Title: Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome

Authors: Liang Yu, Jin Zhao, Lin Gao

PII: S0933-3657(16)30565-6
DOI: http://dx.doi.org/doi:10.1016/j.artmed.2017.03.009
Reference: ARTMED 1515

To appear in: *ARTMED*

Received date: 17-12-2016
Revised date: 6-1-2017
Accepted date: 17-3-2017

Please cite this article as: Yu Liang, Zhao Jin, Gao Lin.Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome.*Artificial Intelligence in Medicine* http://dx.doi.org/10.1016/j.artmed.2017.03.009

Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome

Liang Yu*

School of Computer Science and Technology, Xidian University, Xi'an, 710071, P.R.China

lyu@xidian.edu.cn

*Corresponding author

Jin Zhao

School of Computer Science and Technology, Xidian University, Xi'an, 710071, P.R.China

zhaojing-2108@qq.com

Lin Gao

School of Computer Science and Technology, Xidian University, Xi'an, 710071, P.R.China

lgao@mail.xidian.edu.cn

There are mainly three highlights in our study:

- 1. Considering the tissue specificities of diseases, the incompleteness of data and the triangularly balanced structure between drugs and diseases, we proposed a novel algorithm to predict drug indications for a specific disease.
- 2. Taking breast cancer and hepatocellular carcinoma (HCC) as case studies, 96.9% and 90.3% results in the top-5% predicted associations match with known associations in Comparative Toxicogenomics Database (CTD) for breast cancer and hepatocellular carcinoma respectively.
- 3. We find inspiring results: D05589 (Pralatrexate) may be a potential treatment for breast cancer; D06304 (Vindesine) and D09755 (Cabazitaxel) may be treat HCC; D05589 (Pralatrexate) may be cause HCC. The results are validated by CTD database, literature mining, KEGG pathway enrichment analysis and Clinical verification.

Abstract

Finding new uses for existing drugs has become a new strategy for decades to treat more patients. Few traditional approaches consider the tissue specificities of diseases. Moreover, disease genes, drug targets and protein interaction (PPI) networks remain largely incomplete and the relationships between drugs and diseases conform to the triangularly balanced structure. Therefore, based on tissue specificities of diseases, we apply the triangularly balanced theory and the module distance defined for incomplete interaction networks to build drug-disease associations. Our method is named as TTMD (**T**issue specificity, **T**riangle balance theory and **M**odule **D**istance). Firstly, we combine three different drug similarity networks. Then, in the tissue-specific PPI network of a disease, we calculate its similarities with drugs using module distance. Finally, breast cancer and hepatocellular carcinoma (HCC) are taken as case studies. In the top-5% of predicted associations, 96.9% and 90.3% results match with known associations in Comparative Toxicogenomics Database (CTD) for breast cancer and hepatocellular carcinoma respectively. Clinical verification, literature mining and KEGG pathways enrichment analysis are further conducted for the top-5% newly predicted associations. Overall, TTMD is an effective approach for predicting new drug indications for tissue-specific diseases and provides potential values for the treatments of complex diseases.

Keywords: drug repositioning; triangularly balanced structure; tissue specificity; module distance

Introduction

Complex diseases, such as cancers, diabetes mellitus and cardiovascular disease, are caused by a combination of genetic and environmental factors[1]. By conservative

estimates, it now takes over 15 years[2] and $800 million to $1 billion to bring a new drug to market[3]. There is a pressing need to find effective drugs to treat complex diseases with limited cost and time. Drug repositioning[4] is a strategy to identify new therapeutic applications for existing drugs. It is also known as drug reprofiling, drug redirecting, drug re-tasking and therapeutic switching[5]. However, the new drug indications were found early by accident, such as the use of zinc acetate for Wilson's disease[6], arsenic for acute promyelocytic leukemia[7] and amphotericin B for leishmaniasis[8]. Therefore, it is necessary to propose new computational approaches to predict drug-disease associations.

With the generation of large scale biological data, such as gene expression data, protein interaction data, drug and disease phenotypic data, the number of computational methods for drug repositioning has observably increased. They can be classified into network-based[9,10,11], knowledge-based[12,13], compound structure-based[14,15], side effect-based[16,17,18], signature-based[19,20,21] and target-based[22,23]. These methods focus on different fields and enable researchers to examine potential drug candidates and test on related diseases within significantly shortened time lines and money. Target-based methods are more accurate within the above drug-repositioning methods because most targets linked directly with mechanisms of diseases. For example, Li et al.[24] developed a bipartite drug-target relationship network method to identify new indications of existing drugs through its similar drugs. Wu et al.[25] applied a network clustering algorithm to a drug-disease bipartite network to find closely modules of diseases and drugs, which is used to find potential drug-disease associations. However, it is limited to construct associations between drugs and diseases based on common genes because the incompleteness of data result in many drugs and diseases having few or no known related genes. In 2016, Würth R et al.[26] focused on known non-oncological drugs with new therapeutic applications in oncology to review. This kind of method moved from incompleteness of the knowledge of drug-target interactions and put emphasis on the opportunities for repurposing compounds as cancer therapeutics.

Researchers find that understanding the genetic complex tissues and individual cell-lines are important for developing improved diagnostics and therapeutics. In 2015, Chen B et al.[27] compared the gene expression profiles of 200 HCC tumor samples from The Cancer Genome Atlas (TCGA)[28] and over 1000 cancer cell lines including 25 HCC cancer cell lines from Cancer Cell Line Encyclopedia (CCLE)[29]. They discovered that HCC tumor samples are closely correlated with HCC cell lines. Kosti I et al.[30] analyzed 16,561 genes and the corresponding proteins in 14 tissue types across 200 samples. They found genes and proteins were highly related with cancer tissues compared to normal tissues. Guan Y et al.[31] constructed 107 tissue-specific functional relationship networks through integrating genomic data and tissue-specific gene expression patterns. They applied these tissue-specific networks to predict

phenotypes related to genes and the improved performance demonstrated the importance of tissue specificity. However, few algorithms for predicting new indications of drugs have considered the tissue specificities of diseases and the relationships between drugs and diseases conform to the theory of triangular structural balance[32], i.e. similar drugs may treat similar diseases, and vice versa. Furthermore, high-throughput approaches include less than 20% of all potential protein-protein interactions in the human cell at present[33,34,35], which indicates that we discover drug and disease relationships depending on interactome maps that are 80% incomplete. Also, the knowledge of disease- and drug-associated genes is limited.

Therefore, in this study, we propose a new approach TTMD to predict diseases' potential drugs based on its tissue-specific information and triangular structure balance theory[32] by using module distance[36] in incomplete human interactome. Its framework is shown in Figure 1. First, based on three related data of drugs: chemical structures, gene expressions and ATC codes, we respectively construct CS (Chemical Structure network), GE (Gene Expression network) and ATC (ATC code network) (see part (A) in Figure 1). Second, after filtering possible false positive edges in the above three networks, we further integrate them and get a high quality network, CGA (see part (B) in Figure 1). Third, in the tissue-specific PPI network of disease *D* (see part (C) in Figure 1), we calculate the direct association between disease *D* and drug *d* based on their related gene sets using module distance. And, based on CGA (see part (B) in Figure 1), we calculate the indirect association between disease *D* and drug *d* through *d*'s direct neighbors. The sum of direct and indirect associations is their final association

of disease *D* and drug *d*. Finally, we can get all the final associations between disease *D* and all the drugs in CGA.

Data and Method

Data

Drug-target data: Drugs of human and their corresponding targets are downloaded from KEGG database[37,38] and DrugBank[39]. We merge the two datasets and get 3,885 drugs, 3,906 targets and 14,461 drug-target pairs.

Disease-gene data: The related genes of diseases are downloaded from KEGG database. We take breast cancer and hepatocellular carcinoma (HCC) as tissue-specific disease cases to assess our method.

Tissue-specific PPI Interaction network: We download tissue-specific PPI networks marked as "Top Edges" from GIANT (Genome-scale Integrated Analysis of

gene Networks in Tissues) database[40] (http://giant.princeton.edu/) (2015 version). "Top Edges" represents that the network is filtered to only include edges with evidence supporting a tissue-specific functional interaction. GIANT proposes a tissue-specific benchmark to automatically up-weight datasets relevant to a tissue from a large data of different tissues and cell-types. As cases study, we download the liver tissue-specific network containing 10,328 genes and 785,531 edges, and the mammary gland tissue-specific network containing 15,269 genes and 883,071 edges.

