Abstract: Identification of candidate target genes related to a particular disease is an important stage in drug development. A number of studies have extracted disease-related genes from the biomedical literature. We herein present a novel evaluation measure that identifies disease-associated genes and prioritizes the identified genes as drug target genes in terms of fewer side-effects using the biomedical literature. The proposed measure evaluates the specificity of a gene to a particular disease based on the number of diseases associated with the gene. The specificity of a gene is measured by means of, for example, term frequency-inverse document frequency (tf-idf), which is widely used in Web information retrieval. We assume that if a gene is chosen as a target gene for a disease, then side-effects are more likely to occur as the number of diseases associated with the gene increases. We verified the obtained ranking results by checking the ranks of known drug targets. As a result, 177 known drug targets were found to be ranked within the top 100 genes, and 21 drug targets were top ranked. The results suggest that the proposed measure is useful as a primary filter for extracting candidate target genes from a large number of genes.

Keywords: drug targets, side-effects, gene prioritization, gene-disease association, PubMed abstract

1. Introduction

Recently, a number of studies have investigated the extraction of biological knowledge, particularly gene-disease associations, from the biomedical literature, such as PubMed abstracts [1], [2], [3], [4]. Although there are a number of criteria for evaluating the associations between diseases and genes, most of these criteria depend on the co-occurrence frequency (i.e., the number of documents) of gene and disease terms. For example, Adamic et al. [1] proposed the statistical significance of the occurrence frequency of a particular gene term in documents that contain a particular disease term. Cheng et al. [3] measured the degree of association between two terms based on their co-occurrence frequency with other scoring strategies such as rule-based pattern matching in sentences.

On the other hand, for the purpose of supporting new drug development, it is desirable that only genes that are specifically associated with a particular disease are identified so that drug developers can avoid costly and time-consuming wet experiments with genes which have associations with other diseases, i.e., genes that may have side-effects. Although frequency-based measures can present a sufficient number of good candidates for finding genes related to a particular disease, extracted genes may also have associations with other diseases which are not of interest. In other words, these genes are probably related, but may not be good target genes for the disease. Therefore, in order to ensure that extracted genes can actually be used as target genes, these genes must be verified not to be extracted as candidate target genes for other diseases.

In the present study, we propose another measure for extracting genes specifically associated with a given disease. This enables the identification of associated genes that are expected to have fewer side-effects, which contributes to efficient drug development. The specificity of a gene to a disease is measured by, for example, tf-idf, which is widely used in Web information retrieval.

The proposed measure is different from existing approaches in that it incorporates the number of associated diseases as a factor of specificity, whereas other approaches, such as mutual information based on term occurrence probabilities [5], focus on the association between a particular pair consisting of a disease and a gene and do not distinguish between other associated diseases. We consider that if a gene is chosen as a target gene for a disease, then the likelihood that side-effects will occur increases as the number of diseases associated with the gene increases.

Another approach to extracting gene-disease associations is to use additional data such as known disease genes [6], phenotypes [7], expression data [8], and ontologies [9]. GeneSeeker [8] 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. [9] used the eVOC anatomical ontology [10] and human gene expression data, and evaluated their approach using 17 known disease genes. Moreover, protein-interaction networks can be used to predict gene-disease associations. As an example of such approaches, the method proposed by Özgür et al. [6] first constructs gene net-
works for a disease by literature mining based on dependency trees of sentences and support vector machines which classify sentences based on whether they describe interactions between genes. Central nodes are then identified as candidate genes under the assumption that central genes in the network are likely to be associated with the disease. Yu et al. [11] compared various alternatives to gene prioritization methods, such as the representation of a term vector, a ranking algorithm of associated genes, and available vocabularies. As a result, they concluded that inverted document frequency (idf), 1-SVM, and eVOC and MeSH vocabularies are most effective. Additional vocabularies and sophisticated methods using natural language processing and machine learning techniques improve the precision of association extraction, but require datasets in addition to literature and/or excessive time to analyze the entire set of documents. The proposed approach currently uses documents only, but can be combined as a method with these approaches, where additional data such as pathways are available.

2. Material and Method

We use the co-occurrence frequency of disease and gene names in the literature as evaluation criteria for relationships between human diseases and genes. Considering a gene as a target gene for a particular disease, we evaluate the possibility that the gene causes side effects using the number of distinct diseases associated with the gene. We extract the gene as candidates for drug target genes when the association between the gene and the target disease is strong and associations between the gene and other diseases are weak. We quantify the number of diseases related to the gene by considering the relationship with the target disease.

2.1 Term Dictionaries

Gene dictionary: We downloaded human gene data from the FTP site of the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nlm.gov/gene/DATA/) for November 2011. We then selected Entrez Gene ID, gene symbol, gene synonym, and gene name fields from the data. The gene dictionary contains a total of 117,170 entries, including gene synonyms.

