IDO: Inferring Describable Disease-Gene Relationships Using Opinion Sentences

Jeongwoo Kim
Yonsei University
Department of Computer Science,
Yonsei University, Seoul, Korea
+82-2-2123-7757
jwkim2013@cs.yonsei.ac.kr

Youngmi Yoon
Gachon University
Department of Computer Engineering,
Gachon University, Seongnam, Korea
+82-10-7261-2967
ymyoon@gachon.ac.kr

Sanghyun Park*
Yonsei University
Department of Computer Science,
Yonsei University, Seoul, Korea
+82-2-2123-5714
sanghyun@cs.yonsei.ac.kr

ABSTRACT
Text mining is widely used to infer relationships between biological entities. Most text-mining algorithms utilize a co-occurrence-based approach. The term co-occurrence denotes a relationship between two interesting entities if they appear in the same sentence. Using these approaches current studies have extracted relationships between biological entities such as disease-gene relationships. However, these approaches cannot provide specific information for inferred relationships such as the role of the gene in the disease. To overcome this limitation, we propose a novel approach for inferring disease-gene relationship that provides specific knowledge of the inferred relationships. To implement this method, we first built terms based on text analysis to extract opinion sentences that include disease-gene relationships. We then extracted these opinion sentences and inferred disease-gene relationships by using disease-related and gene-related terms in the opinion sentences. Using these extracted relationships and terms, we inferred disease-related genes and constructed a disease-specific gene network. To validate our approach, we investigated the top k (k = 20) inferred genes for prostate cancer and analyzed the constructed gene network using three network analysis measures. Our approach found more disease-gene relationships than comparable method, and inferred describable disease-gene relationships.

CCS Concepts
• Applied computing ➡ Life and medical sciences ➡ Bioinformatics

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579
E-mail address: sanghyun@cs.yonsei.ac.kr (S. Park)

SAC 2016, April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851616

Keywords
Text-mining; Relationship; Gene; Disease; Network; Analysis

1. INTRODUCTION
Biological research has been driven primarily by an interest in disease processes, and as such, large amounts of literature data have been generated. However, extraction of the knowledge from the data generated is important in biology. One of the best-known methods for extracting data from the literature is text mining. Text mining provides opportunities to reduce time and effort for extracting knowledge from the biological literature.

In biology, text mining can be used to infer relationships between biological entities such as proteins, genes, diseases, and drugs. Their relationships can be useful knowledge to identify the association between biological entities and disease.

Previous studies [2][4][13] have inferred relationships between biological entities by using various text-mining approaches. However, these approaches cannot describe these relationships concretely. For example, these approaches can confirm that the gene may be associated with the disease, but they cannot identify the role of the gene in the extracted relationship.

The goal of this study is to infer describable disease-gene relationships. Here, we propose a novel approach to addressing this goal that utilizes opinion sentences in PubMed [12] literature data. An opinion sentence is a sentence that describes the authors conclusions based on their experimental results. Our assumptions are as follows:

- Opinion sentences include useful knowledge to describe disease-gene relationships.

- Opinion sentences are stated in the conclusion section among the several sections in the literature.

- If the opinion sentence describes disease-gene relationships, the sentence includes disease-related terms (associated with diseases such as metastasis) and gene-related terms (associated with the gene such as mutation) as well as disease name and gene symbol.

Our aim in extracting disease-gene interactions is to identify disease-gene relationships by using disease-related and gene-related terms to confirm specific information. To achieve this, we extracted opinion sentences that contain useful information for disease-gene relationships.

The main contributions of this work include:

- Extraction of opinion sentences that have disease-gene relationships and specific information related to these relationships.

- A novel method for identifying describable disease-gene relationships.

- Construction of a disease-specific gene network based on disease-related and gene-related terms that varies from previous gene networks.

The rest of the paper is organized as follows. Section 2: related work from similar previous studies that use Swanson's ABC model, extracting methods for biological relationships from literature, and biological network construction methods. Section 3: the proposed method of using opinion sentences to identify specific disease-gene relationships. Section 4: experimental data and results based on the application of our approach to prostate cancer literature data. Section 5: a discussion based on these findings. Section 6: conclusions of the study that highlight implications of our findings.

2. RELATED WORK

There have been several text-mining approaches developed in the biomedical field. Among the various text-mining applications, the Swanson's ABC model [16, 17] is well known and many researchers have tried to develop the method. Other approaches have also been presented to extract biological relationships using text-mining. Furthermore, several studies related to network analysis have been introduced because a network can be used to briefly describe biomedical relationships.

