2011 International Conference on Complex, Intelligent, and Software Intensive Systems

Associated Disease Frequency-based Measure for Finding Candidate Target Genes from the Biomedical Literature

Yeondae Kwon*, Hideaki Sugawara†, Shogo Shimizu‡ and Miyazaki Satoru*
*Department of Medicinal and Life Science
Tokyo University of Science, Noda, Chiba 278-8510, Japan
Email: yekwon@rs.noda.tus.ac.jp, smiyazaki@rs.noda.tus.ac.jp
†National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
Email: hsugawar@genes.nig.ac.jp
‡Advanced Institute of Industrial Technology
Tokyo Metropolitan University, Shinagawa, Tokyo 140-0011, Japan
Email: shimizu-syogo@aiit.ac.jp

***Abstract*—The identification of the side-effects of chemicals is the serious and costly stage in the drug development. Most side-effects are caused by their toxicity and also their correlation with other diseases than target diseases. We present a novel measure that identifies disease-associated genes from the biomedical literature in terms of causing less side-effects. This enables the identification of specific disease-associated genes, the decreased expression of which would result in a lower probability of side-effects, thus contributing to efficient drug development. Our method evaluates the specificity of a gene to a particular disease based on the number of associated diseases with the gene. In addition, we consider transitive gene-disease associations, that is, indirect gene-disease associations *via* intermediate genes. Gene-disease associations are extracted from the PubMed abstracts based on term co-occurrences. Also, we discuss the ranking results for Alzheimer disease and various cancers to verify the effectiveness of our measure. Ranking results for other diseases are available at http://www.ps.noda.tus.ac.jp/ddss/.**

***Keywords*-target gene; side-effect; gene-disease association; gene prioritization; PubMed abstract**

I. Introduction

Recently, a lot of works have been done on extracting biological knowledge, especially gene-disease associations, from the biomedical literature such as the PubMed abstracts. Though there are a number of criteria for evaluating association between diseases and genes, most of them depend on the co-occurrence frequency (that is, the number of documents) of gene and disease terms [1], [4]. Adamic *et al.* [1] propose the statistical significance of the occurrence frequency of a particular gene term under documents that contain a particular disease term. Cheng *et al.* [4] measure the degree of association between two terms by their co-occurrence frequency with other scoring strategies such as rule-based pattern matching in sentences.

For supporting new drug development, it is desirable that only specifically associated genes with a particular disease are extracted so that drug developers can avoid cost- and time-consuming wet experiments on genes also associated with other diseases, which may produce unexpected side-effects. Existing co-occurrence frequency-based measures can extract enough good candidates for disease-associated genes, but may not limit results to good target genes for the disease because they evaluate association with a particular disease, but do not explicitly consider the possibility of causing side-effects. Thus, with those measures, drug developers must verify that extracted candidate genes are actually target genes by examining whether those genes are not extracted as candidate genes for other diseases.

In this work, we propose another measure for disease-associated genes which aims at extracting specifically associated genes with a given disease. This enables the identification of associated genes that are more likely to have fewer side-effects, which contributes to efficient drug development. We measure the specificity of a gene to a disease by a tf-idf (term frequency-inverse document frequency) like method, where the number of documents in which the gene and the disease co-occur is used in the tf part and the number of diseases associated with a gene is used instead of the number of documents in the idf part. Furthermore, based on an assumption that a disease is also indirectly associated with genes *via* intermediate genes, we extend the notion of specificity to incorporate indirect gene-disease associations. This idea comes from a heuristics that if two genes co-occur in the same part of a document, then they may belong to a same family or be on a same pathway, and therefore, share same affect to diseases. This type of transitivity is also adopted for mutual information in [15] using common terms.