Disease dictionary: We used disease terms of the Comparative Toxicogenomics Database (CTD) [12] for February 2011 and the Medical Subject Headings (MeSH) database of the National Library of Medicine (NLM) (http://www.nlm.nih.gov/mesh) for November 2011. The CTD provides curated disease names, whereas the MeSH provides numerous synonyms for disease names. In order to receive the benefit of these two databases, we adopt CTD disease names as primary diseases and the MeSH thesaurus as synonyms for CTD disease names. The resulting disease dictionary contains a total of 48,480 entries.

2.2 Term Occurrences

As a collection of documents, we downloaded MEDLINE/PubMed abstracts from NLM (ftp://ftp.nlm.nih.gov/pubmed/) for February 2011 and extracted the PubMed ID, ArticleTitle, and AbstractText fields from each abstract for a total of 19,019,815 documents. First, all of the gene symbol occurrences are extracted from the PubMed data by 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 of the occurrences of synonyms of a gene are normalized into the occurrences of the corresponding single official symbol. For example, gene symbol APP, beta-site amyloid precursor protein, may appear as AAA, ABETA, ABPP, AD1, APPI, CTFgamma, CVAP, or PN2 in documents.

In addition to keyword search, additional checking, referred to as neighbor search, is performed in order to reduce the number of false positives of gene symbol occurrences. Some official symbols and synonyms of short length have the same spellings as general words, such as cell, which is a synonym for carboxyl ester lipase (CEL). Since gene symbols are generally created from acronyms of gene names, some symbols, such as a gene IMPACT (imprinted and ancient gene homolog), have the same spellings as general words. Such symbols may produce numerous false positives in a keyword search. Neighbor search checks whether any constitution word of a symbol appears near the symbol (i.e., in the same sentence). Constitution words of a symbol are created by splitting a 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 the constitution words because they do not, in general, positively support the occurrence of a particular symbol. If any constitution word is found in the same sentence, the occurrence of the symbol is decided to be positive. Given the occurrence of a symbol, whether a 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 the case in which gene symbol ALK, Entrez Gene ID 238, occurs in an abstract. The gene names of ALK 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: anaplastic, lymphoma, receptor, tyrosine, kinase, CD246, antigen, and 2p23. Among these words, receptor, tyrosine, kinase, and antigen are dropped as constitution words because these words are not specific to ALK. Next, the

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

This is an example of neighbor search. From the gene names of ALK, the constitution words are generated: anaplastic, lymphoma, CD246, and 2p23. In the first case (PubMed ID=7772531), anaplastic and lymphoma occurs. Therefore, this occurrence of ALK is considered to be positive. In the second case (PubMed ID=1522609), none of these constitution words appear. Therefore, this occurrence is discarded.
neighbor search attempts to find any occurrence of one of these constitution words in the sentence in which ALK appears. In the first case, as shown in Fig. 1 (PubMed ID=7772531), we can see that the word anaplastic or lymphoma occurs. On the other hand, in the second case (PubMed ID=1522609), none of these words appear. Therefore, this occurrence of ALK is decided to be a false positive and is deleted from the gene occurrence table.

The occurrence table for disease terms is constructed by keyword search. Moreover, synonyms are normalized into the representative disease name. We regard the co-occurrence of gene and disease terms in the same sentence as the 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 string of words of a fixed length. In general, a broad range generates high recall and low-precision results. For the present study, we chose one sentence in the same abstract because it is sufficient to find a small number of candidate genes that are worth being verified for new drug development. By combining the gene occurrence table and the disease occurrence table for the PubMed ID and the sentence number fields, we obtain co-occurrence tables of gene-disease associations.

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

### 2.3 Measuring Specificity

For the purpose of supporting new drug development, it is desirable that only genes specifically associated with a particular disease are identified so that drug developers can avoid conducting costly and time-consuming wet experiments with genes which have associations with other diseases, i.e., genes that may have side-effects. The proposed method of measuring specificity is based on a tf-idf method [13]. The difference from the original definition is that the number of diseases associated with a particular gene is used in the idf definition instead of the number of documents in which a gene appears, because we want to estimate the specificity in terms of gene-disease associations.

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

$$gtf_d(g) = \frac{n_d(g)}{\sum_{g_i} n_d(g_i)}$$

where $n(d,g)$ denotes the number of co-occurrences of $d$ and $g$. In the context of drug development, a gene of high gtf value is more appropriate as a target gene.

Next, we define the associated disease frequency (adf). The adf term evaluates the specificity of co-occurrences between a particular disease and its associated gene. Then, similar to the original idf definition, the adf term of $g$ is defined as follows:

$$adf_g(g) = \log \frac{m}{adf_d(g)}$$

where $m$ is the number of distinct diseases. Since highly related diseases are more likely to have side-effects than non-related diseases, the degree of influence of side-effects of each related disease should be counted differently. Therefore, we define $adf_d(g)$ as the virtual number of diseases expected to be related to $g$.

