Briefings in Bioinformatics, 18(3), 2017, 488–497

doi: 10.1093/bib/bbw030
Advance Access Publication Date: 25 April 2016
Paper

Literature-based discovery of new candidates for drug repurposing

Hsih-Te Yang*, Jiun-Huang Ju*, Yue-Ting Wong, Ilya Shmulevich and Jung-Hsien Chiang

Corresponding author: Jung-Hsien Chiang, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. Tel.: +886-6-2757575 ext.62534; Fax: +886-2747076; E-mail: jchiang@mail.ncku.edu.tw
*These authors contributed equally to this work.

Abstract

Drug development is an expensive and time-consuming process; these could be reduced if the existing resources could be used to identify candidates for drug repurposing. This study sought to do this by text mining a large-scale literature repository to curate repurposed drug lists for different cancers. We devised a pattern-based relationship extraction method to extract disease–gene and gene–drug direct relationships from the literature. These direct relationships are used to infer indirect relationships using the ABC model. A gene-shared ranking method based on drug target similarity was then proposed to prioritize the indirect relationships. Our method of assessing drug target similarity correlated to existing anatomical therapeutic chemical code-based methods with a Pearson correlation coefficient of 0.9311. The indirect relationships ranking method achieved a significant mean average precision score of top 100 most common diseases. We also confirmed the suitability of candidates identified for repurposing as anticancer drugs by conducting a manual review of the literature and the clinical trials. Eventually, for visualization and enrichment of huge amount of repurposed drug information, a chord diagram was demonstrated to rapidly identify two novel indications for further biological evaluations.

Key words: text-mining, ABC model, drug repurposing, information extraction, natural language processing, drug target discovery

Introduction

Drug development is a time-consuming, expensive and high-risk process [1, 2], and this has led to the emergence of drug repurposing (also known as drug repositioning or drug re-profiling), which is the application of already approved drugs to new diseases. The principal advantage of drug repurposing over drug development is that approved drugs have already been through several stages of clinical trials and, therefore, have well-known safety and pharmacokinetic profiles [3]. A well-known example is that of aspirin, which was originally used for pain relief and has since been used to prevent cardiovascular disease and cancer [4, 5]. Sildenafil, initially used to treat high

Hsih-Te Yang is an assistant professor at the Institute of Medical Informatics, Department of Computer Science and Information Engineering, and Institute of Oral Medicine, National Cheng Kung University, Taiwan. His research focuses are mainly on therapeutics discovery, e.g. small molecule drug, biologics and vaccine, by making use of genomic and clinical trial data.
Jiun-Huang Ju was a PhD candidate of Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan. He majored in bioinformatics and biomedical literature mining for his thesis regarding identification of novel protein–protein interactions, document triage for chemical-gene-disease information and drug repurposing.
Yue-Ting Wong was a master student of Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, and was also well trained in Java programming, database management and text-mining technologies.
Ilya Shmulevich is currently a Professor at the Institute for Systems Biology, Seattle, USA. Dr Shmulevich directs a Genome Data Analysis Center as part of The Cancer Genome Atlas project, a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.
Jung-Hsien Chiang is a distinguished professor at the Institute of Medical Informatics, Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan. Dr. Chiang has been working in the fields of text mining and artificial intelligence, and performing outstanding leaderships to promote and carry out interdisciplinary research over the international collaborations.
Submitted: 24 November 2015; **Received (in revised form):** 2 March 2016

Table 1. Comparison of the effectiveness of four different methods of relationship extraction (RE)

RE method	Precision	Recall
Abstract level	0.7666	0.2284
Sentence level	0.8521	0.1341
Our approach (all POS)	0.859	0.1041
Our approach (verb, adj and noun only)	0.8674	0.091

blood pressure, has been repurposed to treat erectile dysfunction [6]. Thalidomide was originally used against nausea and to ease morning sickness in pregnant women but failed because it can cause Phocomelia syndrome. However, thalidomide has recently been found effective for the treatment of dermatological disorders, aphthous stomatitis, as well as multiple myeloma [7, 8]. These repurposed drugs have demonstrated that drug repurposing is a promising way to improve drug discovery.

The identification of disease–gene, gene–drug and disease–drug relationships is key to identifying and curating new candidates for drug repurposing. Although there are databases [9–12] that extensively curate the relationships between various biomedical entities, many other unidentified relationships may be buried in the biomedical literature. As noted in a review and two research articles [13–15], literature-based discovery (LBD) to generate scientific hypotheses for finding new indications of existing drugs seems to be a well-suited strategy. Andronis *et al.* [13] reviewed various LBD approaches showing the detection of hidden connection between biomedical entities is crucial and suggested that visualization techniques could facilitate the detection for scientists. Beyond the review work, Tari *et al.* [14] used a declarative programming language, AnsProlog, to achieve the automated reasoning for the incomplete information of indirect relationships for drug indications. Furthermore, Tari *et al.* [15] introduced several publicly available knowledge resources such as chemical structures, side effects and signalling pathways for identifying alternative drug indications. Therefore, identification of otherwise hidden relationships by text mining biomedical literature repositories could enable the discovery of unidentified relationships for drug repurposing.

Under the assumption that extensive knowledge is hidden in the large-scale literature, we sought to develop a relationship extraction method for the purpose of drug repurposing. Conventional approaches that focused on extracting the existing disease–drug relationships could fail to find potential new disease–drug relationships [16, 17]. In this study, we focused on extracting disease–gene relationships and gene–drug relationships to discover hidden disease–drug relationships and found that useful indirect relationships could be identified using this strategy. We developed a text-mining-based ranking method to allow the detection of indirect relationships that may facilitate the discovery of new candidates supporting the curation for drug repurposing.

Specifically, the aim of this study was to design an intuitive pattern-based learning method to extract relationships from the biomedical literature along with a drug vector space-based ranking method to identify the most promising potential drugs.

Materials and methods

Target document triage

PubMed comprises more than 24 million citations from MEDLINE and other data sources of biomedical literature. We downloaded the MEDLINE database (version 2014) for use as our primary resource. As the database also contained articles that were irrelevant to our work, we first sought to discard the literature that did not concern disease–gene, gene–drug and disease–drug relationships.

Predefined lexicon compiling

The Therapeutic Target Database (TTD; http://bidd.nus.edu.sg/group/cjttd/) provides a great deal of information concerning targets and their corresponding drugs and diseases [11]. Within the context of our work, TTD is a more appropriate resource to generate a list of relevant entity names because it contains sufficient information regarding disease–drug relationships. We gathered 2723 diseases, 3188 targets and 20 043 drugs from TTD (see Supplementary Data for details), and these data were used to generate a list of named entities for filtering the whole MEDLINE database.