Our measure is different from existing co-occurrence frequency-based approaches in that it incorporates the number of associated diseases with a gene as a factor of specificity, while others, such as mutual information based on term occurrence probabilities [13], focus on association only between a particular disease and a gene. That is, other measures do not consider the number of associated diseases

978-0-7695-4373-4/11 $26.00 © 2011 IEEE
DOI 10.1109/CISIS.2011.86

. Other approaches to extract disease-associated genes include using known disease genes [9], phenotypes, expression data [6], ontologies [12], and so on. GeneSeeker [6] collects these data from multiple human and mouse databases and prioritizes candidate genes for a particular disease based on positional, expression and model data. Tiffin *et al.* use the eVOC anatomical ontology and human gene expression data, and evaluate their approach using known 17 disease genes [12]. Özgür *et al.* use protein-interaction networks [9]. Their method first constructs gene networks for a disease by literature mining based on dependency trees of sentences and support vector machines which classify sentences as describing interactions between genes or not. Then, central nodes are identified as candidate genes under the assumption that central genes in the network are likely to be associated with the disease. Yu *et al.* compare various alternatives in gene prioritization methods such as the representation of a term vector, a ranking algorithm of associated genes, and available vocabularies [16]. Though our approach uses only documents, it can be combined with other methods where additional data are available to further improve precision.

II. Material and method

The outline of our method is as follows. First, we create term dictionaries of disease and gene names. Using these term dictionaries, we create occurrence tables of disease and gene names in the collection of PubMed abstracts. By joining the occurrence tables on PubMed IDs, we obtain co-occurrence tables of disease-gene and gene-gene names. For each disease, its associated genes are extracted from the co-occurrence table of disease-gene names and specificity score is assigned to each gene according to our proposed measure. In addition, we incorporate indirect associations between genes and diseases into the specificity score so as not to miss the possibility of implicit side-effects. These indirect associations are extracted from the co-occurrence tables of disease-gene and gene-gene names.

We describe the detail of each step in the following.

A. Term Dictionaries

Gene dictionary: We downloaded human gene data from NCBI (National Center for Biotechnology Information) FTP site (ftp://ftp.ncbi.nlm.gov/gene/DATA/) in March 2010. Then, we select Entrez Gene ID, gene symbol, gene synonym, and gene name fields from the data. The gene dictionary contains a total of 115,624 entries including gene synonyms.

Disease dictionary: We use CTD (Comparative Toxicogenomics Database) disease terms in January 2010 [5] and NLM (National Library of Medicine) MeSH (Medical Subject Headings) database (http://www.nlm.nih.gov/mesh/filelist.html) in March 2010. CTD provides curated disease names, while MeSH provides a lot of synonyms for disease names. To receive benefit from the two databases, we adopt CTD disease names as primary diseases and MeSH thesaurus as synonyms for CTD disease names. The resulting disease dictionary contains a total of 45,522 entries.

Table I
Synonyms for APP and the number of occurrences.

Gene	*occ*	Gene	*occ*	Gene	*occ*
APP	4563	ABPP	14	CTFgamma	6
AAA	1238	AD1	10	CVAP	3
ABETA	6241	APPI	16	PN2	10

B. Term Occurrences

As a collection of documents, we downloaded MEDLINE/PubMed abstracts from NLM (ftp://ftp.nlm.nih.gov/nlmdata/) in January 2010 and select PubMed ID, ArticleTitle, and AbstractText fields from each abstract of total 18,502,912 documents. First, all gene symbol occurrences are extracted from the PubMed data using keyword search. The gene occurrence table consists of Gene ID, PubMed ID, and the sentence number in which a corresponding gene symbol appears in an abstract. All occurrences of synonyms of a gene are normalized into the occurrences of the corresponding single official symbol. For example, gene symbol APP, amyloid beta (A4) precursor protein, may appear as AAA, ABETA, ABPP, AD1, APPI, CTFgamma, CVAP, or PN2 in documents. Table 1 shows a list of synonyms for APP and the number of occurrences of each synonym. The total number of occurrences, in this case 10,656 (excluding duplications), is considered as the number of occurrences of APP.

