

Fig. 7. Precision and recall for the top N genes for five cancers by each method

Figure 7 shows the precision and recall values of inferred genes using each method. The x-axis indicates the number of inferred genes and the y-axis indicates the precision and recall values. For breast, stomach, and colorectal cancers, the IMA method found more disease-related known genes than other methods for all N values. For prostate cancer, IMA had higher precision values than the other methods for the top 10–30 inferred genes. However, when N was larger than 30, DISEASES found more disease-related known genes than our method. In the case of lung cancer, our proposed method showed higher precision for the top 10–40 inferred genes, and for the top 50–100 sections, IMA and DISEASES showed similar precision values. When N was small, IMA found more known disease-related genes than the other methods. This shows that the IMA method is well designed as a ranking algorithm.

4.4 Literature validation

The size of the answer set was too small to validate inferred genes. To address this issue, we conducted literature validation for inferred genes that were not validated by the answer set. Literature validation involves finding evidence for disease-gene relationships from the literature. In this experiment, we used the inferred top 20 genes for each disease and conducted literature validation for candidate disease-related genes. The inferred top 20 genes are shown in Table 3.

Table 3. Inferred top 20 genes by the IMA method

Rank	Prostate cancer	Lung cancer	Breast cancer	Stomach cancer	Colorectal cancer
1	**AR**	**EGFR**	**TP53**	**TP53**	**KRAS**
2	**PTEN**	**KRAS**	**PTEN**	**VEGFA**	BRAF
3	**ERG**	**TP53**	**ERBB2**	**CDH1**	**MLH1**
4	**CDKN1A**	**BRAF**	**BRCA1**	**KRAS**	**CTNNB1**
5	**TP53**	**PIK3CA**	**ATM**	**ERBB2**	**TP53**
6	**NKX3-1**	**MET**	**BRCA2**	MLH1	**PIK3CA**
7	**TMPRSS2**	**ERBB2**	**CHEK2**	**CTNNB1**	**MSH2**
8	**MYC**	**PTEN**	EGFR	**PIK3CA**	PTEN
9	**EZH2**	**STK11**	**PALB2**	**EGFR**	**APC**
10	CASP3	**AKT1**	**CDH1**	**MET**	EGFR
11	**GSTP1**	**ROS1**	**PIK3CA**	PTEN	**NRAS**
12	AKT1	**RET**	**BRIP1**	STAT3	**SMAD4**
13	**CCND1**	**NRAS**	**NBN**	RUNX3	**PMS2**
14	**STAT3**	**MAP2K1**	**CCND1**	CCND1	MGMT
15	FOXA1	MTOR	BLID	CDKN1A	**MSH3**
16	**CDKN1B**	KIF5B	**AKT1**	MTOR	FBXW7
17	MDM2	**EML4**	CDKN1A	MMP2	**AKT1**
18	VEGFA	**ERCC1**	CD44	CASP3	RUNX3
19	E2F1	TTF1	**MYC**	BAX	MET
20	**HOXB13**	NFE2L2	MDM2	PTGS2	IGF2

Table 3 shows the top 20 inferred genes for each disease. The genes with gray background are validated i.e., they have a relationship with the disease by answer set. The others indicate disease-related candidate genes. To validate these genes, we conducted literature validation as described in Table 4.

Table 4. Literature validation of candidate disease-related genes

Gene	Disease	Rank	PMID
Evidence			
CASP3	Prostate cancer	10	26507126
procaspase-3 and cleaved caspase-3 may help to identify prostate cancer patients at risk of progression.			
AKT1	Prostate cancer	12	28363000
Our study found that the variant genotype CT of rs3730358 of AKT1 was associated with a decreased risk of prostate cancer, which suggested that this polymorphism could play an important role in the development of the disease.			
FOXA1	Prostate cancer	15	22138582
These findings suggest FOXA1 overexpression as a novel mechanism inducing castration resistance in prostate cancer.			
MDM2	Prostate cancer	17	22902907
In summary, our meta-analysis showed that the MDM2 309G variant was significantly associated with a decreased PCa risk			
VEGFA	Prostate cancer	18	24435801
Our findings indicate that the VEGF-A ATTGC haplotype may predict clinical recurrence in prostate cancer patients treated with radiotherapy.			
E2F1	Prostate cancer	19	19276347
These studies uncovered a novel mechanism for E2F1-induced suppression of apoptosis in prostate cancer.			
MTOR	Lung cancer	15	25893736
Hence, our study reveals a new dimension in mTOR-rictor relationship, where rictor stands to be a suitable therapeutic target for lung cancer.			
KIF5B	Lung cancer	16	25047660
Our data suggest that KIF5B-RET promotes the cell growth and tumorigenicity of non-small cell lung cancers through multilevel activation of STAT3 signaling, providing possible strategies for the treatment of KIF5B-RET positive lung cancers.			

