BMC
Medical Genomics

RESEARCH   **Open Access**

Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation

Yu-Fen Huang[1†], Hsiang-Yuan Yeh[2*†], Von-Wun Soo[1,2]

Abstract

Background: During the last few years, the knowledge of drug, disease phenotype and protein has been rapidly accumulated and more and more scientists have been drawn the attention to inferring drug-disease associations by computational method. Development of an integrated approach for systematic discovering drug-disease associations by those informational data is an important issue.

Methods: We combine three different networks of drug, genomic and disease phenotype and assign the weights to the edges from available experimental data and knowledge. Given a specific disease, we use our network propagation approach to infer the drug-disease associations.

Results: We apply prostate cancer and colorectal cancer as our test data. We use the manually curated drug-disease associations from comparative toxicogenomics database to be our benchmark. The ranked results show that our proposed method obtains higher specificity and sensitivity and clearly outperforms previous methods. Our result also show that our method with off-targets information gets higher performance than that with only primary drug targets in both test data.

Conclusions: We clearly demonstrate the feasibility and benefits of using network-based analyses of chemical, genomic and phenotype data to reveal drug-disease associations. The potential associations inferred by our method provide new perspectives for toxicogenomics and drug reposition evaluation.

Background

Disease an intricate phenotype is usually caused by congenital disorder or dysfunctions of abnormal genes which induce multi-factor-driven alterations and disrupt functional modules [1]. Drugs achieve their therapeutic effect by changing downstream processes of their targets which contend with the alterations of those abnormal genes. The previous reports also showed that pharmaceutical company takes approximately 15 years and over $1 billion to develop a novel drug into the market and more than 90% of experimental drugs fail to move beyond the early

clinical test stages [2,3]. Because drug discovery is complexity, time-consuming process and there are odds of low therapeutic efficacy and/or unacceptable toxicity [4,5]. With the merits of shorting development time and reducing risk, more and more scientists have been drawn attention to inferring drug-disease associations by computational method. Development of an integrated approach for systematic discovering those associations is necessary.

Several studies investigated some methods to increase the efficacy of drug discovery and they found there are positive and negative relationships between existing drugs and disease phenotypes. The Comparative Toxicogenomics Database (CTD; http://ctd.mdibl.org) is the public database which inferred chemical-disease associations by manually curated chemical-gene interactions, and gene-disease relationships from published literature [6].

* Correspondence: d926708@oz.nthu.edu.tw
† Contributed equally
[2]Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan
Full list of author information is available at the end of the article

Cheng also presented a comprehensive predicted database of chemical -gene-disease associations (PredCTD) by integrating the information from chemical, gene, and disease [7]. Pharmacogenetics and pharmacogenomics knowledge base (PharmGKB) is a repository which contains the relationships between genomics, drug-response and its related phenotype and clinical information [8]. Eichborn developed PROMISCUOUS database which includes network-based resources of protein-protein and protein-drug interactions, side-effects and structural information [9].

The high-throughput microarray technology plays an important role in investigating drug-disease associations by providing a genome-side monitoring of gene expression in the past decade. Some methods aims at restoring the abnormal state to normal state which means the expressions of the transcriptional level induced by drug should reverse those under disease state. On the other hand, if the differential expression profile under drug exposure and disease states is significantly anti-correlated, the drug compounds may have the potential to cure that disease. The Connectivity Map (Cmap) project is one of the most comprehensive and systematic approaches for drug-disease associations [10]. The Cmap provided a reference collection of genome-wide gene expressions profiles among drugs, which were obtained by systematically exposing to few key cell lines [11]. Drug compounds negatively correlated to disease-specific gene signature may be the candidate therapeutic for further investigation. On the other hand, drug compounds positively correlated to gene signature are able to induce the disease phenotype. Li built disease-specific drug-protein associations derived from the Cmap by integrating gene/protein and drug connectivity information based on protein interaction network (PIN) and literature mining from PubMed abstracts [12]. Previous research used the "guilt by association" (GBA) approach, which assumed that when two diseases share similar therapies then the drug treats only one of the two might be also treat another, to predict novel drug-disease associations [13]. With a gold standard set of the drug-disease associations, Gottlieb designed a novel computational method called PREDICT to identify drug-disease associations and also predict new drug indications based on their features including chemical structure, side effects, gene expression profile, and chemical-protein interactome [14]. However, to build an accurate prediction model based on different feature must have the positive and negative data to infer drug-disease associations. There are some technically difficulties to obtain negative data such as non-drug targets due to the lack of value of research. Except learning a classifier to predict the associations between drugs and disease, network-based approach has been widely used to infer the relationships. In genetic and molecular biology, increasing

evidences suggested that common functional modules are not affected by an individual gene but usually are organized by a group of interacting genes underlie similar diseases, which point out the therapeutic importance of those modules [15]. Therefore, the other basic hypothesis is that the mechanism of the drug and disease in the pathological processes may share similar functional modules. Daminelli created a drug-target-disease network and mined the bi-cliques where every drug is linked to every target and disease [16]. If the known data form an incomplete bi-clique, the incomplete relations in the bi-cliques to be identified as predicted links between drugs and diseases. Ye integrated known drug target information and proposed a disease-oriented strategy for evaluating the relationships between drugs and a specific disease based on their pathway profile [17].