In addition to keyword search, additional checking, called neighbor search, is performed to reduce false positives of gene symbol occurrences. Because gene symbols are generally created from acronyms of gene names, some symbols such as IMPACT (imprinted and ancient gene protein homolog) and LARGE (like-acetylglucosaminyltransferase) have the same spells as general words. Such symbols may produce many false positives in keyword search. Neighbor search checks whether any constitution word of a symbol appears near the symbol (in our setting, in the same sentence). Constitution words of a symbol are created by splitting its gene name into a set of words delimited by special signs such as pluses, minuses, parentheses, brackets, hyphens, and spaces. General words such as body, cell, and protein, which are defined manually, are deleted from constitution words because they do not positively support the occurrence of a particular symbol in general. If any constitution word is found in the same sentence, the occurrence of the symbol is decided to be positive. Given an occurrence of a symbol, whether neighbor search is performed is determined by the character length of the symbol and characteristic letters such as digits and hyphens.

Figure 1 shows an example of neighbor search. Consider a case that gene symbol ALK, Entrez Gene ID 238, occurs in

PubMed ID=7772531
Recent molecular characterization of the translocation breakpoint has identified a gene fusion between NPM (nucleophosmin) and ALK (anaplastic lymphoma kinase).

PubMed ID=1522609
There was no histopathological evidence of hepatic damage with ethanol alone, and no effect on hepatic cytochrome P-450 and glutathione levels or on serum levels of alanine aminotransferase (ALT), aspartate aminotransferase (AST), and alkaline phosphatase (ALK).

Figure 1. This is an example of neighbor search. From ALK's gene names, the constitution words are generated: anaplastic, lymphoma, CD246, and 2p23. Because the first sentence (PubMed ID=7772531) contains "anaplastic" and "lymphoma", this occurrence of ALK is considered to be positive. The second sentence (PubMed ID=1522609) does not contain any of the constitution words, and therefore, this occurrence is discarded.

some abstract. ALK's gene names are anaplastic lymphoma receptor tyrosine kinase, tyrosine kinase receptor, CD246 antigen, and 2p23. By splitting these gene names by delimiters, we obtain a set of constitution words as anaplastic, lymphoma, receptor, tyrosine, kinase, CD246, antigen, and 2p23. Among these words, repceptor, tyrosine, kinase, and antigen are dropped from constitution words because they are not specific words to ALK. Next, neighbor search tries to find any occurrence of one of these constitution words in the same sentence where ALK appears. In the first case in Fig. 1 (PubMed ID=7772531), we can see that the word "anaplastic" or "lymphoma" occurs. On the other hand, the second case (PubMed ID=1522609) does not contain any of these words, and therefore, this occurrence of ALK is decided to be a false positive and deleted from the gene occurrence table.

The occurrence table for disease terms is constructed using keyword search. Also, synonyms are normalized into its representative disease name. We regard the co-occurrence of gene and disease terms in a same sentence as association between the two terms. There are other alternatives to the range of co-occurrence of two terms, e.g., one document, one paragraph, and a fixed length of words. In general, a broad range generates high recall and low precision results. Among these, we chose one sentence in a same abstract because it is enough to find a small number of candidate genes that are worth being verified for new drug development. By joining the gene occurrence table and disease occurrence table on the PubMed ID and the sentence number fields, we obtain co-occurrence tables of gene-disease and gene-gene associations. Gene-gene associations are used for incorporating indirect gene-disease associations, which will be described later.

Further refinement methods to extract gene-disease associations are also applicable such as natural language processing and machine learning techniques. However, these methods take a significant amount of time and need a large amount of training data and thus, are not suited for the exhaustive analysis of a large set of documents, especially when data should be updated constantly.

C. Measuring Associations

Specificity is measured based on a tf-idf like method. The difference from the original definition is that the number of diseases associated with a particular gene is used in the idf definition rather than the number of documents in which the gene appears.