2.1 ABC Model

The ABC model is one of the approaches to extract relationships between interesting entities in the biological area. Swanson proposed the concept of discovering new relationships based on existing relationships. Based on this, he developed the ABC model, which determines a relationship between "A" and "C" by using the A-B relationship and the B-C relationship. For instance, if a disease is related to a gene and the gene is related to a drug, then a candidate relationship between the disease and drug is inferred by the ABC model.

Several techniques have been developed based on Swanson's ABC model, such as Petric et al. [11] who attempted to discover new relationships between biomedical concept A and C. In their experiments, they used agents as biomedical concept A and observed phenomena as biomedical concept C to discover relationships between an agent and an observed phenomenon. They extracted 3 terms, namely rare terms, joint terms, and linking terms. The rare terms indicate low frequency terms among the identified terms in the literature related to phenomenon C. These can be considered unusual observations about phenomenon C. The joint terms indicate identified common terms in the literature related to the rare terms. The linking terms are extracted from the literature between those on joint term A and phenomenon C. Using these terms, they calculated correlation similarity between phenomenon C and agent A and extracted relationships between them. They applied the method to autism, and extracted relationships between autism and calcineurin. Baker et al. [1] proposed a methodology to infer drug-disease relationships by using Swanson's ABC model and biological literature data. They reproduced Swanson's discovery of a connection between magnesium and migraine by using only protein as the intermediate B term, and used MeSH protein annotations to extract proteins from the literature. Using this methodology, they predicted zinc and retinoic as candidate chemical compounds associated with migraine.

2.2 Biological Relationship Extraction

Several studies attempted to extract relationships between biological entities. Senger et al. [15] constructed prolific database as a solution to discover potential drug targets, side effects, and protein functions by providing literature that include proteins, homologues, and compounds mentioned in the same sentences. To construct prolific database, they extracted protein-studied compound relationships based on co-occurrence in the same sentence. The extracted relationships were categorized as four types: (i) co-occurrence of protein and compound in the abstract, (ii) co-occurrence in the same sentence, (iii) co-occurrence in the same sentence with "functional process" or "molecular function", (iv) co-occurrence in the same sentence with "relationship" verbs. The "relationship" verbs were derived by analyzing random selected abstracts that include verbs occurring with compounds and proteins in the same sentence. Considering these relationship data, they constructed the prolific database, which allows discovery of potential targets, side effects, and protein functions. Lee et al. [8] attempted to discover drug-disease relationships by using the ABC model and context terms. Context term refers to biological terms co-occurring with the relationships of interest in the literature. In their experiments, they extracted drug-disease relationships based on drug-gene and gene-disease relationships. To calculate the score between drug and disease, they used the similarity between context terms extracted from drug-gene and gene-disease relationships. They applied the method to Alzheimer's disease and extracted relationships between Alzheimer's disease and several drugs.

2.3 Network Analysis in Biology

A network can be used to describe complex relationships between biological entities. In particular, a network provides several scoring measures for analyzing nodes such as degree centrality, closeness centrality, and betweenness centrality. Using these measures, we can determine meaningful relationships between biological entities in the network.

Gottlieb et al. [5] presented the PRINCIPLE tool, which describes gene networks based on the PRINCE [18] algorithm. The PRINCE algorithm method was developed to infer relationships between genes and diseases using network analysis. They used disease-disease similarity and protein-protein interaction data.

Chen et al. [3] presented a method for inferring gene regulatory network using literature and microarray data. They extracted gene interactions from the literature. To enhance the gene interactions, they generated weights based on microarray analysis in the interactions. Using the approach, they demonstrated the advantage of combining gene interactions from the literature data with microarray analysis.

3. METHODS

In this section, we describe the proposed method used to infer describable disease-gene relationships from literature data using opinion sentences. Figure 1 outlines the proposed method.

