journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A literature-driven method to calculate similarities among diseases

Hyunjin Kim[a], *Youngmi Yoon*[b], *Jaegyoon Ahn*[c], *Sanghyun Park*[a,*]

[a] *Department of Computer Science, Yonsei University, South Korea*
[b] *Department of Computer Engineering, Gachon University, South Korea*
[c] *Department of Integrative Biology and Physiology, University of California, Los Angeles, USA*

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form 1 July 2015
Accepted 1 July 2015

Keywords:
Disease network
Disease–disease similarity
Biomedical text mining

ABSTRACT

Background: "Our lives are connected by a thousand invisible threads and along these sympathetic fibers, our actions run as causes and return to us as results". It is Herman Melville's famous quote describing connections among human lives. To paraphrase the Melville's quote, diseases are connected by many functional threads and along these sympathetic fibers, diseases run as causes and return as results. The Melville's quote explains the reason for researching disease–disease similarity and disease network. Measuring similarities between diseases and constructing disease network can play an important role in disease function research and in disease treatment. To estimate disease–disease similarities, we proposed a novel literature-based method.

Methods and results: The proposed method extracted disease–gene relations and disease–drug relations from literature and used the frequencies of occurrence of the relations as features to calculate similarities among diseases. We also constructed disease network with top-ranking disease pairs from our method. The proposed method discovered a larger number of answer disease pairs than other comparable methods and showed the lowest *p*-value.

Conclusions: We presume that our method showed good results because of using literature data, using all possible gene symbols and drug names for features of a disease, and determining feature values of diseases with the frequencies of co-occurrence of two entities. The disease–disease similarities from the proposed method can be used in computational biology researches which use similarities among diseases.

1. Introduction

Diseases are functionally connected to one another. One gene can cause various diseases, or inhibiting protein translation by one miRNA can be a contributing factor to various diseases. Therefore, a person with a certain disease has a higher probability of getting functionally connected disease than normal people. By using the disease connection information, the possibility of a specific disease onset for a person can be predicted and it is a simple example showing how disease–disease similarity can be utilized for disease-related function research. Disease–disease similarity will be of much help to disease research. It can be useful for development

* *Corresponding author*. Tel.: +82 221235714.
E-mail addresses: chriskim@cs.yonsei.ac.kr (H. Kim), ymyoon@gachon.ac.kr (Y. Yoon), jgahn@ucla.edu (J. Ahn), sanghyun@cs.yonsei.ac.kr (S. Park).

http://dx.doi.org/10.1016/j.cmpb.2015.07.001
0169-2607/

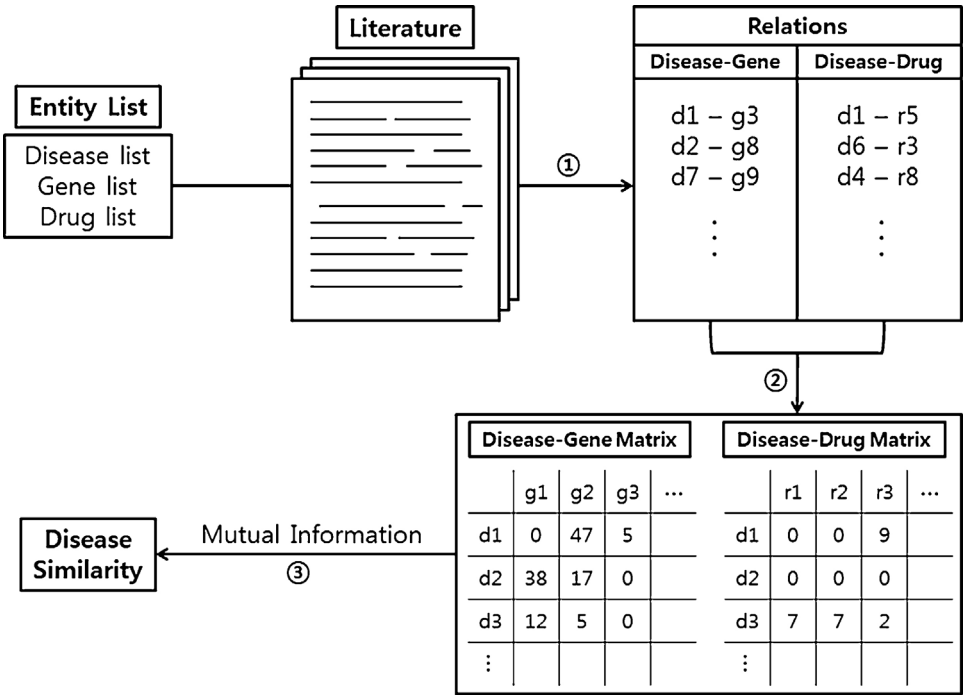

Fig. 1 – Overview of the proposed method. The proposed method consists of three steps. 1. Extracting relations from literature. 2. Constructing the matrices with the extracted relations. 3. Calculating disease similarities using the matrices.

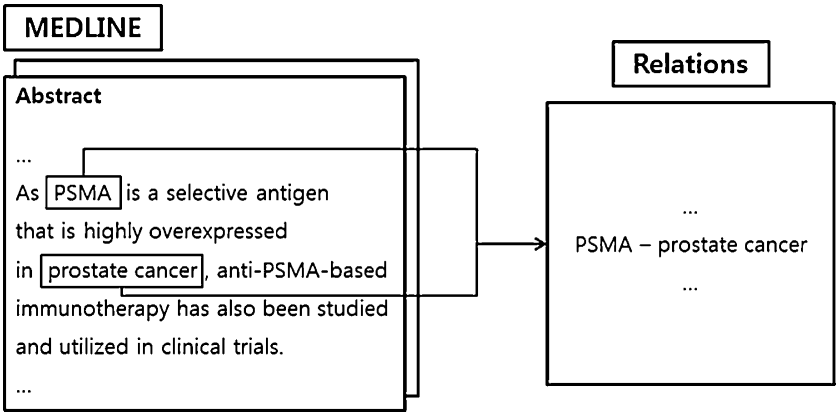

Fig. 2 – An example of extracting relations from the MEDLINE.

of new drug by aiding in drug repositioning, for searching new genes related to disease, and it can increase efficiency of network analysis in disease-related function research by enhancing disease networks.

There are three primary approaches to get disease–disease similarity: function-based approaches [1–3] and semantic-based approaches [4–12], and hybrid approaches of combining previous two approaches [13]. To seek the disease–disease similarity, function-based approaches compare functionally related genes, pathways and biological processes, and semantic-based approaches find similarity between disease terms of ontology related to diseases. Hybrid approaches utilize both functional similarity and semantic similarity. Liu et al. [1] calculated disease–disease similarity using both genetic information from GAD (Genetic Association Database) and environmental etiological factors from MeSH (Medical Subject Headings). Suthram et al. [2] calculated disease–disease similarity using mRNA expression from GEO (Gene Expression Omnibus) database and protein–protein interaction from HPRD (Human Protein Reference Database). Mathur and Dinakarpandian [3] calculated disease–disease similarity using semantic similarity of biological process based on gene ontology. In Li's case [4], a software package for calculating disease–disease similarity was developed using semantic similarities among terms of disease ontology and in the software, 10 methods of seeking semantic similarity are applied to disease ontology in calculating disease–disease similarity. Lastly, Cheng's [13] is a hybrid approach, which first calculates association score utilizing a gene function network and disease-related gene set, and secondly calculates

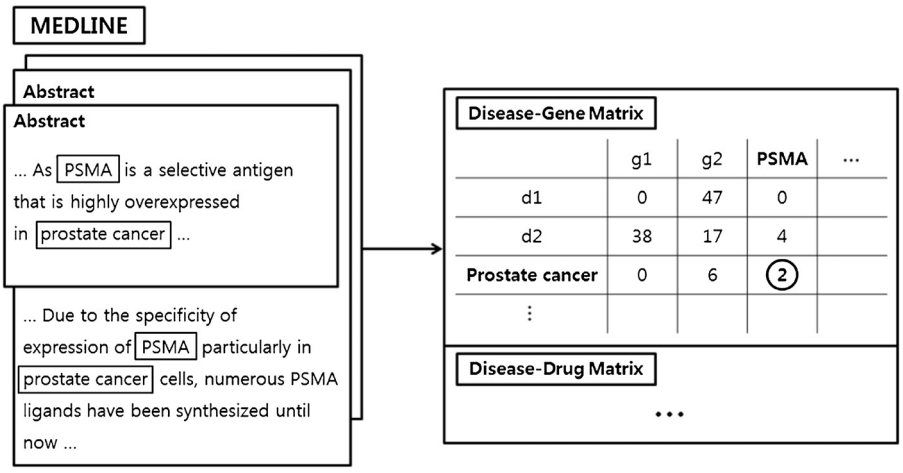

Fig. 3 – An example of constructing the matrices with the extracted relations.