Tissue-specific networks are weighted and the weights on the edges are proportional to the relationships between nodes. In order to apply module distance[36] to calculate the distances between drugs and diseases in incomplete tissue-specific interaction networks, we process the original weights in the networks by formula (1):

$$w' = e^{-w^2} \tag{1}$$

Benchmark of drug-disease associations: All the known associations between drugs and diseases or its descendants are got from Comparative Toxicogenomics Database (CTD) in November 2016 as our benchmark[41]. CTD contains two kinds of chemical–disease associations: curated and inferred. Curated associations are extracted from the published literatures by CTD biocurators and inferred associations are established via CTD–curated chemical–gene interactions. In our study, we extract both curated and inferred associations, which can help researchers develop hypotheses about environmental diseases and their underlying mechanisms.

Method

Triangularly balanced structures

In 1946, Heider[42], who proposed the general field-theoretical approach, considered certain aspects of cognitive fields which contain perceived people and impersonal objects or events. His basic hypothesis affirms that there is a tendency for cognitive units to achieve a balanced state. Figure 2A to 2D represent four basic relationships among three points (marked as d1, d2 and d3). Their connections are labeled by "+1" or "-1" to represent positive or negative correlation. In graph Figure 2A and 2C, the cycles d1d2d3d1 are labeled by (+1, +1, +1) and (-1, +1, -1), respectively. Heider took the two situations as balanced states and in this paper, they are named as triangularly balanced structures.

For the prediction of relationships between drugs and diseases, the assumption is similar drugs may treat same or similar diseases, that is to say, they are all positive correlations. Therefore, the assumption is in accordance with the triangularly balanced structure shown in Figure 2A. We rewrite them in Figure 2E, where d1 and d2

represent two drugs, and D represents a disease. If two positive edges indicate d1 is similar to d2 and d1 can treat D (represented by blue lines in Figure 2E), we can predict the remaining edge (d2, D) is likely to be a new drug-disease relationship (represented by purple line in Figure 2E).

In this paper, we do not consider the triangularly balanced structure shown in Figure 2C because it is considered unlikely to conform to the assumptions of drug repositioning. For example, if drug d1 is not similar to drug d2 and d1 cannot treat disease D, it is hard to say that drug d2 may treat disease D.

Construct three drug similarity networks

In this section, based on chemical structures, related gene expressions and ATC codes of drugs, we construct three drug similarity networks: **c**hemical **s**tructure similarity network (CS), **g**ene **e**xpression similarity network (GE) and **ATC** code similarity network (ATC). The drug chemical structure similarity can be calculated by SIMCOM[43] based on the information of chemical structure of drugs from DRUG and COMPOUND sections in the KEGG LIGAND database[44]. The gene expression data related to drugs is got from Connectivity Map[45,46] (CMap) database. The ATC code similarity between drugs is calculated by a probabilistic model[47]. ATC code is a system of alphanumeric codes developed by the WHO (World Health Organization) for the classification of drugs and other medical products[48]. The similarity score between two ATC codes is calculated by their prior probability (frequency) and the probability of their longest common prefix[49]:

$$S(i,j)=\frac{2*\log(\Pr(pre(i,j)))}{\log(\Pr(i))+\log(\Pr(j))} \quad (2)$$

where $pre(i,j)$ is the longest common prefix of ATC code i and j; $\Pr(i)$ is the frequency of ATC code i in all ATC codes. Because some drugs may have more than one ATC code, we define the maximum ATC code similarity as the similarity (AS) between two drugs, $d1$ and $d2$:

$$\mathrm{AS}(d_1,d_2)=\underset{i\in \mathrm{ATC}(d_1),\, j\in \mathrm{ATC}(d_2)}{Max}(S(i,j)) \quad (3)$$

where $\mathrm{ATC}(d)$ represents all the ATC codes belonging to drug d. In this way, we can get an ATC code similarity network (ATC).

Integrate three drug similarity networks

Filter three drug similarity networks with proper thresholds

In order to obtain a reliably integrated drug similarity network, we first filter the above three drug networks before integrating them. It is very important to select proper thresholds for network filtering. It is hoped that drug pairs with shared targets are more similar to each other[49]. Therefore, we investigate the enrichment of drug pairs with common targets on the above three drug similarity networks. Here, we take GE network as an example, which includes 563 nodes (drugs) and 158,203 weighted edges (drug pairs). In GE network, we get 6,983 drug pairs with common targets and calculate its proportion (P_o) in the total number of edge, i.e. $P_o = 6983/158203 = 0.0441$. Then, we try to choose a proper threshold to discard false positive drug pairs. The potential threshold is increased from 0.1 to 0.95 with step 0.05 (see Figure 3). For each value, we preserve drug pairs with weights higher than the threshold, and compute the proportion (marked as P_n) of drug pairs with common targets in all the preserved drug pairs. Then, we calculate the fold enrichment score, which is defined as P_n/P_o [49]. For example, when the threshold is set to be 0.85 in GE network, P_n is 0.234 and the fold enrichment score is $0.234/0.0441 = 5.3$. Higher P_n/P_o represents more reliable drug pairs[49]. All the fold enrichment scores of three drug networks and a random work are shown in Figure 3.

From Figure 3, we can find the change of threshold has little impact on random network (RN) and for our three networks, with the increase of threshold, the fold enrichment value increases. However, it is clear that as the threshold increases, the number of edges decreases. Less edges and nodes contain less information. Therefore, for the three networks: CS, GE and ATC, the selected thresholds are 0.35, 0.75 and 0.75, respectively. We will verify the effectiveness of the three thresholds in the following section.

Integrate three filtered drug similarity networks

From the above section, we know the threshold of CS is 0.35, which is different from 0.75 of the other two networks. Thus, we firstly normalize the weights of edges in CS into the range of 0.75 to 1.0. Then we integrate the three drug similarity networks into a more reliable one CGA. To reduce the bias of each similarity measurement and standardize the overlapped drug-drug relationships, we use a weighted combination approach to integrate the similarities. For each of the drug-drug pairs, the integrated similarity S_N is defined as equation (4):

$$S_N = (\alpha_1 S_C + \alpha_2 S_G + \alpha_3 S_A) / \sum_{j=1}^{3} \alpha_j \qquad (4)$$

where S_C, S_G and S_A indicate the corresponding similarities in CS, GE and ATC, respectively. Here, we set $\alpha_j = 1 (j = 1,2,3)$ if its corresponding S_C, S_G or S_A exists in networks, or $\alpha_j = 0 (j = 1,2,3)$. Finally, we get an integrated drug similarity network CGA with high quality.

Construct drug-disease associations based on module distance and triangularly balanced structure

Because disease genes, drug targets and protein interaction (PPI) networks remain largely incomplete, Menche et al.[36] proposed a new definition for calculating distance between two modules based on the shortest path in incomplete networks. Figure 4 gives an example to show the calculation process in a weighted PPI network. In Figure 4, disease *A* has four related genes, marked as *a*, *b*, *c* and *d*, and drug *B* has five targets, marked as *c*, *e*, *f*, *g* and *h*. Considering the node *a*, its distance to targets {*c*, *e*, *f*, *g*, *h*} of drug *B* are 0.8, 1.0, 1.1, 1.5 and 1.9 respectively, so its shortest distance to drug *B* is 0.8. In this way, we can obtain the distances between each node in gene set {*a*, *b*, *c*, *d*} of disease *A* and drug *B*, and the distances between each node in target set {*c*, *e*, *f*, *g*, *h*} of drug *B* and disease *A*, shown in Figure 4. Finally, the distance between disease *A* and drug *B*, $d_{A,B}'$, is equals to the sum of all the distances divided by the total number of nodes related to disease *A* and drug *B*.

In order to make the distances be proportional to drug-disease direct correlations, we use formula (5) to normalize the distances:

$$d_{A,B} = \frac{Max_d - d_{A,B}'}{Max_d - Min_d} \qquad (5)$$

where Max_d and Min_d represent the maximum and the minimum of all the drug-disease distances, respectively; $d_{A,B}'$ represents the distance between disease *A* and drug *B*; $d_{A,B}$ represents the direct association between disease *A* and drug *B*.

According to the triangularly balanced structure shown in Figure 2E, we know the relationships between drugs and diseases also rely on the indirect association supplied by drugs' direct neighbors. Therefore, we calculate the relationships between drugs and diseases from two aspects: the direct association and the indirect association.

First of all, the genes related to drugs and disease *D* are mapped to the tissue-specific PPI network of disease *D*. Using formula (5), we can calculate the direct associations between each drug and disease *D*.

In the next step, we map 1,338 approved drugs got from DrugBank database to the integrated CGA and take these drugs as seeds. For each seed drug *d*, we combine their direct and indirect associations with disease *D* by formula (6).

$$Similarity_{d,D} = mean(\sum_{i=1}^{n}(d_{D,d_i} + S_{d,d_i}) + \beta \times (d_{D,d} + S_{d,d})) \qquad (6)$$

Where $Similarity_{d,D}$ represents the final correlation between drug *d* and disease *D*; d_i represents the *i*-th neighbor of drug d in CGA; n represents the number of neighbors of drug d ; β is 1 if drug *d* have genes mapped to the PPI network, otherwise β is 0; $d_{D,d}$ represents the direct association between disease *D* and drug *d* calculated by formula (5); S_{d,d_i} represents the similarity between drug *d* and drug *di* in CGA; $S_{d,d}$ =1 represents the similarity between drug *d* and itself. Finally, we rank drugs by their correlations with disease *D* in descending order and the top drugs are very likely to have therapeutic relationship with disease *D*.