First, we define the gene term frequency (*gtf* in short). The *gtf* part evaluates the frequency of co-occurrences between a particular disease and its associated gene. Similar to the original *tf* definition, the *gtf* part of gene g with respect to disease d is defined as follows:

$$gtf_d(g) = \frac{n(d,g)}{\sum_{g'} n(d,g')},$$

where $n(d,g)$ denotes the number of documents in which d and g co-occur. In the context of drug development, a gene of high *gtf* value is more appropriate for a target gene.

Next, we define the associated disease frequency (*adf* in short). The *adf* part evaluates the specificity of co-occurrences between a particular disease and its associated gene. Let $ad(g)$ be the number of diseases sufficiently associated with gene g. Here, disease d is said to be sufficiently associated with g if $n(d,g) \geq th$, where th is a given threshold. Then, similar to the original idf definition, the *adf* part of g is defined as follows:

$$adf(g) = \log \frac{m}{ad(g)},$$

where m is the number of distinct diseases. In the context of drug development, a gene of high *adf* value has less possibility of side-effects and therefore, can be considered as a good candidate target gene.

The association score of g to d, denoted as $as_d(g)$, is defined as follows:

$$as_d(g) = gtf_d(g) \cdot adf(g).$$

D. Indirect Associations via Intermediate Genes

In addition to direct gene-disease associations, there may be indirect gene-disease associations *via* intermediate genes. The notion of indirect associations is based on the assumption that gene g' has an association with disease d if there is another gene g that co-occurs with d and frequently co-occurs with g' in literature, even if g' does not directly co-occur with d. Two genes that co-occur in a same sentence often appear owing to belonging to a same family or being located in a same pathway. Those genes are expected to share similar affect to a particular disease, since they have similar functions in the former case, and have multiplier effects in the latter case. For example, apolipoprotein E (APOE), which is known as a causal gene for Alzheimer

disease, co-occurs with lipoprotein lipase (LPL) in 74 document abstracts. This implies that diseases associated with LPL such as Cachexia may be associated also with APOE. In fact, we can expect that APOE also has some relationship to Cachexia *via* LPL because APOE and LPL are both lipid metabolism-related genes. According to this assumption, we redefine the *adf* of a gene so that the number of indirect gene-disease associations *via* intermediate genes are taken into account.

First, we define the strength of association between two genes. Again, among various methods for detecting gene relationships from the PubMed abstracts such as natural language processing, statistics [2], and multiple thesauri [11], we adopt the frequency of co-occurrences of gene symbols because of its simplicity and efficiency. Let g_i and g_j be distinct genes. Then, the *similarity* of g_i and g_j, denoted as $sim(g_i, g_j)$, is defined as:

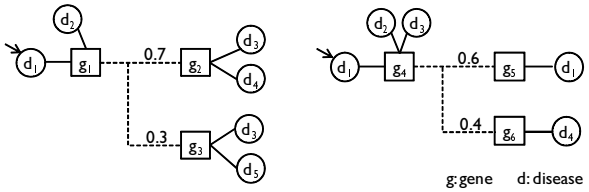
$$sim(g_i, g_j) = c_i \cdot \frac{n(g_i, g_j)}{\sum_{j \neq i} n(g_i, g_j)},$$

where $n(g_i, g_j)$ represents the number of documents in which gene g_i and g_j co-occur and c_i (≤ 1) is a weight that reflects a decline of influence by one gene-gene association. Then, an extended version of the $adf(g)$ definition that incorporates indirect gene-disease associations is defined as follows:

$$adf_I(g_i) = \log \frac{m}{ad(g_i) + \sum_{j \neq i} sim(g_i, g_j) ad_{\neq i}(g_j)},$$

where $ad_{\neq i}$ is the number of diseases other than directly associated diseases with g_i, which is introduced to avoid duplicate count of same diseases. In the above expression, the left term in the denominator corresponds to the contribution of direct gene-disease associations to the specificity and the right term corresponds to that of indirect gene-disease associations. The number of diseases associated with g_j is weighted by $sim(g_i, g_j)$ under the assumption that the probability that two genes share same associated diseases gets higher depending on the similarity of the two genes. Note that so-called hubgenes which have many links with other genes in this gene-gene association network are ranked lower in our measure because of their many indirect associations with other diseases. This is different from existing methods such as [9], where hubgenes in a protein-interaction network are extracted as most associated genes with a disease.