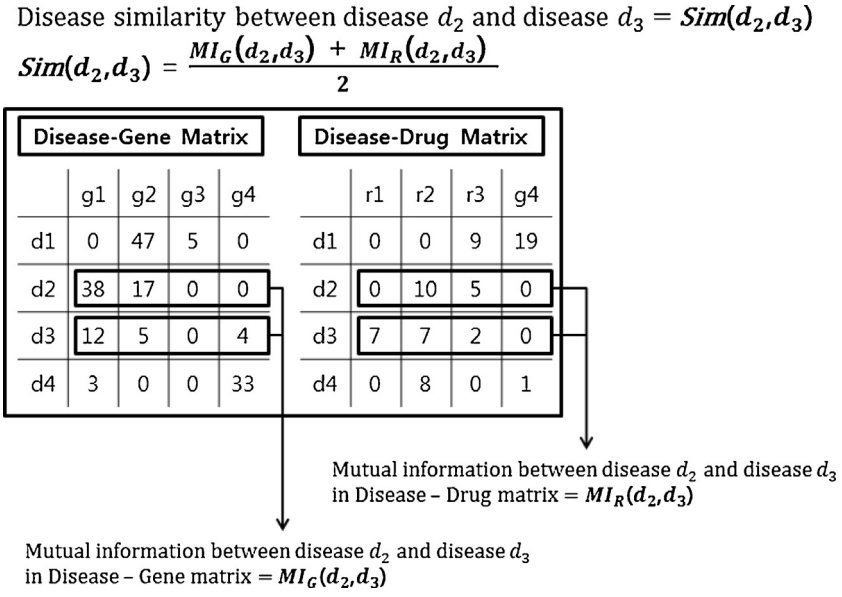

Fig. 4 – An example of calculating disease similarity using the matrices.

Table 1 – Condition information using in the disease similarity equation.

Conditions	$isAllZeroG(d_m)$	$isAllZeroG(d_n)$	$isAllZeroR(d_m)$	$isAllZeroR(d_n)$
A	0	0	0	0
	0	0	0	1
B	0	0	1	0
	0	0	1	1
	0	1	0	0
C	1	0	0	0
	1	1	0	0

* $isAllZeroG(d) = 1$, if disease d has all zero values in disease–gene matrix. 0, otherwise.
* $isAllZeroR(d) = 1$, if disease d has all zero values in disease–drug matrix. 0, otherwise.

semantic score on disease ontology, and finally gets disease–disease similarity adding these two scores.

Many existing methods find disease–disease similarity using genetic information or semantic information on gene ontology but there are also other similar approaches. Goh et al. [14] constructed human disease network with gene–disease associations from OMIM (Online Mendelian Inheritance in Man) database. They made a connection between two diseases if the diseases shared at least one gene. Lee et al. [15] constructed bipartite human disease association network

Table 2 – The statistics of the 3,353,503 similarities from the proposed method.

Min	Max	Median	Mean	Standard deviation
3.84×10^{-10}	2.09×10^{-1}	4.95×10^{-8}	3.27×10^{-4}	1.33×10^{-3}

using shared metabolic pathways from KEGG (Kyoto Encyclopedia of Genes and Genomes) database. Two diseases are linked if mutated enzymes associated with them catalyze adjacent metabolic reactions. Goh's method and Lee's method are related to the proposed method, but they cannot calculate disease-disease similarity. Zhang et al. [16] created feature vectors for disease phenotypes by utilizing phenotype records and calculated cosine similarities among disease phenotypes using the feature vectors, and then developed a disease phenotype network. van Driel et al. [17] employed text mining approach to calculate similarities between diseases. MeSH terms were served as features, and the number of times the term was found in an OMIM record was counted for feature value. They used MeSH hierarchy and inverse document frequency measure to refine feature values. Lastly, similarity between two diseases was computed by the cosine of the angle between their corresponding feature vectors. Hamaneh et al. [18] calculated disease–disease similarity by considering information flow on disease-protein network. The disease–protein network was made by using disease–gene associations from CTD (The Comparative Toxicogenomics Database) database and protein–protein interactions from ppiTrim database. Proteins were treated as features of a disease, and feature value was defined by the expected number of visits by random walker on the disease–protein network. Then disease–disease similarity was calculated by the cosine of the angle similar to van Driel's method.

Biomedical term relations from literature (research papers) can also help calculate disease–disease similarity. We propose a new literature-based method LDDSim (Literature-Driven Disease Similarities) to measure disease similarity. The proposed method extracts disease–gene relations and disease–drug relations from literature, and with the number of those relations, it builds disease–gene matrix and disease–drug matrix. Then the method calculates disease–disease similarity utilizing mutual information between the two diseases. In addition to it, we constructed disease network using the disease similarities.

2. Methods

2.1. *Materials*

In this paper, we proposed a method that calculates disease-disease similarity and developed a disease network using disease pairs having high similarity. We extracted 36,686 disease-gene relations and 25,721 disease–drug relations from 9,803,245 MEDLINE abstracts in between year 1980

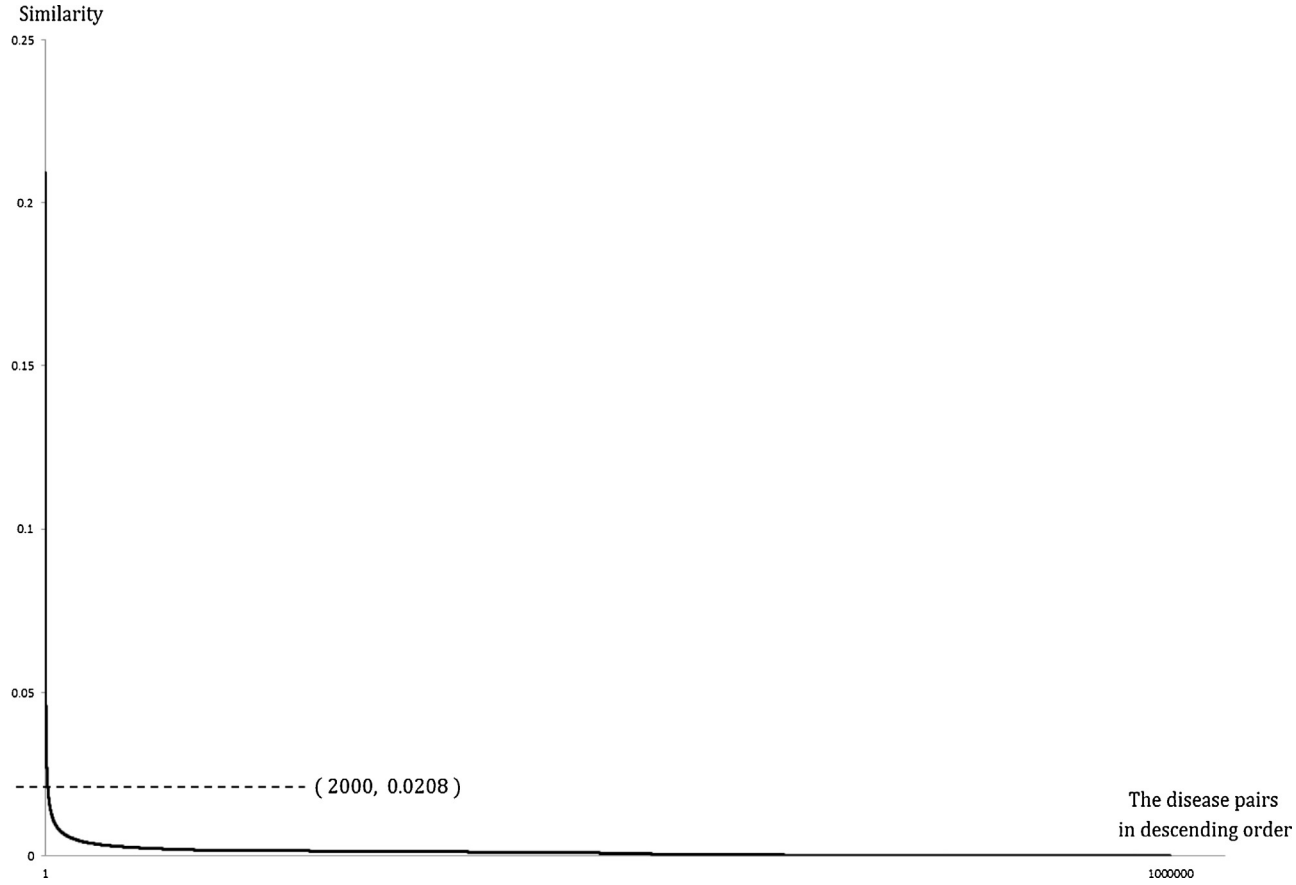

Fig. 5 – The trend of the similarities of disease pairs in descending order.

Table 3 – The number of answer pairs and *p*-value of 50 top-ranking disease pairs.