TTF1	Lung cancer	19	26912193
In summary, this study provides evidence that TTF-1 may reprogram lung cancer secreted proteome into an antiangiogenic state, offering a novel basis to account for the long-standing observation of favorable prognosis associated with TTF-1(+) lung adenocarcinomas.			
NFE2L2	Lung cancer	20	27477511
We previously reported that oncogenic Kras induced the redox master regulator Nfe2l2/Nrf2 to stimulate pancreatic and lung cancer initiation.			
EGFR	Breast cancer	8	27569656
These results suggest that SIAH and EGFR are two prognostic biomarkers in breast cancer with lymph node metastases.			
BLID	Breast cancer	15	24532431
Our finding gives a new clue that BLID might be a potential indicator for progression of breast cancer in the future.			
CDKN1A	Breast cancer	17	24005533
We examined several p21/CIP1 genotypes in the patients with breast cancer and found that there is no significant association of these p21/CIP1 genotypes with the risk of developing breast cancer.			
CD44	Breast cancer	18	25909162
Taken together, it was supposed that CD44 promotes tumorigenesis through the interaction and nuclear-translocation of its intracellular domain and stemness factors.			
MDM2	Breast cancer	20	25326024
Our result revealed that 40-bp ins/del polymorphism in the promoter of MDM2 increased the risk of breast cancer in an Iranian population.			
MLH1	Stomach cancer	6	23098428
This study revealed that hMLH1 hypermethylation is strongly associated with GC and suggested roles for epigenetic changes in stomach cancer causation in the Kashmir valley.			
PTEN	Stomach cancer	11	25823029
Overall, our findings suggest that inhibition of Notch signaling can be employed as a PTEN activator, making it a potential target for gastric cancer therapy.			

STAT3	Stomach cancer	12	25822437
The Stat3 role has been recently highlighted in carcinogenesis of the diffuse type of gastric cancer.			
RUNX3	Stomach cancer	13	24447545
Effective therapy targeting the RUNX3 pathway may help control gastric cancer cell invasion and metastasis by inhibiting the EMT.			
CCND1	Stomach cancer	14	25202078
Cell-cycle regulation may play a role in gastric cancer initiation and development and the CCND1 A870G genotype may be a useful biomarker for detection of early gastric cancer.			
CDKN1A	Stomach cancer	15	24619835
The expression of p21 was an independent prognostic factor for patients with AFP-producing gastric cancer.			
MTOR	Stomach cancer	16	26287940
These findings suggest that potentially functional SNPs of mTOR may individually or collectively contribute to the risk of gastric cancer.			
MMP2	Stomach cancer	17	15929171
MMP-2 may play an important role in the development of invasion and metastasis of gastric cancer.			
CASP3	Stomach cancer	18	25002346
Gene polymorphism and haplotype of caspase 3 can increase gastric cancer risk.			
BAX	Stomach cancer	19	25267570
SNPs located in BAX and CDKN1A genes are closely associated with clinical outcomes in patients with gastric cancer.			
PTGS2	Stomach cancer	20	25339021
This meta-analysis suggested that the -765G>C polymorphism of the COX-2 gene is a potential risk factor for digestive system cancer in Asians and Africans and gastric cancer overall.			
BRAF	Colorectal cancer	2	27034263
BRAF mutation may have different prognostic implications in early- and late-stage colorectal cancer.			
PTEN	Colorectal cancer	8	26403191

Expression of PTEN and RPA1 shows strong correlation in colorectal cancer.			
EGFR	Colorectal cancer	10	23800895
EGFR polymorphisms can serve as prognostic predictors for CRC patients receiving 5-FU-based chemotherapy.			
MGMT	Colorectal cancer	14	27006309
In conclusion, MGMT was an important in vitro predictor of TMZ activity in CRC.			
FBXW7	Colorectal cancer	16	19739118
FBXW7 expression provides a prognostic factor for patients with CRC.			
RUNX3	Colorectal cancer	18	22234069
The serum RUNX3 promoter hypermethylation may be a promising biomarker for the early diagnosis of ESCC, GC and CRC, which was further confirmed by combining with CEA and CA19-9.			
MET	Colorectal cancer	19	26459369
c-MET is a new promising target that may help in understanding the pathogenesis of CRC, and to be used as independent prognostic biomarker to predict local disease recurrence in CRC.			
IGF2	Colorectal cancer	20	27337954
This study supports the concept of direct autocrine/paracrine tumor cell activation through IGF2 and a shows role of IGF2 in CRC proliferation, adhesion and, to a limited extent, apoptosis.			

Table 4 shows candidate genes, rank, PMID, and key sentences. PMID indicates the reference number generated by PubMed used to access the literature. Evidence comprises a sentence describing the relationship between gene and disease. These sentences are extracted from the literature using the PMID. For example, we confirmed that the CASP3 gene can be used to identify progression of prostate cancer. As shown in Table 4, we found evidence for all disease-related candidate genes which were extracted by the IMA method. These results showed that the proposed approach is useful to infer disease-related candidate genes. Through literature validation, we presented 34 meaningful disease-related candidate genes.

5. Discussion

In this section, we illustrate the gene network for each disease and present a summary of the experimental results. We constructed five disease-specific gene networks using MeSH terms and association rules. These are described in the network visualization section. We present various experimental results for inferred genes. We have also summarized the experimental data for inferred genes in the results summary section.

5.1. Network visualization

To visualize the constructed gene network, we used a Gephi [3] network visualization tool. The size of the top 20 inferred genes (nodes) are proportional to eigenvector centrality and the other nodes are presented smaller for visibility. In this visualization, we only considered the node property. The networks are shown in Figure 8 and the properties for each network are described in Table 5.