Results

Verify the chosen thresholds for three drug networks

In order to verify the rationality of selected thresholds, we predict drug-disease associations using different thresholds in CS, GE and ATC networks. We take GE network as an example and select breast cancer as our case here. In Figure 5, for each given threshold, the precision of our method is calculated by formula (7),

$$precision = \frac{P_{CTD}}{P} \qquad (7)$$

where *P* represents the number of predicted drug-disease pairs; *PCTD* represents the number of drug-disease pairs, which can be found in CTD[41] database with reference score over 20. Reference score represents the number of references that

mention the curated and inferred associations. Here, reference score is set to be over 20 because the change of reference score will not affect the trends of precision curve. Moreover, 20 is the mean of all reference scores. Figure 6 and 7 show the precision curves of CS and ATC network with different thresholds. As shown in Figure 5, 6 and 7, when to choose 0.35, 0.75 and 0.75 for CS, GE and ATC respectively, their precisions are better than other thresholds.

Choose breast cancer and hepatocellular carcinoma as cases

Breast cancer is an uncontrolled growth of breast cells. Worldwide, breast cancer is the leading type of cancer in women, accounting for 25% of all cases[50]. In 2012, it resulted in 1.68 million cases and 522,000 deaths[50]. Hepatocellular carcinoma (HCC) is a type of liver cancer, which is one of the most common malignant tumors with a high rate of morbidity and mortality. Therefore, in our study, we choose breast cancer and hepatocellular carcinoma as tissue-specific diseases. The drug-breast cancer associations and drug-hepatocellular carcinoma associations are ranked in descending order according to their scores (see Supplementary Tables S1 and S2). In Figure 8, we give the precision curves of predicted drug-breast cancer pairs and drug-hepatocellular carcinoma pairs at different top-x%. From the figure, we find the higher the associations ranking, the higher the accuracy. Hence, for the two tissue-specific diseases, we choose their top-5% for further analysis.

CTD benchmark verification and clinical evaluation

Case study: Breast cancer

The top 5% results including 32 drugs related to breast cancer are shown in Table 1. Firstly, we manually validate the 32 drugs by CTD database and find 20 (62.5%) of them are marked as “therapeutic (T)”, “marker/mechanism and therapeutic (M|T)” or “marker/mechanism (M)”, which means they are highly confident associations with breast cancer. Moreover, we find the other 11 drugs also have connections with breast cancer in CTD database with inference score[51] over 0 and they are marked as “Ref” in Table 1. The inference score[51] reflects the degree of similarity between CTD chemical-gene-disease networks and a similar scale-free random network, which is computed as shown below:

$$Y = -ln\left[P\left(G \text{ associated with both } C \text{ and } D \mid k, n_G\right) P\left(\text{no other } G \text{ connects } C \text{ and } D \mid k, n_G\right)\right] \quad (8)$$

where Y represents the inference score; P represents the probability that a vertex in a large network interacts with another vertex decays according to a power law[52]; G,

C, and *D* represent a gene, chemical, and disease respectively; *k* represents the number of connection between *G*, *C*, or *D*; *nG* represents a gene set. The higher the *Y* score, the more likely the inference network has atypical connectivity[53].

That is to say, there are 96.9% (31 of 32) drugs are recorded in CTD database and only one drug D05589 (drug name="Pralatrexate", marked as boldface in Table 1) cannot be found related with breast cancer in CTD database at present.

Therefore, we further make use of the ClinicalTrials.gov (https://clinicaltrials.gov/) for

Ranked by drug's similarity score. Direct Evidence has five values: M(marker/mechanism), T(therapeutic), M|T(marker/mechanism and therapeutic), Ref (inferred by genes) and None(no record in CTD database). Inference Score represents the score for the inference based on the topology of the network consisting of the chemical, disease, and one or more genes used to make the inference. Reference Score represents the number of reference(s) that mention(s) the curated and inferred associations.

verifying our prediction. ClinicalTrials.gov is a web-based resource that provides patients, their family members, health care professionals, researchers, and the public with easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions. Currently, it lists 230,894 studies with locations in all 50 states and in 193 countries (November 29, 2016). Through ClinicalTrials.gov, we find a clinical study of pralatrexate (D05589) in 22 female patients with previously-treated breast cancer. Its ClinicalTrials.gov identifier is NCT01118624. The purpose of this study is to determine the efficacy (ability to provide a beneficial treatment of the disease) of pralatrexate for the treatment of female patients with advanced or metastatic breast cancer who have failed prior chemotherapy. At the same time, patients will receive vitamin B12 and folic acid supplementation.

Case study: Hepatocellular carcinoma

Hepatocellular carcinoma is the most common type of liver cancer. Its top 5% related drugs are shown in Table 2. There are 31 drugs in all in the table and 9 of them are known as "therapeutic (T)", "marker/mechanism and therapeutic (M|T)" or "marker/mechanism(M)". In addition, the other 19 drugs are inferred by related genes in CTD database and they are marked as "Ref" in Table 2. In other words, 90.3% (28 of 31) predicted drugs can be found in CTD database with inference score over 0. Three drugs, D06304 (Vindesine), D05589 (Pralatrexate) and D09755 (Cabazitaxel) (marked as boldface in Table 2), currently have no records in CTD database. However, though D06304 (Vindesine) has no direct relationship with hepatocellular carcinoma in CTD database, it is related to liver neoplasms by BCL2 gene[54] and it results in increased

Ranked by drug's similarity score. Direct Evidence has five values: M(marker/mechanism), T(therapeutic), M|T(marker/mechanism and therapeutic), Ref (inferred by genes) and None(no record in CTD database). Inference Score represents the score for the inference based on the topology of the network consisting of the chemical, disease, and one or more genes used to make the inference. Reference Score represents the number of reference(s) that mention(s) the curated and inferred associations.

expression of BCL2 mRNA. BCL2 gene encodes an integral outer mitochondrial membrane protein that blocks the apoptotic death of some cells such as lymphocytes[55].

For D05589 (Pralatrexate), it has relationship with hepatocellular carcinoma inferred via CASP3 gene[56]. Pralatrexate results in increased activity of CASP3 protein, which has been found to be necessary for normal brain development as well as its typical role in apoptosis, where it is responsible for chromatin condensation and DNA fragmentation[57].

Based on ClinicalTrials.gov database, we find two studies about drug D09755 (Cabazitaxel) acting on hepatocellular carcinoma. One is a phase II study of cabazitaxel in patients with urothelial carcinoma who have disease progression following platinum-based chemotherapy (ClinicalTrials.gov Identifier: NCT01437488) and the other is phase I safety and pharmacokinetic study of XRP6258 (Cabazitaxel) in advanced solid tumor patients with varying degrees of hepatic impairment (ClinicalTrials.gov Identifier: NCT01140607).

KEGG pathway functional enrichment analysis and literature verification

In the above section, the top-5% results are validated by CTD benchmark and Clinical database. We mainly analyze two tissue-specific diseases: breast cancer and hepatocellular carcinoma. After our analysis, we obtain one potential drug D05589 (Pralatrexate) for breast cancer, and three potential drugs, D06304 (Vindesine), D05589 (Pralatrexate) and D09755 (Cabazitaxel), for hepatocellular carcinoma. By ClinicalTrials.gov database, we can verify that D05589 (Pralatrexate) is likely to treat breast cancer and D09755 (Cabazitaxel) is a possible treatment of hepatocellular carcinoma.

In this section, we will further perform KEGG pathway enrichment analysis on potential drugs and their related diseases. KEGG (http://www.kegg.jp/ or http://www.genome.jp/kegg/) is an encyclopedia of genes and genomes[58]. Its primary objective is assigning functional meanings to genes and genomes both at the molecular and higher levels. Hence, drugs or diseases can be connected to some

pathways through their related genes. If a drug has overlapped KEGG pathways with a disease, the drug and the disease may have great relevance. That is to say, the drug is likely to treat or cause the disease through acting on the overlapped pathways.

Case studies: breast cancer and hepatocellular carcinoma

We apply the functional annotation tool of DAVID[59,60] to carry out KEGG pathway enrichment analysis. DAVID provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to visualize genes on BioCarta & KEGG pathway maps, identify enriched biological themes, particularly GO terms, and so on. Therefore, we use DAVID to find overlapped KEGG pathways between potential drugs and two tissue-specific diseases. The *p*-value is set to be lower than 0.05.