Based on the assumption that related diseases are associated with similar genes, the relationship between two diseases $d_1$ and $d_2$ is defined as follows:

$$drel(d_1, d_2) = \frac{|G(d_1) \cap G(d_2)|}{|G(d_1) \cup G(d_2)|}$$

where $G(d)$ is a set of genes related to disease $d$ and $|G|$ denotes the cardinality of $G$. Then, $adf_d(g)$ is defined as the sum of $drel(d, d_i)$ for every $d_i$, which is decided to be related to $g$. In the context of drug development, a gene of high adf has a lower possibility of side-effects and therefore can be considered to be a good target gene. Finally, the association score of $g$ to $d$, denoted as $as_g(d)$, is defined as follows:

$$as_g(d) = gtf_d(g) \cdot adf_g(g)$$

### 3. Results

We verify the effectiveness of the proposed measure by examining whether known drug targets and causative genes are ranked higher by the proposed measure.

#### 3.1 Ranks of Known Drug Targets

In order to examine comprehensively the ranks of known drug targets by the proposed measure, we investigated the relationships between known drugs, known targets, and diseases using data in DrugBank [14] and PharmGKB [15] for March 2012. As a result, we obtained 9,360 associations between 389 diseases, 407 drugs, and 540 target genes. We checked whether these associations are obtained by the proposed measure, and checked the ranks of the 540 target genes in our ranking results. As a result, we found that the proposed measure extracts 2,982 associations between 253 diseases, 335 drugs, and 349 targets, and 177 target genes are ranked within the top 100 genes, which is 33% of all known drug targets.
| Target | Disease                                      | Drug                                      |
|--------|---------------------------------------------|-------------------------------------------|
| ABL1   | Leukemia                                    | imatinib                                  |
|        | Philadelphia-positive myeloid leukemias      | dasatinib, imatinib, nilotinib            |
|        | Myeloid leukemias                           | imatinib                                  |
| ACE    | Angioedema                                   | captopril, enalapril, fosinopril, lisinopril |
|        | Heart failure                               | perindopril                               |
|        | Hypertension                                | benazepril, captopril, enalapril, fosinopril, lisinopril, perindopril, ramipril |
|        | Myocardial infarction                       | captopril, enalapril, fosinopril, lisinopril |
| AR     | Gynecomastia                                | spironolactone                            |
| CBS    | Homocystinuria                              | pyridoxine                                |
| CHRNA5 | Tobacco use disorder                        | nicotine                                  |
| DHFR   | Plasmodium vivax malaria                    | proguanil                                  |
| DRD4   | Psychotic disorders                         | aripiprazole, clozapine, olanzapine, quetiapine, risperidone, ziprasidone |
| EGFR   | Adenocarcinoma                              | erlotinib                                  |
|        | Non-small-cell lung carcinoma               | cetuximab, erlotinib, gefitinib, lapatinib, panitumumab, trastuzumab |
|        | Exanthema                                   | cetuximab, erlotinib, gefitinib, panitumumab |
|        | Glioblastoma                                | cetuximab, erlotinib, gefitinib           |
|        | Head and neck neoplasms                     | erlotinib                                  |
|        | Lung neoplasms                              | erlotinib, gefitinib, lapatinib, panitumumab |
| ERBB2  | Breast neoplasms                            | lapatinib, trastuzumab                    |
| HTR2C  | Medication-induced dyskinesias              | risperidone                                |
| KCNH2  | Torsades de pointes                         | amidoradine, casprade, dofe til de, halofantrine, ibutide, pimoa zide, quinidine, sertindole, terfenadine |
| KCNQ1  | Long QT syndrome                           | bepridil                                   |
| KIT    | Gastrointestinal stromal tumors             | imatinib, sorafenib, sunitinib            |
| MS4A1  | Non-Hodgkin’s lymphoma                      | rituximab                                  |
|         | Waldenstrom Macroglobulinemia               | rituximab                                  |
| MTHFR  | Hyperhomocysteinemia                        | cyanocobalamin                             |
| OPRM1  | Opioid-related disorders                    | buprenorphine, methadone                  |
| PAH    | Phenylketonurias                            | tetrahydrobiopterin                       |
| PTGS2  | Stomach neoplasms                           | ibuprofen                                  |
| RYR1   | Malignant hyperthermia                      | caffeine                                   |
| SLC6A3 | Attention deficit disorder with hyperactivity | venlafaxine                             |
|        | Cocaine-related disorders                   | cocaine                                    |
|        | Anxiety disorders                           | citalopram, duloxetine, fluoxetine, nefazodone, paroxetine, sertraline, venlafaxine |
| TNF    | Arthritis                                   | adalimumab, etanercept, infliximab        |
|        | Rheumatoid arthritis                        | adalimumab, etanercept, infliximab        |
|        | Inflammatory bowel diseases                 | infliximab                                 |
|        | Falciparum malaria                          | chloroquine                                |