Irrelevant document filtering

Apache Lucene (http://lucene.apache.org/), a high-performance search engine, was used to quickly search the large document collection. We queried the whole MEDLINE database using the Apache Lucene's search libraries to remove the documents with no mention of the disease, gene or drug entities identified using the predefined lexicon. We collected 5.3 million disease-related documents, 7.1 million gene-related documents and 5.5 million drug-related documents. We only wanted to retain documents that mentioned at least two different entities. These kept documents are therefore referred to as 'target documents'. However, relying only on the predefined lexicon to gather the target documents could lead to false negatives. We learned that capturing true positives as accurate as possible is quite crucial for the detection of indirect relationships though it also leads to a lower recall in our experiment (Table 1). A paper of Tari *et al.* [15] appears to have better results for recall, whereas we particularly aimed at the precision of prediction because of the huge amount of citations used. As shown in Figure 1A, the documents expressing '(Disease∩Gene)∪(Gene∩Drug)∪(Disease∩Drug)' defined our target documents. In total, we identified 5.4 million target documents.

Relationship extraction

In general, co-occurrence-based relationship extraction is a straightforward approach although it may lead to the extraction of incorrect relationships. In this study, we used a semantic pattern-based relationship extraction method to improve the precision of our method owing to the importance of correct relationships. Although the recall can be expected to be somewhat lower, the corpus was large enough (5.4 million) for tackling this problem. As shown in Figure 1B, the relationship extraction method includes the three stages discussed below.

Named entity recognition

Named entity recognition (NER) seeks to locate and classify elements into predefined categories such as disease, gene or drug, and it can be simply divided into rule-based, dictionary-based and machine learning-based approaches. In this stage, we developed a dictionary-based NER system that identifies elements by matching the named entities extracted from the predefined lexicon (Section Predefined lexicon compiling). The named entities were then used to determine whether a sentence contains enough information for being retained.

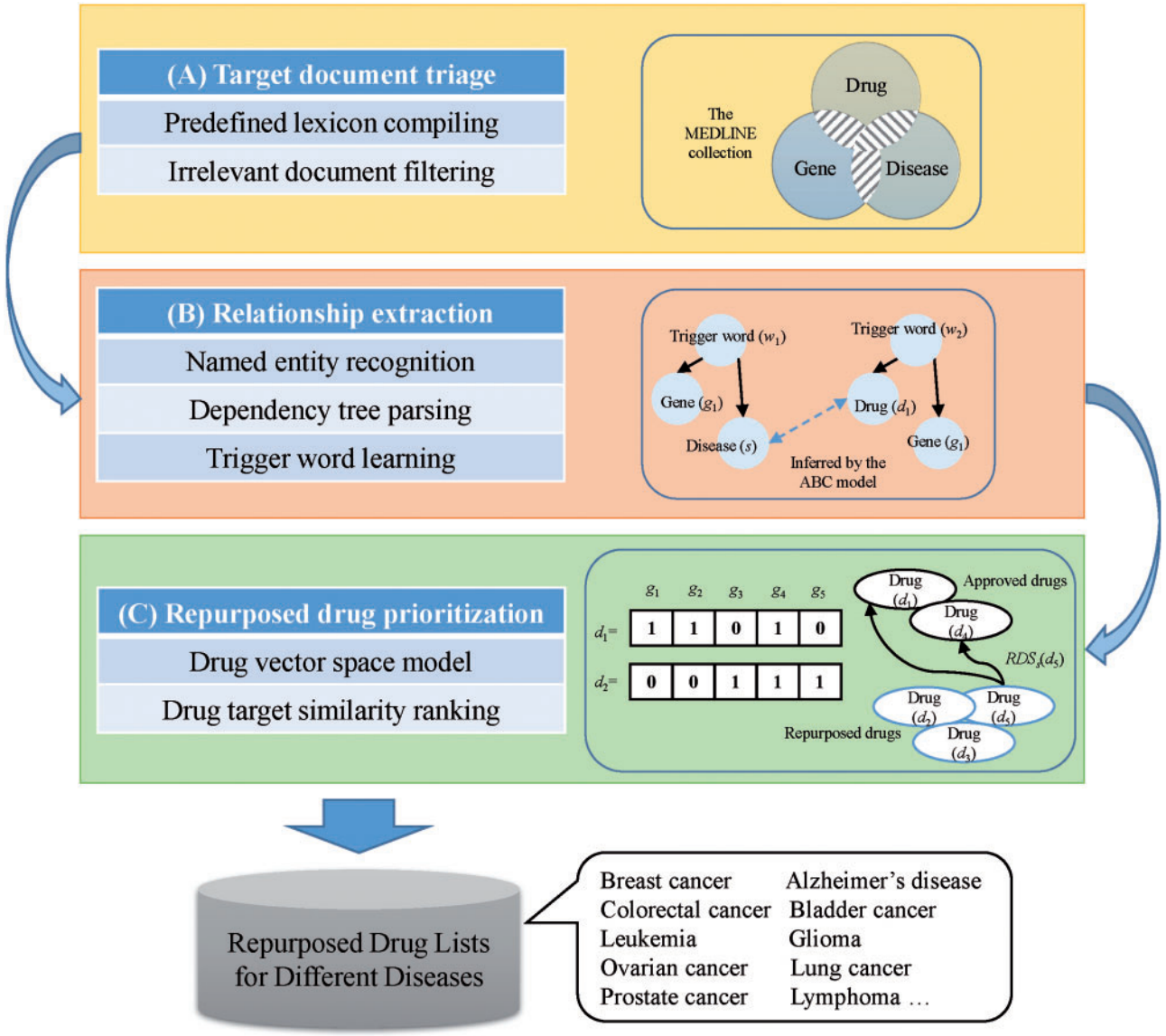

Figure 1. Overview of the method for identifying new candidates for drug repurposing by text mining the medical literature. (**A**) The entire MEDLINE collection is reduced. (**B**) The remaining articles are analysed for recognizing biomedical entities and inferring drugs. (**C**) Inferred drugs are represented by drug vector spaces and ranked by a drug prioritization. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Dependency tree parsing

Dependency grammar represents sentences with a syntactic tree and focuses on the underlying relationships between words [18]. In dependency grammar, verbs usually act as the structural root of complete clauses that consist of a subject and an object. Other words are either directly or indirectly dependent on the root. The Stanford parser (http://nlp.stanford.edu/software/lex-parser.shtml) is a natural language parser tool that was developed by the Stanford Natural Language Processing Group [19]. It can be used to determine which words are the subject or object in a sentence. For example, the sentence 'Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas' is converted to the root part 'were submitted', the subject noun part 'Senator Brownback, Republican of Kansas' and the object noun part 'Bills on ports and immigration'. Some studies have demonstrated that using the smallest common subtree of the subject part and the object part for relation extraction could emphasize the local characteristics and reduce noise of relations [20, 21]. In this study, we converted the sentences of the target documents to the smallest common subtrees in dependency format (hereafter referred to as 'triplet containers') for distinguishing the root part, the subject noun part and the object noun part of the sentences that contain two named entities in different categories (i.e. disease, gene, drug).