We find D06304 (Vindesine) and D09755 (Cabazitaxel) both have two overlapped KEGG pathways with hepatocellular carcinoma, shown in Table 3. Specially, for vindesine (D06304), its two overlapped pathways have very small p-values: 5.01E-16 and 4.96E-18. Vindesine (D06304) is a medication used to treat a number of types of cancer including: Hodgkin's lymphoma, non-small cell lung cancer, bladder cancer, brain cancer, and testicular cancer among others[61]. “Hsa04540: Gap junction” is one pathway overlapped between vinblastine (D06304) and hepatocellular carcinoma. In fact, protein connexin 43 (Cx43), a part of intercellular gap junctions, is frequently down-regulated in tumors[62]. Studies have demonstrated that gap junctions (GJs) composed of connexin (Cx) proteins have the potential to modulate drug chemosensitivity in multiple tumor cells[63]. Furthermore, the combination of VV-SMAC and vinblastine (D06304) may provide a new avenue in treatment of HCC[64] and the temsirolimus/vinblastine combination induce a significant and sustained antitumor activity[65].

Cabazitaxel (D09755) is a semi-synthetic derivative of the natural taxoid 10-deacetylbaccatin III with potential antineoplastic activity (http://www.cancer.gov/drugdictionary/?CdrID=534131). It binds to and stabilizes tubulin, resulting in the inhibition of microtubule depolymerization and cell division, cell cycle arrest in the G2/M phase, and the inhibition of tumor cell proliferation[66]. It is a FDA approved drug for the treatment of hormone-refractory prostate cancer on June 17, 2010[67]. As shown in Table 3, KEGG enrichment analysis finds there are two overlapped pathways between cabazitaxel (D09755) and hepatocellular carcinoma: “hsa04540: Gap junction” and “hsa05130:Pathogenic Escherichia coli infection”. For the pathway “hsa05130: Pathogenic Escherichia coli infection”, Escherichia coli (E. coli) infection may be instrumental for the breakdown of

tolerance to mitochondrial and nuclear autoantigens in patients with primary biliary cirrhosis (PBC)[68,69]. And the recent development of credible animal models of PBC will give us the opportunity to investigate the role of E. coli infection as a pathogenic trigger of PBC[69]. PBC is a chronic cholestatic liver disease and usually consists of slow disease progression, which can result in liver cirrhosis, liver failure, and eventually require liver transplantation[70]. Hepatocellular carcinoma (HCC) occurs with increased frequency in patients with primary biliary cirrhosis (PBC)[71,72]. The exact frequency is unknown but is estimated to be between 0.7% and 16%[73,74].

As for pralatrexate (D05589), because it only has two targets: DHFR and TYMS and it has few related KEGG pathways, pralatrexate has no overlapped KEGG pathways with the two diseases at present. Pralatrexate is an anti-cancer medication[75]. It is the first drug approved as a treatment for patients with relapsed or refractory peripheral T-cell lymphoma. From the website of drugs.com (https://www.drugs.com/), we find that *in vitro*, pralatrexate is a substrate for the breast cancer resistance protein (BCRP), MRP2, multidrug resistance-associated protein 3 (MRP3), and organic anion transport protein 1B3 (OATP1B3) transporter systems at concentrations of pralatrexate that can be reasonably expected clinically. The relationship between pralatrexate and hepatocellular carcinoma may be negative because pralatrexate has been reported to increase certain liver enzymes, which could be a sign of liver damage[76] and it can cause hepatic toxicity and liver function test abnormalities[77].

All in all, the predicted drug-disease associations could provide valuable information for finding potential treatment of diseases.

Comparisons based on tissue-specific and general PPI networks

In order to illustrate the importance of tissue specificity for drug repositioning, we also perform our approach TTMD in a general PPI network without tissue-specific information and compare its results with ours. The general protein interaction network is got from ref 36, which integrates seven different interactions. It includes 13,460 human genes and 141,296 edges. We map all the drug targets and the disease genes related to breast cancer and hepatocellular carcinoma to this network. Then in this network, we apply our method TTMD to predict potential drugs for breast cancer and hepatocellular carcinoma. For the predicted results, if breast cancer-drug or hepatocellular carcinoma-drug pairs are found in CTD database, they are considered as true positive relationships. Then, we can plot two kinds of ROC curves in Figure 9: one for the tissue-specific PPI network (green color), the other for the general PPI network (red color). In Figure 9A, the two ROC curves correspond to breast cancer:

the AUC of green curve for tissue-specific network is 0.8413 and that of red curve for general PPI network is 0.7447. The AUC is the area under the ROC curve, which represents the accuracy of predicted drug-disease associations here. In Figure 9B, the two ROC curves are for hepatocellular carcinoma and the AUCs for tissue-specific and general PPI networks are 0.8209 and 0.7281, respectively. From Figure 9, we can see the tissue-specific PPI networks perform better than the general PPI networks in predicting potential relationships between drugs and diseases. The results further demonstrate the accuracy and robustness of our approach based on tissue-specific PPI networks.

Discussions and conclusions

Drug repositioning is one of the essential strategies for improving drug discovery and solving the threat of complex disease. In this study, we propose a new method TTMD to predict drug indications for a specific disease, which consider the tissue specificities of diseases, the incompleteness of data and the triangularly balanced structure between drugs and diseases. We firstly integrate three filtered drug similarity networks into a reliable network, named as CGA. Then, based on the tissue-specific PPI network of a specific disease and the drug similarity network CGA, we calculate the association scores between drugs and the specific disease based on module distance and drug's direct neighbors. High score suggests high possibility to be potential association. Finally, we choose breast cancer and hepatocellular carcinoma as cases to validate our approach TTMD. We evaluate the top-5% drugs by their overlaps with drug indications that are reported in literatures and CTD database, and also make clinical verifications and KEGG pathway enrichment analysis on the potential drugs for diseases. The results of our experiments are inspiring: D05589 (Pralatrexate) may be a potential treatment for breast cancer; D06304 (Vindesine) and D09755 (Cabazitaxel) may be treat HCC; D05589 (Pralatrexate) may be cause HCC.

Moreover, our method TTMD can be expanded to other tissue-specific diseases easily.

Competing interests

The authors declare that they have no competing interests.

In the future, with the constant improvement of the data, more biological data such as the highest level clinical phenotypes, RNA data and DNA methylation data can be integrated to predict new indications for drugs and more accurate methods will be proposed to deal with the treatments of complex disease. Although computational approaches are far from the animal tests and clinical trials, we still trust that network-based methods will finally change our understanding of the interaction mechanism between complex diseases and drugs and lead to practical applications in drug discovery.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 61672406, 61532014, 91530113, 61502363 and 61402349), the Natural Science Basic Research Plan in Shaanxi Province of China (Nos. 2016JQ6057, 2015JM6283).

Author contributions

L.Y. designed experiments, analyzed data and wrote the paper; J.Z. performed experiments and analyzed data; L.G. modified the paper.

Reference

[1] Schork N J. Genetics of complex disease: approaches, problems, and solutions[J]. American journal of respiratory and critical care medicine, 1997, 156(4): S103-S109.

[2] DiMasi J A. New drug development in the United States from 1963 to 1999[J]. Clinical Pharmacology & Therapeutics, 2001, 69(5): 286-296.

[3] Adams C P, Brantner V V. Estimating the cost of new drug development: is it really $802 million?[J]. Health affairs, 2006, 25(2): 420-428.

[4] Ashburn T T, Thor K B. Drug repositioning: identifying and developing new uses for existing drugs[J]. Nature reviews Drug discovery, 2004, 3(8): 673-683.

[5] Dudley J T, Deshpande T, Butte A J. Exploiting drug–disease relationships for computational drug repositioning[J]. Briefings in bioinformatics, 2011: bbr013.

[6] Scheindlin S. Rare diseases, orphan drugs, and orphaned patients[J]. Molecular interventions, 2006, 6(4): 186.

[7] Soignet S L, Maslak P, Wang Z G, et al. Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide[J]. New England Journal of Medicine, 1998, 339(19): 1341-1348.

[8] Yardley V, Croft S L. Activity of liposomal amphotericin B against experimental cutaneous leishmaniasis[J]. Antimicrobial agents and chemotherapy, 1997, 41(4): 752-756.

[9] Iorio F, Saez-Rodriguez J, Di Bernardo D. Network based elucidation of drug response: from modulators to targets[J]. BMC systems biology, 2013, 7(1): 1.

[10] Li J, Lu Z. Pathway-based drug repositioning using causal inference[J]. BMC bioinformatics, 2013, 14(16): 1.

[11] Zhao H, Jin G, Cui K, et al. Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases[J]. Cancer research, 2013, 73(20): 6149-6163.

[12] Bisgin H, Liu Z, Kelly R, et al. Investigating drug repositioning opportunities in FDA drug labels through topic modeling[J]. BMC bioinformatics, 2012, 13(15): 1.