Figure 2 shows an example of adf_I calculation when incorporating indirect gene-disease associations. A number on a line between genes represents a *sim* value between the two genes. Assume that we want to find genes specific to disease d_1. A gene directly associated with d_1 in the left (and respectively, right) network of the figure is g_1 (and respectively, g_4). With direct associations, g_1 is more specific to d_1 than g_4 because g_1 has an association with only one



Number of directly associated diseases $= d_1 + d_2 = 2d$
Number of indirectly associated diseases $= (0.7 + 0.3)d_3 + 0.7d_4 + 0.3d_5 = 2d$
Number of associated diseases $= 2d + 2d = 4d$

Number of directly associated diseases $= d_1 + d_2 + d_3 = 3d$
Number of indirectly associated diseases $= 0.4d_4$
Number of associated diseases $= 3d + 0.4d = 3.4d$

Figure 2. Indirect assocations are incorporated into association scores. Numbers on lines between genes represent *sim* values between them. In the left network of the figure, $adf_I(g_1) = 2 + (0.7 * 2 + 0.3 * 2) = 4$. In the right network of the figure, $adf_I(g_4) = 3 + 0.4 = 3.4$. Since d_1 is already counted in the calculation of direct associations, its weight is omitted in the calculation of indirect associations. As a result, g_4 is more specific to d_1 in the context of less side-effects.

disease d_2 other than d_1. However, with direct and indirect associations, g_4 is more specific to d_1 because g_4 has less indirect associations with other diseases *via* intermediate genes. Thus, in the context of drug development, it is effective to set g_4 as a candidate target gene because g_4 is expected to have less possibility of side-effects than g_1.

The above definition is given for indirect gene-disease associations *via* one intermediate gene, but can be naturally extended for an arbitrary number of intermediate genes. To compute specificity values, it is sufficient to create an adjacency matrix, where an element of the i-th row and the j-th column is $sim(g_i, g_j)$ and obtain indirect associations *via* n intermediate genes by multiplying the matrix n times.

III. Results and Discussion

To verify the effectiveness of our measure, we experimented with well-studied Alzheimer disease (AD) by checking whether drug targets and known associated genes are ranked higher than results by the co-occurrence frequency-based measure including mutual information.

Table II shows the top-20 results for AD. APP, PSEN1, PSEN2, and APOE, which are emphasized in Table II, are known causal genes for AD [14]. Many current drug therapies of AD such as donepezil, rivastigmine and galantamine use acetyl cholinesterase (ACHE) inhibitors to reduce the rate at which acetylcholine is broken down. PharmGKB [8], a curated database of gene-drug-disease relationships, enumerates butyryl cholinesterase (BCHE) other than ACHE as drug targets of rivastigmine. ACHE and BCHE are ranked at the 5th and 20th positions by *gtf* values, respectively. However, ACHE co-occurs with over 500 disease names including AD, brain neoplasms, breast neoplasms, colonic neoplasms, intestinal neoplasms, lung neoplasms, ovarian neoplasms, stomach neoplasms, and thyroid neoplasms,

Table II
TOP-20 ASSOCIATED GENES FOR ALZHEIMER DISEASE.