Trigger word learning

A triplet container, considered as a predicate argument structure (PAS), was shown to produce high-quality information extraction results [22]. However, the lack of suitable and off-the-shelf roots (hereafter referred to as 'trigger words') of triplet containers has hindered the development of extraction of correct relations. The Comparative Toxicogenomics Database (CTD; http://ctdbase.org/) currently contains >1 million chemical–gene, gene–disease, chemical–disease relationships that have been manually extracted from >100 000 documents [12]. We therefore collected trigger words by extracting the root parts

of the CTD relationships, which were also converted to the triplet containers. For example, a chemical–gene pair (C1305-PARP1) was extracted from PMID: 15231658, and the document states '... mouse cells lacking PARP-1 are extremely sensitive to C-1305, a new topoisomerase II inhibitor'. By converting the sentence to the triplet container, a trigger word 'sensitive' was derived as it belongs to the root part. The parts of speech (POS) of the trigger words were restricted to verb, adjective and noun, which have been studied in some previous works [20, 23, 24]. We parsed 94 513 documents from CTD and obtained about 11 000 unique trigger words. However, the obtained trigger words contained many words that were effectively noise. To guarantee the reliability of the trigger words, only the 1896 trigger words that appeared >10 times were used. Finally, we manually filtered some noise words (e.g. common words, digits and meaningless symbols) and retained the top 1000 most frequently occurring trigger words (see Supplementary Data for details). In summary, we converted the sentences of the target documents to triplet containers, and the relationships of the triplet containers were extracted if the following requirements were satisfied: (1) the subject part and the object part matched the predefined lexicon; (2) the root part matched the trigger words. We finally collected 114 381 disease–gene pairs, 176 219 gene–drug pairs and 88 573 disease–drug pairs from the target documents (in which all the pairs are direct relationships). In the next step, the pairs were used to find potential repurposed drugs using the ABC model.

2.3 Repurposed drug prioritization

The ABC model has been successful in explaining how two relationships are linked by an intermediate for drug discovery [25, 26]. Specifically, from a direct disease–gene relationship and a direct gene–drug relationship, one could infer an indirect disease–drug relationship. A drug in an indirect disease–drug relationship is one of the repurposed drugs for the disease, because it is indirectly linked to the disease by common genes. By linking 114 381 direct disease–gene relationships and 176 219 direct gene–drug relationships, we inferred 1 812 775 indirect disease–drug relationships based on the ABC model. However, many indirect disease–drug relationships may be spurious. We therefore developed a drug vector space representation and a drug target similarity ranking to prioritize the repurposed drugs.

Drug vector space model

We proposed a drug vector space to represent drugs by the linkage of genes to calculate the similarity between two drugs. Figure 1C shows the process that was used to transform drugs into a bit vector representation; the values of elements of a drug vector were determined based on connections to the linked genes. Denoting the whole gene set as $G = \{g_1, g_2, \ldots, g_n\}$ and the whole drug set as $D = \{d_1, d_2, \ldots, d_m\}$, the gene–drug relationships of the indirect disease–drug relationships with the same disease can be represented by an $n \times m$ matrix with each element $a_{ij} = 1$ if gene g_i links to drug d_j; otherwise, $a_{ij} = 0$. As a consequence, a drug d_j can be represented by the vector $[a_{1j}, a_{2j}, \ldots, a_{nj}]$. This drug vector space representation could also reveal similarities between drugs via clustering of drugs based on their linkage to genes. For example, if two drugs are linked to most of the same genes (i.e. their bit vectors are similar), this may imply that the two drugs are similar and have the same mechanism of action [27]. In this study, we prioritized repurposed drugs based on their similarity to approved drugs using the drug vector space model.

Drug target similarity ranking

The key concepts of the method are depicted in Figure 1C. If a drug is similar to an approved drug, then it may be a promising repurposing candidate. The drugs of the indirect disease–drug pairs of a given disease were split into two categories: approved drugs *AD* (if indicated as the targeted drugs of the disease) and repurposed drugs *RD* (if indicated as not the targeted drugs of the disease). The records of the targeted drugs for a disease were retrieved from TTD. The similarity of each repurposed drug to approved drugs was quantified using the Jaccard index, which measures the ratio of the sizes of the intersection of two sets to their union, and was scored using a summation method (some other methods such as maximum, geometric mean and average were also tested as discussed in Section Drug repurposing evaluation) over all approved drugs to get a final value of drug target similarity. We note that the binary drug vectors can be thought of as indicator vectors of their respective sets (of genes). Thus, the Jaccard index *JI* is

$$JI(d_p, d_q) = \frac{|d_p \cap d_q|}{|d_p \cup d|} \tag{1}$$

where d_p denotes the bit vector $[a_{1p}, a_{2p}, \ldots, a_{np}]$ of drug d_p, and d_q denotes the bit vector $[a_{1q}, a_{2q}, \ldots, a_{nq}]$ of drug d_q; the symbols $\cap$ and $\cup$ are interpreted as bit-wise AND OR, respectively, and $|\cdot|$ is the Hamming weight (number of 1s) of the vector. The repurposed drug score (*RDS*) for a given diseases *s* is then

$$RDS_s(RD) = \frac{\left\{\sum_{c=1}^{k} JI(RD, AD_c)\right\} - \min(RDS_s)}{\max(RDS_s) - \min(RDS_s)} \tag{2}$$

where *k* indicates the number of the approved drugs for disease *s*, AD_c is the *c*-th approved drug in all of *k* approved drugs, $\max(RDS_s)$ and $\min(RDS_s)$ represent the highest and lowest RDSs of the given disease *s*, respectively.

Results

We designed evaluation methods to verify the accuracy of our relationship extraction, drug similarity calculation and drug repurposing assessment. Our relationship extraction evaluation was used to verify whether the extracted relationship was reliable; the drug similarity evaluation was to verify whether the drug vector space was suitable for calculating drug similarity; and the drug repurposing evaluation was to verify whether our drug target similarity ranking prioritized the repurposed drugs in a suitable manner.

Relationship extraction evaluation

If our relationship information is incorrect, it could lead to unsuitable drug repurposing owing to wrong inference of indirect relationships. Our approaches were compared with co-occurrence methods at the abstract level and sentence level. The abstract-level method treats two named entities in the same abstract as a relationship. The method that works at the sentence level only treats two named entities in the same sentence as a relationship. In general, working at the sentence level was more precise than working at the abstract level.