[13] An S M, Ding Q P, Li L. Stem cell signaling as a target for novel drug discovery: recent progress in the WNT and Hedgehog pathways[J]. Acta Pharmacologica Sinica, 2013, 34(6): 777-783.

[14] Novick P A, Ortiz O F, Poelman J, et al. SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery[J]. PloS one, 2013, 8(11): e79568.

[15] Wang Y, Chen S, Deng N, et al. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data[J]. PloS one, 2013, 8(11): e78518.

[16] Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects[J]. PloS one, 2011, 6(12): e28025.

[17] Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning[J]. PloS one, 2014, 9(2): e87864.

[18] Bisgin H, Liu Z, Fang H, et al. A phenome-guided drug repositioning through a latent variable model[J]. BMC bioinformatics, 2014, 15(1): 1.

[19] Haeberle H, Dudley J T, Liu J T C, et al. Identification of cell surface targets through meta-analysis of microarray data[J]. Neoplasia, 2012, 14(7): 666-669.

[20] Sanseau P, Agarwal P, Barnes M R, et al. Use of genome-wide association studies for drug repositioning[J]. Nature biotechnology, 2012, 30(4): 317-320.

[21] Qu X A, Rajpal D K. Applications of Connectivity Map in drug discovery and development[J]. Drug discovery today, 2012, 17(23): 1289-1298.

[22] Swamidass S J. Mining small-molecule screens to repurpose drugs[J]. Briefings in bioinformatics, 2011, 12(4): 327-335.

[23] Li J, Lu Z. Pathway-based drug repositioning using causal inference[J]. BMC bioinformatics, 2013, 14(16): 1.

[24] Li J, Lu Z. A new method for computational drug repositioning using drug pairwise similarity[C]//Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference On. IEEE, 2012: 1-4.

[25] Wu C, Gudivada R C, Aronow B J, et al. Computational drug repositioning through heterogeneous network clustering[J]. BMC systems biology, 2013, 7(Suppl 5): S6.

[26] Würth R, Thellung S, Bajetto A, et al. Drug-repositioning opportunities for cancer therapy: novel molecular targets for known compounds[J]. Drug discovery today, 2016, 21(1): 190-199.

[27] Chen B, Sirota M, Fan-Minogue H, et al. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research[J]. BMC medical genomics, 2015, 8(2): 1.

[28] Weinstein J N, Collisson E A, Mills G B, et al. The cancer genome atlas pan-cancer analysis project[J]. Nature genetics, 2013, 45(10): 1113-1120.

[29] Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity[J]. Nature, 2012, 483(7391): 603-607.

[30] Kosti I, Jain N, Aran D, et al. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues[J]. Scientific reports, 2016, 6.

[31] Guan Y, Gorenshteyn D, Burmeister M, et al. Tissue-specific functional networks for prioritizing phenotype and disease genes[J]. PLoS Comput Biol, 2012, 8(9): e1002694.

[32] Cartwright D, Harary F. Structural balance: a generalization of Heider's theory[J]. Psychological review, 1956, 63(5): 277.

[33] Mosca R, Pons T, Céol A, Valencia A, Aloy P. Towards a detailed atlas of protein-protein interactions. Curr. Opin. Struct. Biol. 2013; 23(6):929-40.

[34] Mohammadi S, Grama A. A convex optimization approach for identification of human tissue-specific interactomes. Bioinformatics. 2016; 32(12):i243-i252.

[35] Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? Genome Biol. 2006;7(11):120.

[36] Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome[J]. Science, 2015, 347(6224): 1257601.

[37] Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation[J]. Nucleic acids research, 2015: gkv1070.

[38] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30.

[39] Wishart D S, Knox C, Guo A C, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration[J]. Nucleic acids research, 2006, 34(suppl 1): D668-D672.

[40] Greene C S, Krishnan A, Wong A K, et al. Understanding multicellular function

[41] Davis A P, Grondin C J, Lennon-Hopkins K, et al. The Comparative

[42] Heider F. Attitudes and cognitive organization[J]. The Journal of psychology, 1946, 21(1): 107-112.

[43] Hattori M, Okuno Y, Goto S, et al. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways[J]. Journal of the American Chemical Society, 2003, 125(39): 11853-11865.

[44] Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG[J]. Nucleic acids research, 2006, 34(suppl 1): D354-D357.

[45] Lamb J, Crawford E D, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease[J]. science, 2006, 313(5795): 1929-1935.

[46] Lamb J. The Connectivity Map: a new tool for biomedical research[J]. Nature Reviews Cancer, 2007, 7(1): 54-60.

[47] Lin D. An information-theoretic definition of similarity[C]//ICML. 1998, 98: 296-304.

[48] Nacher J C, Schwartz J M. A global view of drug-therapy interactions[J]. BMC pharmacology, 2008, 8(1): 5.

[49] Zhao S, Li S. Network-based relating pharmacological and genomic spaces for drug target identification[J]. PloS one, 2010, 5(7): e11764.

[50] World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 1.1. ISBN 92-832-0429-8.

[51] Li, H., Liang S. Local network topology in human protein interaction data predicts functional association. PLoS ONE 4, e6410 (2009).

[52] Barabasi, A.L., Albert, R. Emergence of scaling in random networks. Science 286, 509-512 (1999).

[53] King, B.L., Davis, A.P., Rosenstein, M.C., Wiegers, T.C. & Mattingly, C.J. Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. PLoS One 7, e46524 (2012).

[54] Liu ZH, et al. The growth-inhibition effect of tamoxifen in the combination chemotherapeutics on the human cholangiocarcinoma cell line QBC939. Mol Biol Rep. 2010 Jul;37(6):2693-701.

[55] Tsujimoto Y, Finger L R, Yunis J, et al. Cloning of the chromosome breakpoint of neoplastic B cells with the t (14; 18) chromosome translocation[J]. Science, 1984, 226(4678): 1097-1099.

[56] Favier RP, et al. COMMD1-deficient dogs accumulate copper in hepatocytes and provide a good model for chronic hepatitis and fibrosis. PLoS One. 2012;7(8):e42158.

[57] Porter A G, Jänicke R U. Emerging roles of caspase-3 in apoptosis[J]. Cell death and differentiation, 1999, 6(2): 99-104.

[58] Minoru Kanehisa1, Miho Furumichi1 Mao Tanabe, Yoko Sato and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucl. Acids Res. (2016) doi: 10.1093/nar/gkw1092.

[59] Huang DW, Sherman BT, Lempicki RA.Systematic and integrative analysis of

[60] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths

[61] Beard Jr E L. The American Society of Health System Pharmacists[J]. JONA'S healthcare law, ethics and regulation, 2001, 3(3): 78-79.

[62] Radić J, Krušlin B, Šamija M, et al. Connexin 43 Expression in Primary Colorectal Carcinomas in Patients with Stage III and IV Disease[J]. Anticancer research, 2016, 36(5): 2189-2196.

[63] Yang Y, Zhu J, Zhang N, et al. Impaired gap junctions in human hepatocellular carcinoma limit intrinsic oxaliplatin chemosensitivity: A key role of connexin 26[J]. International journal of oncology, 2016, 48(2): 703-713.

[64] Pan Q, Huang Y, Chen L, et al. SMAC-armed vaccinia virus induces both apoptosis and necroptosis and synergizes the efficiency of vinblastine in HCC[J]. Human cell, 2014, 27(4): 162-171.

[65] Zhou Q, Lui V W Y, Lau C P Y, et al. Sustained antitumor activity by co-targeting mTOR and the microtubule with temsirolimus/vinblastine combination

[66] Gerwing M, Jacobsen C, Dyshlovoy S, et al. Cabazitaxel overcomes cisplatin resistance in germ cell tumour cells[J]. Journal of Cancer Research and Clinical Oncology, 2016: 1-16.

[67] Armstrong A J, Carducci M A. New drugs in prostate cancer[J]. Current opinion in urology, 2006, 16(3): 138-145.

[68] Agmon-Levin N, Katz BS, Shoenfeld Y. Infection and primary biliary cirrhosis. Isr Med Assoc J. 2009, 11(2):112-5.

[69] Bogdanos DP, Baum H, Vergani D, Burroughs AK. The role of E. coli infection in the pathogenesis of primary biliary cirrhosis. Dis Markers. 2010, 29(6):301-11.

[70] Silveira MG, Suzuki A, Lindor KD. Surveillance for hepatocellular carcinoma in patients with primary biliary cirrhosis. Hepatology. 2008, 48(4):1149-56.

[71] Piscaglia F, Sagrini E. Malignancies in primary biliary cirrhosis. Eur J Gastroenterol Hepatol, 2008, 20(1):1-4.

[72] Shibuya A1, Tanaka K, Miyakawa H, Shibata M, Takatori M, Sekiyama K, Hashimoto N, Amaki S, Komatsu T, Morizane T. Hepatocellular carcinoma and survival in patients with primary biliary cirrhosis. Hepatology. 2002, 35(5):1172-8.