Gene	$n(d,g)^{\dagger}$	$ad(g)^{\ddagger}$	Rank by					
			gtf	*adf*	adf_I	*as*	as_I	*MI**
MAPT	135	117	3	195	210	1	1	22
PSEN1	73	68	4	124	156	2	3	10
APP	543	331	1	298	306	3	2	15
PSEN2	31	34	8	71	113	4	5	8
BACE1	32	42	6	86	104	5	4	16
SORL1	14	23	11	45	51	6	6	4
APOE	243	451	2	327	330	7	7	29
CHAT	21	106	10	176	309	8	8	82
APBB1	7	18	28	34	46	9	9	20
NCSTN	4	6	40	10	14	10	11	6
TARDBP	9	57	17	109	112	11	10	23
SLC6A3	32	344	6	304	305	12	12	78
BCHE	8	60	20	115	153	13	17	61
A2M	9	75	17	135	146	14	13	26
IDE	6	35	33	73	76	15	14	34
CST3	4	12	41	25	33	16	18	9
PRNP	14	177	11	239	241	17	15	130
HTT	14	181	11	242	243	18	16	109
CDK5	8	86	21	156	198	19	23	76
ACHE	38	508	5	340	338	20	19	108

† The number of documents in which the gene and Alzheimer Disease co-occur. ‡ The number of diseases associated with the gene. * Mutual Information
A2M alpha-2-macroglobulin; *ACHE* acetyl cholinesterase (Yt blood group); *APBB1* amyloid beta A4 precursor protein-binding family B member 1; *APOE* apolipoprotein E; *APP* amyloid beta (A4) precursor protein; *BACE1* beta-site APP-cleaving enzyme 1; *BCHE* butyryl cholinesterase; *CDK5* cyclin-dependent kinase 5; *CHAT* choline acetyltransferase; *CST3* cystatin C; *HTT* huntingtin; *IDE* insulin-degrading enzyme; *MAPT* microtubule-associated protein tau; *NCSTN* nicastrin; *PRNP* prion protein; *PSEN1* presenilin 1; *PSEN2* presenilin 2 (Alzheimer disease 4); *SORL1* sortilin-related receptor, L(DLR class) A repeats-containing; *SLC6A* solute carrier family 6 member 3;

Table III
RELATED DRUGS AND TARGET GENES TO CANCERS.

Disease	Drug†	Target†	Rank‡	
			as	as_I
Breast neoplasms	capecitabine	DPYD	NR	NR
	cetuximab	EGFR	52	NR
Colonic neoplasms	anastrozole	CYP19A1	11	17
	acetaminophen	PTGS2	28	50
	cetuximab	EGFR	37	51
	lapatinib	ERBB2	86	NR
Colorectal neoplasms	capecitabine	DPYD	9	7
	cetuximab	EGFR	37	49
	bevacizumab	VEGFA	89	91
Gastrointestinal neoplasms	capecitabine	DPYD	5	1
Head and neck neoplasms	capecitabine	DPYD	43	36
	cetuximab	EGFR	1	2
	docetaxel	BCL2	25	27
Kideney neoplasms	bevacizumab	VEGFA	39	42
Liver neoplasms	celecoxib	PTGS2	65	NR
	paclitaxel	BCL2	79	NR
Lung neoplasms	cetuximab	EGFR	7	11
	lapatinib	ERBB2	18	24
	gemcitabine	RRM1	44	26
	topotecan	ABCG2	62	46
	rifampin	ABCB1	65	93
	docetaxel	BCL2	97	NR
Ovarian neoplasms	lapatinib	ERBB2	6	10
	rifampin	ABCB1	37	64
Pancreatic neoplasms	capecitabine	DPYD	88	69
	lapatinib	ERBB2	22	29
	cetuximab	EGFR	25	32
	celecoxib	PTGS2	70	NR
Prostatic neoplasms	testosterone	AR	12	8
	docetaxel	BCL2	99	NR
Stomach neoplasms	capecitabine	DPYD	21	18
	lapatinib	ERBB2	23	28
	docetaxel	BCL2	72	84
Thyroid neoplasms	sorafenib	BRAF	4	4
Uterine neoplasms	trastuzumab	EGFR	27	28

