Discovering Hidden Connections among Diseases, Genes and Drugs Based on Microarray Expression Profiles with Negative-Term Filtering

Jain-Shing Wu[1], **E-Fong Kao**[2], **Chung-Nan Lee**[1,2]*

[1] Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan, [2] Department of Medical Imaging and Radiological Sciences, Kaohsiung Medical University, Kaohsiung, Taiwan

Abstract

Microarrays based on gene expression profiles (GEPs) can be tailored specifically for a variety of topics to provide a precise and efficient means with which to discover hidden information. This study proposes a novel means of employing existing GEPs to reveal hidden relationships among diseases, genes, and drugs within a rich biomedical database, PubMed. Unlike the co-occurrence method, which considers only the appearance of keywords, the proposed method also takes into account negative relationships and non-relationships among keywords, the importance of which has been demonstrated in previous studies. Three scenarios were conducted to verify the efficacy of the proposed method. In Scenario 1, disease and drug GEPs (disease: lymphoma cancer, lymph node cancer, and drug: cyclophosphamide) were used to obtain lists of disease- and drug-related genes. Fifteen hidden connections were identified between the diseases and the drug. In Scenario 2, we adopted different diseases and drug GEPs (disease: AML-ALL dataset and drug: Gefitinib) to obtain lists of important diseases and drug-related genes. In this case, ten hidden connections were identified. In Scenario 3, we obtained a list of disease-related genes from the disease-related GEP (liver cancer) and the drug (Capecitabine) on the PharmGKB website, resulting in twenty-two hidden connections. Experimental results demonstrate the efficacy of the proposed method in uncovering hidden connections among diseases, genes, and drugs. Following implementation of the weight function in the proposed method, a large number of the documents obtained in each of the scenarios were judged to be related: 834 of 4028 documents, 789 of 1216 documents, and 1928 of 3791 documents in Scenarios 1, 2, and 3, respectively. The negative-term filtering scheme also uncovered a large number of negative relationships as well as non-relationships among these connections: 97 of 834, 38 of 789, and 202 of 1928 in Scenarios 1, 2, and 3, respectively.

Citation: Wu J-S, Kao E-F, Lee C-N (2014) Discovering Hidden Connections among Diseases, Genes and Drugs Based on Microarray Expression Profiles with Negative-Term Filtering. PLoS ONE 9(6): e98826. doi:10.1371/journal.pone.0098826

Editor: Francisco J. Esteban, University of Jaén, Spain

Received February 7, 2013; **Accepted** May 6, 2014; **Published** June 10, 2014

Funding: No current external funding sources for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cnlee@mail.cse.nsysu.edu.tw

These authors contributed equally to this work.

Background

In recent years, a considerable volume of biomedical literature has been published, covering a range of research topics including the functional genome, epigenetic modifications, and mRNA and protein expression profiles. This research outlines the functions, mechanisms, and behaviors of genes in regular cellular processes. A considerable quantity of useful data has been obtained; however, still more information awaits discovery. For example, one study may identify genes related to colon cancer from a microarray expression profile while other researchers observe changes in the gene expression value through a study on drug response. Thus, the same genes may possess hidden connections based on the specifics of the disease and drugs in question. Swanson [1] referred to this situation as undiscovered public knowledge (UPK) residing in two previously published collections of complementary literature. UPK can provide useful new information of scientific interest. Swanson [1,3] combined information related to target terms (A), intermediate concepts (B), and other terms (C) from two discrete sources, thereby providing new directions from which to obtain hidden relationships (A->B, B->C => A->C).

A variety of methods have been developed for the extraction of UPK from previously published studies. Spasser [2] applied the concepts proposed by Swanson in other areas, such as library informatics. Lamb [4] proposed a systematic method for the construction of connectivity maps comprising a reference collection of gene expression data from human cells treated with bioactive small molecules. However, the gene-expression signature (one file of up-regulated genes and one file of down-regulated genes derived from a transcriptional profiling experiment) must be prepared in advance.

Butte and Kohane [5] used the unified medical language systems (UMLS) meta thesaurus, Gene Expression Omnibus (GEO) files, and publically available gene expression profiles to construct connectivity maps which link biologically significant terms (phenotypes, diseases, environmental and experimental conditions) with related genes. They also performed a t-test to determine the relationship between datasets and concepts, the values of which were used to identify genes of significance from within the gene expression profiles. However, the integration of data can be difficult when experiments are performed in different laboratories, processing different samples, on different experimental platforms [6]. For example, the results obtained using cDNA microarrays cannot be integrated using GEO files. Furthermore, it is difficult to judge relationships when two conceptually similar experiments provide opposing relationships via the expression profiles.

A set of interactive software and database search strategies called ARROWSMITH [7] has been developed to reveal plausible hypotheses-linking findings from across a range of specialties. This system collects studies from MEDLINE, the titles of which contain term A, which are stored as File A. In the same manner, studies are collected with titles containing term C, which are stored as File C. A list is then compiled from Files A and C using common term B (words or phrases). A 5000 item stop-list is then used to filter out uninteresting terms from B before scrutinizing list B for plausible links between A and C. For each B term, the system generates AB files with titles containing A and B terms as well as BC files with titles containing B and C terms. This tool focuses on titles containing A or C terms, which are inherently limited with regard to the number of words employed. In addition, the system only finds co-occurrences of A and B or B and C in the title. Finally, to generate AB or BC files, ARROWSMITH employs manual editing in its search for plausible links between A and C via B terms. This is a novel approach to discover relationships between clinical conditions and physiological states through the analysis of MEDLINE document titles.

Homayouni et al. [8] proposed the use of latent semantic indexing (LSI), a vector space model capable of automatically identifying relationships among genes from titles and abstracts listed as MEDLINE citations. A semantic gene organizer (SGO) first retrieves LocusLink PMID abstracts from which it extracts gene-related documents. The SGO then creates terms using a gene-document matrix followed by singular value decomposition (SVD) for the creation of a low-rank approximation matrix. A vector space model is then used by the SGO to identify relationships among genes or between genes and keywords. However, when a large number of abstracts have been retrieved, the gene-document matrix can become unwieldy. Determining appropriate restrictions to control the size of the dictionary has proven difficult.

Another common approach to identify the relationships among biomedical concepts is the co-occurrence method, based on the theory that if two biomedical concepts occur in the same study, they might be related.

Frijters et al. [9] proposed a co-occurrence method called CoPub, which uses multiple genes of humans and mice to generate lists of keywords from several biomedical databases. These lists link to MEDLINE abstracts, in which the keywords are highlighted. In 2010, Frijters et al. [10] applied CoPub to produce connectivity maps by employing A to B and B to C relationships to find hidden relationships A to C. However, unlike ARROWSMITH, A, B and C represent biological terms such as diseases, drugs, and biological processes. Using this approach, a collection of abstracts related to the biological terms A and B is first obtained, and then a second collection from B and C is also generated. The relationship between the two collections is then calculated and a connectivity map is then produced to display the results. This is a powerful method of revealing hidden connections. CoPub uses catalog terms instead of original terms to retrieve related studies; however, this approach occasionally results in high recall but low precision. For example, terms such as “acute lymphoblastic leukemia” and “acute myeloid leukemia” are classified using simplified terms, such as “leukemia”, despite the fact that studies related to

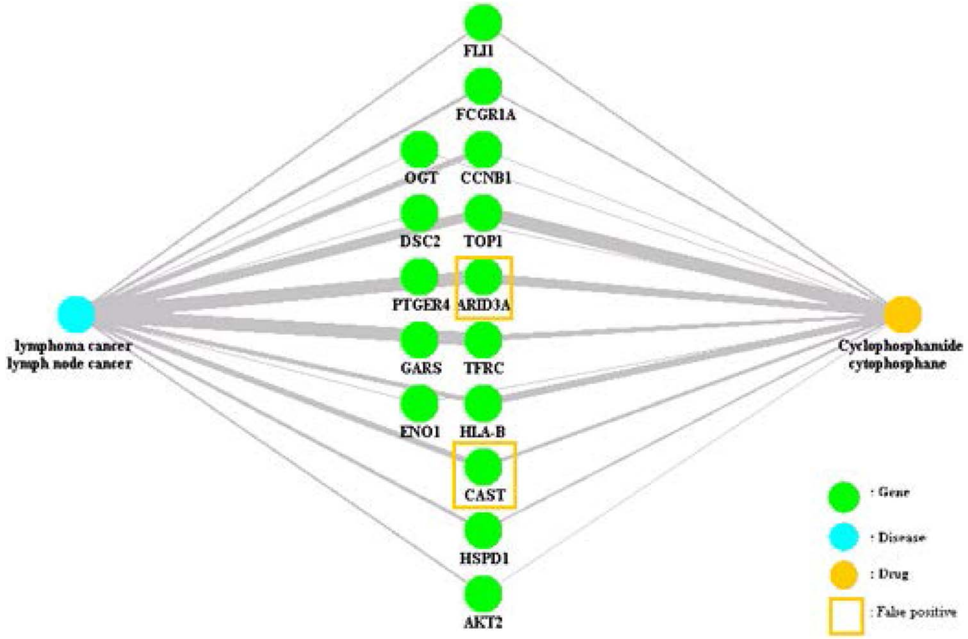


Figure 1. Connectivity map of the disease “lymphoma cancer, lymph node cancer” and the drug “cyclophosphamide, cytophosphane” obtained using the proposed method. The node marked in blue represents the disease node; the nodes marked in green represent the gene nodes; the node marked in yellow represents the drug node. The thicker the edge is connecting diseases and genes or drugs and genes, the stronger the relationship. The genes that are obtained using the proposed method are connected to both diseases and drugs. The genes bounded in a yellow box are the false-positive genes.
doi:10.1371/journal.pone.0098826.g001