**Table 1** Target genes ranked at the top.

**Figure 2** shows the number of known target genes ranked within the top 100 genes. Each bar in the histogram represents the number of target genes ranked within each interval and may include multiple occurrences of the same target gene because we counted the number of target genes of prescription drugs depending on each disease. For example, acetylcholinesterase (ACHE), a target gene of galantamine, is counted six times within the top 100 gene. Galantamine is used as a treatment for dementia, Alzheimer’s disease, arteriosclerotic dementia, apraxia, neurodegenerative disease, and aphasia. ACHE is ranked 6th, 7th, 7th,
65th, 75th, and 82nd as a gene related to these diseases, respectively, and thereby is counted six times. Similarly, the leftmost bar in the histogram represents that the number of target genes is ranked within the top 10 is 204. Excluding duplicates, 68 out of 4,402 genes are used as targets for these prescription drugs. In the proposed measure, there are 4,402 genes associated with breast cancer, and 75 genes are used as targets for these prescription drugs. In the proposed measure, there are 4,402 genes associated with breast cancer, and 75 genes are used as targets for these prescription drugs. We found many known drug targets to be top ranked by the proposed measure.

3.2 Ranks of Target Genes to Cancers

Table 2 shows the target genes ranked in the top most position among the target genes used in prescription drugs for each cancer containing neoplasms. For example, 58 prescription drugs for breast cancer exist, and 75 genes are used as targets for these prescription drugs. In the proposed measure, there are 4,402 genes associated with breast cancer, and 75 genes are used as targets for these prescription drugs. In the proposed measure, there are 4,402 genes associated with breast cancer, and 75 genes are used as targets for these prescription drugs. From Table 2, we found that most of the known target genes involved in cancers are ranked within the top 100 genes by the proposed measure.

3.3 Ranks of Alzheimer’s Disease

Table 3 shows the associated genes ranked higher for Alzheimer’s disease. The proposed measure extracts 733 genes associated with Alzheimer’s disease and ranks beta-site amyloid precursor protein (APP), apolipoprotein E, presenilin 1, and presenilin 2 as causative genes for Alzheimer’s disease [18]. Among these causal genes, KCNJ11 and ABC288 mutations are recognized as the major causes of TNDM and PNDM, and PPAR is known as a clinical risk factor for type 2 diabetes mellitus [19]. As shown in Table 4, these genes are ranked 22nd, 49th, and 10th, respectively. Moreover, these genes are drug targets for glimepiride, gliclazide, and repaglinide, for example.

3.4 Ranks for Diabetes Mellitus, Type 2

Neonatal diabetes mellitus (NDM) is classified clinically into a transient form (TNDM) and a permanent form (PNDM), and its causal genes are identified [18]. Among these causal genes, KCNJ11 and ABC288 mutations are recognized as the major causes of TNDM and PNDM, and PPAR is known as a clinical risk factor for type 2 diabetes mellitus [19]. As shown in Table 4, these genes are ranked 22nd, 49th, and 10th, respectively. Moreover, these genes are drug targets for glimepiride, gliclazide, and repaglinide, for example.

4. Discussion

We evaluated the proposed measure by checking whether drug targets were listed at higher ranks as disease-associated genes. We found many known drug targets to be top ranked by the proposed measure.

We summarized the ranks of drug targets to cancers in Table 2 in terms of the top targets. Most of the drug targets are ranked within the top 20, but B-cell CLL/lymphoma 2 (BCL2) is ranked 86th as a target gene for docetaxel prescribed to treat nasopharyngeal cancer. We may guess whether docetaxel is an effective drug and/or whether a more effective drug exists for nasopharyngeal cancer. As shown in Table 5, docetaxel is also used to treat

### Table 2 Ranks of target genes to cancers.

| Cancer               | Drug                  | Target Rank |
|----------------------|-----------------------|-------------|
| Breast neoplasms     | lapatinib, trastuzumab| ERBB2 1     |
| Colonic neoplasms    | celecoxib             | PGI2S 3     |
| Colorectal neoplasms | cetuximab, erlotinib  | EGFR 5      |
| Esophageal neoplasms | erlotinib             | EGFR 13     |
| Head and neck neoplasms | erlotinib             | EGFR 1      |
| Kidney neoplasms     | everolimus, temsirolimus | MTOR 7    |
| Liver neoplasms      | pacitaxel             | BCL2 33     |
| Lung neoplasms       | cetuximab, erlotinib  | EGFR 1      |
| Nasopharyngeal neoplasms | docetaxel           | BCL2 86     |
| Ovarian neoplasms    | lapatinib, trastuzumab| ERBB2 4     |
| Pancreatic neoplasms | cetuximab, erlotinib  | EGFR 4      |
| Prostatic neoplasms  | testosterone          | AR 3        |
| Rectal neoplasms     | cetuximab             | EGFR 18     |
| Stomach neoplasms    | ibuprofen             | PTGS2 1     |
| Thyroid neoplasms    | sorafenib             | BRAF 4      |
| Uterine neoplasms    | trastuzumab           | EGFR 41     |