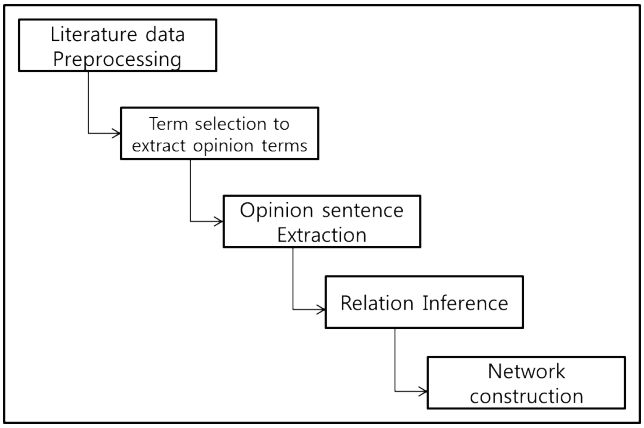

Figure 1. Outline of the proposed method

Our method has five steps. First, the "conclusion sections" in the literature data are gathered to build frequently used terms. In the next step, terms are extracted for extracting opinion sentences. Opinion sentences are then extracted based on the selected terms. Using these opinion sentences, disease-gene relationships are inferred using DRT (disease-related terms) and GRT (gene-related terms). Finally, a gene-gene interaction network is constructed using DRT and GRT.

3.1 Literature Data Preprocessing

To implement our method, we gathered abstracts related to prostate cancer from PubMed. The abstracts were divided into two types. The first type included structured abstracts with several sections such as "background", "methods", and "results". The other type was unstructured. Figure 2 demonstrates these abstract types.

(A)

BACKGROUND. Increased expression of human endogenous retroviruses, especially HERV-K(HML-2) proviruses, has recently been associated with prostate carcinoma progression. In particular, a HML-2 locus in chromosome 22q11.23 (H22q) is upregulated in many cases. We therefore aimed at delineating the extent and repertoire of HML-2 transcription in prostate cancer tissues and cell lines and to define the transcription pattern and biological effects of H22q. **METHODS.** Sanger and high throughput amplicon sequencing was used to define the repertoire of expressed HML-2 in a selected set of samples. qRT-PCR was used to quantify expression of selected proviruses in an extended set of prostate cancer tissues. Transcription factor binding sites (TFBS) were compared bioinformatically using the Transfac database. Expression of H22q was further characterized by siRNA-mediated knockdown, 5′ RACE mapping of transcriptional start sites (TSS) and identification of splice sites. Functional effects of H22q knockdown were investigated by viability and apoptosis assays. **RESULTS.** In addition to H22q, a limited number of other proviruses were found expressed by sequencing. Of these, provirus ERVK-5 and to a lesser degree ERVK-15 were frequently upregulated in prostate cancer. In contrast, expression of ERVK-24, predominant in germ cell tumors, was not detectable in prostatic tissues. While HML-2 LTRs contain binding sites for the androgen receptor and cofactors, no consistent differences in transcription factor binding sites were found between expressed and non-expressed proviruses. The H22q locus contains two 5′-LTRs of which the upstream LTR is predominantly used in prostatic cells, with an imprecise TSS. Splicing of H22q transcripts is complex, generating, among others, a transcript with an Np9-like ORF. Knockdown of H22q did not significantly affect proliferation or apoptosis of prostate cancer cells. **CONCLUSIONS.** Our findings further underline that HML-2 expression is commonly highly tissue-specific. In prostate cancer, a limited number of loci become activated, especially H22q and ERVK-5. As expressed and non-expressed proviruses do not differ significantly in TFBS, tissue- and tumor-specific expression may be governed primarily by chromatin context. Overexpression of HML-2 H22q is more likely consequence than cause of prostate cancer progression. *Prostate*

(B)

Prostate cancer is initially responsive to androgen deprivation, but the effectiveness of androgen receptor (AR) inhibitors in recurrent disease is variable. Biopsy of bone metastases is challenging; hence, sampling circulating tumor cells (CTCs) may reveal drug-resistance mechanisms. We established single-cell RNA-sequencing (RNA-Seq) profiles of 77 intact CTCs isolated from 13 patients (mean six CTCs per patient), by using microfluidic enrichment. Single CTCs from each individual display considerable heterogeneity, including expression of *AR* gene mutations and splicing variants. Retrospective analysis of CTCs from patients progressing under treatment with an AR inhibitor, compared with untreated cases, indicates activation of noncanonical Wnt signaling ($P = 0.0064$). Ectopic expression of Wnt5a in prostate cancer cells attenuates the antiproliferative effect of AR inhibition, whereas its suppression in drug-resistant cells restores partial sensitivity, a correlation also evident in an established mouse model. Thus, single-cell analysis of prostate CTCs reveals heterogeneity in signaling pathways that could contribute to treatment failure.

Figure 2. Two types of abstract are gathered from PubMed.

Type (A) indicates structured abstracts and the type (B) indicates unstructured abstracts.