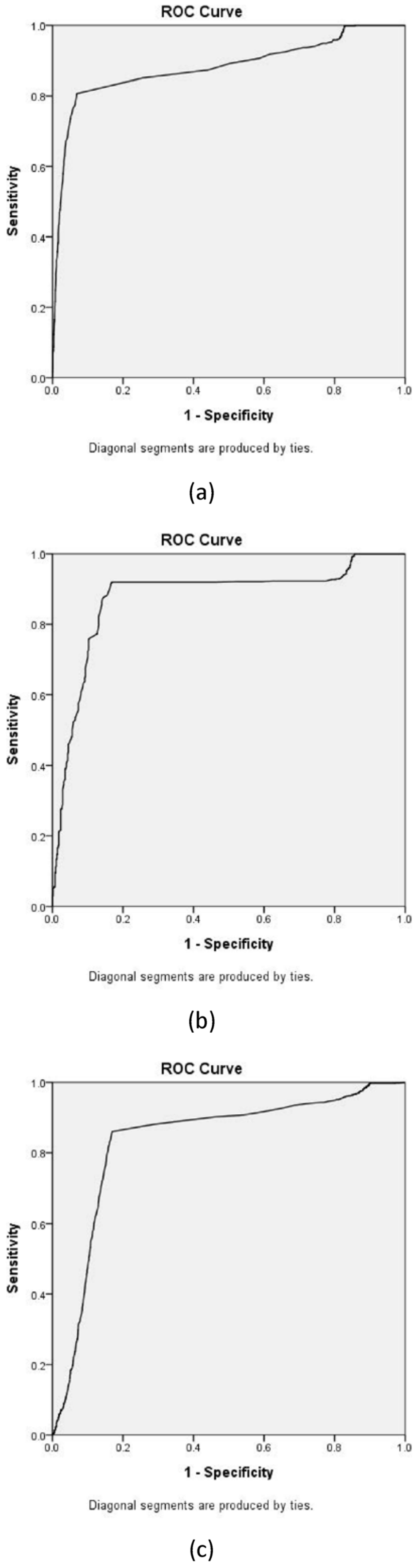


Figure 2. ROC results of Scenario 1, Scenario 2, and Scenario 3. The variable "score" is the related score of the study calculated using the proposed method with the weight function and the negative-term filtering scheme with zero as the threshold value. Figure 2(a) shows the ROC results of Scenario 1. The sensitivity and the specificity of Scenario 1 results are 81.98% and 92.21%, respectively. The area under curve (AUC) in this case is 0.883, and the 95% confidence interval (CI) is between 0.873 and 0.893. (b) The sensitivity and the specificity of Scenario 2 results are 92.54% and 83.27%, respectively. AUC in this case is 0.882, and the 95% CI is between 0.862 and 0.899. (c) The sensitivity and the specificity of Scenario 3 results are 83.78% and 83.82%, respectively. AUC in this case is 0.822, and the 95% CI is between 0.809 and 0.834.
doi:10.1371/journal.pone.0098826.g002

"chronic lymphoblastic leukemia" and "chronic myeloid leukemia" are also retrieved. When focusing on genes related to AML or ALL, studies related to CML and CLL are regarded as false positives and must be filtered out. In addition, terms such as "colorectal cancer" are often classified within a specific catalog such as "colon cancer". CoPub deals only with the term "colon cancer" but not the term "colorectal cancer", which can often appear in the same studies. As a result, many studies that could be highly relevant to the input terms (disease names) are often missed. Furthermore, CoPub deals only with co-occurrences, without any consideration of the status of the relationship (whether it is a positive relationship, negative relationship, or non-relationship).

Li et al. [6] proposed a novel algorithm to build disease-specific drug-protein connectivity maps using molecular interaction networks and PubMed [11]. Proteins were imported as "seeds" for use in the expansion of a protein network list. The list of proteins is used to generate queries for the identification of relevant abstracts as well as a "drug filtering criteria" with which to identify drugs of interest. After combining the disease-related proteins and drug list, cluster analysis is used to generate a disease-specific connectivity map. Unfortunately, this method requires manual intervention for the selection of proteins to seed the list as well as specific drug filter criteria.

The above methods were developed to facilitate the discovery of UPK; however, additional information related to microarray data remains largely unexplored. Moreover, most of these methods are unable to adequately deal with negative relationships or non-relationships among the entities in question. If a relationship between a particular gene and a disease or drug is reported in an abstract, but that relationship is in fact negative, the abstract must be removed from the collection. For example, "hSP (human spasmolytic polypeptide), the domain-duplicated homolog of pS2 protein is co-expressed with pS2 in stomach but not in breast carcinoma" details a negative relationship between "hSP" and "breast carcinoma". Many studies [12–14] have pointed out the importance of effectively dealing with studies that exhibit negative relationships or non-relationships.

This study employed microarray expression profiles for the identification of intermediate terms (genes). In the spotted microarray technique, the genes of interest must be selected first and corresponding probes are then printed to the glass for the experiment [15]. For in situ synthesized microarrays, such as Affymetrix GeneChips [16], microarray chips are produced for the observation of several topics. In this case, probes are designed according to pre-selected genes, thereby considerably narrowing the scope. Microarray expression profiles also provide evidence of the relationship between genes and various topics of interest. The main objective of this paper was to use microarray profiles for the discovery of relationships among diseases, genes, and drugs, whereupon biomedical literature can be used to confirm the obtained results. The proposed system is not intended to be a

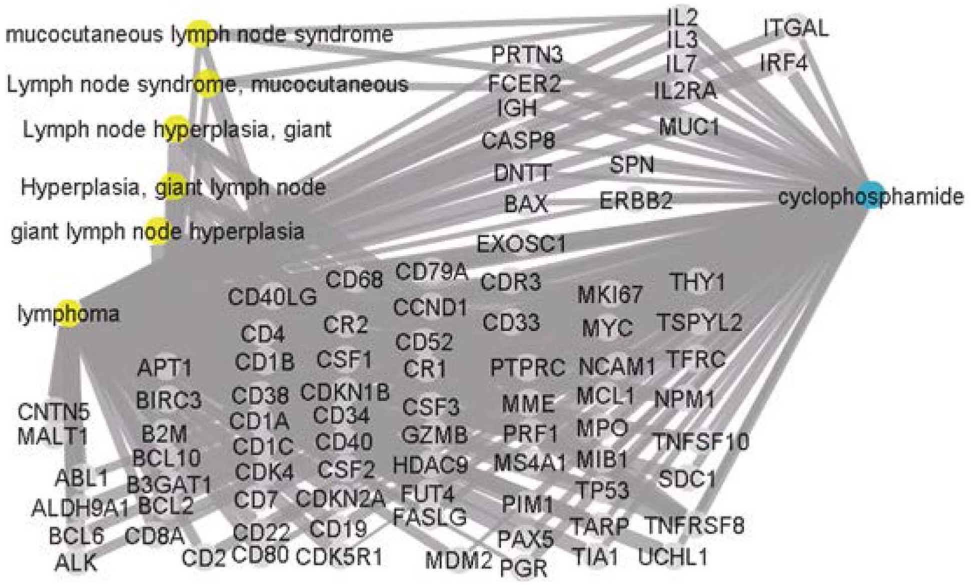


Figure 3. Connectivity map of the disease ''lymphoma cancer, lymph node cancer'' and the drug ''cyclophosphamide, cytophosphane'' obtained from CoPub.
doi:10.1371/journal.pone.0098826.g003

replacement for existing approaches such as CoPub, but to be a complementary approach to obtaining hidden connections as well as an effective means of dealing with documents that exhibit negative relationships or non-relationships.

Results

Three scenarios were used to verify the performance of the proposed method. Scenario 1: Disease- and drug-related gene lists are obtained from disease- and drug-related microarray expression profiles in which the names of diseases and drugs are also adopted. Scenario 2: Various disease- and drug-related microarray expression profiles are used to obtain disease- and drug-related gene lists. Scenario 3: A disease-related gene expression profile is adopted and drug-related gene files are obtained from the PharmGKB website [17] (http:/www.pharmgkb.org/index.jsp).

In the event that GEPs are used, then a gene selection method called the disease oriented gene selection algorithm (DOFA) [18] is first used to obtain disease- or drug-related gene lists. DOFA analyses each gene in the GEP. If all expression values of samples

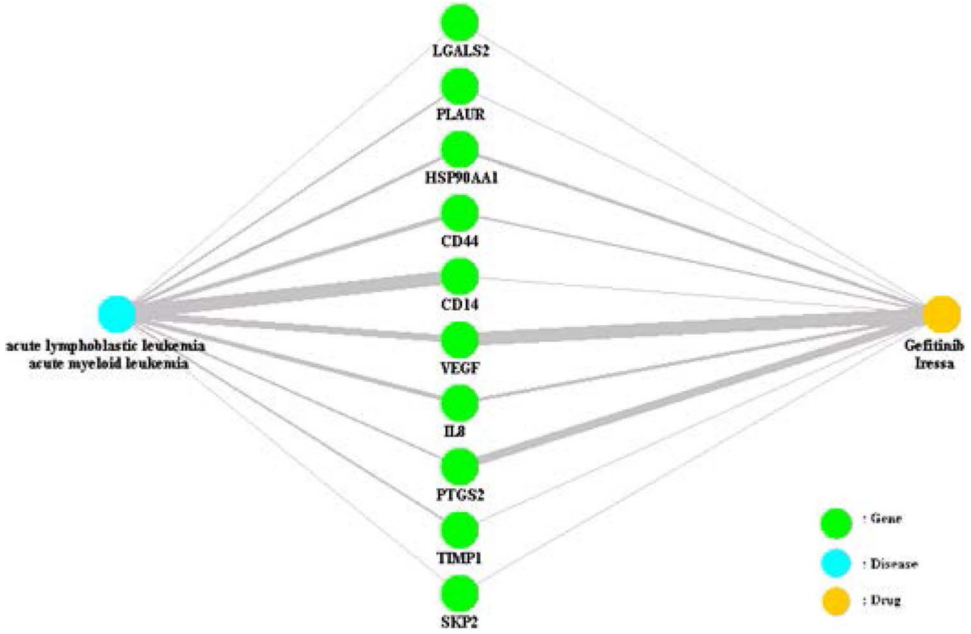


Figure 4. Hidden connections of the disease ''acute lymphoblastic leukemia, acute myeloid leukemia'' and the drug ''Gefitinib, Iressa'' obtained using the proposed method.
doi:10.1371/journal.pone.0098826.g004