### Table 3 Ranks of associated genes for Alzheimer’s disease.

| Gene Description                                      | Rank |
|-------------------------------------------------------|------|
| AR androgen receptor; BCL2 B-cell CLL/lymphoma 2; BCL2 v-raf murine sarcoma viral oncogene homolog B1; EGFR epidermal growth factor receptor; ERBB2 e-erb-b2 erythroblastic leukemia viral oncogene homolog 2; PTG32 prostaglandin-endoperoxide synthase2; RRM1 ribonucleotide reductase M1; VEGFA vascular endothelial growth factor A |      |

### Table 4 Ranks of drug targets for Diabetes Mellitus.

| Gene Description                  | Drug                  | Rank |
|-----------------------------------|-----------------------|------|
| ACE                               | captoril, enalapril    | 9    |
| PPAR                              | glipizide, nateglizide, pioglitazone, repaglinide | 10   |
| KCNJ11                            | glimepiride, verapamil | 22   |
| PPARA                             | fenofibrate           | 28   |
| ABC288                            | chlorpropamide, gliclazide, nateglinide, repaglinide, tolbutamide | 49   |

ACE androgen receptor; PPARA peroxisome proliferator-activated receptor gamma; KCNJ11 potassium inwardly-rectifying channel, subfamily J, member 1; PPARA peroxisome proliferator-activated receptor alpha; ABC288 ATP-binding cassette transporter subfamily C member 8

### Table 5 Ranks of associated genes for Diabetes Mellitus.

| Gene Description                  | Drug                  | Rank |
|-----------------------------------|-----------------------|------|
| KCNJ11                            | glimepiride, verapamil | 22   |
| PPARA                             | fenofibrate           | 28   |
| ABC288                            | chlorpropamide, gliclazide, nateglinide, repaglinide, tolbutamide | 49   |

ACE angiotensin I converting enzyme; PPARA peroxisome proliferator-activated receptor gamma; KCNJ11 potassium inwardly-rectifying channel, subfamily J, member 1; PPARA peroxisome proliferator-activated receptor alpha; ABC288 ATP-binding cassette transporter subfamily C member 8
other cancers. Each figure in the second column represents the rank of BCL2 among genes associated with each cancer in the first column. Since BCL2 is ranked the lowest for nasopharyngeal cancer, we may guess that docetaxel may be not an effective drug for nasopharyngeal cancer.

Similarly, we can anticipate that one of the genes ranked within the top 85 genes may be a new target gene based on our ranking results. As shown in Table 6, stanniocalcin-1 (STC1) is ranked at the top for nasopharyngeal cancer. The roles of this gene in apoptosis have recently been reported [20], [21], [22]. Lai et al. [20] and Ching et al. [22] reported studies on STC1 expression in apoptotic human nasopharyngeal cancer cells. This implies the possibility that STC1 may be one of the new targets for nasopharyngeal cancer. With a background in drug discovery, a new target gene for a new drug could be easily found from our ranking lists.

Also, gap junction alpha-1 protein (GJA1) ranked 16th has been used as a target gene for carvedilol. Carvedilol has been prescribed as a treatment for many diseases including cardiac arrhythmias, cardiomyopathy, and heart failure, which seem likely to be related to heart. Alajez et al. [23] reported that GJA1 is a target gene for underexpression of a microRNA called miR-218 in nasopharyngeal carcinoma tissues. This implies the possibility that carvedilol can be used as a new treatment for nasopharyngeal cancer.

We showed the top-32 associated genes for Alzheimer’s disease in Table 3. For microtubule-associated protein tau (MAPT) ranked 2nd, Spillantini et al. [24] reported that mutations in this gene cause neurodegenerative diseases, many of which are frontotemporal dementias. This implies that MAPT is likely to be one of target genes for Alzheimer’s disease. Also, beta-site APP-cleaving enzyme 1 (BACE1), which is ranked 6th, has been developed as a drug target for Alzheimer’s disease [25]. BACE1 cleaves APP and produces beta-amyloid (Aβ) peptides that are the main component of the amyloid plaques deposits in the brains of Alzheimer’s disease patients [26]. This implies that genes ranked higher by the proposed measure could become target genes in the future.

Moreover, we found that solute carrier family 6 member 3 (SLC6A3) ranked 9th for Alzheimer’s disease is ranked at the top for attention deficit hyperactivity disorder (ADHD). SLC6A3 has used as a target gene for venlafaxine, which is prescribed as a treatment for diseases related to nerve including ADHD [27]. This implies that venlafaxine can be used as one of the drug therapies for Alzheimer’s disease.