Precision and recall are often used for evaluating an information retrieval system. Precision is defined as the number of correctly retrieved relationships divided by the total number of retrieved relationships. Recall is defined as the number of

correctly retrieved relationships divided by the total number of correct relationships possible.

Our experimental data set was downloaded from the CTD Web site, which describes curated relationships from documents. In total, 94 513 biomedical documents were collected from which we randomly selected 10 000 biomedical documents as our data set. We used 10-fold cross-validation to test the performance of the methods. During the 10-fold cross-validation process, the data set was equally divided into 10 splits (subsets), with one subset used for testing and the remaining nine used for generating the list of trigger words. Table 1 shows that our approach has the highest precision (0.8674) of all methods (ranging from 0.7666 to 0.8521).

Drug similarity evaluation

To verify whether our drug similarity method was feasible, we investigated the correlation between the similarity computed by the drug vector space model and the similarity calculated by the anatomical therapeutic chemical (ATC) classification system. The ATC classification system is used for the classification of drugs and was considered as a ground truth in our study. According to the organ on which the drugs act as well as their therapeutic, pharmacological and chemical properties, they are classified into several different groups at five different levels and are assigned to the so-called ATC codes [28].

For example, the drug metformin is first classified into one of the 14 anatomical main groups—specifically group 'A', which is for alimentary tract and metabolism. Subsequently, at the second level, metformin is classed as '10' among the therapeutic subgroups as it is a drug that is used to treat diabetes. The third and fourth levels are the pharmacological subgroup and the chemical subgroup, respectively. The classifications are 'B' for the third level—indicating drugs used to lower glucose—and 'A' for the fourth level—indicating biguanides. Finally, the fifth level indicates the chemical substance, which for metformin is '02'. Therefore, the whole ATC code for metformin is 'A-10-B-A-02'.

Cheng *et al.* proposed a drug therapeutic similarity for calculating drug similarity using the ATC code [29]. The i-th level drug similarity (DS_i) between two drugs d_p and d_q is defined as:

$$DS_i(d_p, d_q) = \frac{ATC_i(d_p) \cap ATC_i(d_q)}{ATC_i(d_p) \cup ATC_i(d_q)} \quad (3)$$

where $ATC_i(d_p)$ and $ATC_i(d_q)$ represent all ATC codes at the i-th level of drug d_p and drug d_q, respectively. $ATC_i(d_p) \cap ATC_i(d_q)$ denotes the number of identical ATC codes of d_p and d_q at all levels up to the i-th level. $ATC_i(d_p) \cup ATC_i(d_q)$ denotes the number of unique ATC codes of d_p and d_q at all levels up to the i-th level. The similarity between the two drugs d_p and d_q is then averaged across all levels as follows:

$$Similarity_{ATC}(d_p, d_q) = \frac{\sum_{i=1}^{5} DS_i(d_p, d_q)}{5} \quad (4)$$

For example, for a drug pair with ATC code 'A-10-B-A-02' (metformin) and 'A-10-B-F-01' (acarbose), the first-, second- and third-level codes are the same, but the fourth and fifth codes are different. The ATC similarity is calculated as shown below:

$$Similarity_{ATC}(\text{Metformin}, \text{Acarbose}) = \frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{3}{5} + \frac{3}{7}}{5} = 0.805$$

Table 2. A comparison of the correlations between the proposed drug target similarity and the ATC code-derived similarity

Sematic similarity	Pearson correlation	Spearman correlation
Jaccard index	0.9311	0.9
Dice coefficient	0.8625	0.7143
Cosine similarity	0.8577	0.8857
Overlap coefficient	0.5707	0.7364

Table 2 shows a comparison of the correlations between the proposed drug target similarity and the ATC code-derived similarity. Several semantic similarities such as the Jaccard index, the Dice coefficient, the cosine similarity and the overlap coefficient were taken into account to calculate the similarity between two drugs. Numbers obtained using the Jaccard index are strongly positively correlated with ATC similarity (Pearson correlation of 0.9311 and Spearman correlation of 0.9). Furthermore, we tested the hypothesis that the shared target genes of a pair of drugs might significantly co-express under the conditions of treating the same disease cell line with these two drugs. As shown in Supplementary Table S1, the higher Jaccard index (>0.5) a pair of drugs reveals the more significant correlation of gene expression signature and structure similarity these two drugs manifest in treating the same particular disease. Therefore, the method using the Jaccard index could be suitable for use in our drug vector space to calculate drug similarity.

Drug repurposing evaluation

If a new indication for a drug is identified using our method and is described in a later article (chronologically), it implies that the indirect relationship has been validated and our method was successful. The MEDLINE literature was split into two parts based on the publication year: articles published before the cut-off point and articles published after the cut-off point. The first part was used as the analysis set for finding indirect relationships using our approach, while the second part was used as a validation set to assess the indirect relationships discovered using the first set. Direct disease–drug relationships that were present in the validation set but not in the analysis set were treated as the candidates, which could be located by a successful system. For example, as illustrated in Figure 2, the idea of using magnesium to treat migraine was mentioned in 1990 and not before. If an indirect relationship between migraine and magnesium was found in the analysis set, this would show the effectiveness of our methodology.

We used mean average precision (MAP) to evaluate our system in drug repurposing ranking. The time frame of the validation set was set from 2008 to 2014 for our experiment because 6.5 years were suggested to be a suitable duration for evaluating whether a repurposed indication is accurate [30].

Equation (5) was used to determine average precision, where k is the rank in the sequence of repurposed drugs, n is the number of repurposed drugs and $P(k)$ is the precision at position k in the list; $rel(k)$ is equal to 1 if a direct disease–drug relationship of the repurposed drug at position k is found in the validation set, and 0 otherwise. The MAP (6) for a set of diseases D is the mean of the average precision scores.

$$\text{AveP} = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{n} \quad (5)$$

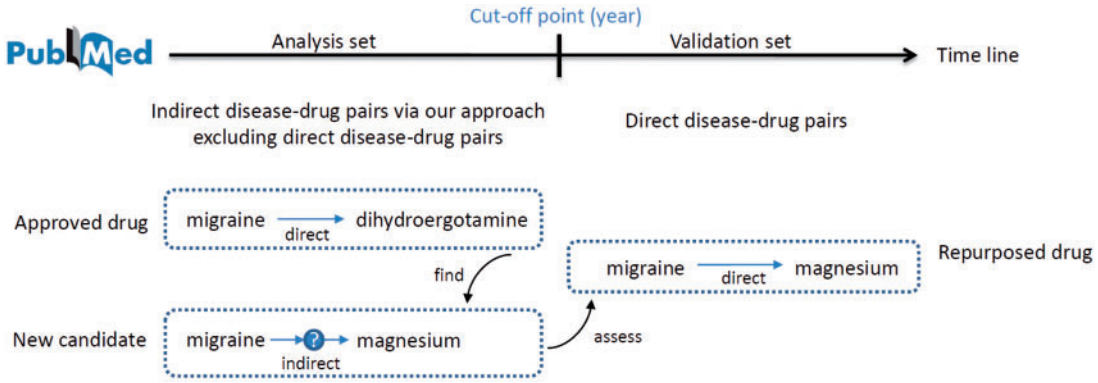

Figure 2. Overall design of method for evaluating our drug repurposing assessment.