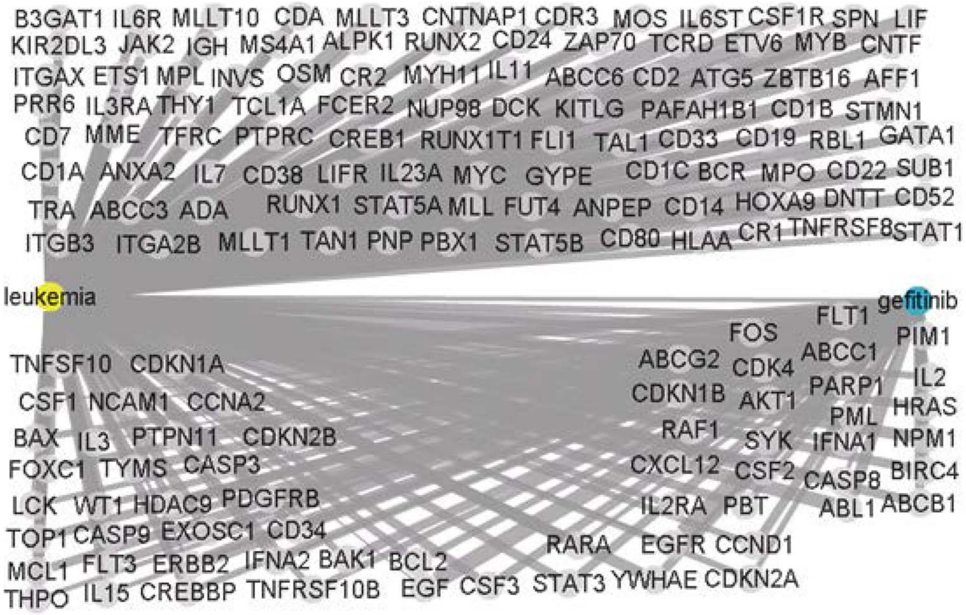

B3GAT1 IL6R MLLT10 CDA MLLT3 CNTNAP1 CDR3 MOS IL6ST CSF1R SPN LIF
KIR2DL3 JAK2 IGH MS4A1 ALPK1 RUNX2 CD24 ZAP70 TCRD ETV6 MYB CNTF
ITGAX ETS1 MPL INVS OSM CR2 MYH11 IL11 ABCC6 CD2 ATG5 ZBTB16 AFF1
PRR6 IL3RA THY1 TCL1A FCER2 NUP98 DCK KITLG PAFAH1B1 CD1B STMN1
CD7 MME TFRC PTPRC CREB1 RUNX1T1 FLI1 TAL1 CD33 CD19 RBL1 GATA1
CD1A ANXA2 IL7 CD38 LIFR IL23A MYC GYPE CD1C BCR MPO CD22 SUB1
TRA ABCC3 ADA RUNX1 STAT5A MLL FUT4 ANPEP CD14 HOXA9 DNTT CD52
ITGB3 ITGA2B MLLT1 TAN1 PNP PBX1 STAT5B CD80 HLAA CR1 TNFRSF8 STAT1
leukemia
gefitinib
FLT1
FOS
PIM1
ABCC1
TNFSF10 CDKN1A
ABCG2 CDK4
IL2
PARP1
CSF1 NCAM1 CCNA2
CDKN1B AKT1
HRAS
PML
BAX IL3 PTPN11 CDKN2B
RAF1
SYK IFNA1 NPM1
FOXC1 TYMS CASP3
CXCL12 CSF2
BIRC4
CASP8
LCK WT1 HDAC9 PDGFRB
IL2RA PBT
ABL1 ABCB1
TOP1 CASP9 EXOSC1 CD34
MCL1 FLT3 ERBB2 IFNA2 BAK1 BCL2
RARA EGFR CCND1
THPO IL15 CREBBP TNFRSF10B EGF CSF3 STAT3 YWHAE CDKN2A

Figure 5. Connectivity map of the disease ''acute lymphoblastic leukemia, acute myeloid leukemia'' and the drug ''Gefitinib, Iressa'' obtained from CoPub.
doi:10.1371/journal.pone.0098826.g005

at one gene are close to the mean value of their classes and far from those of other classes, then this gene is regarded as important. If the expression values of samples at one gene make it difficult to distinguish the classes, then this gene is viewed as unimportant. In this manner, DOFA filters out unimportant genes and then uses a genetic algorithm (GA) for the selection of genes capable of precise classification of the samples. These genes are deemed to be related to the disease or drugs in question. Further details of DOFA are provided in Algorithm S1. After obtaining gene lists related to the diseases or drugs in question, synonyms of the gene names in the lists are obtained from the Genebank. These names and the input terms (disease or drug names) are then used to fetch abstracts and titles from PubMed, which are evaluated according to relevance using a weight function. The abstracts in PubMed are fetched using a search engine based on the co-occurrence method; therefore, they do not always display positive relationships. Thus, the proposed method examines all retrieved abstracts and removes those that are unrelated or evidence a negative relationship. Finally, the remaining abstracts are used to identify relationships indicative of hidden connections among genes, diseases, and drugs.

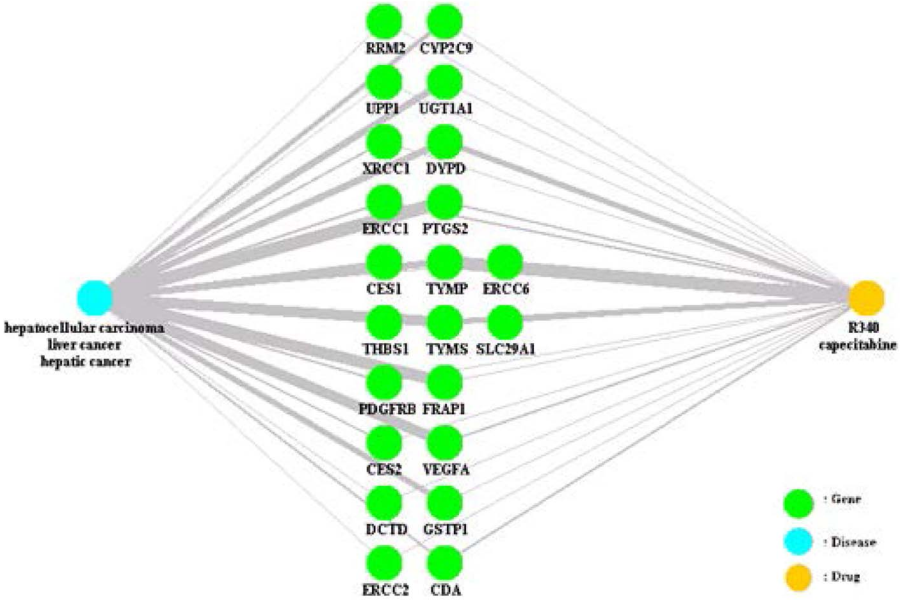


Figure 6. Hidden connections of the disease ''liver cancer, hepatic cancer, hepatocellular carcinoma'' and the drug ''R340, capecitabine'' obtained using the proposed method.
doi:10.1371/journal.pone.0098826.g006

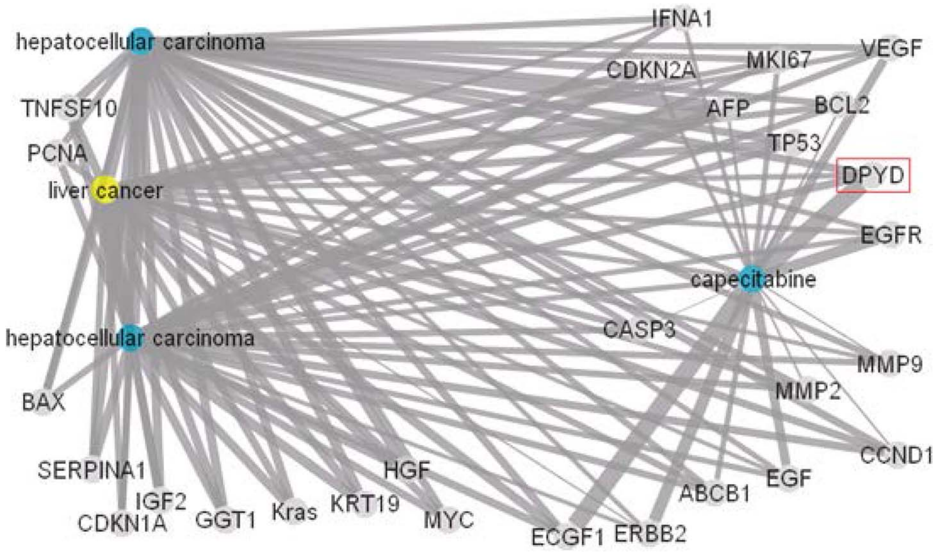


Figure 7. Hidden connections of the disease ''liver cancer, hepatic cancer, hepatocellular carcinoma'' and the drug ''R340, capecitabine'' obtained from CoPub.
doi:10.1371/journal.pone.0098826.g007

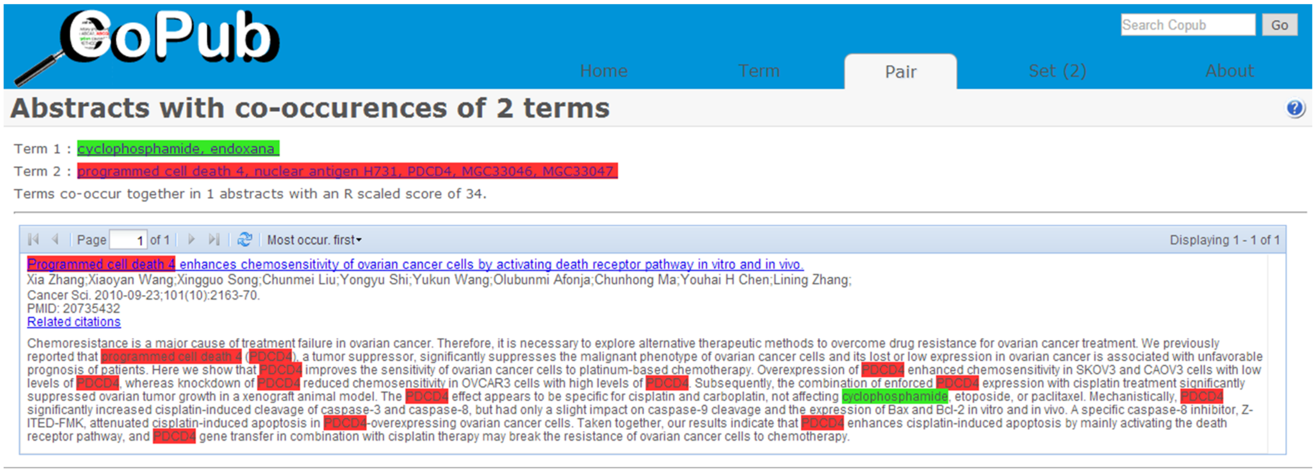
Further details of the proposed method are outlined in the Methods section, and details of the experiments are presented in the following. And all related softwares, data, analysis script and user's manual can be found in [40].

Scenario 1

DOFA was used for the selection of dominant genes from the disease- and drug-related GEPs. This scenario included ''lymphoma cancer'' and ''lymph node cancer'' as the target disease terms and ''cyclophosphamide'' and ''cytophosphane'' as the target drug terms. The weight function was applied to the disease and drug terms to evaluate the relevance of the retrieved abstracts. The lymphoma cancer dataset reported by Shipp et al. [19] was selected as the disease-related GEP. This dataset contains 77 samples, each of which contains 7129 genes. We also adopted the drug-related GEP in [19], containing 58 samples, each of which contains 7129 genes.

DOFA identified 54 genes from the disease-related GEP and 33 genes from the drug-related GEP, which were used to retrieve related abstracts from the PubMed website. DOFA identified 15 genes with hidden connections between the input disease terms ''lymphoma cancer, lymph node cancer'' and the input drug terms ''cyclophosphamide, cytophosphane'', as shown in Fig. 1. The node marked in blue represents the disease node; the node marked in green represents the gene nodes; and the node marked in yellow represents the drug node.

The gene ''TOP1'' is an intermediate node between the disease and drug of interest. Kancherla et al. [20] claimed that ''the authors report Phase I data of topotecan and etoposide combination for patients with recurrent or refractory non-Hodgkin lymphoma and correlation of topoisomerase-DNA complex formation to clinical response''. In addition, Minagawa et al. stated, ''In 8 cases whose samples could be obtained before and after cyclophosphamide (CAP), topoisomerase I (topo I) activity significantly increased after CAP therapy'' [21]. These



Abstracts with co-occurences of 2 terms

Term 1 : cyclophosphamide, endoxana
Term 2 : programmed cell death 4, nuclear antigen H731, PDCD4, MGC33046, MGC33047
Terms co-occur together in 1 abstracts with an R scaled score of 34.