Table 5. Properties for gene networks

	Prostate cancer	Lung cancer	Breast cancer	Stomach cancer	Colorectal cancer
# node	1,227	351	584	1001	416
# edge	3,316	562	994	2122	757

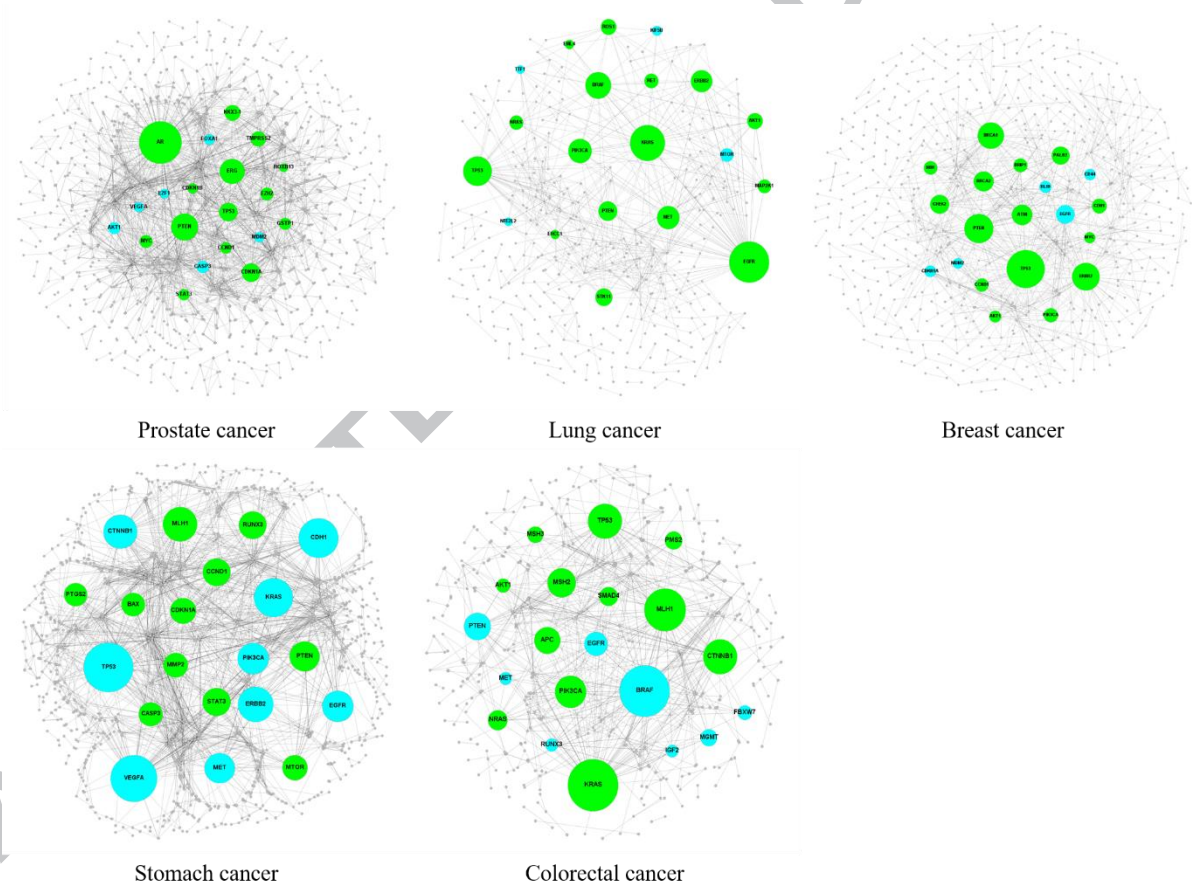

Fig. 8. Gene network for five diseases

In Figure 8, the green nodes indicate the known genes and the blue nodes indicate confirmed candidate genes. The number of nodes and edges are described in Table 5. As shown in Figure 8 and Table 5, the IMA method constructed a small gene-gene interaction network. The network also has meaningful disease-related genes as hub nodes. In the case of co-occurrence based text mining, the size of the constructed gene network is too large to present candidate genes, because when the number of genes exceeds 30,000 it generates a vast number of

edges. From the perspective of network size, networks constructed using the IMA method are more useful for inferring candidate disease-related genes than conventional text-mining approaches.

5.2. Results summary

To present the experimental results briefly, we have summarized the results in tables. Table 6 indicates terms and definitions and Table 7 is a summary of the top 20 inferred genes.

Table 6. Terms definition

Term	Description
Inferred genes	Inferred genes by the IMA method
Known genes	Known disease-related genes included in the answer set
Candidate genes	Inferred genes not validated by the answer set
Confirmed candidate genes	Candidate genes validated by literature validation
Percentage of known genes	(the number of known genes) / (the number of inferred genes)
Percentage of candidate genes	(the number of candidate genes) / (the number of inferred genes)
Percentage of confirmed candidate genes	(the number of confirmed candidate genes) / (the number of candidate genes) * 100

Table 6 defines terms that are used in Table 7. Inferred genes consist of known genes and candidate genes. If the candidate genes are validated by literature validation, the genes are considered as confirmed candidate genes.