The huge amount of chemical, genomic and disease phenotype data is rapidly accumulated, but the drug-diseases associations are still not clear. For this purpose, we design a method of inferring drug- protein/gene-disease phenotype relationships with a network propagation model, where genes with similar functional modules are related to not only drugs but also the disease phenotype.

Methods

We demonstrate the integrated network including three heterogeneous networks of the phenotype, drug, protein homo-networks and two hetero-networks capture interactions between two different homo-networks in Figure 1. The homo-network is defined as an undirected graph $G_i = (V_i, E_i)$ where V_i is the node set and E_i is the edge set in the homo-network i. Connections between two kinds of homo-networks define as hetero-network, where the nodes from different homo-networks are related to each other. The hetero-network is defined as bipartite graph $G_{ij} = (V_i \cup V_j, E_{ij})$. Here, E_{ij} represents the set of edges which connect the nodes in different kinds of the homo-networks i and j.

Construct phenotype homo-network

A node in a phenotype homo-network as a disease phenotype is extracted from Online Mendelian Inheritance in Man (OMIM) database [18]. We use a scoring schema of phenotypic similarities as edges that quantitatively measures the phenotypic overlap of OMIM records constructed by van Driel [19] using text mining techniques. If the similarity score of two diseases falls in the range [0.6, 1], it means informative similarity which indicates potentially relevant phenotypic similarity. On the other hand, if the similarity score falls within [0, 0.3], it indicates non-informative similarity. Therefore, we apply a logistic function from [20] to convert the phenotypic similarity scores among diseases into a value as close to 1 as possible while the non-informative score into a value as close

Figure 1 The idea of our proposed method.
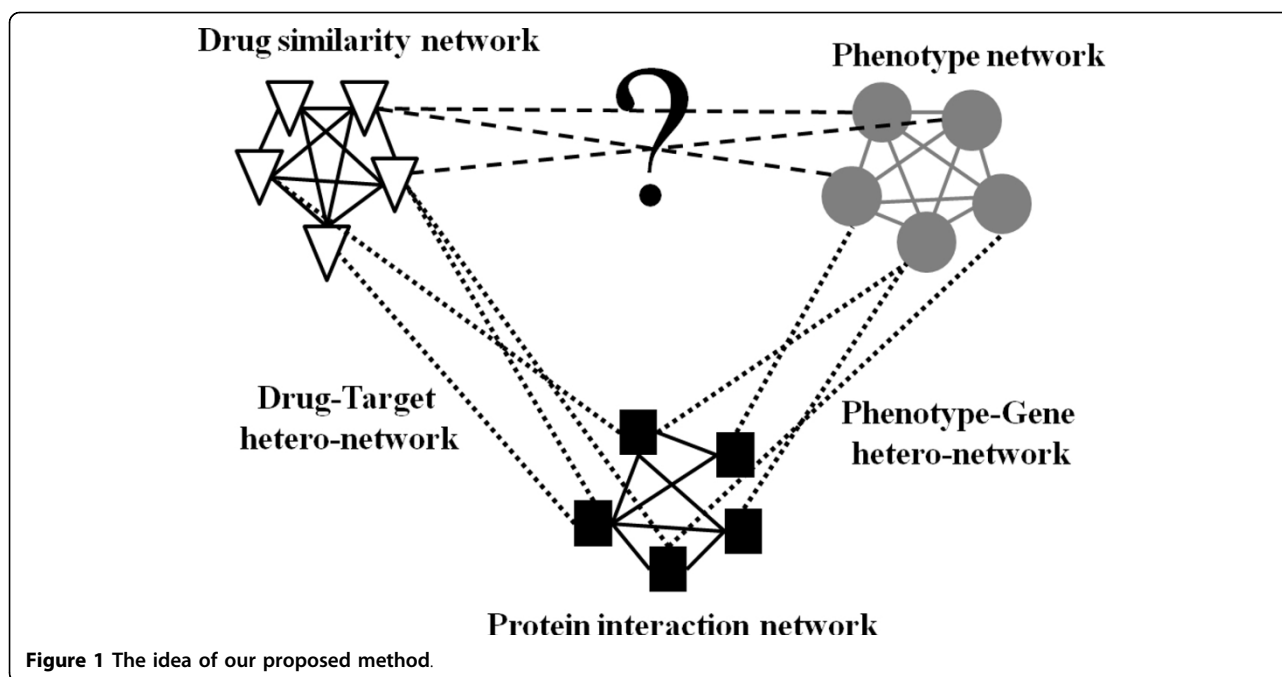
to 0 possible over all the entries in the phenotype similarity score matrix. The symmetric similarity matrix $W_p(p_i p_j)$ in phenotype homo-network denotes the phenotypic similarity score between phenotypes p_i and p_j.