Programmed cell death 4 enhances chemosensitivity of ovarian cancer cells by activating death receptor pathway in vitro and in vivo.
Xia Zhang;Xiaoyan Wang;Xingguo Song;Chunmei Liu;Yongyu Shi;Yukun Wang;Olubunmi Afonja;Chunhong Ma;Youhai H Chen;Lining Zhang;
Cancer Sci. 2010-09-23;101(10):2163-70.
PMID: 20735432
Related citations

Chemoresistance is a major cause of treatment failure in ovarian cancer. Therefore, it is necessary to explore alternative therapeutic methods to overcome drug resistance for ovarian cancer treatment. We previously reported that programmed cell death 4 (PDCD4), a tumor suppressor, significantly suppresses the malignant phenotype of ovarian cancer cells and its lost or low expression in ovarian cancer is associated with unfavorable prognosis of patients. Here we show that PDCD4 improves the sensitivity of ovarian cancer cells to platinum-based chemotherapy. Overexpression of PDCD4 enhanced chemosensitivity in SKOV3 and CAOV3 cells with low levels of PDCD4, whereas knockdown of PDCD4 reduced chemosensitivity in OVCAR3 cells with high levels of PDCD4. Subsequently, the combination of enforced PDCD4 expression with cisplatin treatment significantly suppressed ovarian tumor growth in a xenograft animal model. The PDCD4 effect appears to be specific for cisplatin and carboplatin, not affecting cyclophosphamide, etoposide, or paclitaxel. Mechanistically, PDCD4 significantly increased cisplatin-induced cleavage of caspase-3 and caspase-8, but had only a slight impact on caspase-9 cleavage and the expression of Bax and Bcl-2 in vitro and in vivo. A specific caspase-8 inhibitor, Z-ITED-FMK, attenuated cisplatin-induced apoptosis in PDCD4-overexpressing ovarian cancer cells. Taken together, our results indicate that PDCD4 enhances cisplatin-induced apoptosis by mainly activating the death receptor pathway, and PDCD4 gene transfer in combination with cisplatin therapy may break the resistance of ovarian cancer cells to chemotherapy.

Figure 8. Example abstract exhibiting a negative relationship.
doi:10.1371/journal.pone.0098826.g008

Table 1. Summaries of performances of the negative-term filtering scheme of three scenarios.

	Scenario 1	Scenario 2	Scenario 3
Total # of fetched document	4028	1216	3791
# of remained documents after filtering using weight function	834	789	1928
# of related documents among the remained documents (gold standard)	500	657	1335
# of unrelated documents among the remained documents (gold standard)	334	132	593
# of true positive of remained documents judged by the negative-term filtering scheme	471	606	1272
# of false positive of remained documents judged by the negative-term filtering scheme	237	94	391
# of true negative of remained documents judged by the negative-term filtering scheme	97	38	202
# of false negative of remained documents judged by the negative-term filtering scheme	29	51	63

doi:10.1371/journal.pone.0098826.t001

statements attest to the relationship between the diseases and genes as well as between the drugs and genes.

In contrast, the gene “CAST” was a false positive because the term “cast” is a commonly used verb. The gene “ARID3A” was also a false positive because its alias “bright” is a commonly used adjective.

A receiver operating characteristic (ROC) curve was used to evaluate the ability of the proposed method to uncover hidden connections among previously published studies. All retrieved abstracts are manually checked for relationships between diseases and genes or drug and genes to provide a gold standard with which to evaluate the results obtained after using the weight function and negative-term filtering scheme. Finally, an ROC curve is drawn with zero as the threshold value. To draw the ROC curves, this study used the statistical software, Statistical Product and Service Solutions (SPSS), which is commonly used for statistical calculation. Figure 2 (a) presents the ROC results of Scenario 1, the sensitivity and specificity of which were 80.7% and 93.10%, respectively. The area under the curve (AUC) in this case was 0.881 with a 95% confidence interval (CI) of between 0.862 and 0.9. Data of three scenarios for drawing three ROC curves are provided in Data S1 and also available in [40].

We also compared our results with those obtained while using CoPub. Figure 3 lists CoPub results with the disease node marked in yellow, the gene node marked in gray, and the drug node marked in blue. A total of 86 genes were selected by CoPub, none of which were obtained using the proposed method.

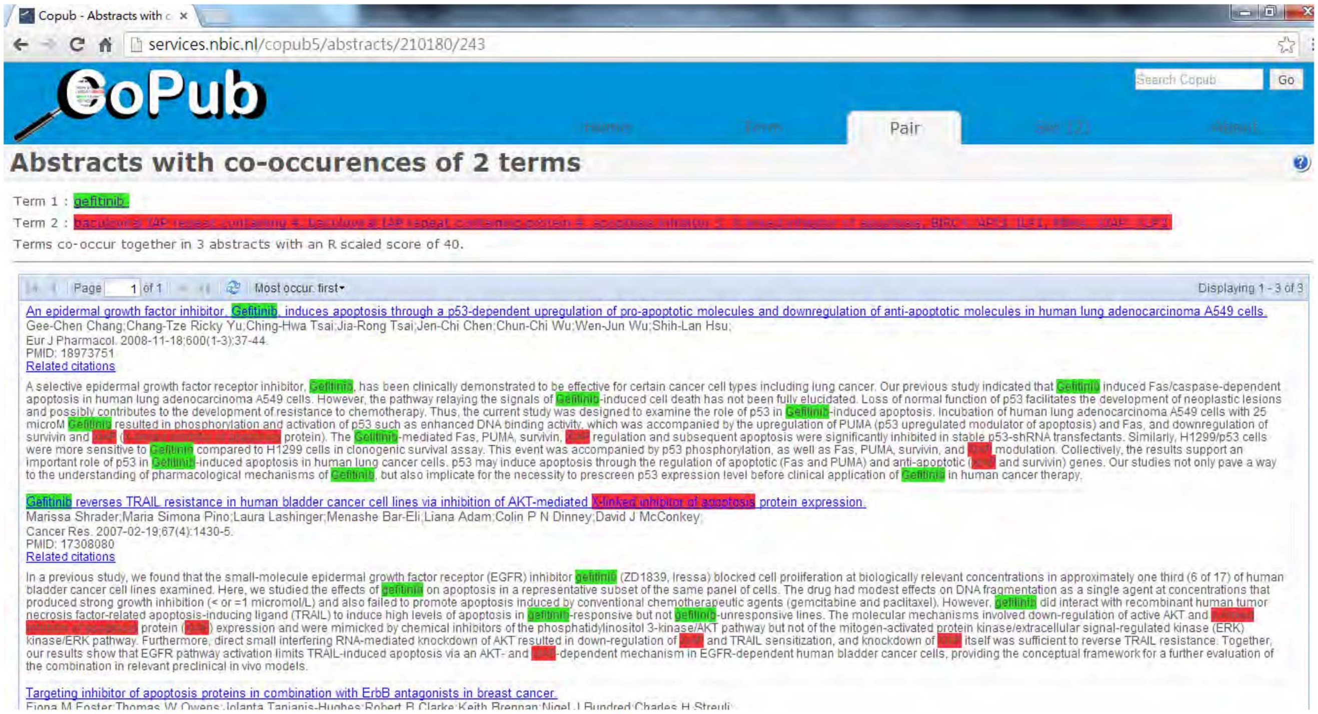
Scenario 2

This scenario included “acute lymphoblastic leukemia, acute myeloid leukemia” as the target disease terms and “Gefitinib, Iressa” as the target drug terms. The AML-ALL dataset reported by Golub et al. [22] was used as the disease-related GEP. This dataset contains 72 samples, each of which contains 7129 genes. We employed the drug-related GEP reported by Stegmaier et al. [23], which includes 12 samples, each containing 22283 genes.

Table 2. Comparison between only using the weight function and using both the negative-term filtering scheme and the weight function of three scenarios.

	Scenario 1	Scenario 2	Scenario 3
Total # of fetched document	4028	1216	3791
# of related documents among the fetched documents (gold standard)	584	659	1479
# of unrelated documents among the fetched documents (gold standard)	3444	557	2312
# of true positive judged by using **only the weight function**	500	657	1335
# of false positive judged by using **only the weight function**	334	132	593
# of true negative judged by using **only the weight function**	3110	425	1719
# of false negative judged by using **only the weight function**	84	2	144
# of true positive judged by **the weight function and the negative-term filtering scheme**	471	606	1272
# of false positive judged by **the weight function and the negative-term filtering scheme**	237	94	391
# of true negative judged by **the weight function and the negative-term filtering scheme**	3207	463	1921
# of false negative judged by **the weight function and the negative-term filtering scheme**	113	53	207

doi:10.1371/journal.pone.0098826.t002



Abstracts with co-occurences of 2 terms

Term 1 : gefitinib

Term 2 : baculoviral IAP repeat-containing 4, baculoviral IAP repeat containing protein 4, apoptosis inhibitor 3, X-linked inhibitor of apoptosis, BIRC4, API3, ILP1, MIHA, XIAP, ILP2

Terms co-occur together in 3 abstracts with an R scaled score of 40.

Page 1 of 1

An epidermal growth factor inhibitor, Gefitinib, induces apoptosis through a p53-dependent upregulation of pro-apoptotic molecules and downregulation of anti-apoptotic molecules in human lung adenocarcinoma A549 cells.
Gee-Chen Chang;Chang-Tze Ricky Yu;Ching-Hwa Tsai;Jia-Rong Tsai;Jen-Chi Chen;Chun-Chi Wu;Wen-Jun Wu;Shih-Lan Hsu;
Eur J Pharmacol. 2008-11-18;600(1-3):37-44.
PMID: 18973751
Related citations

A selective epidermal growth factor receptor inhibitor, Gefitinib, has been clinically demonstrated to be effective for certain cancer cell types including lung cancer. Our previous study indicated that Gefitinib induced Fas/caspase-dependent apoptosis in human lung adenocarcinoma A549 cells. However, the pathway relaying the signals of Gefitinib-induced cell death has not been fully elucidated. Loss of normal function of p53 facilitates the development of neoplastic lesions and possibly contributes to the development of resistance to chemotherapy. Thus, the current study was designed to examine the role of p53 in Gefitinib-induced apoptosis. Incubation of human lung adenocarcinoma A549 cells with 25 microM Gefitinib resulted in phosphorylation and activation of p53 such as enhanced DNA binding activity, which was accompanied by the upregulation of PUMA (p53 upregulated modulator of apoptosis) and Fas, and downregulation of survivin and XIAP (X-linked inhibitor of apoptosis protein). The Gefitinib-mediated Fas, PUMA, survivin, XIAP regulation and subsequent apoptosis were significantly inhibited in stable p53-shRNA transfectants. Similarly, H1299/p53 cells were more sensitive to Gefitinib compared to H1299 cells in clonogenic survival assay. This event was accompanied by p53 phosphorylation, as well as Fas, PUMA, survivin, and XIAP modulation. Collectively, the results support an important role of p53 in Gefitinib-induced apoptosis in human lung cancer cells. p53 may induce apoptosis through the regulation of apoptotic (Fas and PUMA) and anti-apoptotic (XIAP and survivin) genes. Our studies not only pave a way to the understanding of pharmacological mechanisms of Gefitinib, but also implicate for the necessity to prescreen p53 expression level before clinical application of Gefitinib in human cancer therapy.