	Random	Resnik	Lin	Relevance	GIC	simIC	Wang	Modified Resnik	Modified Lin	Ours (LDDSim)
The number	2.50	11.49	17.95	17.86	12.02	17.86	11.00	12.83	17.87	23.79
p-value	1	4.14×10^{-3}	3.89×10^{-8}	4.92×10^{-8}	8.23×10^{-5}	4.62×10^{-8}	2.67×10^{-4}	2.03×10^{-4}	3.43×10^{-8}	4.33×10^{-12}

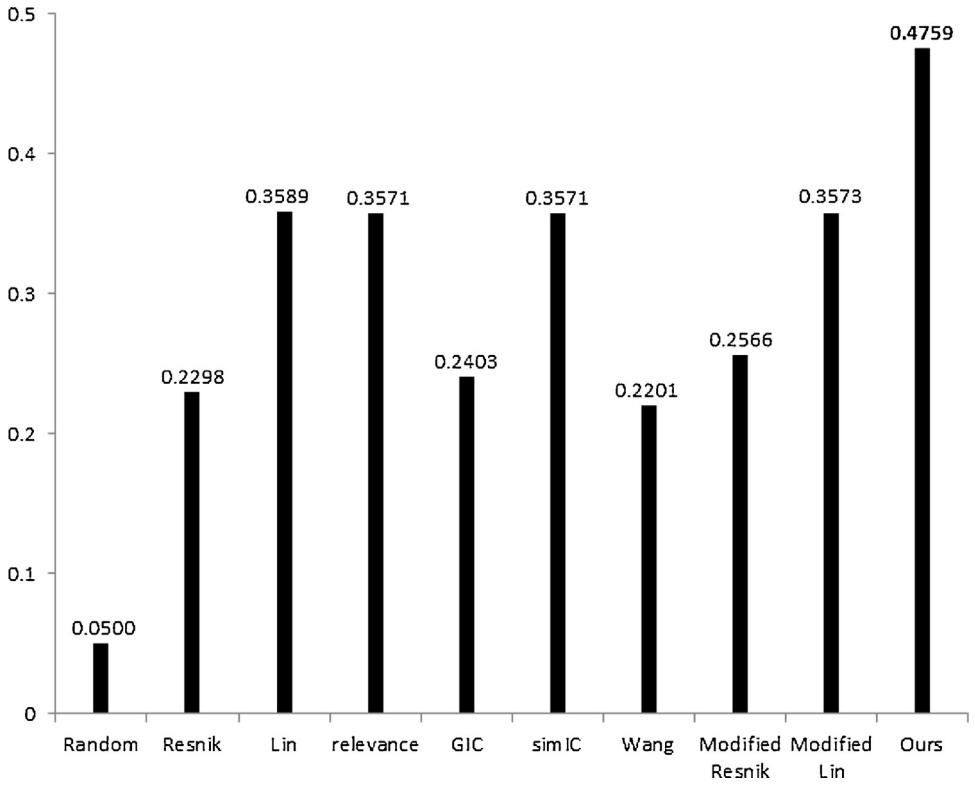

Fig. 6 – Precision and recall comparison of the experiment.

Table 4 – The result of (ours – van Driel's) experiment.

Method names	The number of answer pairs	*p*-value	Precision and recall
Random	1.05	1	0.0500
van Driel's (mimMiner)	2.98	1.23×10^{-1}	0.1421
Ours (LDDSim)	9.51	3.55×10^{-5}	0.4530

Table 5 – The result of (ours – Hamaneh's) experiment.

Method names	The number of answer pairs	*p*-value	Precision and recall
Random	0.45	1	0.0500
Hamaneh's	1.81	1.72×10^{-1}	0.2009
Ours (LDDSim)	3.96	2.43×10^{-2}	0.4400

and 2012 using 27,850 disease names, 61,304 gene symbols and 9388 drug names from PharmGKB database. After constructing disease–gene matrix and disease–drug matrix with these relations, similarities of 3,353,503 disease pairs were calculated using mutual information between diseases. We uploaded source code and data files of the proposed method to (http://embio.yonsei.ac.kr/files/hjkim/LDDSim/Kim_LDDSim.zip).

2.2. *Overview*

The proposed method has three steps to calculate disease–disease similarities (Fig. 1). First step is to extract disease–gene relations and disease–drug relations from literature using disease list, gene list, and drug list. Second step is to make disease–gene matrix and disease–drug matrix utilizing disease–gene relations and disease–drug relations which are the outcomes of the previous step. In the third step, disease–disease similarities are calculated by using mutual information among diseases from disease–gene matrix and from disease–drug matrix.

2.3. *Extracting relations*

The purpose of the first step is to extract disease–gene relations and disease–drug relations from literature which have high possibility to have actual relations. If a disease name and a gene symbol or a disease name and a drug name are co-occurred in a sentence of literature, then we can assume

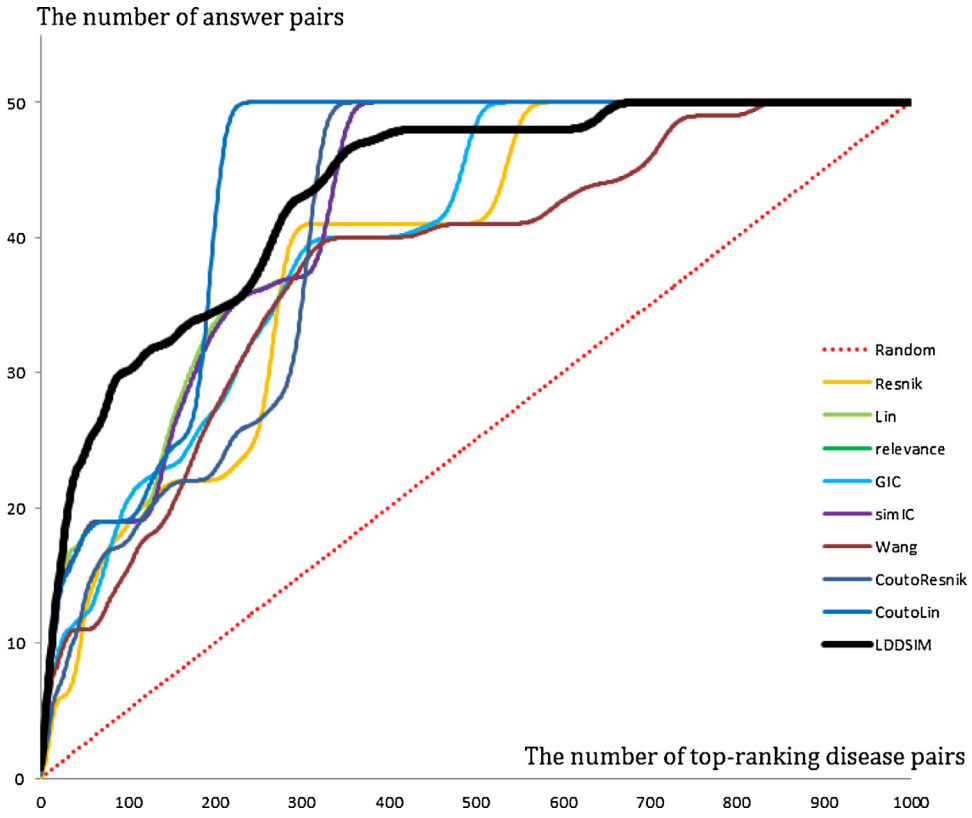

Fig. 7 – The number of answers with varying the number of top-ranking disease pairs.

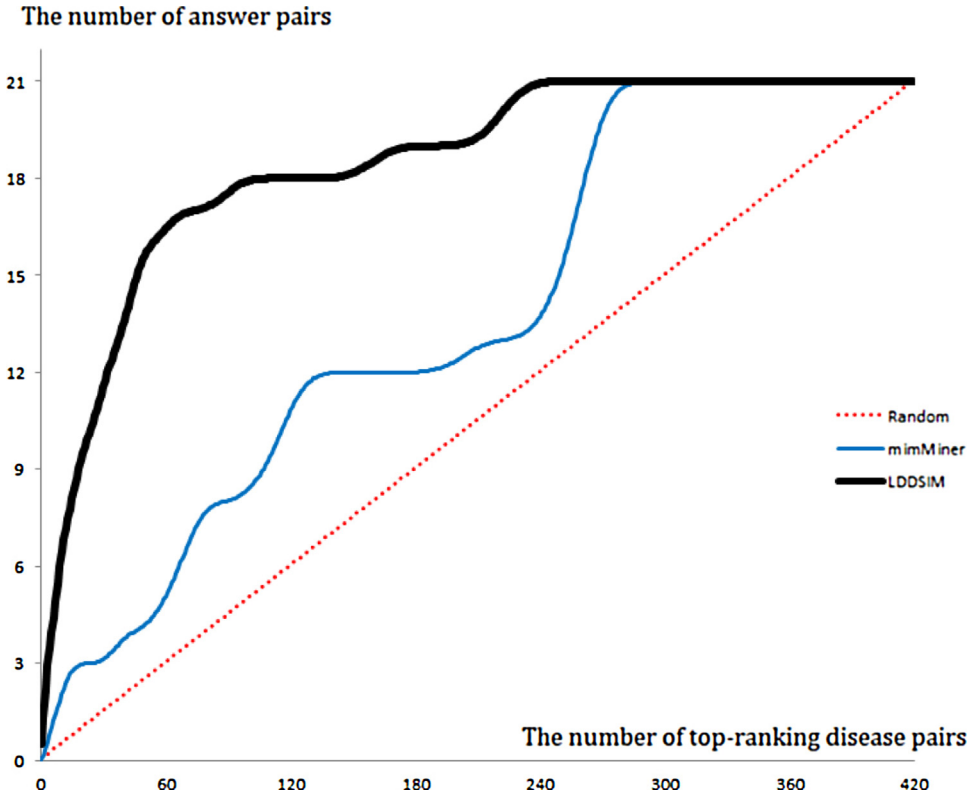

Fig. 8 – The number of answers with varying the number of top-ranking disease pairs (ours – mimMiner).