† This field is cited from PharmGKB. ‡ NR denotes that the corresponding target is not ranked in top-100 results. *ABCB1* ATP-binding cassette, sub-family B, member 1; *ABCG2* ATP-binding cassette, sub-family G, member 2; *AR* androgen receptor; *BCL2* B-cell CLL/lymphoma 2; *BRAF* v-raf muring sarcoma viral oncogene homolog B1; *CYP19A1* cytochrome P450, family 19, subfamily A, polypeptide 1; *DPYD* dihydropyrimidine dehydrogenase; *EGFR* epidermal growth factor receptor; *ERBB2* e-erb-b2 erythroblastic leukemia viral oncogene homolog 2; *PTGS2* prostaglandin-endoperoxide synthase2; *RRM1* ribonucleotide reductase M1; *VEGFA* vascular endothelial growth factor A

while BCHE co-occurs with only 60 disease names including AD and lung neoplasms. BCHE does not co-occur with other neoplasms other than lung neoplasms, and thus, BCHE is ranked higher than ACHE by *as* values. Actually, PharmGKB enumerates a lot of side-effects of donepezil such as severe nausea, vomiting, salivation, sweating, bradycardia, hypotension, respiratory depression, collapse and convulsions as symptoms of overdose.

Next, we examine the effectiveness of indirect associations *via* intermediate genes. There are over 100 genes associated with BCHE including APOE. According to an article of PubMed ID 15519745, there is a synergic association between butyrylcholinesterase-K variant (BChE-K) and apolipoproteinE-epsilon 4 (ApoE-epsilon 4) to promote risk for AD. APOE co-occurs with 451 diseases such as atherosclerosis and hypercholesterolemia. Thus, one of the reasons that the rank of BCHE gets lower by as_I than by *as* is that a number of diseases associated with APOE are considered to be transitively associated with BCHE in the as_I measure.

Table III shows a summary of the ranking results for cancers that contain "neoplasms." The 2nd and 3rd columns are related drugs and target genes to cancers, where their field values are cited from PharmGKB. The 4th and 5th columns represent the ranks of the corresponding target genes by *as* and as_I measures, respectively.

Capecitabine, which is emphasized in Table III, is a drug given as a treatment for many types of cancers, including breast cancer, colorectal cancer, gastrointestinal cancer, head and neck cancer, pancreatic cancer, and stomach cancer, and has side effects such as fatigue, diarrhea, constipation headaches, conjunctivitis, and anorexia [10]. Dihydropyrimidine dehydrogenase (DPYD), a target gene of capecitabine, is ranked at the 5th and the 1st positions for gastrointestinal neoplasms by *as* and as_I, respectively, but is ranked much lower than for other cancers and does not appear in even top-100 results for breast cancer in both measures. Therefore, we can expect that DPYD is the most specific gene to gastrointestinal cancer because the gene is strongly associated with gastrointestinal cancer and capecitabine that uses DPYD as

a target gene is more effective to gastrointestinal cancer than other cancers. Furthermore, DPYD is more associated with colorectal cancer, gastrointestinal cancer and stomach cancer than with other cancers such as breast cancer and pancreatic cancer. This implies that capecitabine is effective to intestine- and stomach-related cancers.

Cetuximab, which is underlined in Table III, is also a drug given as a treatment for many types of cancers. A target gene of cetuximab, epidemal growth factor receptor (EGFR), is ranked at the 1st position by *as* for head and neck cancer. This implies that EGFR is deeply associated with the two cancers compared to other cancers. Thus, we can expect that cetuximab is more effective to head and neck cancer than other cancers, as shown in [3], [7].

From this experiment, it was found that there is a possibility that better results can be obtained by our measure. Also, it can be expected that this method is effective as a filtering step for identifying candidate target genes.