Gefitinib reverses TRAIL resistance in human bladder cancer cell lines via inhibition of AKT-mediated X-linked inhibitor of apoptosis protein expression.
Marissa Shrader;Maria Simona Pino;Laura Lashinger;Menashe Bar-Eli;Liana Adam;Colin P N Dinney;David J McConkey;
Cancer Res. 2007-02-19;67(4):1430-5.
PMID: 17308080
Related citations

In a previous study, we found that the small-molecule epidermal growth factor receptor (EGFR) inhibitor gefitinib (ZD1839, Iressa) blocked cell proliferation at biologically relevant concentrations in approximately one third (6 of 17) of human bladder cancer cell lines examined. Here, we studied the effects of gefitinib on apoptosis in a representative subset of the same panel of cells. The drug had modest effects on DNA fragmentation as a single agent at concentrations that produced strong growth inhibition (< or =1 micromol/L) and also failed to promote apoptosis induced by conventional chemotherapeutic agents (gemcitabine and paclitaxel). However, gefitinib did interact with recombinant human tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) to induce high levels of apoptosis in gefitinib-responsive but not gefitinib-unresponsive lines. The molecular mechanisms involved down-regulation of active AKT and X-linked inhibitor of apoptosis protein (XIAP) expression and were mimicked by chemical inhibitors of the phosphatidylinositol 3-kinase/AKT pathway but not of the mitogen-activated protein kinase/extracellular signal-regulated kinase (ERK) kinase/ERK pathway. Furthermore, direct small interfering RNA-mediated knockdown of AKT resulted in down-regulation of XIAP and TRAIL sensitization, and knockdown of XIAP itself was sufficient to reverse TRAIL resistance. Together, our results show that EGFR pathway activation limits TRAIL-induced apoptosis via an AKT- and XIAP-dependent mechanism in EGFR-dependent human bladder cancer cells, providing the conceptual framework for a further evaluation of the combination in relevant preclinical in vivo models.

Targeting inhibitor of apoptosis proteins in combination with ErbB antagonists in breast cancer.
Fiona M Foster;Thomas W Owens;Jolanta Tanianis-Hughes;Robert B Clarke;Keith Brennan;Nigel J Bundred;Charles H Streuli;

Figure 9. The abstracts retrieved by Copub.
doi:10.1371/journal.pone.0098826.g009

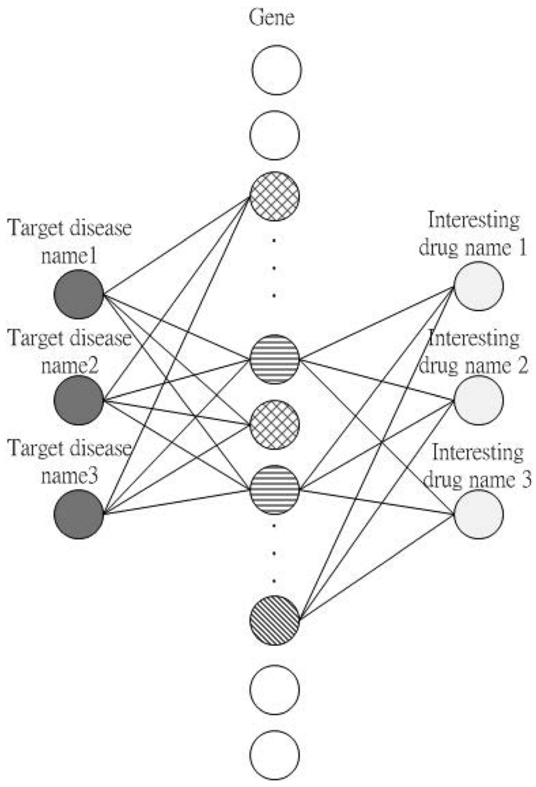


Figure 10. Concept behind the proposed method. Given names of target diseases and drugs and the related microarray expression profile, the proposed method searches the abstracts of genes related to the diseases and drugs, finds common parts, and provides the user a visual report.
doi:10.1371/journal.pone.0098826.g010

The proposed method identified 61 dominant genes from the disease-related GEP and 191 genes from the drug-related GEP, which were then used in a search for abstracts in relation to "acute lymphoblastic leukemia, acute myeloid leukemia" in PubMed.

In this experiment, we identified 10 genes with hidden connections between the disease terms "acute lymphoblastic leukemia, acute myeloid leukemia" and the drug "Gefitinib, Iressa", as shown in Fig. 4.

Gene CD14 represents an intermediate node between the disease and the drug. Yu et al. [24] claimed that "the findings suggest that CD14-260C/T polymorphism can contribute to B-ALL risk in a Chinese population". In addition, Noh et al. stated that "Gefitinib induced the expression of differentiation markers including CD11b and CD14 in ATO-treated NB4 cells and facilitated Arsenic trioxide (ATO)-induced morphologic changes and reactive oxygen species (ROS) generation" [25]. These statements attest to the relationships between the diseases and genes as well as to that between the drugs and genes.

Figure 5 presents the results obtained by inputting the same disease and drug terms into CoPub. It should be noted that the terms "acute lymphoblastic leukemia, acute myeloid leukemia" are classified by CoPub under the catalog term "leukemia". As a result, the disease node in Fig. 5 is labelled "leukemia". A total of 159 genes were selected by CoPub; however, none of these match those obtained using the proposed method.

Figure 2 (b) presents the ROC results of Scenario 2 with sensitivity and specificity of 92.0% and 83.1%, respectively. AUC in this case was 0.877 with a 95% CI of between 0.856 and 0.899.

Scenario 3

Not every drug has GEP; therefore, this scenario simulated the situation in which only the disease-related GEP is used to obtain the list of disease-related dominant genes using DOFA. The drug-related gene list was compiled from the PharmGKB website,

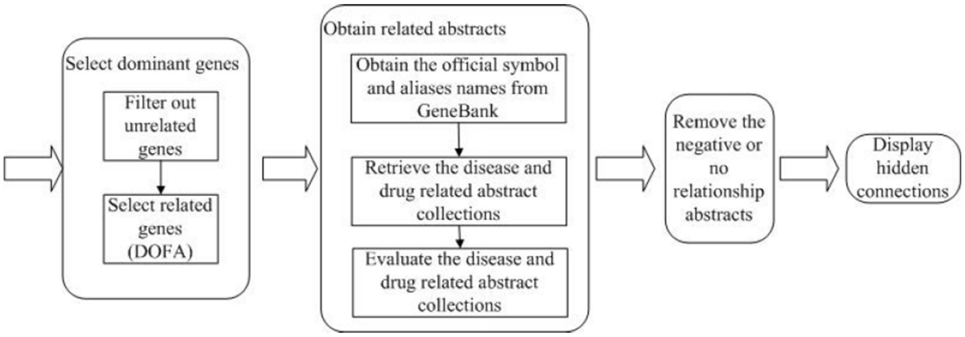

Select dominant genes
Filter out unrelated genes
Select related genes (DOFA)
Obtain related abstracts
Obtain the official symbol and aliases names from GeneBank
Retrieve the disease and drug related abstract collections
Evaluate the disease and drug related abstract collections
Remove the negative or no relationship abstracts
Display hidden connections

Figure 11. Flowchart of the proposed method.
doi:10.1371/journal.pone.0098826.g011

which provides a great deal of drug-related and gene-related information. This scenario used the terms "liver cancer", "hepatic cancer", and "hepatocellular carcinoma" as the target disease terms and "R340" and "capecitabine" as the target drug terms. The dataset provided by Chen et al. [26] was used as the input GEP. This dataset contains 156 samples, each comprising 3964 genes.

The proposed method identified 65 genes from the GEP and 30 genes from the PharmGKB website, which were used to identify target drug-related abstracts from the PubMed website. Figure 6 presents the 22 genes with hidden connections among the input diseases and drugs.

The gene "thymidine phosphorylase" (TYMP) represents an intermediate node between the disease and the drug. Ezaki et al. [27] claimed that "the measurement of thymidine phosphorylase (TP) activity in normal liver tissue adjacent to hepatocellular carcinoma (HCC) may predict multicentric recurrence a long time after an operation." Ko et al. [28] also stated, "thymidine phosphorylase (TP) is the rate-limiting enzyme for the activation of capecitabine (pro-drug of fluorouracil) and a useful predictor of tumor response to capecitabine-based chemotherapy." Clearly, the gene "TYMP" represents a hidden connection between the disease and the drug.

Figure 7 presents the results of inputting the same terms into CoPub. It should be noted that the term "liver cancer" is classified by CoPub as a disease term; however, "hepatocellular carcinoma" is classified as "tissue". Hence, in Figure 7, the disease node labelled "liver cancer" is marked in yellow, "hepatocellular carcinoma" is marked in blue, and the drug node "capecitabine" is also marked in blue. A total of 28 genes were selected by CoPub, but only one of these was obtained using the proposed method.

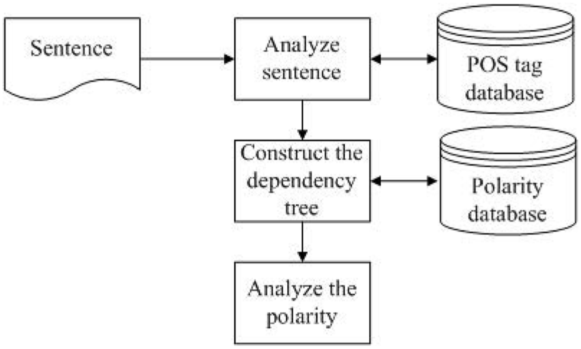


Figure 12. Flowchart of the negative-term filtering scheme.
doi:10.1371/journal.pone.0098826.g012

Figure 2 (c) shows the ROC results of Scenario 3 with sensitivity and specificity of 86.0% and 83.1%, respectively. AUC in this case was 0.833 with a 95% CI of between 0.819 and 0.848.

Discussion

Uncovering hidden connections among diseases, genes, and drugs revealed that many of the retrieved abstracts were reported as negative relationships or non-relationships. For example, the co-occurrence method retrieved one study [29] pertaining to the relationship between the gene "PDCD4" and the drug "cyclophosphamide", as shown in Figure 8. Although this study contains both of these terms, it describes a non-relationship in which the gene is not directly influenced by the drug.

The three scenarios presented a large number of documents providing negative relationships or non-relationships. Scenario 1 presented 17 hidden connections before negative-term filtering and 15 hidden connections after filtering. In Scenario 1, the proposed algorithm fetched 4028 documents (abstracts and titles). To evaluate the performance of the negative-term filtering scheme, we used the proposed weight function to verify whether these fetched documents were related to the input terms (gene or disease name), resulting in the retrieval of 834 from the 4028 documents. Then the negative-term filtering scheme removed 126 of these 834 documents. Among the removed documents, 97 were deemed to be correctly identified as "negative", demonstrating the high performance of the negative-term filtering scheme. One example of false positives is presented by documents indicating relationships between the genes "PDCD4"and "STAT1" and the drug "cyclophosphamide", which were subsequently removed. Another example involves "STAT1", which recovered an abstract reporting the following: "Cs-A and IFN-gamma, but not glucocorticoids, **cyclophosphamide**, or azathioprine, inhibited TGF-beta-induced signaling, as assessed by luciferase reporter gene assays, and collagen deposition. In contrast, the effects of IFN-gamma required signal transducer and activator of transcription **(STAT)-1**" [30]. Clearly, no direct relationship is indicated between the drug "cyclophosphamide" and the gene "STAT1". This abstract implied a non-relationship and was therefore removed.