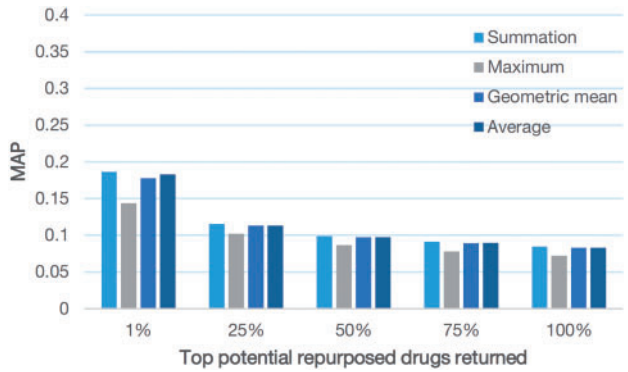

Figure 3. MAP score calculated using four different methods across different fractions of 790 diseases. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

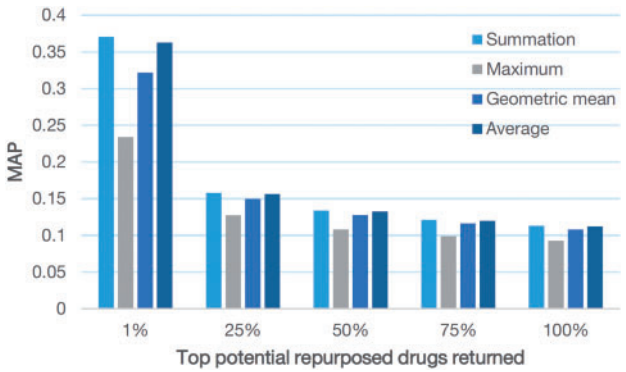

Figure 4. MAP score calculated using four different methods across different fractions of 100 diseases. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

$$\mathrm{MAP} = \frac{\sum_{d=1}^{|D|} AveP(d)}{|D|} \quad (6)$$

We identified 790 disease names in the target documents and calculated the MAP scores of the top 1%, 25%, 50%, 75% and 100% potential repurposed drugs (ranked by the *RDS*) for each disease. As shown in Figure 3, we obtained the most accurate ranking results (MAP of 0.1864) when using the summation method that combines all the diseases in the top 1%. When we reduced the 790 diseases to the 100 most frequently mentioned diseases, the MAP score of top 1% potential repurposed drugs using the summation method increased from 0.1864 to 0.3706 as shown in Figure 4.

Discussion

We have evaluated our system by using the MEDLINE database to select five different cancer diseases as examples. A drug treemap was prepared for selecting various repurposed drugs of a specific disease as illustrated in Figure 5. The candidates for drug repurposing identified in our study include *celecoxib* for treating ovarian cancer (*RDS* = 0.9535) or breast cancer (*RDS* = 0.9709), *raloxifene* for treating prostate cancer (*RDS* = 0.9938), *erlotinib* for treating colorectal cancer (*RDS* = 0.9232) and *rapamycin* for treating leukaemia (*RDS* = 0.9613). Moreover, we have compiled the repurposed drugs (*RDS* > 0.8) of several diseases for further studies (the repurposed drug lists are available in the Supplementary Data). The lists described the repurposed drugs, which were considered as high-potential new drugs of different diseases and the original indications of those repurposed drugs. The repurposed drug lists curated by our work could facilitate the screening of more than a thousand potential drugs of a disease of interest.

Literature review

Table 3 shows the details of five indirect disease–drug relationships identified by our methodology. Evidence for the suitability of the identified repurposed drugs is discussed below.

Celecoxib is a selective cyclooxygenase-2 (COX-2) inhibitor. Kim *et al.* showed that combining celecoxib with paclitaxel might be an effective treatment for ovarian cancer [31] because celecoxib may regulate paclitaxel-induced apoptosis in ovarian cancer cell line OVCAR-3 via down-regulation of nuclear factor kappa B and Akt activation (PMID: 24520227). Taurin *et al.* demonstrated the potential for selective estrogen receptor modulators to be used for the treatment of castrate-resistant prostate cancer [32]. They used poly(styrene-co-maleic acid) micelles to encapsulate raloxifene and increase its efficiency (PMID: 24689036). Li *et al.* [33] indicated that co-administration of erlotinib and rapamycin—inhibitors of epidermal growth factor receptor and mammalian target of rapamycin, respectively—can inhibit the growth of colorectal carcinoma cells (PMID: 22552366). Li *et al.* [34] also demonstrated that rapamycin plus celecoxib could induce cell cycle arrest and apoptosis, and decrease the expressions of mammalian target of rapamycin, 4E-BP1 and p70S6K; this led to improved effectiveness against chronic myelogenous leukaemia cells (PMID: 24682932). Preclinical results published by Kumar *et al.* [35] indicated that