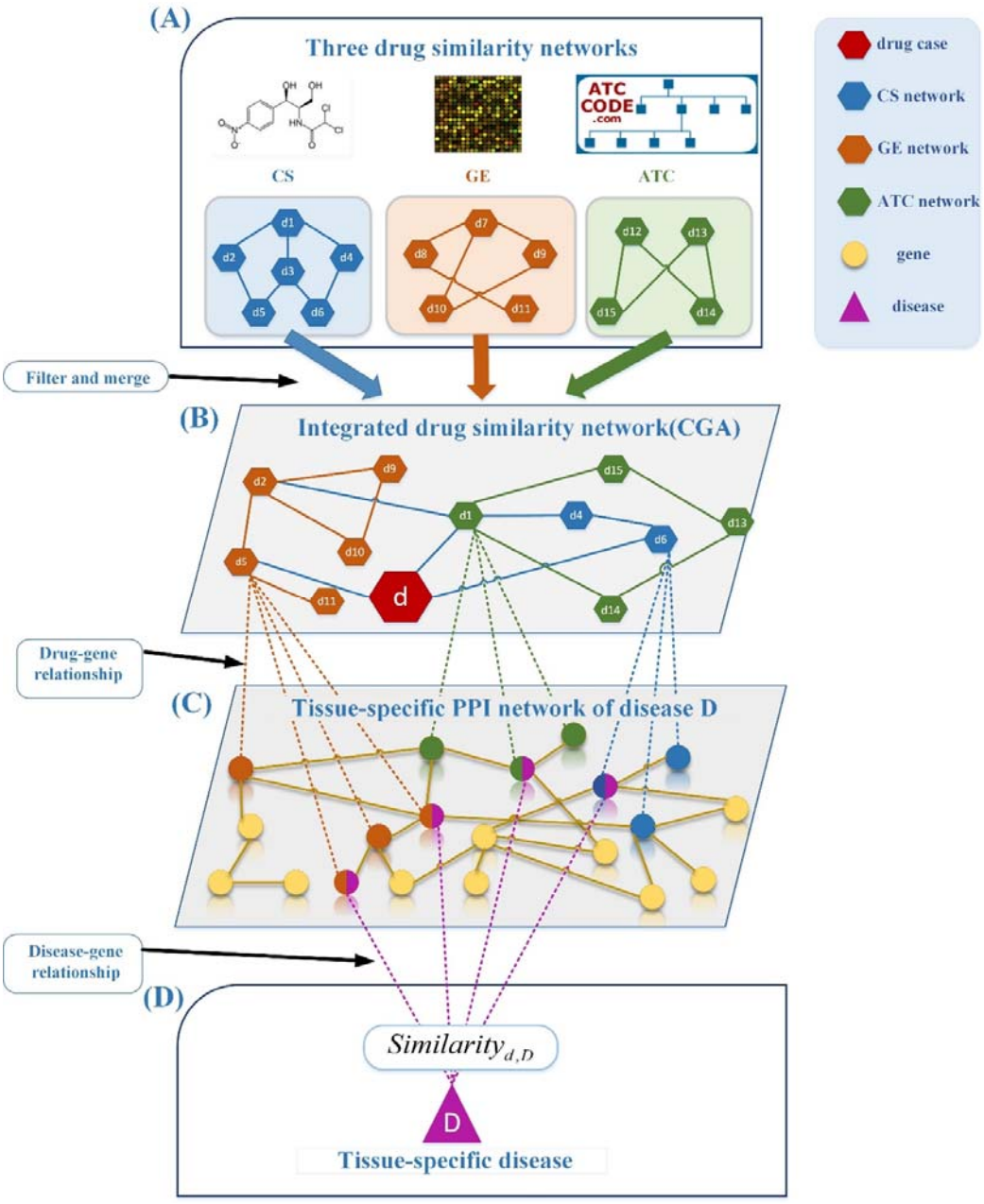
[73] Deutsch M, Papatheodoridis GV, Tzakou A, Hadziyannis SJ. Eur J Gastroenterol Hepatol. Risk of hepatocellular carcinoma and extrahepatic malignancies in primary biliary cirrhosis. 2008, 20(1):5-9.

[74] Farinati F, Floreani A, De Maria N, Fagiuoli S, Naccarato R, Chiaramonte M. Hepatocellular carcinoma in primary biliary cirrhosis. J Hepatol. 1994, 21(3):315-6.

[75] O'Connor O A. Pralatrexate: an emerging new agent with activity in T-cell lymphomas[J]. Current opinion in oncology, 2006, 18(6): 591-597.

[76] Folotyn. Westminster, CO: Allos Therapeutics, Inc.;2011 January.

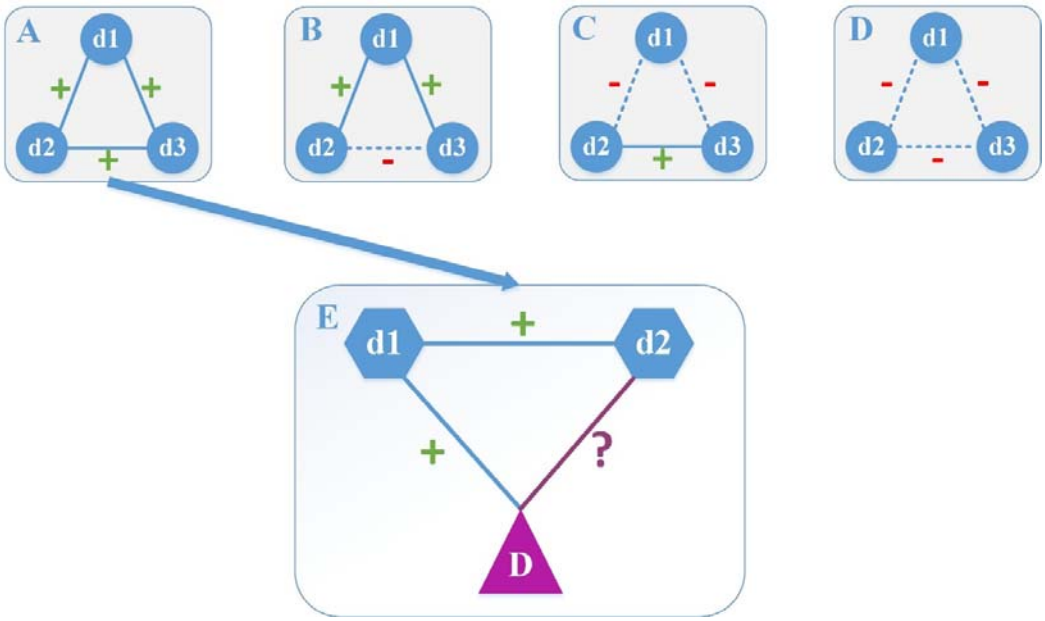
[77] American Society of Health-System Pharmacists. ASHP guidelines on handling hazardous drugs. Am J Health-Syst Pharm. 2006, 63:1172-1193.



<InlineImage1>

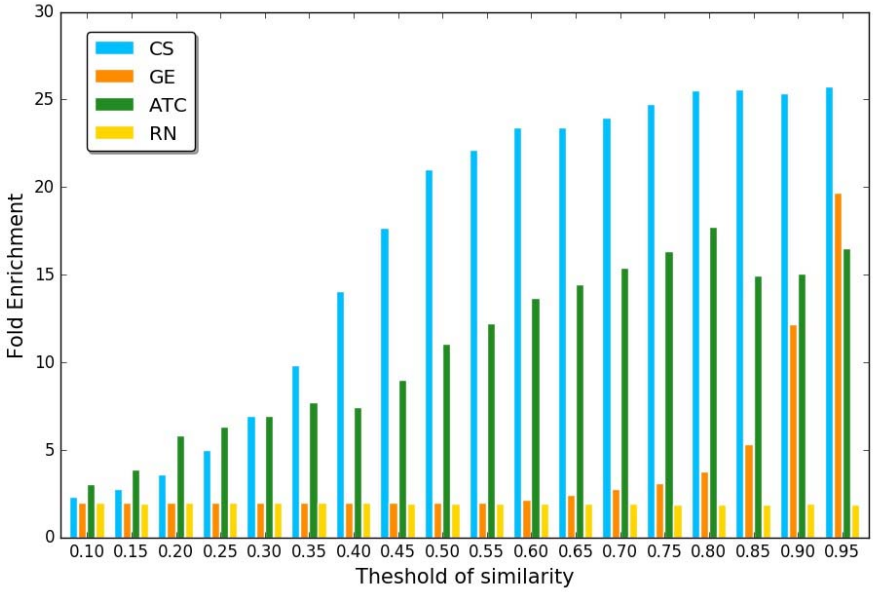
Figure 1. The framework of TTMD. (A**)** Based on data of drugs: chemical structures, gene expressions and ATC codes, three drug similarity networks are constructed: CS (marked as blue), GE (marked as orange) and ATC (marked as green). (B) After filtering CS, GE and ATC, we combine the three networks into one more reliable drug similarity network, named as CGA. If one edge exists in multiple networks, its weight is their mean. Here, *d* is a drug case to illustrate our framework. (C) In the tissue-specific PPI network of disease *D*, the targets of drugs in CGA are mapped to the PPI network and the targets have the same color as drugs. At the same time, we map the related genes of *D* to the PPI network. Nodes with multiple colors indicate they relate to drugs and disease *D* simultaneously. Yellow nodes represent the background. (D) Based on the tissue-specific PPI network

and CGA, we calculate the association between each drug and disease *D*. In this figure, $Similarity_{d, D}$ represents the association between drug *d* and disease *D*.