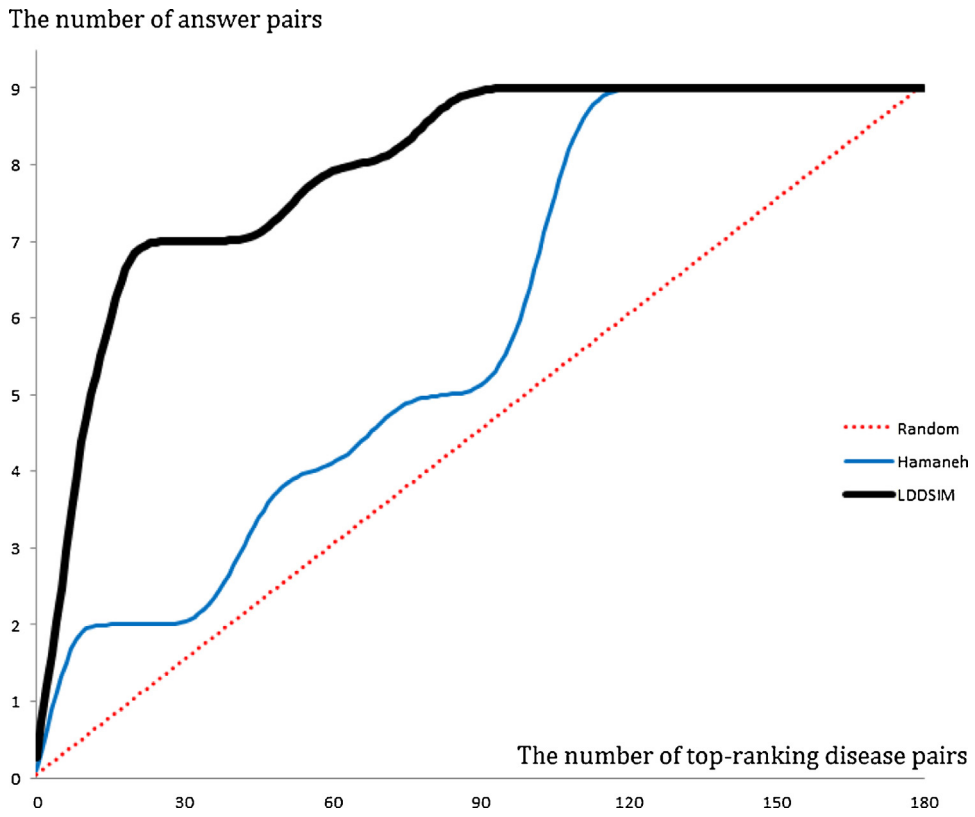

Fig. 9 – The number of answers with varying the number of top-ranking disease pairs (ours – Hamaneh's).

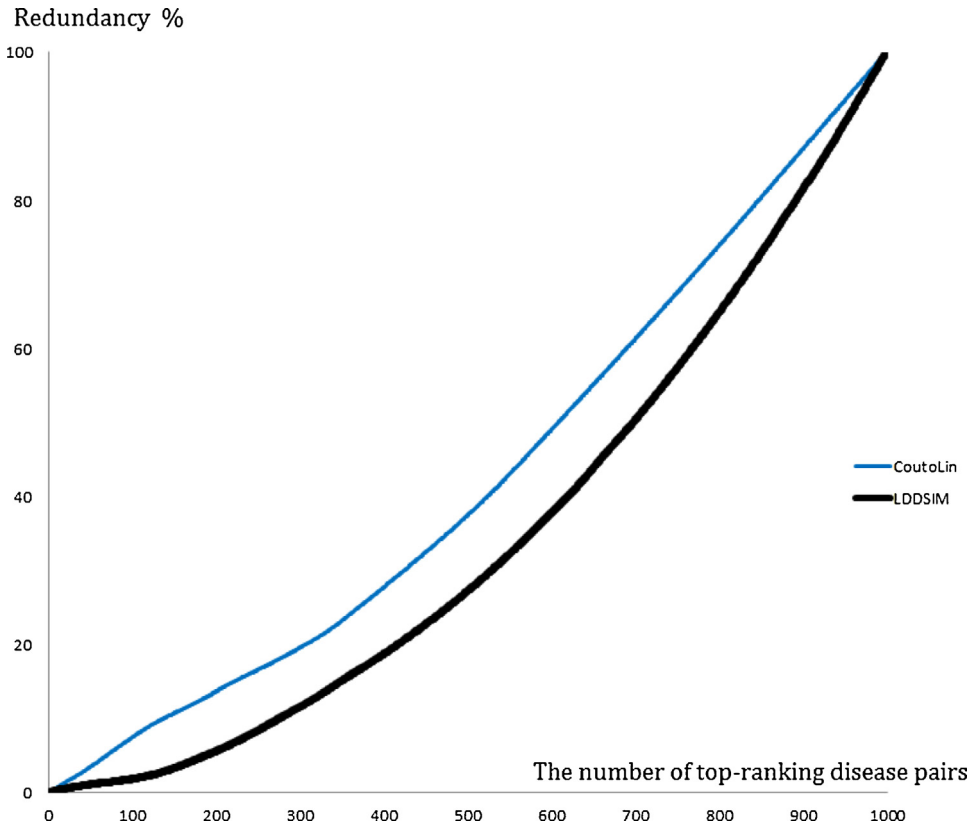

Fig. 10 – Redundancy of top-ranking disease pairs in comparing with the other 7 methods.

that there is a relation between two entities (disease–gene or disease–drug). We extracted disease–gene relations and disease–drug relations using disease names, gene symbols, and drug names from PharmGKB (The Pharmacogenomics Knowledgebase). The two entities of relations are co-occurred in abstracts of MEDLINE (Medical Literature Analysis and Retrieval System Online) (Fig. 2). The extracted relations are used in the second step to get frequencies of occurrence in the literature of the certain relations.

2.4. Constructing matrices

After extracting disease–gene relations and disease–drug relations, the next step is making disease-gene matrix and disease–drug matrix with the relations. A value of each cell in the matrices can be computed by counting the number of co-occurrences of corresponding two entities in the abstracts of MEDLINE (Fig. 3). In the disease–gene matrix, the rows represent diseases and the columns represent genes. Similarly in the disease–drug matrix, the rows represent diseases and the columns represent drugs. Because a large value of two entities in a matrix indicates a high frequency of co-occurrences in literature, we can assume that the larger the value of two entities, the closer the relation of the two entities is. With the values of the matrices, we can quantitatively estimate which gene or drug has a close relation with the certain disease and we can utilize the values as the features to calculate disease–disease similarities.

2.5. Calculating disease similarities

The third step is a phase to compute closeness among diseases. We used mutual information [19] between every two diseases in the matrices to calculate disease–disease similarities (Fig. 4). There are a few methods to calculate similarity (Mutual information, Pearson correlation, Spearman correlation, etc.). We assumed the values of diseases can have non-linear relations, therefore we used mutual information because it can generally find non-linear relations as well as linear relations [20,21]. When calculating disease–disease similarity, we consider both mutual information from the disease–gene matrix and mutual information from the disease–drug matrix. But some diseases can have all zero values in a matrix which indicates that the disease did not co-occur with any genes or drugs in the literature. In this case, mutual information between those two diseases is not applicable, and thus we only used available mutual information. In other words, we calculated mutual information between two diseases when the both diseases do not have all zero values in a matrix (Table 1). Similarity between disease d_m and disease d_n can be described as below.

$$Sim(d_m, d_n) = \begin{cases} \dfrac{MI_G(d_m, d_n) + MI_R(d_m, d_n)}{2} & , \text{if Condition} = A \\ MI_G(d_m, d_n) & , \text{if Condition} = B \\ MI_R(d_m, d_n) & , \text{if Condition} = C \\ N/A & , \text{otherwise} \end{cases}$$

$MI_G(d_m, d_n)$ = mutual information between disease d_m and disease d_n in Disease–Gene matrix. $MI_R(d_m, d_n)$ = mutual information between disease d_m and disease d_n in Disease–Drug matrix. Condition information is described in Table 1.

When calculating similarity between two diseases, we consider both mutual information from the disease–gene matrix and mutual information from the disease–drug matrix. However, if mutual information between two diseases in one matrix is not computable, then we can only use mutual information between the two diseases in the other matrix. If both mutual information from the two matrices are all available, the average value of the two mutual information is used for normalization.

3. Results

The proposed method can calculate similarities of 3,353,503 disease pairs. We evaluated the statistics of our similarities with mean, median, min, max, and standard deviation (Table 2).

The statistics indicate that the similarities are generally very low and there are high similarity-outliers because the mean is much larger than the median. For that reason, we can assume that the high similarity-outliers are significant. We investigated the trend of the similarities to get the outliers (Fig. 5).

We arranged similarities from our method in descending order. The curve in the Fig. 5 shows that there are stiffly large differences in high-ranking disease pairs and small differences in low-ranking disease pairs. We concluded that approximately 2000 disease pairs out of 3,353,503 disease pairs are high similarity-outliers and they are considered to be significant. The sorted list of all 3,353,503 disease–disease similarities from the proposed method can be downloaded via (http://embio.yonsei.ac.kr/files/hjkim/LDDSim/LDDSim_list.txt).

Two validations were performed to verify the effectiveness of the proposed method. The first one is a test on how accurately the method predicts disease–disease similarity. For the second, as manually checking over top 30 disease–disease similarities, we analyzed whether there is a real relation between two diseases.

3.1. Validation of the proposed method

To verify how accurately the proposed method can predict disease–disease similarity, we prepared a test set that is a mix of answer disease pairs and random disease pairs, and checked how many answer disease pairs the top-ranking disease pairs can find by ranking in descending order based on the proposed disease similarities. Then the result was compared with other 8 comparable methods (Resnik [5], Lin [7], modified Resnik [8], modified Lin [8], Relevance [9], Graph information content [10], Wang [11], Information coefficient similarity [12]). These 8 comparable methods were performed by using “DOSim” R package (http://cran.r-project.org/web/packages/DOSim/) of CRAN (The Comprehensive R Archive Network). The PharmGKB database has 61,304 disease names and they can produce 1,879,059,556 disease pairs. However, the proposed method can not calculate the similarities of all the disease pairs since its calculation is based on the relations in the

Table 6 – Literature summary of the manually checked top 30 disease pairs.