Table 7. Summary for top 20 inferred genes

	Prostate cancer	Lung cancer	Breast cancer	Stomach cancer	Colorectal cancer
Inferred genes	20	20	20	20	20
Known genes	14	16	15	9	12
Candidate genes	6	4	5	11	8
Confirmed candidate genes	6	4	5	11	8
Percentage of known genes	0.70	0.80	0.75	0.45	0.60
Percentage of candidate genes	0.30	0.20	0.25	0.55	0.40
Percentage of confirmed candidate genes	100 %	100 %	100%	100%	100%

Table 7 shows the top 20 inferred genes extracted using the IMA method. In the case of prostate cancer, we found 16 disease-related genes among the top 20 inferred genes. We confirmed that six candidate genes have evidence that confirms their relationships with prostate cancer. These results showed that the proposed method found disease-related genes with high precision and inferred meaningful candidate disease-related genes. For prostate cancer, lung cancer, and breast cancer, the size of the answer set was larger than those of the other two cancers. For this reason, the percentage of known genes are higher for these cancers than for stomach cancer and colorectal cancer. Through top 20 gene analysis, we presented 34 potential candidate disease-related genes.

5.3. Meaning of MeSH terms

MeSH terms comprise a set vocabulary controlled by the NLM, which is used to index journal articles. MeSH has a hierarchical structure and is well organized to extract biological information. Most articles in PubMed include related MeSH terms. Biomedical researchers generate MeSH terms, and due to this, they can be used to obtain accurate and useful information for biomedical text. Because of their advantages, several studies have reported the development of tools based on MeSH terms.

Polysearch [7] and Polysearch2 [21] are popular biomedical text mining tools that find related biological knowledge (such as diseases, drugs, genes, and MeSH terms) for a given query. To find relationships between MeSH terms and prostate cancer, one can pose a query such as “Given Diseases prostate cancer, Find MeSH Terms”, and obtain prostate-related MeSH terms with associated hit scores. In this study, MeSH terms were used as a biological entity. By using these tools, we can extract a vast number of biological relationships between biological entities.

Because MeSH terms are manually curated, there can be delays in allocating MeSH terms to published literature. To address this issue, Cha et al. [5] proposed a method (called GRiD) to extract more biomedical literature data using MeSH terms. They developed a classifier by analyzing literature that included MeSH terms, and based on their classification rules, they were able to extract data from additional biomedical literature that did not include MeSH terms.

Polysearch and Polysearch2 use MeSH terms as one of several biological entities, and GRiD utilizes MeSH terms to classify biomedical literature. These methods use MeSH terms without filtering for term types, such as genes or drugs. Conversely, our method analyzes MeSH terms to specifically extract genes among them. After extracting the genes, it identifies gene-gene interactions using association rules and constructs gene-gene interaction networks. It also uses several measures to score genes and gene-gene interactions, such as lift value and eigenvector centrality.

The proposed IMA method is more reliable and concise than existing text-mining methods. As professional researchers validate MeSH terms, they can be considered highly reliable, having been verified once more than general biomedical text. Analysis of MeSH terms is also an advantage for mining procedures. Without analyzing biomedical text, such as abstracts and titles, we can obtain key biological terms related to the literature. Furthermore, by using MeSH terms, we can extract genes without considering gene synonyms.

When constructing gene-gene interaction networks, we used only MeSH terms as data sources. However, the IMA method showed better performance for five cancers than existing methods. These results demonstrate that IMA is a powerful method to identify disease-related genes.

5.4. Comparison for IMA and co-occurrence approach

The proposed method used association rule mining to extract gene-gene interactions from MeSH terms. By using support and confidence values, our method filtered out statistically meaningless relationships. The lift value is also used in the weight of extracted relationships. An advantage of the lift value is that although a relationship may not appear frequently, it can obtain high weight if the relationships are statistically meaningful. To demonstrate these advantages, we have presented results comparing our method with a co-occurrence-based approach, which are widely used to extract relationships from the literature. The concept behind co-occurrence-based approaches is that if two terms appear in the same location (such as sentence, paragraph, or article), these terms are considered to have a relationship. Therefore, this concept can be used as a direct way to extract relationships. Generally, in the co-occurrence concept, the frequency, which indicates the number of locations that include both MeSH terms in a relationship, is used to calculate the weight of gene-gene interactions. Using MeSH terms and the co-occurrence approach, we constructed gene-gene interaction networks for five diseases, and presented comparative results for precision and network properties. In this experiment, the co-occurrence approach also used eigenvector centrality as a measure to rank genes.