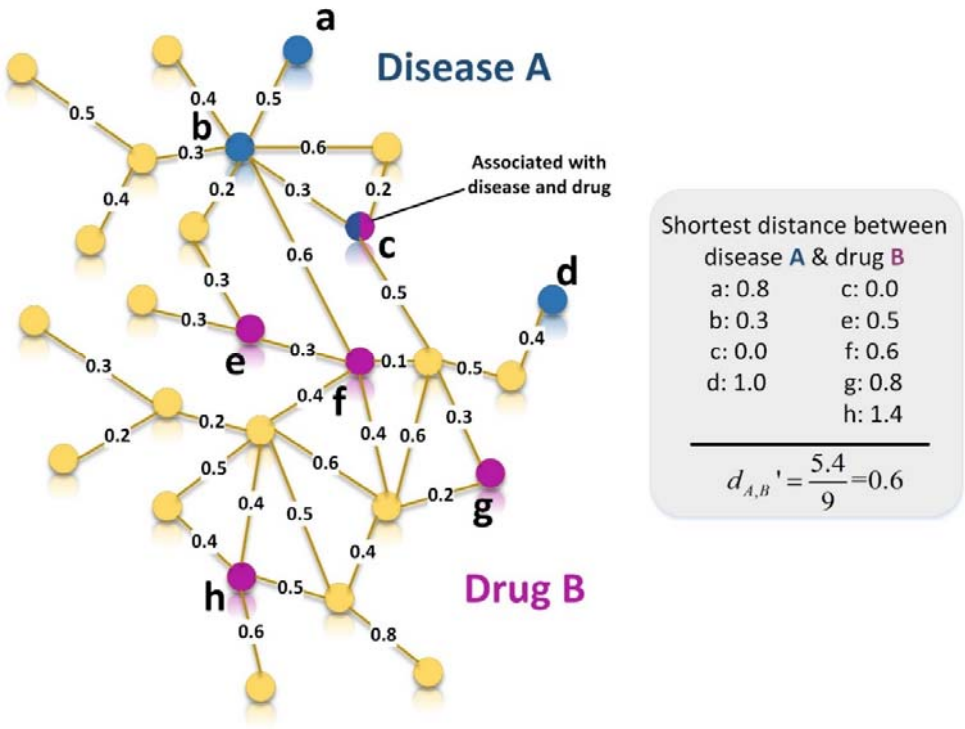
<InlineImage2>

Figure 2. The Triangular structures. A to D. Four types of relationship among three points. Positive relationships are marked as "+" and negative ones are marked as "-". Triangular structures in A and C are balanced and the remaining two structures in B and D are not balanced. E. If drug d1 is similar to drug d2 and drug d1 can treat disease D, we can infer that drug d2 may treat disease D, too.



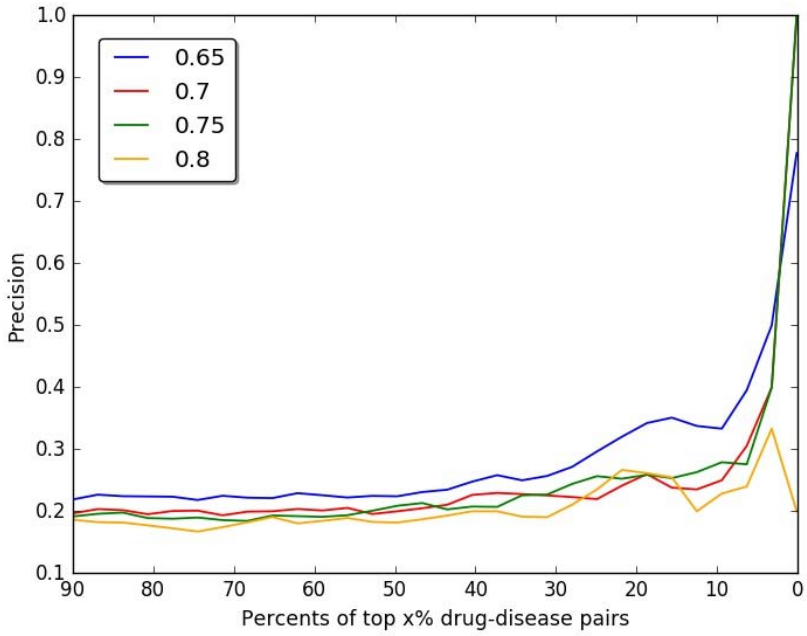
<InlineImage3>

Figure 3. Enrichment analysis for drug pairs with common targets. The blue bar, orange bar, green bar and yellow bar represent CS, GS, AS and Random network (RN).



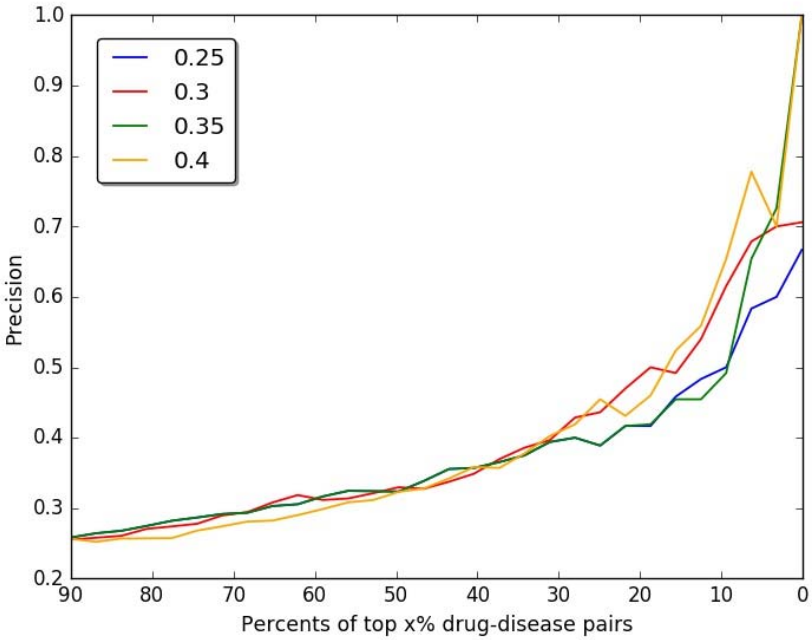
<InlineImage4>

Figure 4. An example for calculating the distance between gene set of disease *A* and target set of drug *B*. Blue and purple nodes represent genes related to disease *A* and drug *B*, respectively. Node *c* is a shared node, so it is marked by two colors.



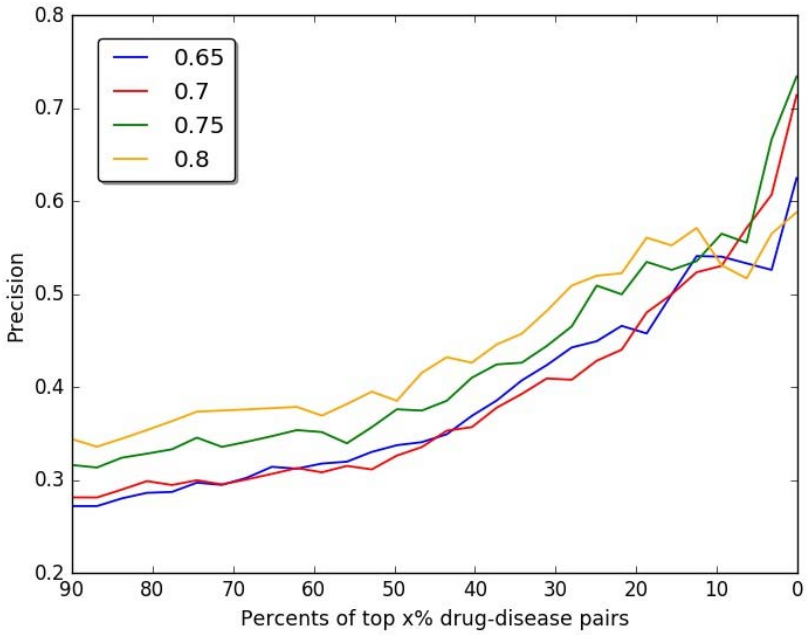
<InlineImage5>

Figure 5. Verification of parameter choice in GE network based on CTD database with reference score over 20.