IV. Conclusions

We proposed a novel measure for identifying candidate target genes that incorporates direct and indirect disease-gene associations in literature and experimented with some diseases to examine the effectiveness of the measure.

Although the way of measuring the possibility of side-effects depends on the associated disease frequency defined simply by co-occurrences in literature, the experimental results show that incorporating associations with other diseases can better filter candidates which may cause side-effects.

As a future work, we plan to perform further detailed analysis and extend our method to combine other fact data such as pathways with the literature analysis. Developing a proper test set for target genes with less side-effects is also required.

Acknowledgment

This work was supported by Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST).

References

[1] L.A. Adamic, D. Wilkinson, B.A. Huberman, and E. Adar, *A literature based method for identifying gene-disease connections*, *Proc. IEEE Conf. Bioinformatics*, 109–117, 2002.

[2] M.A. Andrade and A. Valencia, *Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families*, *Bioinformatics*, 14(7):600–607, 1998.

[3] J.A. Bonner, P.M. Harari, J. Giralt, R.B. Cohen, C.U. Jones, R.K. Sur, D. Raben, J. Baselga, S.A. Spencer, J. Zhu, J. Youssoufian, E.K. Rowinsky, and K.K. Ang, *Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival*, *Lancet Oncol.*, 11(1):21–28, 2010.

[4] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. Wishart, *PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites*, *Nucleic Acids Res.*, 36:W399–W405, 2008.

[5] A.P. Davis, C.G. Murphy, C.A. Saraceni-Richards, M.C. Rosenstein, T.C Wiegers, and C.J. Mattingly, *Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks*, *Nucleic Acids Res.*, 37:D786–D792, 2009.

[6] M.A. van Driel, K. Cuelenaere, P.P.C.W. Kemmeren, J.A.M. Leunissen, H.G. Brunner, and G. Vriend, *GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases*, *Nucleic. Acids Res.*, 33:W758–W761, 2005.

[7] R.S. Herbst and C.J. Langer, *Epidermal growth factor receptors as a target for cancer treatment: the emerging role of IMC-C225 in the treatment of lung and head and neck cancers*, *Semin. Oncol.*, 29:27–36, 2002.

[8] T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman, *Integrating Genotype and Phenotype Information: An Overview of the PharmGKB Project*, *The Pharmacogenomics Journal*, 1:167–170, 2001.

[9] A. Özgür, T. Vu, G. Erkan, and D.R.Radev, *Identifying gene-disease associations using centrality on a literature mined gene-interaction network*, *Bioinformatics*, 24(13):i277–i285, 2008.

[10] M.W. Saif, *Targeting cancers in the gastrointestinal tract: role of capacitabine*, *Onco Targets and Therapy*, 2:29–41, 2009.

[11] M. Stephens, M. Palakal, S. Mukhopadhyay, and R. Raje, *Detecting gene relations from Medline abstracts*, *Pac. Symp. Biocomput.*, 6:483–496, 2001.

[12] N. Tiffin, J.F. Kelso, A.R. Powell, H. Pan, V.B. Bajic, and W.A. Hide, *Integration of text- and data-mining using ontologies successfully selects disease gene candidates*, *Nucleic. Acids Res.*, 33(5):1544–1552, 2005.

[13] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, *FACTA: a text search engine for finding associated biomedical concepts*, *Bioinformatics*, 24(21):2559–2560, 2008.

[14] S.C. Waring and R.N. Rosenberg, *Genome-wide association studies in Alzheimer disease*, *Arch. Neurol.*, 65(3):329–334, 2008.

[15] J.D. Wren, *Extending the mutual information measure to rank inferred literature relationships*, *BMC Bioinformatics*, 5(1):145, 2004.

[16] S. Yu, S.V. Vooren, L.C. Tranchevent, B.D. Moor, and Y. Moreau, *Comparison of vocabularies, reprensetations and ranking algorithms for gene prioritization by text mining*, *Bioinformatics*, 24(16):i119–i125, 2008.