Structured abstracts were used as training data set to build terms to extract opinion sentences, whereas the unstructured abstracts were used as test data to extract opinion sentences based on built terms (Figure 2).

We analyzed sentences included in the "conclusion" section of structured abstracts to extract opinion sentences based on our assumption that this section contains several opinion sentences, including the authors' opinion of their experimental results.

To extract opinion sentences that include disease-gene relationships, we first extracted gene-related sentences by analyzing whether the sentence included a gene symbol. We then built words with frequency, which is the number of appearances of a word in sentences. Among the built terms, we selected terms having frequency of >10 in order to exclude rare terms. The selected words were then used in the "term selection to extract opinion sentences" step.

3.2 Term Selection to Extract Opinion Sentences

The terms extracted from the previous step were categorized into three types, DRT, GRT, and OT (opinion terms), which were curated manually.

Disease-related terms refer to those related to diseases such as malignant, metastatic, and therapy. Gene-related terms are those related to genes such as activation, mutation, and overexpression. Opinion terms are terms that demonstrate evidence confirming the opinion of authors on their experiment results, such as "finding", "suggest", and "conclude". Using biological literature data associated with prostate cancer, we selected several terms to extract opinion sentences.

Table 1. Built terms associated with prostate cancer

	Disease-related terms	Gene-related terms	Opinion terms
The number of category	27	18	26
The number of terms	56	50	86

Table 1 shows the results of the built terms from the biological literature associated with prostate cancer. Among the built terms, synonym and verb conjugation were included in same category. Based on these terms, we extracted opinion sentences. If the sentence contained a disease name, gene symbol, DRT, GRT, and OT, then the sentence was considered an opinion sentence indicating a disease-gene relationship.

3.3 Opinion Sentence Extraction

Using the terms built in the previous step and gene symbols provided from the HUGO Gene Nomenclature Committee (HGNC) [6] gene database, we extracted opinion sentences from prostate cancer related unstructured abstracts.

The results suggest that some VDR gene polymorphisms in Korean men might not only be associated with prostate cancer risk but also significantly related to prostate cancer-related risk factors such as PSA level , tumor stage , and Gleason score .

Figure 3. Example of an extracted opinion sentence

Figure 3 demonstrates an example sentence that was extracted based on built terms in the literature. In the demonstrated sentence, "suggest" was extracted as an OT, "polymorphisms" was extracted as a GRT, "risk" was extracted as a DRT, and "VDR" indicates the gene symbol (Figure 3). If the extracted sentence has several DRT, we extracted only one DRT located close to the disease word. The disease word is that which describes the disease such as "prostate cancer". Similarly, in the case of GRT, we selected only one GRT located near the gene symbol. As shown in Figure 3, we extracted opinion sentences that included the disease word, gene symbol, OT, GRT, and DRT.

3.4 Relationship Inference

Based on the extracted opinion sentences, we inferred disease-gene relationships with DRT and GRT. In the example demonstrated in Figure 3, we inferred a relationship between "VDR" and "prostate cancer" with "polymorphisms" and "risk". Based on the GRT and DRT, we can conclude that "VDR" "polymorphisms" are associated with the "risk" of prostate cancer. In this process, we determined that the OT is not important to present disease-gene relationship. Therefore, we did not use the OT for our analyses.

3.5 Network Construction

In the network construction step, we constructed disease-specific gene networks using DRT and GRT. Among the inferred relationships, if two genes had the same DRT and GRT then we created an edge between the two genes. The generated edges included weight, which was calculated based on the frequency. The frequency between two genes indicates the number of relationships that have the same DRT and GRT.

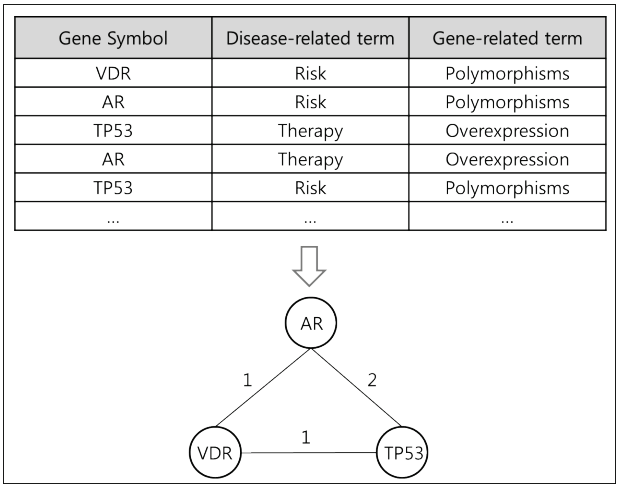

Figure 4. Network construction using DRT and GRT

Figure 4 demonstrates an example of the network construction process. "AR" has the same DRT and GRT as "VDR". Therefore, an edge was generated between "AR" and "VDR" (Figure 4). In addition, "AR" has another edge with "TP53" because they had the same DRT and GRT. The number in the network indicates frequency.