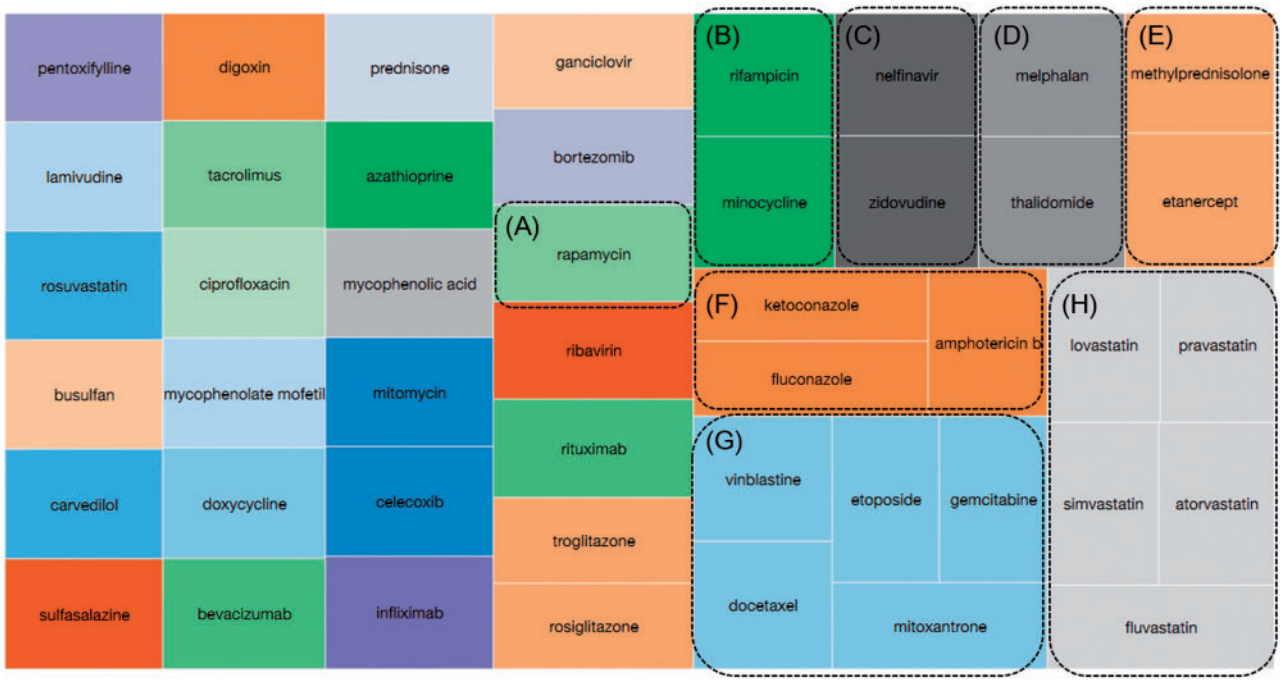

Figure 5. Drug Treemap for visualizing the repurposed drugs (RDS > 0.8) of leukaemia. The RDS of a repurposed drug determines the area size of a block. Drugs with the same original indication are grouped. The original indications (from TTD) of marked groups are (A) 'organ rejection'; (B) 'bacterial infections'; (C) 'HIV infection'; (D) 'multiple myeloma'; (E) 'rheumatoid arthritis'; (F) 'fungal infections'; (G) 'cancers'; (H) 'hypercholesterolemia'. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Table 3. The indirect relationships determined using the ABC model

Disease	Drug	Intermediate
Ovarian cancer	Celecoxib	Cyclooxygenase-2
Prostate cancer	Raloxifene	Estrogen receptor
Colorectal cancer	Erlotinib	Epidermal growth factor receptor
Leukaemia	Rapamycin	Mammalian target of rapamycin
Breast cancer	Celecoxib	Vascular endothelial growth factor

co-administration of tamoxifen with celecoxib is a potential treatment of breast cancer owing to its suppression of vascular endothelial growth factor and vascular endothelial growth factor receptor 2 expression (PMID: 23731702).

ClinicalTrials.gov review

ClinicalTrials.gov (https://clinicaltrials.gov/) is a large database that describes medical studies and clinical trials involving human volunteers. It is a useful resource because, before any trial, the use of the drugs has to be approved by the Food and Drug Administration (FDA). This gives clear indications of the safety of the drugs in question.

Clinical trials are classified into one of five categories by the FDA. Phase 0 is the first-in-human trial to explore if and how a new drug may work. Phase 1 is designed to evaluate the safety and side effects in a small group of 20–100 individuals. Phase 2 studies are used to gather preliminary data regarding the effectiveness and safety of the drug in a group of 100–300 participants. Phase 3 is designed to assess the effectiveness and safety in combination with other drugs in large groups of 1000–2000 patients. Phase 4, also known as post-marketing, occurs after FDA approval and is designed to gather extra information about a drug's safety and effectiveness in the long term.

Table 4. The suitability of our repurposed drug candidates verified using ClinicalTrials.gov

Identifier	Condition	Intervention	Phase
NCT01124435	Ovarian neoplasms	Celecoxib	2
NCT01050842	Prostate cancer	Raloxifene	0
NCT00116506	Colorectal cancer	Erlotinib	2
NCT00795886	Acute lymphoblastic leukaemia	Rapamycin	2
NCT01695226	Breast cancer	Celecoxib	2

Table 4 shows our repurposed drug candidates in the ClinicalTrials.gov database and confirms the suitability of our methodology. Celecoxib and carboplatin have been examined in heavily pre-treated patients with recurrent ovarian cancer (NCT01124435). An early phase 0 trial examines raloxifene in the management of castrate-resistant prostate cancer (NCT01050842). This trial is investigating the co-administration of bicalutamide and raloxifene to treat hormone-refractory prostate cancer. Erlotinib is another targeted agent that is under assessment as part of a multi-component treatment option for colorectal cancer (NCT00116506). Rapamycin is being investigated as a means of preventing graft versus host disease use following stem cell transplantation in patients with acute lymphoblastic leukaemia (NCT00795886). The COX-2 inhibitor celecoxib is being investigated as a potential breast cancer treatment in a randomized controlled phase 2 trial (NCT01695226).

Visualization of new indications with repurposed drugs

As thousands of repurposed drugs and corresponding new indications had been revealed by text mining in this study, a single plot is required to integrate and summarize all heterogeneous information, e.g. drugs, diseases and repurposed relationships. To facilitate