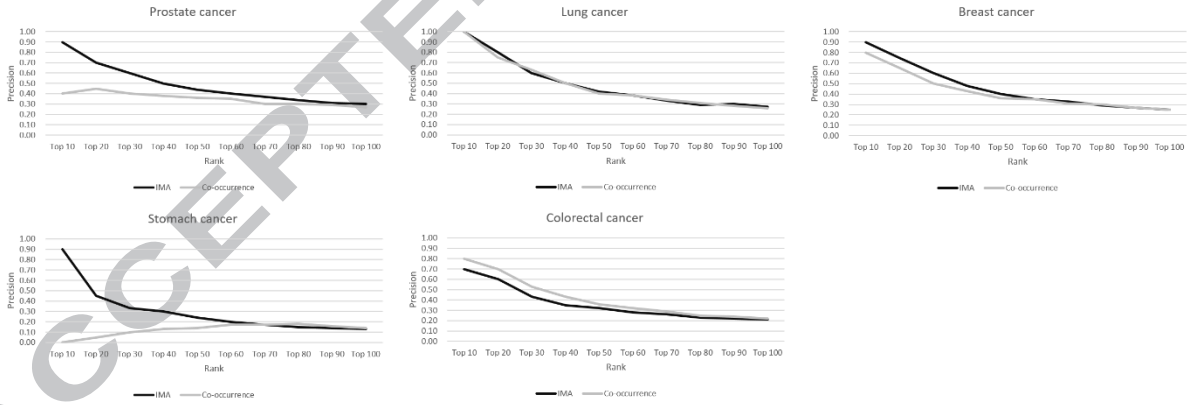

Fig. 9. Comparison of IMA and co-occurrence approach

Figure 9 shows the precision for the Top N inferred genes for five diseases. In prostate, lung, breast, and stomach cancers, the IMA method had high precision when N was small. These results demonstrate that the IMA method is more useful in extracting disease-related genes than the co-occurrence approach. In the case of colorectal cancer, the co-occurrence approach showed higher precision than the IMA method. Therefore, among the five diseases, our method showed better precision in four cases. These results demonstrate that the proposed method is more robust than the co-occurrence approach for diseases. Our method also constructs useful gene-gene interaction networks by filtering meaningless relationships. To illustrate this advantage, we present the network properties for the five diseases in Table 8.

Table 8. Network properties of the IMA and co-occurrence based approach

Disease	Prostate cancer		Lung cancer		Breast cancer		Stomach cancer		Colorectal cancer	
Method	IMA	CO	IMA	CO	IMA	CO	IMA	CO	IMA	CO
Nodes	1,227	2,313	351	3,193	584	3,784	1,001	1,677	416	2,767
Edges	3,316	7,119	562	11,206	994	16,192	2,122	4,253	757	9,378
Proportion of known genes (%)	6.28	3.63	11.68	2.19	5.31	1.43	2.00	1.25	6.01	1.37

Co, co-occurrence-based approach

Table 8 shows the numbers of nodes and edges in each network. The proportion of known genes in each network was calculated by dividing the number of known genes by the number of nodes. For all diseases, the IMA had a higher proportion of known genes than the co-occurrence-based approach. Furthermore, the IMA constructed concise gene-gene interaction networks in terms of the number of nodes and edges, and showed better precision than the co-occurrence approach. These results demonstrate that the IMA method is sufficiently powerful to extract disease-related genes, and that association rule mining is a more useful tool to extract gene-gene interactions from MeSH terms than a co-occurrence approach.

5.5. Comparison of IMA and 2 × 2 association analysis

The 2 × 2 association analysis is an approach that uses a 2 × 2 contingency table and gene-gene interactions to calculate statistical measures for pair-wise associations. Using the 2 × 2 contingency table, we could obtain several statistics describing relationships such as support, lift, odds ratio, relative risk, chi-squared, and p-values to identify meaningful relationships. Furthermore, the 2 × 2 association analysis also filtered meaningless relationships using a support value and calculated the weight of relationships using the lift value.

In this experiment, we extracted gene-gene interactions using the co-occurrence approach from MeSH term and filtered the extracted interactions using a support value (which is the same as the IMA). Then, we constructed the gene-gene interaction network using the lift value as a weight of relationship and analyzed the network using eigenvector centrality.

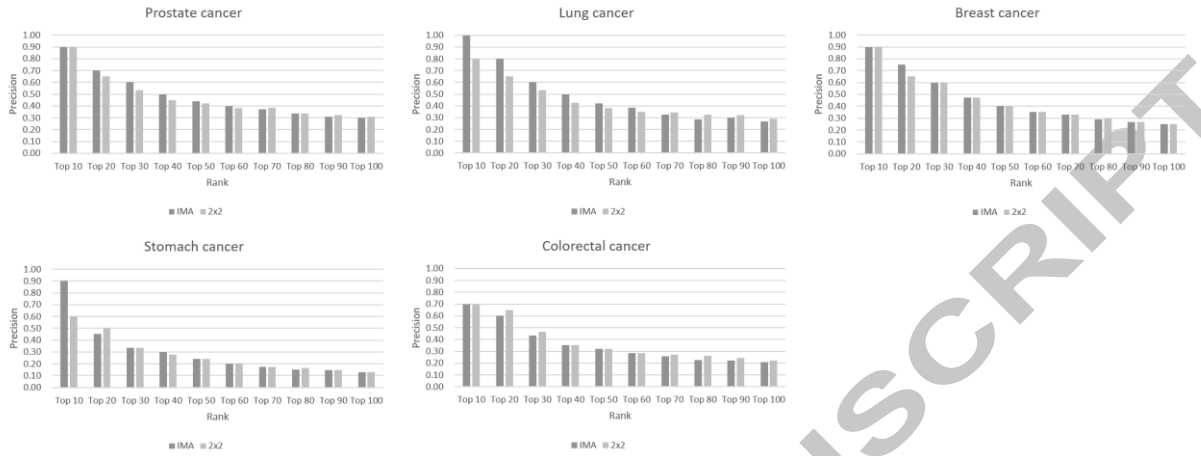

Fig. 10. Comparison of IMA and 2 × 2 association analysis

Figure 10, which shows the precision for the Top N inferred genes for five diseases, illustrates that the IMA method had a comparable or higher precision than that of the 2 × 2 association analysis when N was small (except for the Top 20 of colorectal cancer and Top 20 of stomach cancer). These results demonstrate that the IMA method was more useful for extracting disease-related genes than the 2 × 2 association analysis. Both the IMA method and the 2 × 2 association analysis are similar in that gene pairs are extracted using statistical measures. However, the IMA method generates a large item set, which has support value more than threshold dose, then it extracts all possible gene pairs in this set so that a higher weight can be ascribed to the interactions derived from this item set. Therefore, the IMA method can measure the weight of interaction more significantly than the 2 × 2 association analysis can, including all gene pairs that can be extracted using the 2 × 2 association analysis, which summarizes the difference between the IMA method and 2 × 2 association analysis.