We also identified which drug can be expected to be the most effective therapy for each cancer at the present time. As shown in Table 7, capecitabine is used 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 [28]. Dihydropyrimidine dehydrogenase (DPYD), a target gene of capecitabine, is ranked 8th for gastrointestinal neoplasms but is ranked much lower for other cancers. Therefore, we can expect that DPYD is the gene most specific to gastrointestinal cancer because the gene is strongly associated with gastrointestinal cancer, and capecitabine, which uses DPYD as a target gene, is more effective for treating gastrointestinal cancer than other cancers. Furthermore, DPYD is associated more strongly with colorectal cancer, gastrointestinal cancer, rectal cancer, and stomach cancer than with other cancers, such as breast cancer and pancreatic cancer. This implies that capecitabine is effective for treating intestine- and stomach-related cancers [29].

Similarly, some drugs including cetuximab are used to treat

---

Table 5  Cancers that use docetaxel as a prescription drug.

| Cancer                        | Rank* |
|-------------------------------|-------|
| Head and neck neoplasms       | 9     |
| Prostatic neoplasms           | 10    |
| Breast neoplasms              | 11    |
| Lung neoplasms                | 14    |
| Non-small cell lung carcinomas| 14    |
| Stomach neoplasms             | 15    |
| Ovarian neoplasms             | 24    |
| Nasopharyngeal neoplasms      | 86    |

* Ranks of BCL2

Table 6  Ranks of associated genes for nasopharyngeal cancer.

| Gene   | Description                        | Rank |
|--------|------------------------------------|------|
| STC1   | stanniocalcin-1                    | 1    |
| SLC13A5| sodium-dependent citrate transporter, member 5 | 2    |
| SLC13A2| sodium-dependent dicarboxylate transporter, member 2 | 3    |
| BRCA2  | breast and ovarian cancer          | 4    |
| CROCC  | ciliary rootlet coiled-coil, rootletin pseudogene 3 | 5    |
| CROCC  | ciliary rootlet coiled-coil, rootletin pseudogene 3 | 6    |
| RAB12FIP3| RAB11 family interacting protein 3 | 7    |
| FAM184A| family with sequence similarity 184, member A | 8    |
| SRRM4  | serine/arginine repetitive matrix 4 | 9    |
| MCM9   | minichromosome maintenance complex component 9 | 10   |
| PRDM13 | PR domain containing 13            | 11   |
| JKAMP  | JNK-associated membrane protein     | 12   |
| RAB18  | RAB18, member RAS oncogene family  | 13   |
| RHF20  | PHD finger protein 20              | 14   |
| ADAMTS9| a disintegrin and metallopeptidase with thrombospondin type 1 motif, 9 | 15   |
| GJA1   | gap junction alpha-1 protein       | 16   |
| BCL2   | B-cell CLL/lymphoma 2              | 86   |

Table 7  Ranks of DPYD and EGFR to cancers.

| Cancer                      | DPYD | EGFR |
|-----------------------------|------|------|
| Breast neoplasms            | 626  | 4    |
| Colonic neoplasms           | 419  | 6    |
| Colorectal neoplasms        | 44   | 5    |
| Esophageal neoplasms        | 149  | 13   |
| Gastrointestinal neoplasms  | 8    | NR   |
| Head and neck neoplasms     | 71   | 1    |
| Lung neoplasms              | 507  | 4    |
| Ovarian neoplasms           | NR   | 6    |
| Pancreatic neoplasms        | 1,088| 4    |
| Rectal neoplasms            | 74   | 18   |
| Stomach neoplasms           | 112  | 14   |
| Uterine Neoplasms           | NR   | 14   |

* Ranks in our proposed measure. NR represents that the corresponding target is not ranked. DPYD dihydropyrimidine dehydrogenase; EGFR epidermal growth factor receptor
many types of cancers, such as breast neoplasms, colonic neo-
plasms, colorectal neoplasms, head and neck neoplasms, lung
neoplasms, pancreatic neoplasms, and uterine neoplasms. A
target gene of these drugs, epidermal growth factor receptor
(EGFR), is ranked within the top 20 genes in most cases for each
cancer. This implies that drugs using EGFR as a target gene can
be used in a variety of cancers. Thus, the proposed measure also
is effective for determining not only novel target genes for new
drugs but also known drugs that can treat other diseases.

5. Conclusions

We proposed a new measure that identifies disease-associated
genes and prioritizes the identified genes as drug target genes in
terms of fewer side-effects using the biomedical literature.
The proposed measure can be used as a primary filtering to narrow
candidate target genes of new drugs. We verified the obtained
ranking results by checking the ranks of known drug targets.
Moreover, we demonstrated that known drugs can be used as pre-
scription drugs for another disease for which drugs for treatment
have not yet been developed.

In the future, we intend to perform a more detailed analysis
and to extend the proposed measure to combine other data, such
as pathways, with the literature analysis.

Acknowledgments

A part of this work was supported by Insti-
tute for Bioinformatics Research and Development (BIRD),
Japan Science and Technology Agency (JST).