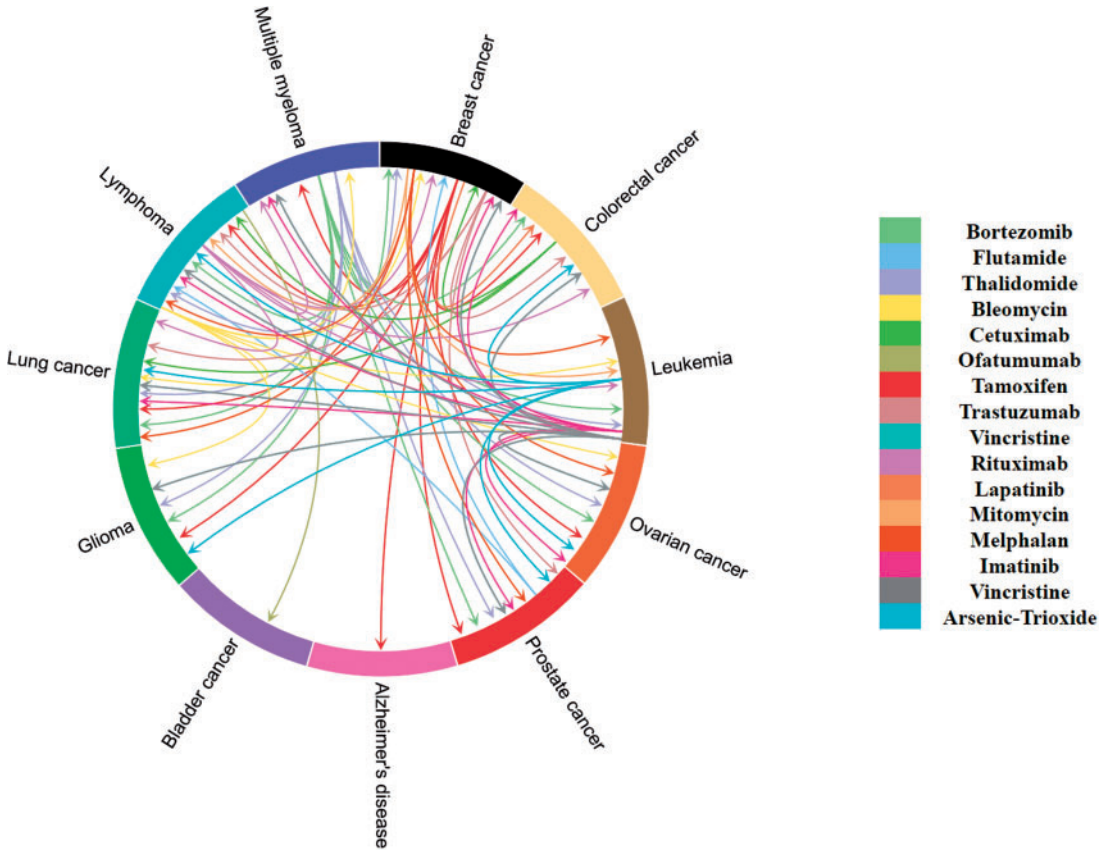

Figure 6. The wheel of drug repurposing driven by text-mining approach. A chord diagram was shown to illustrate the relationships between original and new indication(s) for 16 different repurposed drugs that only have been indicated for treating one disease. Eleven diseases with hundreds of repurposed drugs are enriched to be depicted in a single plot with a global visualization for drug repurposing. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

the discovery of new potential indications for old drugs by screening of thousands of repurposed among multiple diseases, we uniquely adopted a chord diagram (Figure 6) to visualize the both potential 16 drug candidates and 11 diseases (indications) information in the same figure that illustrates the repurposed drugs transiting from one original indication to new indication(s) (Supplementary data). Among these candidates shown in the Figure 6, the intended use of a drug in a particular cancer is most likely repositioned to treat another cancer. This is understandable, and consistent with that, most of the cancers possess common drug targets in the same molecular pathways of different tissues. Intriguingly, two disease–drug relationships were rapidly revealed with less connectivity that might indicate potential repurposed drugs for advanced literature reviews. One demonstrates that Lymphoepithelioma-like carcinoma of the urinary bladder, a type of rare bladder cancer, is infiltrated with abundant of lymphocytes by over-expressing CD20 [36], which is the drug target of Ofatumumab. The other is that Tamoxifen might slow cognitive decline in human studies [37] and prevent memory loss in amyloidosis mouse model [38], a typical animal model of Alzheimer's disease. These two aforementioned repurposed drug–new indication relationships might show novel insights, and be worthy to be further validated by biological experiments.

Miscellaneous

We wanted to determine potential repurposed drugs for specific diseases. Each indirect relationship consists of a disease–gene and gene–drug relationship. We presented the results of five candidates for drug repurposing—celecoxib (ovarian and breast cancer), raloxifene (prostate cancer), erlotinib (colorectal cancer) and rapamycin (leukaemia)—and directly validated their suitability using literature and clinical trial data. Some results were consistent with the results obtained by Gupta *et al.* in an earlier work [39]. In short, our system may hasten the identification of candidates for drug repurposing.

Our system has room for improvement since the success of the system is highly dependent on the lexicon being sufficient. In future work, we will improve named entity recognition and mention normalization algorithm of biomedical entities. Additionally, because this work ignored aliases, important disease–drug relationships may have been missed. For example, acetylcholinesterase (AChE) is an important target for Alzheimer's disease treatment. However, as our lexicon does not include the synonym AChE, our rankings will be adversely affected. Therefore, a comprehensive lexicon is required for higher precision. Many tools for the normalization of gene and disease targets have been developed [40–42] and using these would remove issues related to named entity synonyms thereby making our drug space model more effective.

Different types of intermediate terms lead to different results in drug repurposing, and our work used target and biomarker as the intermediate term. Target identification is important in drug repurposing and determines whether a drug may have a new indication. Biomarkers are often measured and evaluated to examine pharmacologic responses to a therapeutic intervention. However, many intermediate terms can be used for assessing candidacy for drug repurposing. Wu *et al.* used

shared genes, shared biological processes, shared pathways and shared phenotypes as features to establish indirect relationships [43]; we could also adopt such approaches in our system.

Additionally, our study does not consider the possibility of negative relationships that could lead to false positives. For example, nicotine places in the top 30 results for relationships to lung cancer; however, it is obviously a cause of, and not a solution to, lung cancer [44]. Several studies have investigated the detection of negative regulation events in the literature [45–47]. In the future, we could use machine learning-based and rule-based methods that differentiate trigger words and related treatment words to exclude false positives.

Conclusion

This study aimed to use text-mining technology to identify potential candidates for drug repurposing in the large-scale biomedical literature. To achieve our aim, we used a PAS pattern and learning of predicates using CTD-curated literature to extract relationships. This method had a precision score of 0.8674. We also proposed a similarity-based ranking method wherein each indirect relationship was ranked using a drug similarity assessment based on the ABC model.

Text mining is useful for investigating drug repurposing. We can investigate the growth of a repurposed drug over time and this, along with newly curated relationships or approved drugs, can be used to improve the effectiveness of our text mining strategies.

Key Points

- We developed a new text-mining approach that encompasses a pattern-based relationship extraction method and a gene-shared ranking method for measuring drug similarity. To confirm the robustness and accuracy of our approach, these two methods were further evaluated using CTD database and ACT code, respectively.
- The indirect relationships between drugs and new indications, not co-occurred in the same paper but revealed by ABC model, were of particular interest and inferred as new candidates for drug repurposing and subsequently validated as direct relationships by chronologically evaluating later published papers.
- The suitability of new candidates (repurposed drugs) were scientifically reviewed according to PubMed literature and resources of clinicaltrial.gov. These huge amount of evidenced-based repurposed drugs with their corresponding new indications was comprehensively integrated and summarized by a single plot, i.e. chord diagram, which was first used to demonstrate the wheel of drug repurposing driven by text-mining approach.

Supplementary Data

Supplementary data are available at http://tinyurl.com/phapjcd.

Acknowledgements

We would like to thank Dr Brady Bernard for his valuable comments and Dr Jang-Yang Chang for his helpful discussion of the manual review.

Funding

This work was supported by the Ministry of Science and Technology of Taiwan (MOST103-2221-E-006-254-MY2 and MOST 104-2923-E-006-003-MY3).