This difference could eventually have a positive effect on identifying gene-gene interactions related to diseases or biological processes. Because many biological processes or disease-related genetic factors can be explained by the interaction of two or more genes, the method for exacting genetic interactions should adequately model these characteristics. Therefore, we concluded that the IMA method was more suitable than the 2 × 2 association analysis to reflect the various factors affecting biological functions or disease relevance.

5.6. Comparison of methods based on publication date

To increase the fairness of method comparison, in this section, we present comparative results based on the publication dates of previously described methods. PRINCE was published in 2010, and RWRHN and DISEASES were published in 2015. Therefore, for PRINCE, we used literature data up to 2009, and in the other cases, we used literature data up to 2014. These results are shown in Figures 10 and 11.

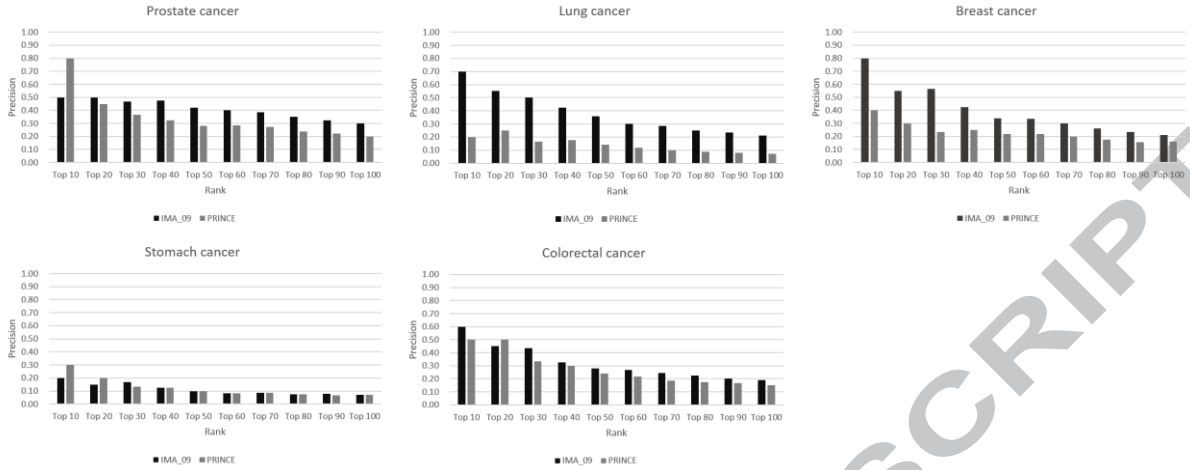

Fig. 11. Comparison results for IMA (up to 2009) and the PRINCE algorithm

Figure 11 shows the precision of each method for each disease. IMA_09 indicates that literature data up to 2009 was used. For lung, breast, and colorectal cancers, the proposed method had higher precision values in the overall ranks, except for colorectal cancer when the Top 20 genes were compared. For prostate cancer, IMA showed higher precision, except when the Top 10 genes were compared. For stomach cancer, IMA showed equal or higher precision except when the Top 10 and 20 genes were compared. These results demonstrate that the proposed method is more useful in identifying disease-related genes, although in some cases, the IMA method showed lower precision. However, when the amount of data is large, the IMA method was more powerful than previous algorithms. This issue can be addressed by using a large amount of literature data.

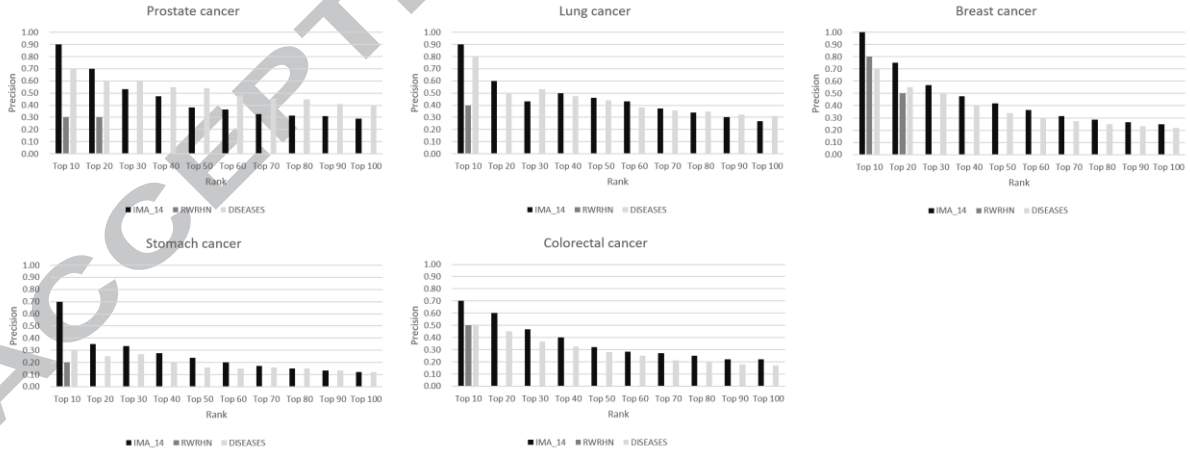

Fig. 12. Comparison results for IMA (up to 2014), RWRHN, and DISEASES

Figure 12 shows the precision of each method for each disease. For RWRHN, we present experimental results for the Top 10–20 genes extracted directly from the research paper. Our method showed higher precision than RWRHN for all diseases studied. Overall, our method also showed higher precision than DISEASES. These results demonstrate that the proposed method is more powerful than existing methods to identify disease-related genes.