References

[1] Adamic, L., Wilkinson, D., Huberman, B. and Adar, E.: A literature
based method for identifying gene-disease connections, Proc. IEEE
Comput. Soc. Bioinform. Conf., pp.109–117 (2002).
[2] Al-Mubaid, H. and Singh, R.K.: A new text mining approach for find-
ing protein-to-disease associations, J. Amer. Soc. Inf. Sci. Tech., Vol.1,
No.3, pp.145–152 (2005).
[3] Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S. and
Wishart, D.: Polysearch: A web-based text mining system for extract-
ing relationships between human diseases, genes, mutations, drugs
and metabolites, Nucleic Acids Res., Vol.36, pp.W399–W405 (2008).
[4] Plake, C., Royer, L., Winnenburg, R., Hakenberg, J. and Schroeder,
M.: GoGene: Gene annotation in the fast lane, Nucleic Acids Res.,
Vol.37, pp.W300–W304 (2009).
[5] Tsuruoka, Y., Tsujii, J. and Ananiadou, S.: FACTA: A text search
engine for finding associated biomedical concepts, Bioinformatics,
Vol.24, No.21, pp.2559–2560 (2008).
[6] Ozgür, A., Yu, T., Erkan, G. and Radle, D.: Identifying gene-disease
associations using centrality on a literature mined gene-interaction
network, Bioinformatics, Vol.24, No.13, pp.277–285 (2008).
[7] Freydenberg, J. and Propping, P.: A similarity-based method for
gene-wide prediction of disease-relevant human genes, Bioin-
formatics, Vol.18, pp.S110–S115 (2002).
[8] van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A.,
Brunner, H.G. and Friend, G.: GeneSeeker: Extraction and integra-
tion of human disease-related information from web-based genetic
databases, Nucleic Acids Res., Vol.33, pp.W758–W761 (2005).
[9] Tiffen, N., Kelso, J.F., Power, A.R., Pan, H., Bajic, A.B. and Hide,
W.A.: Integration of text- and data-mining using ontologies success-
fully selects disease gene candidates, Nucleic Acids Res., Vol.33, No.5,
pp.1544–1552 (2005).
[10] Kelso, J., Visagie, J., Thieker, G., Christoffers, A., Bardien, S.,
Smedley, D., Otaiga, D., Greyling, G., Jongeneel, C.V., McCarthy,
M.L., Hide, T. and Hide, W.: eVOC: A controlled vocabulary for unify-
ing gene expression data, Genome Res., Vol.13, pp.1222–1220 (2003).
[11] Yu, S., Vooren, S., Trancheon, L., Moor, B.D. and Moreau, Y.: Com-
parsion of vocabularies, representations and ranking algorithms for
gene prioritization by text mining, Bioinformatics, Vol.24, pp.116–125
(2008).
[12] Davis, A.P., Murphy, C.G., Johnson, R., Lay, J.M., Lennon-Hopkins,
K., Saraceni-Richards, C., Sciaky, D., King, B.L., Rosenstein, M.C.,
Wiegars, T.C. and Mattingly, C.J.: The comparative toxicogenomics
database: Update 2013, Nucleic Acids Res., Vol.41, pp.D1104–D1114
(2013).
[13] Hersh, W.R.: Information Retrieval: A Health and Biomedical Per-
spective, Springer, pp.174–179 (2003).
[14] Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shirivastava, S., Tsur,
D., Gautam, B. and Hassanali, M.: DrugBank: A knowledgebase for
drugs, drug actions and drug targets, Nucleic Acids Res., Vol.36,
pp.D901–D906 (2008).
[15] Klein, T.E., Chang, J.T., Cho, M.K., Easton, K.L., Fergerson, R.,
Hewett, M., Lim, Z., Liu, Y., Liu, S., Oliver, D.E., Rubin, D.L., Shafa,
F., Stuart, J.M. and Altman, R.B.: Integrating genotype and phenotype
information: An overview of the PharmGKB project, The Pharma-
genomics J., Vol.1, No.3, pp.167–170 (2001).
[16] Waring, S.C. and Rosenburg, R.N.: Genome-wide association studies
in Alzheimer disease, Arch. Neurol., Vol.65, No.3, pp.329–334 (2008).
[17] Lancôt, K.L., Herrmann, N., Yau, K.K., Khan, L.R., Liu, B.A., Loulou,
M.M. and Einarson, T.R.: Efficacy and safety of cholinesterase inhibitors in Alzheimer’s disease: A meta-analysis, Can. Med. Assoc. J., Vol.169, No.6, pp.557–564 (2003).
[18] Suzuki, S., Makita, Y., Mukai, T., Matsuo, K., Ueda, O. and Fujieda,
K.: Molecular basis of neonatal diabetes in Japanese patients, J. Clin.
Endocrinol. Metab., Vol.92, No.10, pp.3979–3985 (2007).
[19] Lyssenko, V., Jonsson, A., Ahmgren, P., Pulzini, N., Isomaa, B.,