References

1. Dimasi JA. Risks in new drug development: approval success rates for investigational drugs. *Clin Pharmacol Ther* 2001;**69**:297–307.
2. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really $802 million? *Health Aff* 2006;**25**:420–8.
3. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**:673–83.
4. Wolff T, Miller T, Ko S. Aspirin for the primary prevention of cardiovascular events: an update of the evidence for the US preventive services task force. *Ann Intern Med* 2009;**150**:405–72.
5. Algra AM, Rothwell PM. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *Lancet Oncol* 2012;**13**:518–27.
6. Goldstein I, Lue TF, Padma-Nathan H, *et al*. Oral sildenafil in the treatment of erectile dysfunction. Sildenafil Study Group. *N Engl J Med* 1998;**338**:1397–404.
7. Calabrese L, Fleischer AB. Thalidomide: current and potential clinical applications. *Am J Med* 2000;**108**:487–95.
8. Palumbo A, Facon T, Sonneveld P, *et al*. Thalidomide for treatment of multiple myeloma: 10 years later. *Blood* 2008;**111**:3968–77.
9. Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**:D109–14.
10. Whirl-Carrillo M, McDonagh EM, Hebert JM, *et al*. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012;**92**:414–17.
11. Zhu F, Shi Z, Qin C, *et al*. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;**40**:D1128–36.
12. Davis AP, Murphy CG, Johnson R, *et al*. The comparative toxicogenomics database: update 2013. *Nucleic Acids Res* 2013;**41**:D1104–14.
13. Andronis C, Sharma A, Virvilis V, *et al*. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform* 2011;**12**:357–68.
14. Tari L, Vo N, Liang S, *et al*. Identifying novel drug indications through automated reasoning. *PLoS One* 2012;**7**:e40946.
15. Tari LB, Patel JH. Systematic drug repurposing through text mining. *Methods Mol Biol* 2014;**1159**:253–67.
16. Chen ES, Hripcsak G, Xu H, *et al*. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* 2008;**15**:87–98.
17. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. *BMC Bioinformatics* 2013;**14**:181.
18. Debusmann R, Kuhlmann M. Dependency grammar: Classification and exploration. Resource-adaptive cognitive processes. Berlin: Springer, 2010, 365–88.
19. De Marneffe M-C, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In: *Proceedings of LREC*. 2006, Genoa, Italy, pp. 449–54.

20. Culotta A, Sorensen J. Dependency tree kernels for relation extraction. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, Barcelona, Spain, pp. 423.
21. Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, Vancouver, Canada, pp. 724–31.
22. Surdeanu M, Harabagiu S, Williams J, *et al*. Using predicate-argument structures for information extraction. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*-Volume 1. Association for Computational Linguistics, 2003, Sapporo, Japan, pp. 8–15.
23. Chiang JH, Liu HH, Huang YT. Condensing biomedical journal texts through paragraph ranking. *Bioinformatics* 2011;**27**:1143–9.
24. Chowdhury FM, Lavelli A, Moschitti A. A study on dependency tree kernels for automatic extraction of protein-protein interaction. In: *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, 2011, Portland, Oregon, USA, pp. 124–133.
25. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**:7–18.
26. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* 1988;**31**:526–57.
27. Iorio F, Bosotti R, Scacheri E, *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;**107**:14621–26.
28. WHO. *Anatomical Therapeutic Chemical (ATC) Classification Index with Defined Daily Doses (DDDs)*. Oslo: WHO Collaborating Centre for Drug Statistics Methodology, 2000.
29. Cheng F, Li W, Wu Z, *et al*. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J Chem Inf Model* 2013;**53**:753–62.
30. Frijters R, van Vugt M, Smeets R, *et al*. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010;**6**:e1000943.
31. Kim HJ, Yim GW, Nam EJ, *et al*. Synergistic effect of COX-2 inhibitor on paclitaxel-induced apoptosis in the human ovarian cancer cell line OVCAR-3. *Cancer Res Treat* 2014;**46**:81–92.
32. Taurin S, Nehoff H, van Aswegen T, *et al*. A novel role for raloxifene nanomicelles in management of castrate resistant prostate cancer. *Biomed Res Int* 2014;**2014**;323594.
33. Li B, Gao S, Wei F, *et al*. Simultaneous targeting of EGFR and mTOR inhibits the growth of colorectal carcinoma cells. *Oncol Rep* 2012;**28**:15–20.
34. Li J, Xue LY, Hao HL, *et al*. Rapamycin combined with celecoxib enhanced antitumor effects of mono treatment on chronic myelogenous leukemia cells through downregulating mTOR pathway. *Tumor Biol* 2014;**35**:6467–74.
35. Kumar BN, Rajput S, Dey KK, *et al*. Celecoxib alleviates tamoxifen-instigated angiogenic effects by ROS-dependent VEGF/VEGFR2 autocrine signaling. *BMC Cancer* 2013;**13**:273.
36. Yoshino T, Ohara S, Moriyama H. Lymphoepithelioma-like carcinoma of the urinary bladder: a case report and review of the literature. *BMC Res Notes* 2014;**7**:779.
37. Legault C, Maki PM, Resnick SM, *et al*. Effects of tamoxifen and raloxifene on memory and other cognitive abilities: cognition in the study of tamoxifen and raloxifene. *J Clin Oncol* 2009;**27**:5144–52.
38. Pandey D, Banerjee S, Basu M, *et al*. Memory enhancement by tamoxifen on amyloidosis mouse model. *Horm Behav* 2016;**79**:70–3.
39. Gupta SC, Sung B, Prasad S, *et al*. Cancer drug discovery by repurposing: teaching new tricks to old dogs. *Trends Pharmacol Sci* 2013;**34**:508–517.
40. Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. *Bioinformatics* 2011;**27**:1032–33.
41. Wei CH, Kao HY. Cross-species gene normalization by species inference. *BMC bioinformatics* 2011;**12** (Suppl 8):S5.
42. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 2013;**29**:2909–2917.
43. Wu C, Gudivada RC, Aronow BJ, *et al*. Computational drug repositioning through heterogeneous network clustering. *BMC Systems Biology* 2013;**7**:S6.
44. Grando SA. Connections of nicotine to cancer. *Nat Rev Cancer* 2014;**14**:419–429.
45. Kim J-D, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Association for Computational Linguistics, 2009, Boulder, Colorado, pp. 1–9.
46. Sarafraz F, Nenadic G. Using SVMs with the command relation features to identify negated events in biomedical literature. In: *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*. Association for Computational Linguistics, 2010, Uppsala, Sweden, pp. 78–85.
47. Cruz Díaz NP, Maña López MJ, Vázquez JM, *et al*. A machine-learning approach to negation and speculation detection in clinical texts. *J Am Soc Inf Sci Technol* 2012;**63**:1398–1410.