5.7. Validation of candidate genes with DAVID

To validate candidate genes, we used DAVID [15, 16], which provides functional annotation tools to understand the biological meaning of genes. By inserting genes as an input and selecting options for select identifiers (official_gene_symbol) and species (homo sapiens), we can obtain several pieces of information for input genes. Among them, we used "Disease" analysis, and three types of results (GAD_DISEASE, GAD_DISEASE_CLASS, and OMIM_DISEASE) were obtained. We used GAD [4] analysis to validate candidate genes, as the OMIM database was used in our answer sets. GAD was constructed by analyzing genetic association studies for many common disease types. Since it contains disease-gene relationships with genomic and molecular information, we considered that it could be used as an alternative answer set for evaluation.

Table 9. GAD validation for candidate genes.

IMA_disease	Gene symbol	GAD_disease	GAD_disease_class
Prostate cancer	*CASP3*		Cancer
Prostate cancer	*AKT1*	Prostate cancer	Cancer
Prostate cancer	*FOXA1*		Cancer
Prostate cancer	*MDM2*	Prostate cancer	Cancer
Prostate cancer	*VEGFA*	Prostate cancer	Cancer
Prostate cancer	*E2F1*		Cancer
Lung cancer	*MTOR*	Lung cancer	Cancer
Lung cancer	*KIF5B*		
Lung cancer	*TTF1*		Cancer
Lung cancer	*NFE2L2*	Lung neoplasms, Lung injury	Cancer
Breast cancer	*EGFR*	Breast cancer, Breast neoplasms	Cancer
Breast cancer	*BLID*		
Breast cancer	*CDKN1A*	Breast cancer	Cancer
Breast cancer	*CD44*	Breast cancer	Cancer
Breast cancer	*MDM2*	Breast cancer, Breast neoplasms	Cancer
Stomach cancer	*MLH1*	Stomach cancer, Stomach neoplasms	Cancer

Stomach cancer	*PTEN*	Stomach neoplasms	Cancer
Stomach cancer	*STAT3*		Cancer
Stomach cancer	*RUNX3*	Stomach neoplasms	Cancer
Stomach cancer	*CCND1*	Stomach cancer, Stomach neoplasms	Cancer
Stomach cancer	*CDKN1A*	Stomach cancer	Cancer
Stomach cancer	*MTOR*		Cancer
Stomach cancer	*MMP2*	Stomach cancer	Cancer
Stomach cancer	*CASP3*	Stomach cancer	Cancer
Stomach cancer	*BAX*		Cancer
Stomach cancer	*PTGS2*	Stomach cancer, Stomach neoplasms	Cancer
Colorectal cancer	*BRAF*	Colorectal cancer, Colorectal neoplasms	Cancer
Colorectal cancer	*PTEN*	Colorectal cancer, Colorectal neoplasms	Cancer
Colorectal cancer	*EGFR*	Colorectal cancer, Colorectal neoplasms	Cancer
Colorectal cancer	*MGMT*	Colorectal cancer, Colorectal neoplasms	Cancer
Colorectal cancer	*FBXW7*		Cancer
Colorectal cancer	*RUNX3*		Cancer
Colorectal cancer	*MET*		Cancer
Colorectal cancer	*IGF2*		Cancer

Table 9 shows validation results for candidate genes based on GAD analysis in DAVID. "IMA_disease" indicates candidate gene-related diseases inferred by the IMA method. "GAD_disease" indicates the candidate gene-related diseases by GAD analysis, and "GAD_disease_class" indicates the candidate gene-related disease class by GAD analysis. For example, *AKT1* was associated with prostate cancer, and cancer as a disease category in GAD analysis. A summary of the GAD validation is described in Table 10.

Table 10. Summary of GAD validation for candidate genes

	Prostate cancer	Lung cancer	Breast cancer	Stomach cancer	Colorectal cancer	Total
Candidate genes	6	4	5	11	8	34

Matching genes by disease	3	2	4	8	4	21
Matching genes by category (cancer)	6	3	4	11	8	32
Percentage of matching disease	0.50	0.50	0.80	0.73	0.50	0.62
Percentage of matching category	1.00	0.75	0.80	1.00	1.00	0.94

Table 10 shows a summary of GAD validation of candidate genes for each disease. “Matching genes by disease” indicates the number of candidate genes that were matched to their associated disease by GAD. “Matching genes by category” indicates the number of candidate genes that were matched to the disease category cancer by GAD. In the case of prostate cancer, we inferred 6 candidate genes and 3 were validated by GAD analysis. The other candidate genes were not directly validated as prostate cancer-related by GAD analysis; however, GAD analysis of disease categories indicated that they are related to cancer. Among the 34 inferred candidate genes, 21 were related to their target diseases, and 32 were related to cancer.