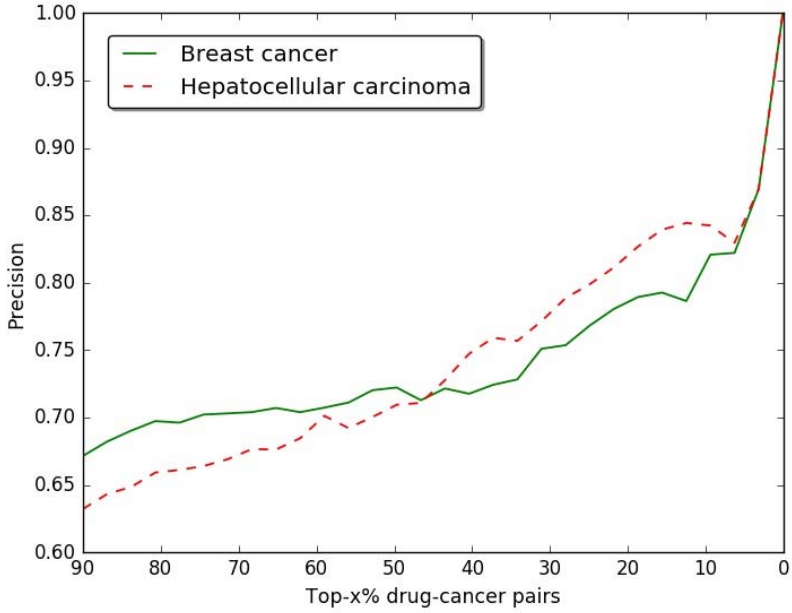
<InlineImage6>

Figure 6. Verification of parameter choice in CS network based on CTD database with reference score over 20.



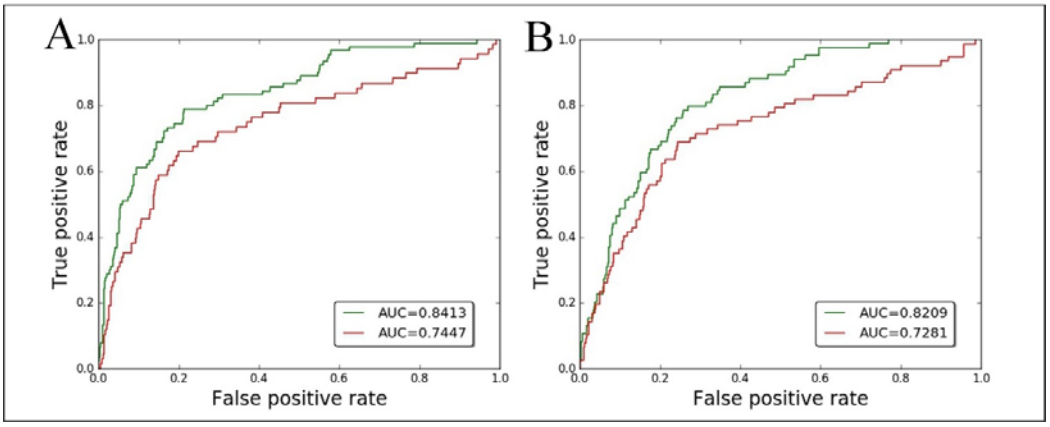
<InlineImage7>

Figure 7. Verification of parameter choice in ATC network based on CTD database with reference score over 20.



<InlineImage8>

Figure 8. The precision of our predictions at different top-x% drug-cancer pairs.



<InlineImage9>

Figure 9. Performance comparison based on tissue-specific and general PPI networks. Green and red curves are for tissue-specific and normal PPI networks, respectively. **A**. ROC curves correspond to breast cancer. **B**. ROC curves correspond to hepatocellular carcinoma.

Table 1. The top-5% drugs related to Breast cancer

KEGG ID	Drug name	Direct Evidence	Inference Score	Reference Score	Similarity Score
D03899	Doxorubicin	M|T	181.12	339	0.94842
D07472	Pemetrexed	T	12.38	23	0.94645
D08958	Ospemifene	Ref	6.07	5	0.94353
D08224	Mitoxantrone	T	24.57	84	0.94319
D05589	**Pralatrexate**	**None**	**None**	**None**	0.94092
D08465	Raloxifene	T	114.42	168	0.9379
D04070	Conjugated estrogens	T	27.34	37	0.92748
D01161	Fulvestrant	T	231.4	186	0.92669
D04066	Estramustine	Ref	4.84	6	0.92112
D07966	Fludarabine	Ref	50.45	51	0.91955
D03546	Clofarabine	Ref	18.04	20	0.91893
D01370	Cladribine	Ref	10.34	13	0.91501
D00292	Dexamethasone	T	156.38	186	0.91149
D00327	Fluoxymesterone	T	0	2	0.91143
D00105	Estradiol	M|T	207.96	305	0.91125
D00067	Estrone	Ref	53.83	48	0.91106
D06320	Vorinostat	T	124.65	164	0.91013
D00950	Levonorgestrel	Ref	20.22	37	0.90554
D02367	Desogestrel	Ref	5.51	15	0.90546
D00951	Depo-provera	M|T	63.78	82	0.90497
D08675	Vinblastine	T	39.84	78	0.89975
D00472	Prednisolone	T	18.15	29	0.89918
D00088	Hydrocortisone	T	52.55	59	0.89613
D08559	Tamoxifen	M|T	195.79	430	0.89577
D04931	Mercaptopurine	Ref	32.09	82	0.89537
D06304	Vindesine	T	2.52	8	0.89262
D08680	Vinorelbine	T	4.27	31	0.89262
D08679	Vincristine	T	48.73	94	0.8923
D08086	Irinotecan	Ref	79.54	114	0.89046
D00898	Dienestrol	Ref	11.13	17	0.8891
D00066	Progesterone	M|T	166.14	236	0.88697
D02368	Gemcitabine	T	105	116	0.88482

Table 2. The predicted top-5% drugs related to Hepatocellular carcinoma

KEGG ID	Drug name	Direct Evidence	Inference Score	Reference Score	Similarity
D07776	Daunorubicin	T	70.24	43	0.94233
D03899	Doxorubicin	T	134.83	97	0.93376
D08224	Mitoxantrone	Ref	22.73	22	0.9296
D06320	Vorinostat	Ref	67.41	71	0.92232
D08679	Vincristine	Ref	34.08	54	0.91831
D08675	Vinblastine	Ref	21.59	29	0.91811
D00184	Cyclosporine	Ref	62.81	111	0.91715
D06304	**Vindesine**	**None**	**None**	**None**	0.91631
D08680	Vinorelbine	Ref	5.03	9	0.91631
D01370	Cladribine	Ref	8.9	4	0.91408
D00491	Paclitaxel	T	76.63	67	0.91224
D07866	Docetaxel	T	53.52	42	0.91224
D03546	Clofarabine	Ref	17.46	14	0.91002
D07966	Fludarabine	Ref	45.26	29	0.90969
D07472	Pemetrexed	Ref	11.38	17	0.90945
D00424	Rifabutin	Ref	27.77	13	0.90091
D08556	Tacrolimus	Ref	40.74	43	0.90061
D05589	**Pralatrexate**	**None**	**None**	**None**	0.88987
D00292	Dexamethasone	T	131.69	97	0.88955
D08086	Irinotecan	T	58.74	48	0.88843
D09755	**Cabazitaxel**	**None**	**None**	**None**	0.88799
D05480	Pimecrolimus	Ref	2.43	1	0.88713
D00545	Isoflurane	Ref	27.8	30	0.88639
D02368	Gemcitabine	T	85.03	55	0.8818
D08559	Tamoxifen	M|T	65.04	96	0.88067
D00543	Enflurane	Ref	2.13	1	0.87848
D00546	Desflurane	Ref	4.71	2	0.87743
D04931	Mercaptopurine	Ref	10	16	0.87376
D00544	Methoxyflurane	Ref	2.69	1	0.87347
D00547	Sevoflurane	Ref	19	23	0.87344
D00472	Prednisolone	M	18.93	25	0.87201

Table 3. Overlapped KEGG pathways between potential drugs and hepatocellular carcinoma

Drug Name (KEGG ID)	Overlapped enriched pathways	*p*-value
Vindesine (D06304)	hsa04540:Gap junction	5.01E-16
	hsa05130:Pathogenic Escherichia coli infection	4.96E-18
Cabazitaxel (D09755)	hsa04540:Gap junction	0.0127
	hsa05130:Pathogenic Escherichia coli infection	0.0074