No.	Disease 1	Disease 2	Similarity	Description
1	HIV	AIDS	0.20910003	HIV(Human Immunodeficiency Virus) causes AIDS(Acquired Immunodeficiency Syndrome). A person infected with the HIV resulting in the development of AIDS [23,24].
2	HIV	Cystic fibrosis	0.13913876	–
3	HIV	Systemic lupus erythematosus	0.13084353	HIV and SLE show similar characteristics in their immune response and they can sometimes be indistinguishable [25–27].
4	HIV	Hepatotoxicity	0.12266148	Hepatotoxicity can be a side effect of HIV medicines [28,29], "Side Effects of HIV Medicines" from AIDSinfo (http://aidsinfo.nih.gov/education-materials/fact-sheets/22/67/hiv-and-hepatotoxicity)].
5	HIV	Tumor	0.11890388	People infected with HIV have a high risk of malignant tumor [30,31].
6	Cancer	Tumor	0.11835699	Tumor has two types, benign and malignant. Cancer is known as a malignant tumor ["Defining Cancer" from National Cancer Institute (http://www.cancer.gov/cancertopics/cancerlibrary/what-is-cancer)].
7	HIV	Cancer	0.11504229	Same description to the #5 since "cancer" is a term used as "malignant tumor".
8	HIV	Diabetic retinopathy	0.11204839	–
9	HIV	Parkinson's disease	0.11009388	Parkinson's disease is the most common movement disorder in HIV-infected patients [32]. HIV can cause Parkinson's syndrome [33].
10	Suicide	HIV	0.10651195	There have been studies to find relationship between HIV and suicide [34,35]. HIV makes immune cells commit suicide [36,37].
11	Systemic lupus erythematosus	Cystic fibrosis	0.10355820	–
12	Cystic fibrosis	Chronic obstructive pulmonary disease	0.10052666	CF and COPD are associated with chronic airway inflammation [38]. A drug developed to treat CF may be useful in the treatment of COPD [39].
13	Retinitis pigmentosa	Cystic fibrosis	0.10029667	–
14	Infection	HIV	0.09516106	HIV is an infectious disease [24].
15	Suicide	Parkinson's disease	0.09422568	There have been studies to find relationship between Parkinson disease and suicide [40,41]. Apoptosis (cell suicide) plays a role in the pathogenesis of Parkinson's disease [42,43].
16	Cystic fibrosis	Tumor	0.09411687	CF patients have a high risk of malignant tumor [44,45].
17	Nephrotoxicity	Cystic fibrosis	0.09347217	CF and drug-induced nephrotoxicity are closely related [46–48].
18	HIV	Retinitis pigmentosa	0.09168938	–
19	Nephrotoxicity	Hepatotoxicity	0.09131222	Nephrotoxicity and hepatotoxicity are commonly co-occurred side effects of many chemical compounds [49–51].
20	HIV	Chronic obstructive pulmonary disease	0.08950009	An association between HIV infection and COPD has been observed in several studies [52–54].
21	Nephrotoxicity	HIV	0.08947959	HIV-related drugs cause nephrotoxicity [55–57].
22	AIDS	Cystic fibrosis	0.08890070	–
23	Cystic fibrosis	Diabetic retinopathy	0.08689588	There was a high prevalence of diabetic retinopathy in patients with cystic fibrosis-related diabetes [58].
24	Cystic fibrosis	Parkinson's disease	0.08656773	–
25	Toxicity	Cancer	0.08583446	Many cancer treatments cause toxicity [59].
26	Hepatotoxicity	Parkinson's disease	0.08351760	Hepatotoxicity is found among people with Parkinson's disease [60,61], "Hepatotoxicity in Parkinson's Disease" from eHealthMe (http://www.ehealthme.com/cs/parkinson's+disease/hepatotoxicity)].
27	Hepatotoxicity	Cystic fibrosis	0.08321186	Hepatic toxicity is induced by toxic drugs from cystic fibrosis treatment [62,63].
28	Systemic lupus erythematosus	Tumor	0.08273616	There is increasing evidence of an association between systemic lupus erythematosus and cancer [64,65].
29	Cystic fibrosis	Cancer	0.08261583	Same description to the #16 since "cancer" is a term used as "malignant tumor".
30	Hepatotoxicity	AIDS	0.08219926	Same description to the #4 since "AIDS" is a disease caused by "HIV".

literature. Likewise, the comparable methods also cannot calculate the similarities of all the disease pairs either because their calculation is based on disease ontology. Hence, the test was performed with 311,790 disease pairs of which all the proposed method and the comparable methods can calculate the similarities in common. We can only use those shared disease pairs in comparison. Answer disease pairs are 70 disease pairs which were manually checked to have a high similarity [3,22]. Cheng [13] also verified his method using the 70 disease pairs as a benchmark set. In this experiment, among the manually checked 70 disease pairs, we chose 50 disease pairs of which all the proposed method and the comparable methods can calculate the similarities in common, and used them as the answer pairs. In other words, we have 70 manually checked answer pairs, and 50 disease pairs out of 70 answer pairs were found from 311,790 disease pairs. Therefore, we used the 50 disease pairs as answer pairs and made a test set with 1000 disease pairs. Except the 50 answer pairs, 950 disease pairs were selected randomly from 311,790 disease pairs to produce total 1000 disease pairs (50 answer pairs + 950 random pairs) for a test set. Then each method carries the similarity calculation on the test set and we check how many answer disease pairs are included among top 50 disease pairs. Since the result can vary depending on the 950 randomly selected pairs, above process was performed over 10,000 times and we computed the average. Compared with the 50 top-ranking disease pairs of the comparable methods, the proposed method has the largest number of answer disease pairs and the lowest p-value (Table 3), and has the highest precision and recall (Fig. 6). When calculating precision and recall, top 50 disease pairs are classified as positives and the other 950 disease pairs are classified as negatives. Therefore, precision is equal to recall because the number of false positives is equal to the number of false negatives.

Since the number of answer disease pairs is 50 among 1000 test disease pairs, the expected number of answer disease pairs becomes 2.50 if 50 pairs are chosen randomly from the whole test pairs. The proposed method found 23.79 answer disease pairs which is the largest number, and the method showed 4.33×10^{-12} as a p-value which is the lowest value compared with the other comparable methods. Moreover, the proposed method showed 0.4759 as precision and recall, and it is higher than any other precision and recall of the other comparable methods.

We also did additional experiments to compare our method with van Driel's method [17] and Hamaneh's method [18]. Since the three methods use different databases, it is impossible to directly compare the three methods but to carry out the comparison with shared similarities between them. Our method can calculate similarities of 3,353,503 disease pairs, van Driel's method can calculate similarities of 12,900,660 disease pairs, and Hamaneh's method can calculate similarities of 3,039,345 disease pairs. Similarities of 87,300 disease pairs are shared between our method and van Driel's method, and similarities of 23,644 disease pairs are shared between our method and Hamaneh's method. We can only use those shared similarities to compare the three methods. We have 70 manually checked answer pairs, and 21 disease pairs out of 70 answer pairs were found from 87,300 disease pairs (ours – van Driel's), 9 disease pairs out of 70 answer pairs were found from 23,644 disease pairs (ours – Hamaneh's). We produced 10,000 test sets to evaluate the three methods. Each test set consists of 420 disease pairs (21 answer pairs + 399 random pairs from 87,279 pairs) for the (ours – van Driel's) experiment and 180 disease pairs (9 answer pairs + 171 random pairs from 23,635 pairs) for the (ours – Hamaneh's) experiment. Then we checked how many answer disease pairs are included among top disease pairs, and also checked p-value and precision (or recall) in 10,000 test sets similar to the previous experiment (Tables 4 and 5).

The results show that the proposed method can discover the larger number of answer disease pairs than van Driel's method and Hamaneh's method. The proposed method also showed the lowest p-value and the highest precision (or recall) among the three methods.

In the first experiment, among 1000 disease pairs of a test set, we only searched answer pairs on 50 top-ranking disease pairs. That is because we tried to validate our method in a similar way to classification problems. There are 50 answer pairs in 1000 disease pairs and our method prioritizes the disease pairs. We can classify that 50 top-ranking pairs are answers and the other 950 pairs are automatically classified to 'predicted to be not related'. Under this circumstance, we can get the number of true positives, false positives, true negatives and false negatives. Moreover, precision and recall can also be calculated as an index for validation. But we decided searching answers on every number of disease pairs is needed. We searched answers on every number of disease pairs of a test set [1–1000 for the first experiment, 1–420 for (ours – mimMiner) experiment, 1–180 for (ours – Hamaneh's) experiment] (Figs. 7–9).

The results of (ours – mimMiner) and (ours – Hamaneh's) show that our method LDDSim discovered the largest number of answer pairs in all the conditions. The result of the first experiment shows LDDSim is superior to the other methods when searching answers on 1 to nearly 200 top-ranking disease pairs but not in 200–700 intervals. Our method calculates disease–disease similarities and prioritizes the disease pairs. In prioritization problems, well-prioritizing in high position is important because we can find all the answers in low position eventually.

We also computed redundancy rate to compare CoutoLin and LDDSim because CoutoLin showed much better performance than LDDSim in 200–400 intervals in the previous experiment (Fig. 10).

Through the evaluation, we wanted to show LDDSim is a novel approach and it can discover different types of disease pairs when compared to CoutoLin. We checked redundancy of top-ranking disease pairs in comparing with the other seven comparable methods (Resnik, Lin, relevance, GIC, simIC, Wang, and CoutoResnik). The disease pairs from CoutoLin are more redundant with respect to the other seven methods than LDDSim. It indicates that LDDSim finds more novel disease pairs than CoutoLin.