In Scenario 2, 11 hidden connections were obtained prior to negative-term filtering. The weight function removed all but 789 documents of the 1216 documents initially retrieved and negative-term filtering removed an additional 89 documents. Among these 89 documents, 38 were deemed to be correctly identified. One example of false positives in Scenario 2 is presented by documents indicating relationships between the gene "CCNE2" and the drug "gefitinib." These were judged to evidence a negative relationship

Table 3. The policy for combining the nodes used in the negative-term filtering scheme.

	Parent node (Positive)	Parent node (Negative)	Parent node (Neural)
Child node (Positive)	Positive	Neutral	Neutral
Child node (Negative)	Neutral	Negative	Negative
Child node (Neural)	Positive	Negative	Neutral
Child node (Anti-)	Negative	Positive	Neutral

doi:10.1371/journal.pone.0098826.t003

or non-relationship, and thus the connection was removed. For example, one abstract stated the following: "combination studies revealed that the response of EGFSP genes to luteolin was not affected by **gefitinib**, even though the two compounds were additive with respect to their abilities to inhibit CCNA2, **CCNE2**, CDC25A and PCNA" [31]. Clearly, this does not indicate a direct relationship between the gene "CCNE2"and the drug "gefitinib".

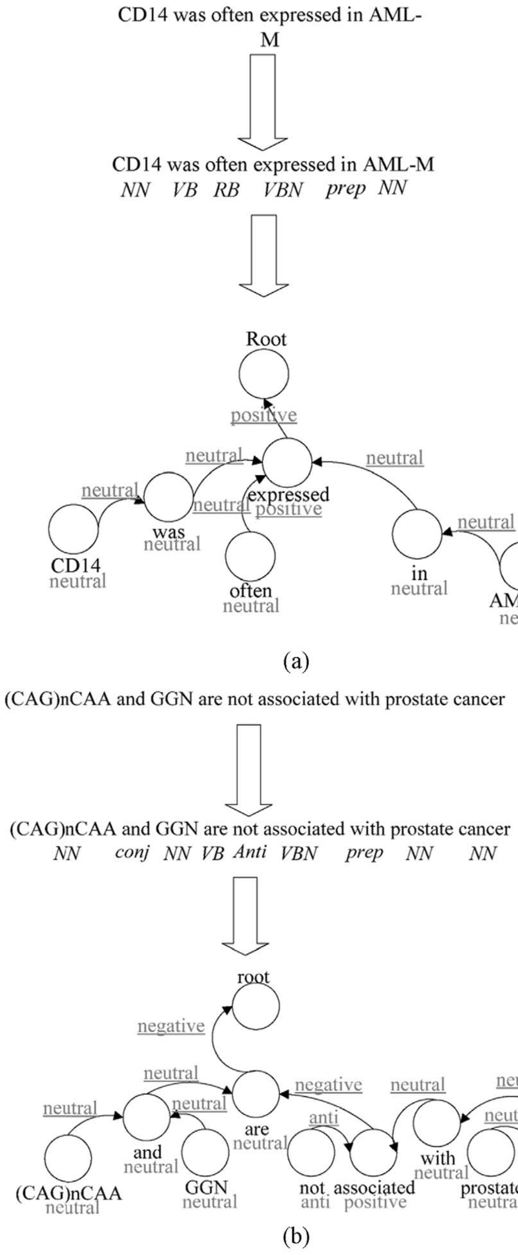


Figure 13. Examples of the negative-term filtering scheme.
doi:10.1371/journal.pone.0098826.g013

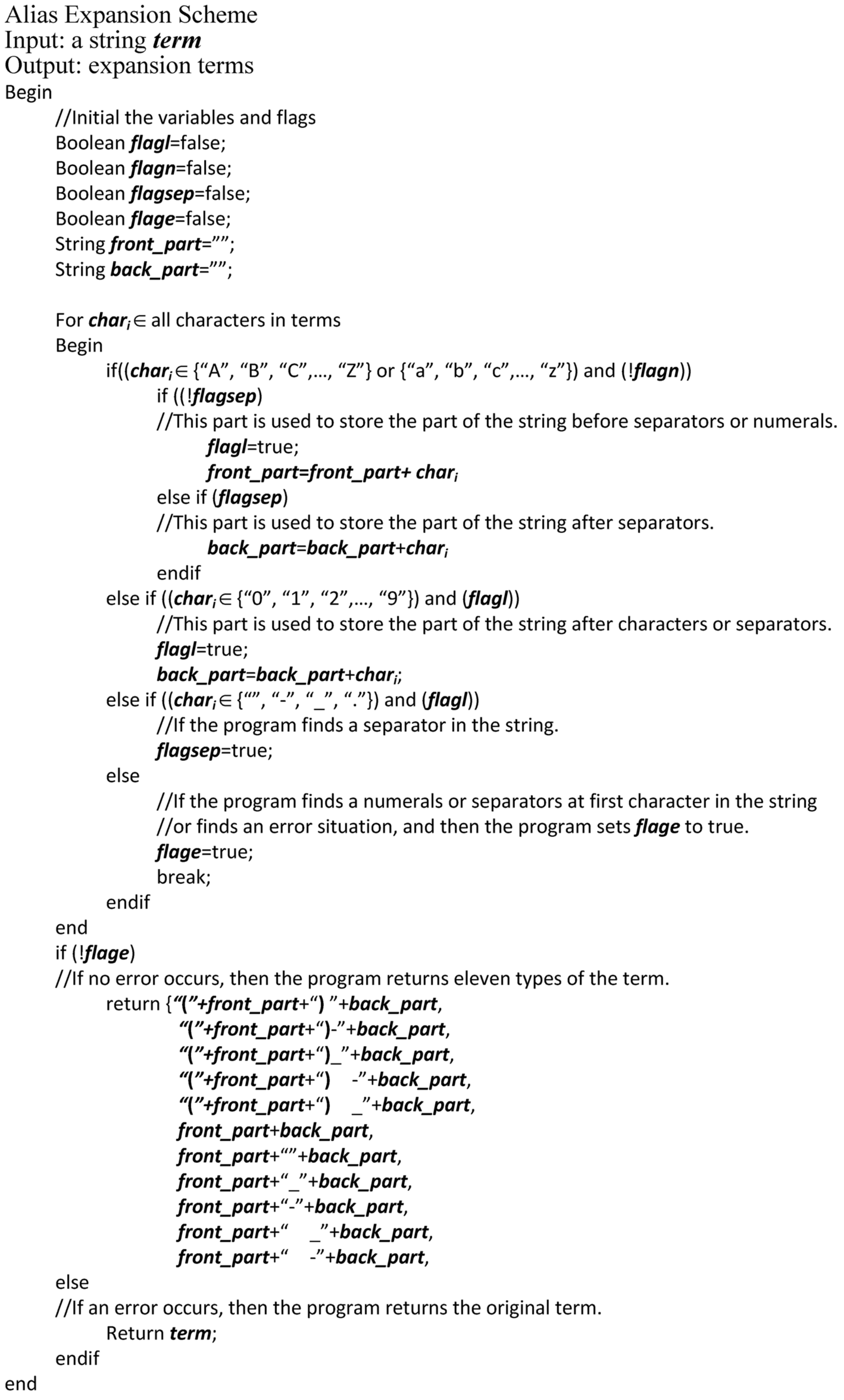
In Scenario 3, 23 hidden connections were identified prior to negative-term filtering. The weight function removed all but 1928 documents of the 3791 documents initially retrieved and negative-term filtering removed an additional 265 documents. Among these 265 documents, 202 were deemed to be correctly identified. The abstracts indicating a relationship between the input terms were judged as negative relationships or non-relationships by the negative-term filtering scheme and subsequently removed. For example, one abstract stated the following: "Taq1 digestion of PCR products revealed that both alleles were transcribed in all samples where both were presented at the genomic level, indicating that the **RRM1** locus is not subjected to imprinting in Wilms' tumour or **hepatoblastoma**" [32]. This clearly indicates that RRM1 is not subject to imprinting in hepatoblastoma, a type of liver cancer. The results of the three scenarios are summarized in Table 1.

The first row presents the total number of retrieved documents; the second row lists the documents remaining following application of the weight function; the third and fourth rows list the number of related and unrelated documents checked by manual; and the last four rows list the number of true positives, false positives, true negatives, and false negatives, respectively. In this case, a true positive refers to a document correctly judged to be related by the negative-term filtering scheme. A false positive represents a document which is deemed to be related by the negative-term filtering scheme, but is in fact unrelated. A true negative represents a document which is correctly deemed to be unrelated by the negative-term filtering scheme. A false negative represents a document which is judged to be unrelated by the negative-term filtering scheme but is in fact related.

We compared the performance of the algorithm using the weight function and the negative-term filtering scheme. Table 1 lists only the references determined by the weight function. The negative-term filtering scheme was then performed for a second check to obtain the following results: true positive, false positive, true negative, and false negative.

Table 2 lists all retrieved references acquired using 1) the weight function and 2) using both the weight function and negative-term filtering.

While performing these experiments, one problem was observed in the calculation of similarity scores among the genes and biological terms. Although synonymous gene names were obtained from GeneBank, still others were encountered due to the idiosyncrasies of individual authors' writing styles. For example, a number of authors referred to the gene "IL8" as "IL-8", others wrote "IL 8", and still others wrote "(IL) 8". To compensate for this kind of variation, we developed an alias expansion scheme to accommodate a variety of styles.



```
Alias Expansion Scheme
Input: a string term
Output: expansion terms
Begin
    //Initial the variables and flags
    Boolean flagl=false;
    Boolean flagn=false;
    Boolean flagsep=false;
    Boolean flage=false;
    String front_part="";
    String back_part="";

    For $char_i$ ∈ all characters in terms
    Begin
        if(($char_i$ ∈ {"A", "B", "C",..., "Z"} or {"a", "b", "c",..., "z"}) and (!flagn))
            if ((!flagsep)
            //This part is used to store the part of the string before separators or numerals.
                flagl=true;
                front_part=front_part+ $char_i$
            else if (flagsep)
            //This part is used to store the part of the string after separators.
                back_part=back_part+$char_i$
            endif
        else if (($char_i$ ∈ {"0", "1", "2",..., "9"}) and (flagl))
            //This part is used to store the part of the string after characters or separators.
            flagl=true;
            back_part=back_part+$char_i$;
        else if (($char_i$ ∈ {"", "-", "_", "."}) and (flagl))
            //If the program finds a separator in the string.
            flagsep=true;
        else
            //If the program finds a numerals or separators at first character in the string
            //or finds an error situation, and then the program sets flage to true.
            flage=true;
            break;
        endif
    end
    if (!flage)
    //If no error occurs, then the program returns eleven types of the term.
        return {"("+front_part+") "+back_part,
              "("+front_part+")-"+back_part,
              "("+front_part+")_"+back_part,
              "("+front_part+")    -"+back_part,
              "("+front_part+")    _"+back_part,
              front_part+back_part,
              front_part+""+back_part,
              front_part+"_"+back_part,
              front_part+"-"+back_part,
              front_part+"    _"+back_part,
              front_part+"    -"+back_part,
    else
    //If an error occurs, then the program returns the original term.
        Return term;
    endif
end
```

Figure 14. Pseudo-code of the alias expansion scheme.
doi:10.1371/journal.pone.0098826.g014

For example, the alias expansion scheme enables the generation of eleven different styles for the gene "IL8": "(IL) 8", "IL 8", "(IL)-8", "IL8", "IL-8", "(IL) -8", "(IL)-8", "(IL) _8", "IL-8", "IL_8" and "(IL)_8". The details of the alias expansion scheme are described in Scheme S1.