4. RESULTS

For validation, we ranked relationships based on frequency and validated for top N disease-gene relationships. After ranking, we compared our findings to an existing method. Additionally, we constructed disease-specific gene networks using inferred relationships with DRT and GRT, and analyzed the constructed gene network based on three network analysis measures. Furthermore, to identify more meaningful candidate disease-related genes, we conducted our method excluding known genes obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [7] database. These known genes have already been shown to be related to the disease.

4.1 Experimental Data

We gathered abstracts from PubMed and divided them into two types. One type was used to build terms, and the other was used to extract opinion sentences based on the built terms. The extracted opinion sentences were used to infer disease-gene relationships with DRT and GRT. Additionally, we constructed a disease-specific gene network using DRT and GRT (Table 2).

Table 2. Data and gene network properties

Structured abstract	Unstructured abstract	Opinion sentences	Gene	Relation
17,237	43,575	2,514	603	15,314

In Table 2, the "Structured abstract" was used to build frequently used terms to extract opinion sentences and the "Unstructured abstract" was used to extract opinion sentences based on built terms. "Opinion" indicates the number of opinion sentences extracted from the "Unstructured abstract", whereas "Gene" and "Relation" indicates the number of nodes and edges in the gene network.

4.2 Relation Inference and Top N Validation

We extracted opinion sentences based on identification of our built terms in the abstract data. The extracted opinion sentences include DRT and GRT, and frequency was calculated based on these terms. We inferred the top N disease-gene relationships based on the frequency and validated the inferred relationships by comparing to an existing method.

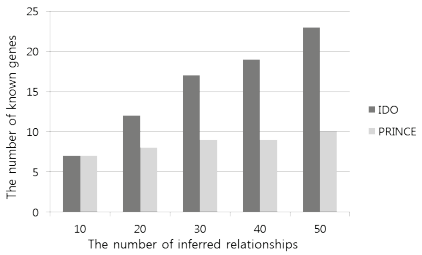

Figure 5. The number of known genes based on the IDO and PRINCE algorithm. The IDO indicates a proposed method and, the PRINCE indicates a comparable method. The x-axis indicates the number of inferred relationships, and the y-axis indicates the number of known genes i.e. those already known be related to prostate cancer.

To validate inferred relationships, we confirmed the number of known genes among the inferred genes using disease-gene

databases. The following databases were used: the human prostate gene database (PGDB) [9], dragon database of genes associated with prostate cancer (DDPC) [10], KEGG, and Sanger [14]. These databases provide a prostate cancer related gene list. The PRINCE algorithm is a method used to infer disease-gene relationships. As shown in Figure 5, the proposed method identified more disease-gene relationships than the PRINCE algorithm.

Additionally, we conducted an experiment using a random value to confirm the suitability of DRT and GRT as ranking features for determining relationships between a disease and gene.

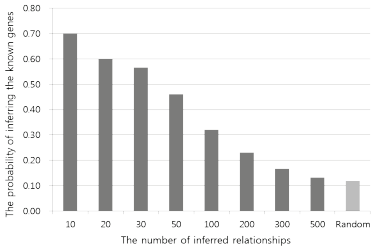

Figure 6. Comparison using a random value.

The x-axis indicates the number of inferred relationships, and the y-axis indicates the probability of inferring the known genes. The probability is calculated as the number of inferring known genes over the number of inferred relationships. "Random" refers to the probability that a randomly selected gene will be related to the disease.

Figure 6 demonstrates that the proposed method identified more known genes than a random value for the entire section (from 10 to 500). The data also indicates that DRT and GRT can be used as suitable ranking features to infer disease-gene relationships.

4.3 Network Analysis

We constructed a prostate cancer-specific gene network based on DRT and GRT.