The proposed method can calculate similarities of 3,353,503 disease pairs. Among them, top 30 disease pairs were manually checked to find whether they are actually related in literature (Table 6).

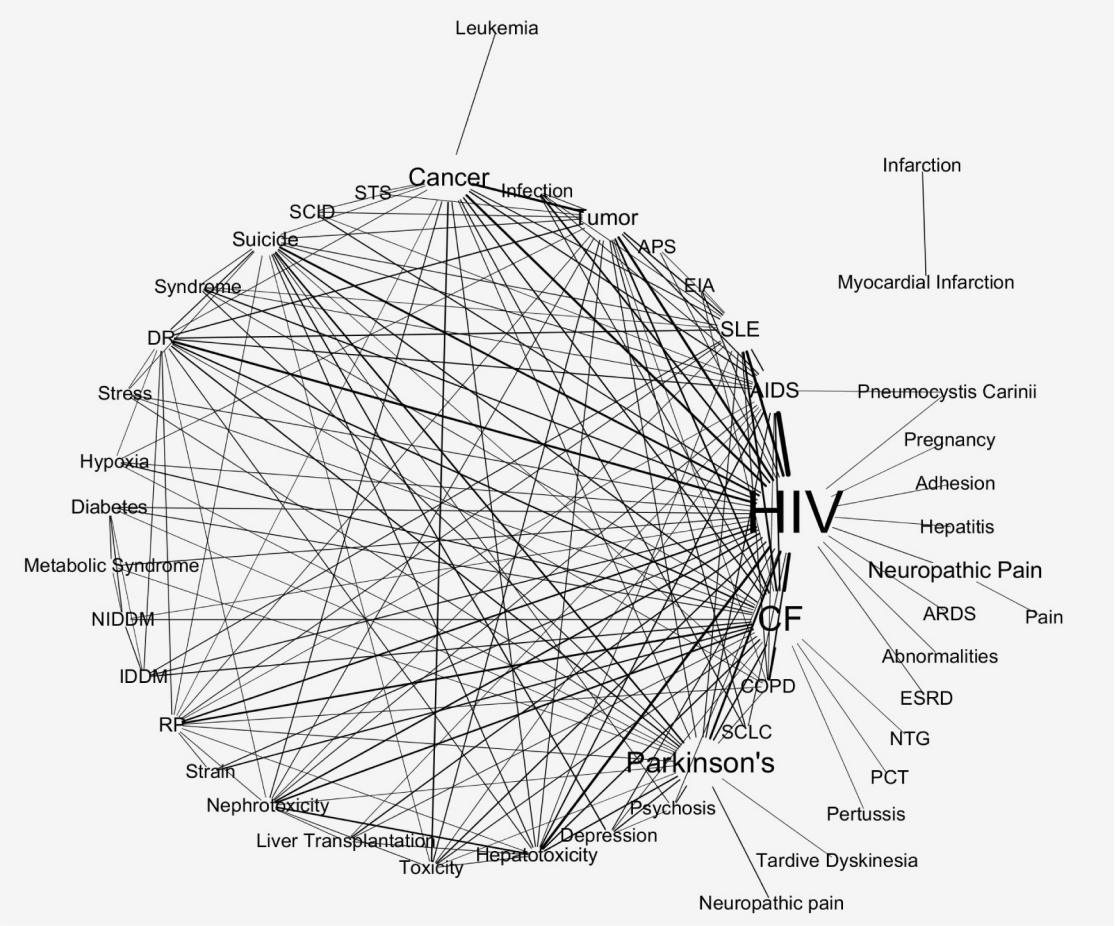

Fig. 11 – Disease network of the top 167 disease pairs from the proposed method. The size of a disease label describes betweenness centrality of the label and the width of an edge represents edge strength between the two diseases.

Table 7 – The 77 disease pairs out of the top 167 disease pairs which were not found by direct text mining.

(HIV, Hepatotoxicity), (Suicide, HIV), (Suicide, Parkinson's), (Nephrotoxicity, CF), (Nephrotoxicity, Hepatotoxicity), (Nephrotoxicity, HIV), (Hepatotoxicity, Parkinson's), (Hepatotoxicity, CF), (CF, Cancer), (Hepatotoxicity, AIDS), (Suicide, Depression), (Suicide, CF), (HIV, Liver Transplantation), (Parkinson's, Psychosis), (Suicide, DR), (Suicide, Hepatotoxicity), (Neuropathic pain, Parkinson's), (Nephrotoxicity, Cancer), (SLE, Parkinson's), (Hepatotoxicity, Cancer), (Tumor, Parkinson's), (SLE, COPD), (Hepatotoxicity, Tumor), (Hepatotoxicity, DR), (CF, NIDDM), (HIV, SCLC), (Metabolic Syndrome, HIV), (Suicide, AIDS), (Suicide, SLE), (Infection, Tumor), (Nephrotoxicity, AIDS), (Suicide, Tumor), (Metabolic Syndrome, Diabetes), (Nephrotoxicity, Tumor), (Suicide, Psychosis), (Suicide, Syndrome), (Neuropathic Pain, HIV), (RP, Parkinson's), (HIV, Hypoxia), (Nephrotoxicity, Toxicity), (Hepatotoxicity, Toxicity), (Liver Transplantation, CF), (Parkinson's, COPD), (HIV, Psychosis), (Hepatotoxicity, SLE), (Liver Transplantation, Hepatotoxicity), (RP, COPD), (RP, Strain), (RP, Tumor), (Stress, CF), (Nephrotoxicity, SLE), (Suicide, RP), (Depression, CF), (DR, COPD), (Nephrotoxicity, Liver Transplantation), (Suicide, Stress), (Nephrotoxicity, RP), (Hypoxia, Tumor), (Nephrotoxicity, DR), (Hepatotoxicity, RP), (Suicide, Cancer), (Tumor, COPD), (Suicide, Toxicity), (Hypoxia, Parkinson's), (Metabolic Syndrome, Parkinson's), (Nephrotoxicity, Parkinson's), (HIV, Pneumocystis Carinii), (Neuropathic Pain, Pain), (AIDS, Pneumocystis Carinii), (CF, Pertussis), (Metabolic Syndrome, NIDDM), (Tardive Dyskinesia, Parkinson's), (Liver Transplantation, AIDS), (Suicide, COPD), (Depression, Psychosis), (NTG, CF), (Toxicity, Parkinson's)

Table 8 – The relative importance of disease–gene matrix and disease–drug matrix.

Proposed method types	The number	p-value	Precision and recall
(Disease–gene)	11.49	4.05×10^{-3}	0.2298
(Disease–drug)	11.47	4.06×10^{-3}	0.2294
(Disease–gene) + (Disease–drug)	23.79	4.33×10^{-12}	0.4759

Table 9 – Network statistics of the disease network.

Number of nodes	Number of edges	Clustering coefficient	Network centralization	Average degree of nodes	Network density	Network heterogeneity
48	167	0.541	0.667	6.958	0.148	1.125

Literature search showed that 23 disease pairs among the top 30 disease pairs in the similarity are actually related. In addition, since the rest 7 disease pairs ("HIV, Cystic fibrosis", "HIV, Diabetic retinopathy", "Systemic lupus erythematosus, Cystic fibrosis", "Retinitis pigmentosa, Cystic fibrosis", "HIV, Retinitis pigmentosa", "AIDS, Cystic fibrosis", "Cystic fibrosis, Parkinson's disease") have the high similarity in the proposed method, they are worth researching.

3.2. *Literature-driven disease network*

We assumed that the top disease pairs from the proposed method have high possibility to have actual relations. The top 167 disease pairs from our method which are the top 0.005% of the all 3,353,503 disease pairs are chosen and used to construct a disease network (Fig. 11). The edge strength among diseases and the betweenness centrality of diseases are described in the figure. The edge strength indicates closeness between two diseases and the betweenness centrality is an indicator of a disease's centrality in the network. And we also checked disease pairs from our method which cannot be found by direct text mining (Table 7). The proposed method employs disease–gene relations and disease–drug relations to calculate disease–disease similarity. We extracted disease–disease relations directly from the MEDLINE and compared the relations with the top 167 disease pairs from our method. The 90 disease pairs were found by direct text mining and the other 77 disease pairs were proved to be novel.

4. Discussion

The proposed method extracts disease–gene relations and disease–drug relations from literature to get feature values of diseases utilizing frequency of occurrences of the relations. Then disease–disease similarities can be calculated by using the feature values. The proposed method discovered a larger number of answer disease pairs than other comparable methods and also found many actual disease pairs when manually checking the top-ranking disease pairs. We presume that our method showed good results for three reasons. First, we used literature data. Biomedical literatures contain a huge number of disease-related researches. When selecting two entities in a sentence of literature, it is very likely that the two entities are actually related. Our method also extracts disease–gene relations and disease–drug relations when two entities are co-occurred in a sentence, and therefore we believe it affects the performance of our method in a good way. Second, we used all possible gene symbols and drug names for features of a disease. The number of features and how to set up the features are crucial factors since the proposed method calculates similarities among diseases based on feature values of the diseases. If we use limited gene symbols or drugs or particular user-defined features, the results will be biased. Third, we determined feature values of diseases using the frequencies of co-occurrence of two entities which are disease and gene or disease and drug. We assumed using the frequencies of occurrence of a relation in the whole literatures as a feature value can make better results than using the number of a certain keyword in the specific disease-related literatures as a feature value.