Nonetheless, there remain a number of situations that the alias expansion scheme is incapable of resolving. For example, in the sentence "we have assigned the most probable six-locus haplotypes determined by HLA-A, -Cw, -B and -DRB1 highly polymorphic genes", we find a term "HLA-A" before the term "HLA-B", in which both terms share "HLA"; however," HLA-B" is represented as "-B". This situation can result in an erroneous classification of the relationship between the gene "HLA-B" and the disease or the drug. This complication will have to be resolved if the hit rate is to be improved.

The results obtained using the proposed method differ considerably from those obtained using CoPub. CoPub uses six thesauri containing human genes, Gene Ontology biological processes, liver pathologies, diseases, pathways, and drugs to search MEDLINE XML files containing the title, abstract and substances. CoPub stores the context and PubMed identifiers (PMIDs) of the abstracts in which the keywords are identified and then computes the strength of the hidden relationship between A and C using the R-scaled scores between A and B, and between B and C. An R-scaled score above a set threshold is regarded as biologically significant. In contrast, the proposed method uses GEPs manufactured using intermediate nodes (genes) identified by experts through experimentation.

The proposed method also handles keywords in a different manner. CoPub uses catalog terms (such as "Leukemia") instead of the original term "acute lymphoblastic leukemia" as well as the full name of the gene, symbols, and aliases. The proposed method uses the original terms (disease or drug names) with the full name of the gene, symbols, and aliases. Although the search range of catalog terms appears to be wide, the fact is that some important information is missing. For example, CoPub uses the catalog term "colon cancer" instead of the original term "colorectal cancer" as a disease term, leading to a failure on the part of CoPub to include abstracts containing the input term "colorectal cancer".

In summary, these two approaches differ with regards to the source of the intermediate nodes as well as the queries. Consequently, the results also differ. Furthermore, CoPub uses only co-occurrences without considering the effect of negative relationships. For example, the abstracts [33–35] retrieved by Copub indicate a negative relationship or non-relationship between the gene "BIRC4" and "gefitinib" in Figure 9.

A number of the hidden connections in the connective map generated by the proposed method are thin, due to the small number of papers that mention this connection. Nonetheless, these connections may still possess biological or medical significance. In Scenario 3, the proposed method identified a hidden connection between the disease "hepatocellular carcinoma" and the drug "capecitabine", which is an important relationship confirmed in [36].

Regardless of the method used, only a portion of the hidden connections can be identified. Both methods are based on studies performed by experts; however, if the basic relationships between A and B or between B and C are not established, neither method can provide very much in the way of helpful information regarding the hidden connections.

Methods

As illustrated in Fig. 10, the objective behind the proposed methodology was the discovery of hidden connection between diseases and drugs. In Fig. 10, the left area provides target disease names such as "colon cancer" and "colorectal cancer", and the right part lists drug names such as "5-fluorouracil" and "5-FU". If one gene is related to a particular disease, then a line is connected from the disease to the gene. Similarly, if one gene is related to a drug, then a line is connected from the drug to the gene. Any gene related to both a disease and drug is considered a hidden connection between the disease and drug.

Figure 11 presents a flowchart of the proposed method divided into four stages: feature selection, collection of related information, relationship filtering, and display of results. Feature selection (performed by the disease-oriented feature selection algorithm, abbreviated as DOFA) [18] is used for the selection of disease- or drug-related gene lists followed by a search for aliases and official symbols of the related genes from within the GeneBank database (http://www.ncbi.nlm.nih.gov/sites/entrez?db = gene). These official names and alias names are used to generate corresponding queries comprising Boolean operators with genes, diseases, and drug names. These queries are used to retrieve abstracts and titles related to these genes and the specific disease or drug names via a search on PubMed. The proposed method uses a weight function to evaluate the retrieved abstracts in order to remove those that are unrelated. Differences in the writing styles of authors can lead to discrepancies in the notation of many genes. Differences in writing style are easily overlooked if only the relationships between the retrieved abstracts and official names and aliases are considered. Therefore our proposed method includes an alias expansion scheme. In the third stage, a negative-term filtering scheme is used to verify whether the retrieved abstracts present a positive relationship, negative relationship, or non-relationship. After obtaining a collection of abstracts with gene/disease links and gene/drug links, genes occurring in two collections of abstracts are output as hidden connections. The weight function, alias expansion scheme, and negative-term filtering scheme are outlined in the following.

The Weight Function

The abstracts retrieved in any search are not necessarily related to the disease, gene, or drugs in question. For example, when searching through documents related to "chronic myeloid leukemia (CML)", the results may include documents related to "chronic lymphocytic leukemia (CLL)". This makes it necessary to evaluate the abstracts collected. To calculate the relevance of the abstract, the specific disease names, and the commonplace terms (names or aliases of the specific gene) related to the specific disease, a weight function $W(D_i,g_j,d_x,T_s,S_t)$ is used to calculate the related score among abstract D_i, specific name d_x, and gene g_j, defined as follows:

$$W(D_i,g_j,d_x,T_s,S_t)$$
$$= \begin{cases} f(g_j,D_i) \times \left(\sum\limits_{T_s \in Dic, S_t \in D} L\left(T_s,g_j,S_t\right) + 1 \right) & \\ \times f(d_x,D_i) \times \left(\sum\limits_{T_s \in Dic, S_t \in D} L(T_s,d_x,S_t) + 1 \right) & \text{if } f(g_j,D_i) > 0 \text{ and } f\ (d_x,D_i) > 0 \\ -f(g_j,D_i) \times \left(\sum\limits_{T_s \in Dic, S_t \in D} L\left(T_s,g_j,S_t\right) + 1 \right) & \text{if } f(g_j,D_i) > 0 \text{ and } f\ (d_x,D_i) = 0 \\ -f(d_x,D_i) \times \left(\sum\limits_{T_s \in Dic, S_t \in D} L(T_s,d_x,S_t) + 1 \right) & \text{if } f(g_j,D_i) = 0 \text{ and } f\ (d_x,D_i) > 0 \\ -1000000 & \text{otherwise} \end{cases}$$

where the term frequency function $f(t, D_i)$ returns the number of times the variable t appears in abstract D_i. Variable t can be the name of a gene or its alias, a drug name, or a disease name. Function $L(T_s,\ t,\ S_t)$ represents the weight of the term t in the sentence S_t and is denoted as follows:

$$L(T_s,t,S_t)=\begin{cases}1, & T_s \text{ and } t \text{ appear in the same sentence } S_t\\ 0, & \text{otherwise}\end{cases}.$$

This function is used to enforce the weight of term t when this term and a special term T_s of the special biological dictionary *Dic* appear in the same sentence S_t. The special biological dictionary *Dic* is a collection of words from ten thousand cancer-related studies with the common words removed. If there is no special term T_s in sentence S_t, but only term t, then $L(T_s, t, S_t)$ would return zero. If all $L(T_s, t, S_t)$ return zero, then the weight functions would return zero without any special term. Hence, we add one to the weight function. After obtaining the related score via the weight function $W(D_i,g_j,d_x,T_s,S_t)$, the unrelated literature is removed from the collection of abstracts. If the related score of a document calculated using the weight function is greater than zero, then this document is regarded as related; otherwise, this document is regarded as unrelated and filtered out. For example, document D reports the following: "Two **HLA-B** mismatches were also *significantly associated* with **lymphoma** in the kidney (HR 2.82, P = 0.009)". Sentence S_t is the same as D, since there is only one sentence in D. Gene g "HLA-B" appears once in D, hence, the $f(g, D)$ returns one; for the specific name d "lymphoma", $f(d, D)$ gives one. Term Ts "significantly" and "associated" appear once in S_t. Hence, the weight score of this document can be calculated as follows:

$$\begin{aligned}&W(D,g,d,T_s,S_t)\\&=f(g,D)\times(L(''\text{significantly}'',g,S_t)+L(''\text{associated}'',g,S_t)+1)\\&\times f(d,D)\times(L(''\text{significantly}'',d,S_t)+L(''\text{associated}'',d,S_t)+1)\\&=1\times(1+1+1)\times1\times(1+1+1)\\&=9\end{aligned}$$

Negative-term Filtering Scheme

The proposed algorithm seeks hidden connections within medical abstracts, yet abstracts containing negative relationships or non-relationships may result in false positives. Abstracts containing negative relationships or non-relationships should be removed to ensure the validity of the hidden connections between diseases and drugs.

To remove documents with negative relationships or non-relationships, we adopted the algorithm proposed for sentiment analysis in [37]. Figure 12 presents a flowchart of the negative-term filtering scheme. First, sentences containing input terms, such as "lymphoma cancer", are parsed into individual words. Each of these words is then checked for the most plausible POS tag from the POS tag database and the input terms are combined into a single term. These combined terms do not necessarily appear in the POS tag database and may refer to disease terms, drug terms, or gene names; therefore, a POS tag is assigned to the input term as "NN". A dependency tree of the sentence is then constructed in accordance with the algorithm proposed by McDonald et al. [38]. The words in each sentence are treated as vertices with the potential dependency of each word providing edges. Each edge is assigned a weight based on the possibility of two words appearing immediately after, which is used in the construction of graph G, representing the sentence. Some of the words may be found in multiple lexical categories (noun, adjective, adverbs); therefore, this graph is used to create the maximum spanning tree (MST) by maximizing number of ways that each term could be used in the sentence. This can be accomplished by determining the probability that a given term will appear in other locations, in accordance with the lexical categories to which it may belong. We then assign the corresponding polarity of each word in the resulting dependency tree via the polarity database from the leaf nodes to the root node. The policy for combining the nodes is included in Table 3.

In this process, the term "anti-" inverts the polarity of the term to follow. For example, the two words "not elicit" imply "anti-" and "positive". The combined polarity of these two words is therefore negative. The proposed algorithm repeats this process of combination until reaching the root node which contains a single final polarity. This final polarity represents the polarity of this sentence with regard to the input term (disease or drug name).