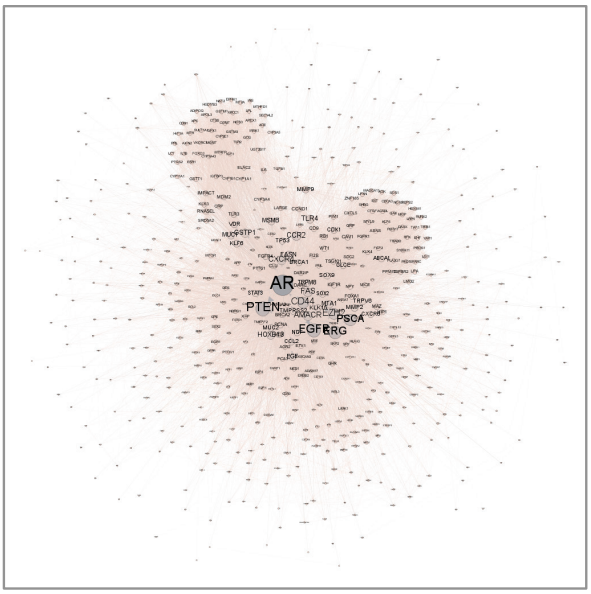

Figure 7. Prostate cancer-specific gene network based on DRT and GRT. In the network, the node size and label is proportional to the node degree. The gene network has 603 nodes and 15,314 edges.

To analyze the network, we used the following network analysis measures: degree, closeness, and betweenness centrality. Centrality identifies the most important nodes within a network, whereas degree centrality was defined as the number of links within a node. Closeness centrality was calculated using the shortest paths from all other nodes, and betweenness centrality was defined as the number of times a node acts as a bridge along the shortest path between two other nodes. Using these measures, we inferred the top 10 genes.

Table 3. Top 10 genes by network analysis

Rank	Betweennes s Centrality	Closeness Centrality	Degree Centrality	Common
1	AR	CCL4	AR	AR
2	PTEN	FGF17	PTEN	PTEN
3	ERG	AR	ERG	ERG
4	EGFR	PTEN	EGFR	EGFR
5	EZH2	ERG	CD44	EZH2
6	GSTP1	EGFR	EZH2	PSCA
7	PSCA	EZH2	PSCA	CD44
8	CD44	CD44	AMACR	AMACR
9	AMACR	PSCA	MUC1	
10	HOXB13	AMACR	FAS	

Table 5 demonstrates the top 10 genes inferred from the network analysis. Among the inferred genes, we selected 8 common genes, and validated the common genes based on databases. Among the common genes in Table 5, the gray color indicates known genes. We identified prostate cancer related genes with 75% probability (Table 5).

5. DISCUSSION

We attempted to extract opinion sentences that include a disease-gene relationship. Based on these opinion sentences, we identified disease-gene relationships with DRT and GRT. Using frequency as weight, we ranked the inferred relationships using the proposed method. In this section, we present describable disease-gene relationships determined using DRT and GRT.

5.1 Inferring Describable Relationships

Using DRT and GRT, we inferred describable disease-gene relationships. Relationships were ranked according to frequency, in order to infer useful relationships. An example opinion sentence extracted by the proposed method supported the inferred relationships (Figure 3).

Table 4. Top 10 describable disease-gene relationships

Gene	GRT	DRT	Frequency	PMID
Opinion sentence				
AR	expression	progression	32	23658830
Moreover, recent studies indicate that constitutively active AR variants are expressed in primary prostate tumors and may contribute to tumor progression.				
VDR	polymorphism	risk	27	24120391
The results suggest that some VDR gene polymorphisms in				

Korean men might not only be associated with prostate cancer risk but also significantly related to prostate cancer-related risk factors such as PSA level, tumor stage, and Gleason score.				
AR	transcription	progression	20	23887938
Androgen receptor (AR), a ligand-dependent transcription factor, plays a critical role in prostate cancer onset and progression, and its transcriptional function is mediated largely by distinct nuclear receptor co-regulators.				
CYP1A1	polymorphism	risk	20	23475304
The meta-analysis suggests an important role of the CYP1A1 MspI polymorphism in the risk of developing prostate cancer, especially in Asians.				
AR	polymorphism	risk	18	16254899
The A allele of the G1733A polymorphism of the AR gene has been associated with increased risk of prostate cancer.				
AR	expression	CRPC	17	22027692
These findings indicate that downmodulation of AR expression may provide a complementary strategy for treating CRPC.				
AR	expression	proliferation	16	15170865
In addition, AR expression plays an important role in the proliferation of human prostate cancer and confers a better prognosis in breast cancer.				
AR	expression	suppression	16	20587701
The results showed that SSA significantly suppressed the growth of human and mouse prostate cancer cells expressing AR in strong association with G(1) arrest, and decreased AR level and AR-dependent transactivation.				
AR	regulated	therapy	15	22315407
Given that AR-mediated gene regulation is enhanced by AR coregulators, inactivation of those coregulators is emerging as a promising therapy for prostate cancer (PCa).				
AR	expression	therapy	15	22426480
Our findings point to BLT2 as a key regulator of AR expression and will contribute to the development of novel therapies for AR-positive prostate cancers, including androgen-responsive and CR prostate cancers.				