In our method, both disease–gene matrix and disease–drug matrix are used to calculate the disease similarities. Therefore, investigation of the relative importance of the disease–gene and disease–drug co-occurrences is needed. We evaluated our method by using only one of the two matrices and compared the results with the hybrid method (Table 8).

The Table 8 shows that result of using disease–gene relations and result of using disease–drug relations are very similar. They both discovered approximately 11 answer pairs, showed about 0.004 as a p-value, and showed about 0.23 as a precision(or recall). However, when they are used together, the performance is dramatically improved. We cannot tell which one is better to make a good result but we can expect synergy effect when using both relations together. This is why both disease–gene matrix and disease–drug matrix are used together in the proposed method.

We also analyzed the disease network of the top 167 disease pairs from the proposed method (Table 9). The disease network is composed of 48 diseases and the diseases are connected by 167 edges. One disease node has about 7 neighbor disease nodes in average. The disease network is fairly centralized but has low density. It can tell us that there are degree differences within diseases and diseases with high degree are located in the core position of the disease network. We calculated betweenness centrality to find the diseases which have many neighbors and are located in the core position. The betweenness centrality of a node reflects the number of shortest paths from all nodes to all other nodes that pass through the given node and it is an indicator of a node's centrality in a network. Therefore, a disease with high betweenness centrality can be assumed as a highly influential disease. In other words, a patient infected with a disease with high betweenness centrality has high possibility to get infected with other diseases and it indicates that a disease with high betweenness centrality is worth researching. With reference to the disease network of the top 167 disease pairs from our method (Fig. 11), there are three diseases (HIV, Parkinson's disease, cystic fibrosis) which have outstanding betweenness centrality. Based on our common sense, it comes as no surprise that HIV is a high influential disease. HIV causes failure of immune system and allows opportunistic infections to flourish [66],"Opportunistic infections and their relationship to HIV/AIDS" from AIDS.gov (http://www.aids.gov/hiv-aids-basics/staying-healthy-with-hiv-aids/potential-related-health-problems/opportunistic-infections)]. Parkinson's disease is often accompanied by difficulty with eating and swallowing, digestive problems, depression, difficulties with memory and thought processes, muscle rigidity, low bone density, and osteoporosis ["Parkinson's disease" from University of Maryland Medical Center (http://umm.edu/health/medical/reports/articles/parkinsons-disease), "Symptoms & Complications" from Parkinson's NSW (http://www.parkinsonsnsw.org.au/about-parkinsons-disease/symptoms-complications)]. Acid reflux, sinusitis, joint pain and arthritis, low bone density, osteoporosis, and diabetes are possible complications of cystic fibrosis ["Additional complications" from Cystic Fibrosis Trust

(http://www.cysticfibrosis.org.uk/about-cf/living-with-cystic-fibrosis/additional-complications)]. We believe these symptoms increase probability of getting infected with other diseases. In particular, both low bone density and osteoporosis are appeared in patients with HIV, patients with Parkinson's disease and patients with cystic fibrosis. It is assumed that a person with low bone density or osteoporosis has a risk to get infected with diseases.

5. Conclusions

We calculated disease–disease similarity using literature data. Our method discovered a larger number of answer disease pairs than other comparable methods and we manually checked that 15 disease pairs out of the top 20 disease pairs have actual relations. Moreover, we constructed literature-driven disease network with the top 167 disease pairs. We presume that our method showed good results because of using literature data, using all possible gene symbols and drug names for features of a disease, and determining feature values of diseases with the frequencies of co-occurrence of two entities. The disease–disease similarities from the proposed method can be used in computational biology researches which use similarities among diseases [67–69]. The disease network can give an insight of relationship between diseases. However, our method has a limitation on calculating similarities of all possible disease pairs because our method is based on literature data. The disease ontology data or OMIM (Online Mendelian Inheritance in Man) data can be complementary to literature data. Besides, we used a small answer set (the 70 disease pairs) as a gold standard. The 70 answer disease pairs are proved to have actual relationships and a larger verified answer set is not available at this moment. The results of this paper are derived from this small answer set and a larger answer set is required to make more significant results. Finding and utilizing new data sets will be one of our further works.

Conflict of interest

None declared.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001845). We also appreciate Mr. Junsik Kim's proofreading efforts.

REFERENCES

[1] Y.I. Liu, P.H. Wise, A.J. Butte, The "etiome": identification and clustering of human disease etiological factors, BMC Bioinform. 10 (Suppl. 2) (2009) S14.
[2] S. Suthram, J.T. Dudley, A.P. Chiang, R. Chen, T.J. Hastie, A.J. Butte, Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets, PLOS Comput. Biol. 6 (2) (2010) e1000662.
[3] S. Mathur, D. Dinakarpandian, Finding disease similarity based on implicit semantic similarity, J. Biomed. Inform. 45 (2) (2011) 363–371.
[4] J. Li, B. Gong, X. Chen, T. Liu, C. Wu, F. Zhang, C. Li, X. Li, S. Rao, X. Li, DOSim: an R package for similarity between diseases based on disease ontology, BMC Bioinform. 12 (2011) 266.
[5] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 1, 1995, pp. 448–453.
[6] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings of International Conference Research on Computational Linguistics, 1997, pp. 19–33.
[7] D. Lin, An Information-theoretic definition of similarity, in: Proceedings of the 15th International Conference on Machine Learning, 1998, pp. 296–304.
[8] F.M. Couto, M.J. Silva, P.M. Coutinho, Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors, in: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005, pp. 343–344.
[9] A. Schlicker, F.S. Domingues, J. Rahnenfuhrer, T. Lengauer, A new measure for functional similarity of gene products based on gene ontology, BMC Bioinform. 7 (2006) 302.
[10] C. Pesquita, D. Faria, H. Bastos, A.O. Falcao, F.M. Couto, Evaluating GO-based semantic similarity measures, in: Proceedings of the 10th Annual Bio-Ontologies Meeting, 2007, pp. 37–40.
[11] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, C.F. Chen, A new method to measure the semantic similarity of GO terms, Bioinformatics 23 (10) (2007) 1274–1281.
[12] B. Li, J.Z. Wang, F.A. Feltus, J. Zhou, F. Luo, Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins, in: Proceedings of the 11th International Conference on Bioinformatics and Computational Biology, 2010, pp. 166–172.
[13] L. Cheng, J. Li, P. Ju, J. Peng, Y. Wang, SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association, PLOS ONE 9 (6) (2014) e99415.
[14] K. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A. Barabasi, The human disease network, Proc. Natl. Acad. Sci. 104 (21) (2007) 8685–8690.
[15] D. Lee, J. Park, K.A. Kay, N.A. Christakis, Z.N. Oltvai, A. Barabasi, The implications of human metabolic network topology for disease comorbidity, Proc. Natl. Acad. Sci. 105 (29) (2008) 9880–9885.
[16] S. Zhang, C. Wu, X. Li, X. Chen, W. Jiang, B.S. Gong, J. Li, Y.Q. Yan, From phenotype to gene: detecting disease-specific gene functional modules via a text-based human disease phenotype network construction, FEBS Lett. 584 (16) (2010) 3635–3643.
[17] M.A. van Driel, J. Bruggeman, G. Vriend, H.G. Brunner, J.A.M. Leunissen, A text-mining analysis of the human phenome, Eur. J. Hum. Genet. 14 (2006) 535–542.
[18] M.B. Hamaneh, Y. Yu, Relating diseases by integrating gene associations and information flow through protein interaction network, PLOS ONE 9 (10) (2014) e110936.
[19] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Conditional entropy and mutual information, in: Numerical Recipes 3rd Edition: The Art of Scientific Computing, Cambridge University Press, New York, 2007.