Figure 13 presents two examples of the negative-term filtering scheme. In Fig. 13 (a), we used the sentence "CD14 was often expressed in AML-M": CD14 and AML-M are the input terms. The POS tagging results are "NN VB RB VBN prep NN", where NN represents a noun, VB represents a verb, RB represents an adverb, VBN represents the past participle of one verb, and prep represents preposition. The polarities of all words are then combined in accordance with the POS tagging results. In Fig. 13 (a), the polarity term is drawn in gray, and the combined polarity results are underlined. For example, the polarity of "CD14" is neutral and the polarity of "was" is neutral. When combined, these two polarity terms output a new polarity "neutral". The polarity of "expressed" is positive and the polarities of the other three combined results are all neutral, such that the root node outputs a final combined result of "positive". Figure 13(b) presents an example of negative polarity derived from the sentence "(CAG)nCAA and GGN are not associated with prostate cancer".

It should be noted that the two databases used in the negative-term filtering scheme were constructed in advance. The POS tag database was constructed using the renowned Brown corpus [39]. The polarity database was manually constructed by defining the polarities of all words in the POS tagging database. The meaning of each word was looked up in the dictionary and a group of experts was assembled to assign polarity to each word according to its meaning. For example, the word "better", is listed as the comparative of "well" in the dictionary. Thus, its polarity was deemed "positive". Any word without a clear polarity was marked "neutral".

(The database can be downloaded in [40]).

The Brown corpus does not contain the terms "downregulate (VB)", "downregulated (VBD)", "downregulated (VBN)", "upregulated (VBN)", "upregulated (VBD)", and "upregulate (VB)", yet these terms are important in the extraction of relationships from documents; therefore, we manually inserted these terms into the polarity database. These six words also follow the same rule to assign their polarities of words. In addition, terms that were not found in the polarity database were automatically assigned a "neutral" polarity.

Although the proposed negative-term filtering scheme is able to filter out a large number of documents with negative relationships or non-relationships, there remains considerable room for improvement through the inclusion of additional negative-terms and semantic analysis of greater subtlety.

Alias Expansion Scheme

The weight function uses the frequency with which the names of input genes appear in the retrieved abstracts in order to evaluate the relevance of the documents. This necessitated developing the means to generate a comprehensive list of gene formats to accommodate variations in the writing styles of authors and subsequently minimize the possibility of misjudging the weights in the proposed algorithm. Figure 14 shows the pseudo code of the alias expansion scheme, which is presented in further detail in Scheme S1.

Supporting Information

Figure S1 The flowchart of DOFA.
(TIFF)

Figure S2 Some of original data of the Colon dataset.
(TIFF)

Figure S3 Flowchart of the classification pattern learning phase of the proposed algorithm.
(TIFF)

Figure S4 An example of the classification pattern *ID* for class *k*.
(TIF)

Algorithm S1 Disease orient feature selection algorithm (DOFA).
(DOCX)

Data S1 Experimental Results.
(RAR)

Scheme S1 Alias Expansion Scheme.
(DOCX)

Acknowledgments

We would like to thank Dr. Hsueh-Wei Chang in Department of Biomedical Science and Environmental Biology, College of Health Sciences, Kaohsiung Medical University for providing useful discussions in the experiments.

Author Contributions

Conceived and designed the experiments: CNL JSW. Performed the experiments: JSW EFK. Analyzed the data: JSW EFK CNL. Contributed reagents/materials/analysis tools: JSW. Wrote the paper: JSW CNL.

References

1. Swanson DR (1986) Undiscovered public knowledge. The Library Quarterly 56: 103–118.
2. Spasser MA (1998) The enacted fate of undiscovered public knowledge. J AM SOC INF SCI. 48: 707–717.
3. Swanson DR (2001) On the fragmentation of knowledge, the connection explosion, and assembling other people's ideas. Bull Amer Soc Inform Sci Technol 27: 12–14.
4. Lamb J (2007) The connectivity map: A new tool for biomedical research. Nat Rev Cancer 7: 54–60.
5. Butte AJ, Kohane IS (2006) Creation and implications of a phenome-genome network. Nat Biotech 24: 55–62.
6. Li J, Zhu X, Chen JY (2009) Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol 5(7): e1000450.
7. Smalheiser NR, Swanson DR (1998) Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed 57: 149–153.
8. Homayouni R, Heinrich K, Wei L, Berry MW (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics 21: 104–115.
9. Frijters R, Heupers B, van Beek P, Bouwhuis M, van Schaik R, et al. (2008) CoPub: A literature-based keyword enrichment tool for microarray data analysis. Nucleic Acids Res (suppl 2): W406–W410.
10. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, et al. (2010) Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol 6: e1000943.
11. NCBI PubMed. Available: http://www.ncbi.nlm.nih.gov/pubmed/. Accessed 2014 May 12.
12. Buyko E, Faessler E, Wermter J, Hahn U (2009) Event extraction from trimmed dependency graphs. Proc. BioNLP ST'11, 19–27. ACL.
13. Bjrne J, Salakoski T (2011) Generalizing Biomedical Event Extraction. Proc. BioNLP ST'11, 183–191. ACL.
14. Sanchez-Graillet O, Poesio M (2007) Negation of protein-protein interactions: analysis and extraction. Bioinformatics. 23: i424–i432.
15. Tomiuk S, Hofmann K (2001) Microarray probe selection strategies. Brief Bioinform 2: 329–340.
16. Affymetrix GeneChip Available: https://www.affymetrix.com.Accessed 2014 May 12.
17. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, et al. (2002) PharmGKB: The pharmacogenetics knowledge base. Nucleic Acids Res 30: 163–165.
18. Wu JS, Wu KY, Lee CN, Chiang CW (2010) DOFA: A novel disease oriented gene selection algorithm. CIIS Lab Tech Rep 2010. Available: http://edith.cse.nsysu.edu.tw/wu/DOFA.pdf. Accessed 2014 May 12.
19. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8: 68–74.
20. Kancherla RR, Nair JS, Ahmed T, Durrani H, Seiter K, et al. (2001) Evaluation of topotecan and etoposide for non-Hodgkin lymphoma: correlation of topoisomerase-DNA complex formation with clinical response. Cancer 91(3): 463–471.
21. Minagawa Y, Kigawa J, Itamochi H, Terakawa N (2001) The role of topoisomerase I inhibitor in cisplatin-resistant ovarian cancer. Hum Cell. 14(3): 237–243.
22. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286: 531–537.
23. Stegmaier K, Corsello SM, Ross KN, Wong JS, DeAngelo DJ, et al. (2005) Gefitinib induces myeloid differentiation of acute myeloid leukemia. Blood 106: 2841–2848.
24. Yu X, Zhang C, Sun A, Jiang L, Zheng J, et al. (2011) Genetic variations in CD14 promoter and acute lymphoblastic leukemia susceptibility in a Chinese population. DNA Cell Biol. 30: 777–82.
25. Noh E, Kim H, Park MJ, Baek JH, Park J, et al. (2010) Gefitinib enhances arsenic trioxide (AS2O3)-induced differentiation of acute promyelocytic leukemia cell line. Leuk Res 34: 1501–1505.
26. Chen X, Higgins J, Cheung S, Li R, Mason V, et al. (2004) Novel endothelial cell markers in hepatocellular carcinoma. Mod Pathol 17: 1198–1210.
27. Ezaki T, Ikegami T, Maeda T, Yamada T, Ishida T, et al. (2005) Prognostic value of thymidine phosphorylase activity in liver tissue adjacent to hepatocellular carcinoma. Int J Clin Oncol 10: 171–176.
28. Ko J, Tsai M, Kuo Y, Chiu Y, Weng S, et al. (2011) Modulation of Rad51, ERCC1, and thymidine phosphorylase by emodin result in synergistic cytotoxic effect in combination with capecitabine. Biochem Pharmacol 81: 680–690.
29. Zhang X, Wang X, Song X, Liu C, Shi Y, et al. (2010) Programmed cell death 4 enhances chemosensitivity of ovarian cancer cells by activating death receptor pathway in vitro and in vivo. Cancer Sci 101: 2163–2170.
30. Eickelberg O, Pansky A, Koehler E, Bihl M, Tamm M, et al. (2001) Molecular mechanisms of TGF-(beta) antagonism by interferon (gamma) and cyclosporine A in lung fibroblasts. FASEB J 15: 797–806.
31. Markaverich BM, Vijjeswarapu M, Shoulars K, Rodriguez M (2010) Luteolin and gefitinib regulation of EGF signaling pathway and cell cycle pathway genes in PC-3 human prostate cancer cells. J Steroid Biochem Mol Biol. 122: 219–31.
32. Byrne JA, Smith PJ (1993) The 11p15.5 ribonucleotide reductase M1 subunit locus is not imprinted in Wilms' tumour and hepatoblastoma. Hum Genet 91: 275–7.
33. Chang GC, Yu CT, Tsai CH, Tsai JR, Chen JC, et al. (2008) An epidermal growth factor inhibitor, Gefitinib, induces apoptosis through a p53-dependent upregulation of pro-apoptotic molecules and downregulation of anti-apoptotic molecules in human lung adenocarcinoma A549 cells. Eur J Pharmacol. 600: 37–44.
34. Shrader M, Pino MS, Lashinger L, Bar-Eli M, Adam L, et al. (2007) Gefitinib reverses TRAIL resistance in human bladder cancer cell lines via inhibition of AKT-mediated X-linked inhibitor of apoptosis protein expression. Cancer Res. 67: 1430–1435.
35. Foster FM, Owens TW, Tanianis-Hughes J, Clarke RB, Brennan K, et al. (2009) Targeting inhibitor of apoptosis proteins in combination with ErbB antagonists in breast cancer. Breast Cancer Res. 11: R41.
36. Xiao YS, Tang ZY, Fan J, Zhou J, Wu ZQ, et al. (2004) Interferon-alpha 2a up-regulated thymidine phosphorylase and enhanced antitumor effect of capecitabine on hepatocellular carcinoma in nude mice. J Cancer Res Clin Oncol. 130(9): 546–550.
37. Nasukawa T, Yi J (2003) Sentiment analysis: capturing favorability using natural language processing, Proc 2nd int conf Knowledge capture 23–25.
38. McDonald R, Pereira F, Ribarov K, Hajič J (2005) Non-projective dependency parsing using spanning tree algorithms, Proc conf Human Language Technology and Empirical Methods in Natural Language Processing 523–530.

39. Brown corpus. Available: http://archive.org/details/BrownCorpus. Accessed 2014 May 12.
40. figshare (HICONA softwares). Available: http://figshare.com/articles/Discovering_Hidden_Connections_among_Diseases_Genes_and_Drugs_Based_on_Microarray_Expression_Profiles_with_Negative_term_Filtering/915873. Accessed 2014 May 12. figshare (HICONA data). Available: http://figshare.com/articles/Discovering_Hidden_Connections_among_Diseases_Genes_and_Drugs_Based_on_Microarray_Expression_Profiles_with_Negative_term_Filtering/915874. Accessed 2014 May 12. figshare (The polarity database used in this paper). Available: http://figshare.com/articles/The_polarity_database_used_in_the_paper_Discovering_Hidden_Connections_among_Diseases_Genes_and_Drugs_Based_on_Microarray_Expression_Profiles_with_Negative_term_Filtering_/930412. Accessed 2014 May 12.