6. Conclusions

In this study, we developed and tested the IMA method to infer disease-related genes using MeSH terms and association rules. Using MeSH terms, we obtained gene symbols from the literature data. These extracted gene symbols were used to extract gene-gene interactions based on association rules; we then constructed gene networks and analyzed them to infer disease-related genes. Using the proposed method, we inferred the top 20 disease-related genes for five cancers. To demonstrate the merit of the experimental results, we presented comparative data for inferred genes using systems published in previous studies. These results showed that the IMA method found more disease-related genes than the other methods for all five cancers. We also presented literature validation results confirming that the candidate genes are possibly related with these diseases. Through validation, we found evidence that 34 inferred candidate genes are related to diseases. These results showed that the proposed method is useful for inferring disease-related candidate genes. We also constructed a small gene-gene network for each disease. These networks are better for inferring candidate genes than conventional co-occurrence based gene-gene networks.

In this study, we used only gene symbols among several possible MeSH terms. In further work, we will utilize other biological terms, such as drugs, symptoms, and therapies. Furthermore, we will design a method to

analyze the MeSH terms by considering several rules of MeSH. In this study, we applied the IMA method to five cancers. We will apply our method to other genetic diseases, such as Alzheimer's disease and Parkinson's disease, and plan to design various versions of the IMA method based on other data mining techniques that extract relationships.

Acknowledgement

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

References

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[2] Ashburner et al. Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9. Online at Nature Genetics.

[3] Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." ICWSM 8 (2009): 361-362.

[4] Becker, Kevin G., et al. "The genetic association database." Nature genetics 36.5 (2004): 431-432.

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

[6] Chen, Jing, Bruce J. Aronow, and Anil G. Jegga. "Disease candidate gene identification and prioritization using protein interaction networks." BMC bioinformatics 10.1 (2009): 73.

[7] Cheng, Dean, et al. "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites." Nucleic acids research 36.suppl_2 (2008): W399-W405.

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

[9] 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.

[10] 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.

[11] Gridach, Mourad. "Character-Level Neural Network for Biomedical Named Entity Recognition." Journal of Biomedical Informatics (2017).

[12] Harpaz, Rave, Herbert S. Chase, and Carol Friedman. "Mining multi-item drug adverse effect associations in spontaneous reporting systems." BMC bioinformatics 11.9 (2010): S7.

[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] Hoffmann, Robert, and Alfonso Valencia. "A gene network for navigating the literature." Nature genetics 36.7 (2004): 664-664.

[15] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57.

[16] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13

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

[18] Kim, Jeongwoo, et al. "LGscore: A method to identify disease-related genes using biological literature and Google data." Journal of biomedical informatics 54 (2015): 270-282.

[19] Krauthammer, Michael, et al. "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease." Proceedings of the National Academy of Sciences of the United States of America 101.42 (2004): 15148-15153.

[20] 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.

[21] Liu, Yifeng, Yongjie Liang, and David Wishart. "PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more." Nucleic acids research43.W1 (2015): W535-W542.

[22] Lou, Yinxia, et al. "A transition-based joint model for disease named entity recognition and normalization." Bioinformatics (2017): btx172.

[23] LuGend: Lung cancer gene database <www.bioinformatics.org/lugend/>

[24] 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.

[25] Maher, Christopher A., et al. "Transcriptome sequencing to detect gene fusions in cancer." Nature 458.7234 (2009): 97-101.

[26] Montojo, Jason, et al. "GeneMANIA: Fast gene network construction and function prediction for Cytoscape." F1000Research 3 (2014).

[27] Murugesan, Gurusamy, et al. "BCC-NER: bidirectional, contextual clues named entity tagger for gene/protein mention recognition." EURASIP Journal on Bioinformatics and Systems Biology 2017.1 (2017): 7.

[28] Nahar, Jesmin, et al. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40.4 (2013): 1086-1093.

[29] NLM: National library of medicine <https://www.nlm.nih.gov/>

[30] OMIM: Online Mendelian Inheritance in Man. McKusick-Nathans Institute of Genetics Medicine, Johns Hopkins University (Baltimore, MD) <http://omim.org/>

[31] Ortutay, Csaba, and Mauno Vihinen. "Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies." Nucleic acids research 37.2 (2008): 622-628.

[32] Pletscher-Frankild, Sune, et al. "DISEASES: Text mining and data integration of disease–gene associations." Methods 74 (2015): 83-89.

[33] PMC: PubMed Central <www.ncbi.nlm.nih.gov/pmc>

[34] PubMed: MEDLINE Retrieval on the World Wide Web <www.ncbi.nlm.nih.gov/pubmed>

[35] Shim, Jung Eun, et al. "GWAB: a web server for the network-based boosting of human genome-wide association data." Nucleic Acids Research (2017).

[36] TAUB, FLOYD, E., JAMES M. DeLEO, and E. BRAD THOMPSON. "Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs." DNA 2.4 (1983): 309-327.

[37] 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).

[38] Wright, A., et al. "Validation of an association rule mining-based method to infer associations between medications and problems. ppl Clin Inf 2013; 4: 100–109 http://dx. doi. org/10.4338/ACI-2012-12-RA-0051 For personal or educational use only." No other uses without permission. All rights reserved. Downloaded from www. aci-journal. org on 501 (2013): 76802.

Highlights

- We propose a method to identify disease-related genes using MeSH terms and association rules.
- We construct gene-gene interaction networks for each disease.
- We identify disease-related genes and meaningful disease-related candidate genes.

Graphical abstract