[20] F. Rossi, A. Lendasse, D. Francois, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemom. Intell. Lab. Syst. 80 (2) (2006) 215–226.
[21] X. Zhang, X.M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.K. Hao, Z.P. Liu, L. Chen, Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information, Bioinformatics 28 (1) (2012) 98–104.
[22] S. Pakhomov, B. Mclnnes, T. Adam, Y. Liu, T. Pedersen, G.B. Melton, Semantic similarity and relatedness between clinical terms: an experimental study, in: Proceedings of American Medical Informatics Association 2010 Annual Symposium, 2010, pp. 572–576.
[23] R.A. Weiss, How does HIV cause AIDS? Science 260 (5112) (1993) 1273–1279.
[24] D.C. Douek, M. Roederer, R.A. Koup, Emerging concepts in the immunopathogenesis of AIDS, Annu. Rev. Med. 60 (2009) 471–484.
[25] I. Sekigawa, H. Kaneko, T. Hishikawa, H. Hashimoto, S. Hirose, Y. Kaneko, N. Maruyama, HIV infection and SLE: their pathogenic relationship, Clin. Exp. Rheumatol. 16 (2) (1998) 175–180.
[26] I. Sekigawa, M. Okada, H. Ogasawara, T. Naito, H. Kaneko, T. Hishikawa, N. Iida, H. Hashimoto, Lessons from similarities between SLE and HIV infection, J. Infect. 44 (2) (2002) 67–72.
[27] F. Kaliyadan, HIV and lupus erythema tosus: a diagnostic dilemma, Indian J. Dermatol. 53 (2) (2008) 80–82.
[28] M.S. Sulkowski, D.L. Thomas, R.E. Chaisson, R.D. Moore, Hepatotoxicity associated with antiretroviral therapy in adults infected with human immunodeficiency virus and the role of hepatitis C or B virus infection, J. Am. Med. Assoc. 283 (1) (2000) 74–80.
[29] I. Sanne, H. Mommeja-Marin, J. Hinkle, J.A. Bartlett, M.M. Lederman, G. Maartens, C. Wakeford, A. Shaw, J. Quinn, R.G. Gish, F. Rousseau, Severe hepatotoxicity associated with nevirapine use in HIV-infected subjects, J. Infect. Dis. 191 (6) (2005) 825–829.
[30] A.E. Grulich, M.T. van Leeuwen, M.O. Falster, C.M. Vajdic, Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis, Lancet 370 (9581) (2007) 59–67.
[31] E.A. Engels, R.J. Biggar, H.I. Hall, H. Cross, A. Crutchfield, J.L. Finch, R. Grigg, T. Hylton, K.S. Pawlish, T.S. McNeel, J.J. Goedert, Cancer risk in people infected with human immunodeficiency virus in the United States, Int. J. Cancer 123 (1) (2008) 187–194.
[32] R. Bhidayasiri, D. Tarsy, HIV-induced Parkinsonism, in: Movement Disorders: A Video Atlas, Humana Press, 2012, pp. 44–45.
[33] J. Jankovic, E. Tolosa, Secondary Parkinson's syndrome, in: Parkinson's Disease and Movement Disorders, Lippincott Williams & Wilkins, 2007, pp. 216–217.
[34] O. Keiser, A. Spoerri, M.W. Brinkhof, B. Hasse, A. Gayet-Ageron, F. Tissot, A. Christen, M. Battegay, P. Schmid, E. Bernasconi, M. Egger, Suicide in HIV-infected individuals and the general population in Switzerland, 1988–2008, Am. J. Psychiatry 167 (2) (2009) 143–150.
[35] D.W. Capron, A. Gonzalez, J. Parent, M.J. Zvolensky, N.B. Schmidt, Suicidality and anxiety sensitivity in adults with HIV, AIDS Patient Care STDs 26 (5) (2012) 298–303.
[36] A. Cooper, M. García, C. Petrovas, T. Yamamoto, R.A. Koup, G.J. Nabel, HIV-1 causes CD4 cell death through DNA-dependent protein kinase during viral integration, Nature 498 (7454) (2013) 376–379.
[37] G. Doitsh, N.L. Galloway, X. Geng, Z. Yang, K.M. Monroe, O. Zepeda, P.W. Hunt, H. Hatano, S. Sowinski, I. Munoz-Arias, W.C. Greene, Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection, Nature 505 (7484) (2013) 509–514.
[38] O. Eickmeier, M. Huebner, E. Herrmann, U. Zissler, M. Rosewich, P.C. Baer, R. Buhl, S. Schmitt-Grohé, S. Zielen, R. Schubert, Sputum biomarker profiles in cystic fibrosis (CF) and chronic obstructive pulmonary disease (COPD) and association between pulmonary function, Cytokine 50 (2) (2010) 152–157.
[39] P.A. Sloane, S. Shastry, A. Wilhelm, C. Courville, L.P. Tang, K. Backer, E. Levin, S.V. Raju, Y. Li, M. Mazur, S. Byan-Parker, W. Grizzle, E.J. Sorscher, M.T. Dransfield, S.M. Rowe, A pharmacologic approach to acquired cystic fibrosis transmembrane conductance regulator dysfunction in smoking related lung disease, PLOS ONE 7 (6) (2012) e39809.
[40] E.N. Stenager, L. Wermuth, E. Stenager, J. Boldsen, Suicide in patients with Parkinson's disease. An epidemiological study, Acta Psychiatr. Scand. 90 (1) (1994) 70–72.
[41] A. Mainio, K. Karvonen, H. Hakko, T. Särkioja, P. Räsänen, Parkinson's disease and suicide: a profile of suicide victims with Parkinson's disease in a population-based study during the years 1988–2002 in Northern Finland, Int. J. Geriatr. Psychiatry 24 (9) (2009) 916–920.
[42] K.A. Jellinger, Cell death mechanisms in Parkinson's disease, J. Neural Transm. 107 (1) (2000) 1–29.
[43] N. Lev, E. Melamed, D. Offen, Apoptosis and Parkinson's disease, Prog. Neuro-Psychopharmacol. Biol. Psychiatry 27 (2) (2003) 245–250.
[44] J.P. Neglia, S.C. FitzSimmons, P. Maisonneuve, M.H. Schöni, F. Schöni-Affolter, M. Corey, A.B. Lowenfels, The risk of cancer among patients with cystic fibrosis, N. Engl. J. Med. 332 (8) (1995) 494–499.
[45] P. Maisonneuve, B.C. Marshall, E.A. Knapp, A.B. Lowenfels, Cancer risk in cystic fibrosis: a 20-year nationwide study from the United States, J. Natl. Cancer Inst. 105 (2) (2013) 122–129.
[46] C.R. Abramowsky, G.L. Swinehart, The nephropathy of cystic fibrosis: a human model of chronic nephrotoxicity, Hum. Pathol. 13 (10) (1982) 934–939.
[47] C. Godson, M.P. Ryan, D. O'Halloran, S. Bourke, H.R. Brady, M.X. FitzGerald, Investigation of aminoglycoside nephrotoxicity in cystic fibrosis patients, Scand. J. Gastroenterol. Suppl. 143 (1988) 70–73.
[48] M.D. Samaniego-Picota, A. Whelton, Aminoglycoside-induced nephrotoxicity in cystic fibrosis: a case presentation and review of the literature, Am. J. Ther. 3 (3) (1996) 248–257.
[49] R.D. Beger, J. Sun, L.K. Schnackenberg, Metabolomics approaches for discovering biomarkers of drug-induced hepatotoxicity and nephrotoxicity, Toxicol. Appl. Pharmacol. 243 (2) (2010) 154–166.
[50] C.N. Fokunang, A.N. Banin, C. Kouanfack, J.Y. Ngogang, Evaluation of hepatotoxicity and nephrotoxicity in HIV patients on highly active anti-retroviral therapy, J. AIDS HIV Res. 2 (3) (2010) 46–57.
[51] G.P. Patel, C.W. Crank, J.B. Leikin, An evaluation of hepatotoxicity and nephrotoxicity of liposomal amphotericin B (L-AMB), J. Med. Toxicol. 7 (1) (2011) 12–15.
[52] A. Morris, M.P. George, K. Crothers, L. Huang, L. Lucht, C. Kessinger, E.C. Kleerup, HIV and chronic obstructive pulmonary disease: is it worse and why? Proc. Am. Thorac. Soc. 8 (3) (2011) 320–325.
[53] C. Raynaud, N. Roche, C. Chouaid, Interactions between HIV infection and chronic obstructive pulmonary disease: clinical and epidemiological aspects, Respir. Res. 12 (2011) 117.
[54] M. Lipman, J. Brown, HIV-related chronic obstructive pulmonary disease. Are lung CD4 T cells bothered? Am. J. Respir. Crit. Care Med. 190 (7) (2014) 718–720.

[55] M. Rho, M.A. Perazella, Nephrotoxicity associated with antiretroviral therapy in HIV-infected patients, Curr. Drug Saf. 2 (2) (2007) 147–154.
[56] M. Harris, Nephrotoxicity associated with antiretroviral therapy in HIV-infected patients, Expert Opin. Drug Saf. 7 (4) (2008) 389–400.
[57] R. Kalyesubula, M. Perazella, HIV-related drug nephrotoxicity in sub-saharan Africa, Internet J. Nephrol. 6 (1) (2009).
[58] B. Yung, A. Landers, B. Mathalone, K.M. Gyi, M.E. Hodson, Diabetic retinopathy in adult patients with cystic fibrosis-related diabetes, Respir. Med. 92 (6) (1998) 871–872.
[59] I.H. Plenderleith, Treating the treatment: toxicity of cancer chemotherapy, Can. Family Phys. 36 (1990) 1827–1830.
[60] R. Benabou, C. Waters, Hepatotoxic profile of catechol-O-methyltransferase inhibitors in Parkinson's disease, Expert Opin. Drug Saf. 2 (3) (2003) 263–267.
[61] N. Borges, Tolcapone in Parkinson's disease: liver toxicity and clinical efficacy, Expert Opin. Drug Saf. 4 (1) (2005) 69–73.
[62] C. Colombo, Liver disease in cystic fibrosis, Neth. J. Med. 41 (1992) 119–122.
[63] I. Abdulhamid, V.T. Lehr, Hepatotoxicity induced by trimethoprim-sulfamethoxazole in a child with cystic fibrosis, J. Pediatr. Pharmacol. Ther. 19 (1) (2014) 56–59.
[64] S. Bernatsky, et al., An international cohort study of cancer in systemic lupus erythematosus, Arthritis Rheum. 52 (5) (2005) 1481–1490.
[65] S. Bernatsky, R. Ramsey-Goldman, A.E. Clarke, Malignancy in systemic lupus erythematosus: what have we learned? Best practice & research, Clin. Rheumatol. 23 (4) (2009) 539–547.
[66] A.C. Jung, D.S. Paauw, Diagnosing HIV-related disease: using the CD4 count as a guide, J. Gen. Intern. Med. 13 (2) (1998) 131–136.
[67] A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, Mol. Syst. Biol. 7 (2011) 496.
[68] R. Jiang, M. Gan, P. He, Constructing a gene semantic similarity network for the inference of disease genes, BMC Syst. Biol. 5 (Suppl. 2) (2011) S2.
[69] X. Chen, G. Yan, Semi-supervised learning for potential human microRNA-disease associations inference, Sci. Rep. 5 (2014) 5501.