Tuomi, T., Berglund, G., Altsusher, D., Nilsson, P. and Group, L.: Clinical
risk factors, DNA variants, and the development of type 2 diabetes, N. Engl.
J. Med., Vol.359, No.21, pp.2220–2232 (2008).
[20] Lai, K.P., Law, A.Y., Yeung, H.Y., Lee, L.S., Wagner, G.F. and Wong,
C.K.: Induction of stanniocalcin-1 expression in apoptotic human na-
sophyryngeal cancer cells by p53, Biochem. Biophys. Res. Commun.,
Vol.356, No.4, pp.968–975 (2007).
[21] Block, G.J., Ohkouchi, S., Fun, F., Frenkel, J., Gregory, C., Pochampally,
Ra., DiMattia, G., Sullivan, D.E. and Procopk, D.J.: Multi-
tipotent stromal cells are activated to reduce apoptosis in part by up-
regulation and secretion of stanniocalcin-1, Stem Cells, Vol.27, No.3,
pp.670–681 (2009).
[22] Ching, L.Y., Yeung, B.H. and Wong, C.K.: Synergistic effect of p53
on TSA-induced stanniocalcin 1 expression in human nasopharyngeal
carcinoma cells, CNE2, J. Mol. Endocrinol., Vol.48, No.3, pp.241–
250 (2012).
[23] Alajez, N.M., Lenarduzzi, M., Ito, E., Hui, A.B., Shi, W., Bruce, J.,
Yue, S., Huang, S.H., Xu, W., Waldron, J., O’Sullivan, B. and Liu,
F.F.: Mir218 suppresses nasopharyngeal cancer progression through
downregulation of survivin and the SLIT2-ROBO1 pathway, Cancer
Res., Vol.71, No.6, pp.2381–2391 (2011).
[24] Spillantini, M.G., Murrell, J.R., Goedert, M., Farlow, M.R., Klug,
C.K.: Induction of stanniocalcin expression in apoptotic human na-
sopharyngeal carcinoma cells, CNE2, J. Mol. Endocrinol., Vol.24, No.3,
pp.929–934 (1995).
[25] Cole, S.L. and Vassar, R.: The basic biology of BACE1: A key ther-
apeutic target for Alzheimer’s disease, Curr. Genomics, Vol.8, No.8,
pp.509–530 (2007).
[26] Selkoe, D.J.: Translating cell biology into therapeutic advances in
Alzheimer’s disease, Nature, Vol.399, pp.A23–A31 (1999).
[27] Dresler, T., Ehls, A.C., Heinzel, S., Renner, T.J., Reif, A., Baehne,
C.G., Heine, M., Boreatti-Hümmmer, A., Jacob, C.P., Lesch, K.P. and
Fallgatter, A.J.: Dopa mine transporter (SLC6A3) genotype impacts
neuropsychiological correlates of cognitive response control in an adult
sample of patients with ADHD, Nueuropsychopharmacology, Vol.35,
No.11, pp.2193–2202 (2010).
[28] Saif, M.W.: Targeting cancers in the gastrointestinal tract: Role of
capcitabine, Onco Targets and Therapy, Vol.2, pp.29–41 (2009).
[29] Ajani, J.: Review of capcitabine as oral treatment of gastric, gastro-
jejunal, and esophageal cancers, Cancer, Vol.107, No.2, pp.221–
231 (2006).
Yeondae Kwon is an Assistant Professor at Tokyo University of Science. She received her M.S. degree in biochemistry from Pusan National University, Busan, Korea, and Ph.D. degree in information science from Nara Institute of Science and Technology, Nara, Japan, in 2000. Her research interests include text mining of biomedical literature and healthcare data mining. She is a member of IPSJ, ANLP and MBSJ.

Shogo Shimizu is an Assistant Professor at Gakushuin Women’s University. He received his B.E. degree from Osaka University in 1996, and M.E. and Ph.D. degrees from Nara Institute of Science and Technology in 1998 and 2001, respectively. His research interests include databases and privacy-preserving techniques. He is a member of IPSJ.

Hideaki Sugawara is a Professor Emeritus at National Institute of Genetics and Graduate University for Advanced Studies. He received his B.S., M.S. and Ph.D. degrees from the University of Tokyo in 1968, 1970 and 1973, respectively. His research interests include microbial and structural life science data cloud. He is a member of WFCC and MBSJ.

Satoru Miyazaki is a Professor at Tokyo University of Science. He received his B.S., M.S. and Ph.D. degrees from Tokyo University of Science in 1986, 1988 and 1997, respectively. His current research interests are molecular evolution, genome sequence informatics and integration of bio-databases on internet. He is a member of MBSJ, JSIK and BSJ.

(Communicated by Masakazu Sekijima)