Table 4 shows the inferred relationships between disease and gene with DRT, GRT, PMID, and opinion sentence. Describable disease-gene relationships were inferred using DRT and GRT. To support the inferred relationships, we provided opinion sentences that were extracted using the proposed method with PMID. The proposed method allows the extraction of specific information relevant to the relationships such as the role of the gene and its target.

5.2 Inferring Describable Candidate Relationships

In order to infer meaningful candidate disease-gene relationships, we excluded known genes, which were identified using the KEGG database. We identified top 10 genes of high frequency and confirmed a relationship with prostate cancer using the opinion sentences.

Table 5. Top 10 describable candidate disease-gene relationships

Gene	GRT	DRT	Frequency	PMID
Opinion sentence				
VDR	polymorphism	risk	28	24120391
The results suggest that some VDR gene polymorphisms in Korean men might not only be associated with prostate cancer risk but also significantly related to prostate cancer-related risk factors such as				
CYP1A1	polymorphism	risk	20	23475304
The meta-analysis suggests an important role of the CYP1A1 MspI polymorphism in the risk of developing prostate cancer, especially in Asians.				
SRD5A2	polymorphism	risk	14	17136762
Controversy exists over the significance of associations between the SRD5A2 (5alpha-reductase type 2) polymorphisms, A49T and V89L, and risk of prostate cancer.				
MTHFR	polymorphism	risk	12	22296369
We suggest that the heterozygote CT genotype and the 677T allele of the MTHFR polymorphism might be associated with an decreased prostate cancer risk.				
XRCC1	polymorphism	risk	11	17196815
This is the first report on the studies of XPC and XRCC1 Arg194Trp polymorphisms in PC, and our present data suggest that XPC Lys939Gln and the T-A haplotype of XRCC1 Arg194Trp and Arg399Gln may be risk factors for PC in Japanese.				
BRCA2	mutation	risk	11	18577985
Men with BRCA2 mutations have been found to be at increased risk of developing prostate cancer.				
CYP1B1	polymorphism	risk	9	24453031
In conclusion, based on 17 eligible studies, we found that the CYP1B1 Leu432Val polymorphism was associated with an increased risk of prostate cancer, while no association of bladder cancer was observed.				
GPX1	polymorphism	risk	8	18563616
We found an overall protective effect of the variant Leu allele of the GPX1 polymorphism on the prostate cancer risk.				
CHEK2	mutation	risk	8	12533788
Overall, our data suggest that mutations in CHEK2 may contribute to prostate cancer risk and that the DNA-damage-signaling pathway may play an important role in the development of prostate cancer.				
GSTT1	polymorphism	risk	7	20056632
This study showed that the inverse association between glucosinolate intake and prostate cancer risk was modified by NQO1 (C609T) and GSTM1 and GSTT1 deletion polymorphisms.				

Table 5 demonstrates the capacity of the proposed method to infer describable candidate relationships. By excluding known genes,

we inferred various disease-gene relationships with DRT and GRT.

5.3 Network Analysis Results with Opinion Sentences

A gene network was constructed based on GRT and DRT and analyzed using various network analysis measures. For each measure, we ranked genes and extracted 10 genes with high frequency. Among the top 10 genes, we confirmed that the common genes are involved in prostate cancer by opinion sentences.

Table 6. Validation of common genes

Common gene	PMID	Supported sentences
AR	21724752	Androgen and androgen receptors (AR) play critical roles in the proliferation of prostate cancer through transcriptional regulation of target genes.
PTEN	23936141	Aggressive and metastatic prostate cancer (PC) is associated with a reduction or loss of PTEN expression.
ERG	21519790	Additionally, overexpression of ERG is associated with unfavorable prognosis in prostate cancer patients similar to leukemia patients.
EGFR	18692155	Hence, our data suggest that p38MAPK-dependent activation of the mTOR/CD1 pathway may represent a mechanism through which AR and EGFR cross-talk contributes to prostate and lung cancer progression.
EZH2	16575874	The data show that amplification of the EZH2 gene is rare in early prostate cancer, whereas a fraction of late-stage tumors contains the gene amplification leading to the overexpression of the gene, thus indicating the importance of EZH2 in the progression of prostate cancer.
PSCA	20501618	The expression of PSCA is positively correlated with advanced clinical stage and metastasis in prostate cancers and is also associated with malignant progression of premalignant prostate lesions.
CD44	21953074	These findings suggest that CD44 may be a valuable biomarker and a predictor of radiosensitivity in CaP treatment.
AMACR	14612535	Recent work has identified AMACR as a new diagnostic marker for prostate cancer (PCa).

Table 6 shows the common genes by network analysis. To validate the relationships between inferred genes and prostate cancer, we found supported sentences among the opinion sentences. This result demonstrates that inferred common genes are involved in prostate cancer, and opinion sentences can be used as evidence to validate inferred genes.

In three unique experiments, we confirmed that our proposed method inferred meaningful opinion sentences and disease-gene relationships with DRT and GRT. Our method provides an opportunity to more specifically explain disease-gene relationships than previous methods.

6. CONCLUSIONS

In the present study, we attempted to infer meaningful disease-gene relationships from the literature. Terms were built to extract opinion sentences from the literature. Using the built terms, we inferred describable disease-gene relationships, and three distinct experiments demonstrated the effectiveness of the proposed method.

The proposed method extracted DRT and GRT as well as disease-gene relationships. Using these terms, we extracted more specific information on the disease-gene relationship than previous algorithms. Based on the proposed method, we can confirm the role of the genes identified in the disease-gene relationship based on GRT as well as the condition of the disease influenced by the gene, based on DRT. Our method is therefore an effective tool for inferring disease-gene relationships.

In this paper, we used only prostate cancer-related literature. Further studies will involve application of our algorithm to other genetic diseases such as Alzheimer's, diabetes, and other cancers. Furthermore, we will construct disease-related terms and gene-related term databases by integrating biological text, a term database, and dictionary.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Baker, N. C., and Hemminger, B. M. Mining connections between chemicals, proteins, and disease extracted from Medline annotations. *J. Biomed. Inform*., *43*, (2010), 510-519.

[2] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H. P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics, 9*, (2008), 207.

[3] Chen, G., Cairelli M. J., Kilicoglu, H., Shin, D., Rindflesch, T. C. Augmenting Microarray Data with Literature-Based Knowledge to Enhance Gene Regulatory Network Inference. *PLoS Comput. Biol.,* 10, 6 (2014), e1003666.

[4] Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., Hishiki, T., and Tsujii, J. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pac. Symp. Biocomput.,* (2006), 4-15.

[5] Gottlieb, A., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. PRINCIPLE: a tool for associating genes with diseases via network propagation. *Bioinformatics, 27*, 23 (2011), 3325-3326.

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

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

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

[9] Li, L. C, Zhao, H., Shiina, H., Kane, C. J., Dahiya, R. PGDB: a curated and integrated database of genes related to the prostate. *Nucl. Acids. Res., 31*, 1 (2003): 291-293.

[10] Maqungo, M., Kaur, M., Kwofie, S., Radovanovic, A., Schaefer, U., Schmeier, S., Oppon, E., Christoffels, A., and Bajic V. B. DDPC: dragon database of genes associated with prostate cancer. *Nucl. Acids Res.,* (2010). http://dx.doi.org/10.1093/nar/gkq849.

[11] Petric, I., Urbancic, T., Cestnik, B., and Macedoni-Luksic, M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. *J. Biomed. Inform., 42,* (2009), 219-227.

[12] PubMed: MEDLINE Retrieval on the World Wide Web www.ncbi.nlm.nih.gov/pubmed [May 2015].

[13] Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.,* (2000), 517-528.

[14] Sanger: Wellcome Trust Sanger Institute <http://www.sanger.ac.uk>.

[15] Senger, C., Gruning, B. A., Erxleben, A., Doring, K., Patel, H., Flemming, S., Merfort, I., and Gunther, S. Mining and evaluation of molecular relationships in literature. *Bioinformatics, 28*, 5 (2012), 709-714.

[16] Swanson, D. R. Undiscovered public knowledge. *Libr. Quart., 56*, 2 (1986), 103-118.

[17] Swanson, D. R. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc., 78*, 1 (1990), 29-37.

[18] Vannu, O., Magger, O., Ruppin, E., and Shlomi, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol., 6*, 1 (2010).