Construct drug homo-network using chemical similarity

We extract the FDA-approved drugs and their canonical simplified molecular input line entry specification (SMILES) from DrugBank database [21][22]. We calculate the hashed fingerprints using Chemical Development Kit (CDK) [23]. The chemical similarities are calculated by two hashed fingerprints using Tanimoto coefficient [24]. It calculates the size of the common substructures over the union between two fingerprints of the drugs which is defined as $sim(x, x') = |x \cap x'|/|x \cup x'|$ between two chemical structures of drug x and x'. The symmetric chemical structures similarity matrix as the edge weight in drug homo-network is denoted as W_d and each value falls in the range between zero (no bits in common) to unity (all bits the same)

Construct protein interaction homo-network using gene expression data

Protein-protein interactions provided an opportunity on the discovery of relationships among proteins in the mechanism of the drugs and human disease phenotype. However, the PIN has its disadvantages: first, the information in current PIN databases is partly complementary and the combination of the multiple databases could improve the knowledge of the protein interactions [25]. Second, PIN in those databases only provides the functional

relations among the products of the genes and do not provide information about the conditions under which the interactions occur. Genome-wide expression profile can help us to extract the changes of the genes which are involved in the activity of a given disease or drug compound by comparing multiple case-control data sets. The co-expressed genes can reveal specific linkages which are more likely to function together, and we apply Pearson correlation coefficient for every pair-wise relation among genes. The positive correlation indicates an increasing linear relationship and negative correlation indicates a decreasing linear relationship. On the other hand, the correlation approaches to zero shows there would be little or no association. By visualizing gene expression overexpressed and down-expressed functional modules, we take the absolute value of correlation value to capture inhibitory activity (negative correlation) as well as activation activity (positive correlation). We define the weight function as the product of the absolute value of correlation and the sum of the absolute value of differential expression changes between two corresponding genes in control and case samples. The symmetric weighted matrix between all interactions among all gene pairs is denoted as W_g and the higher weight denotes the stronger correlation or larger differential expression exchanges.

$$W_g\left(g_i, g_j\right) = \left|R_{g_i g_j}^d\right| \times \left(\left|E_{g_i}^d - E_{g_i}^n\right| + \left|E_{g_j}^d - E_{g_j}^n\right|\right)$$

where $W_g(g_i g_j)$ denotes the weight function from gene g_i to gene g_j. $\left|R_{g_i g_j}^d\right|$ denotes the absolute value of Pearson

correlation coefficient of the interaction between gene g_i and g_j from case data. $E_{g_i}^d$ and $E_{g_i}^n$ are the average gene expression values of gene g_i in case sample and control sample.

Integrated disease, protein interactions and chemical homo-networks

The gene-phenotype hetero-network shows the relationships between disease phenotype and disease-associated genes extracted from OMIM database [18]. The drug-protein hetero-network denotes the drug and its targets which is obtained from DrugBank database [21]. The asymmetric matrices W_{pg}, W_{dg} represent the adjacency matrices of link structures from phenotype-gene relationships and drug-target protein interactions, respectively. If drug d_i has a target g_j, then $W_{dg}(d_i, g_j) = 1$, otherwise $W_{dg}(d_i, g_j) = 0$. When a drug target or disease-associated gene has no link with other proteins in PIN, we set the probability of connection to any other protein as $1/(n\text{-}1)$, where n is the total number of proteins in PIN. Since n is usually very large, so the probability will be very small. The reason that we use small probability instead of zero probability is to prevent a node in the network becoming a "sink node" in PIN and allows the probability to be propagated through the node.

Network propagation in the integrated network

We identify the inferring drug-disease associations problem as probability propagation over a network which simulates a random walker stochastically move on query phenotype to its immediate neighbors in heterogeneous network [26,27]. We adopt the idea from [28] which developed a label propagation algorithm for an integrated network.

In order to balancing the influence on the nodes which they are connected in the network, we normalize all similarity matrices W to be a transition matrix S by a diagonal matrix $D_r(i,i)$ which indicates the sum of row i while another diagonal matrix $D_c(j,j)$ indicates the sum of column j in the similarity matrix, respectively. If the similarity matrix of a homo-network is symmetric, the diagonal matrix $D_r(i,i)$ is equal to $D_c(j,j)$.

$$S\left(i,j\right) = W\left(i,j\right) / \sqrt{D_r\left(i,i\right) D_c\left(j,j\right)}$$

where i denotes the node in a homo-network and j denotes the node in the other homo-network.

Given transition matrix S, diffusion parameter α and the vector p^0 with the initial probability distribution over nodes, the probability transition process in network propagation method on single network within t steps is defined as:

$$p^t = (1 - \alpha)\, p^0 + \alpha S p^{t-1}$$

The first term denotes the random walker can "restarts" to the initial probability distribution among the nodes with the diffusion parameter $1\text{-}\alpha$. The second term denotes an iterative walk to reach the further nodes in the network based on the transition matrix S and a diffusion parameter α. The random walker will be trapped at initial nodes if α is zero. Let P^t be a probability distribution where a node in the network holds the probability of finding itself in the iterative random walker process up to the step t. After certain steps, the probabilities will reach a steady state which the difference between P^t and P^{t-1} measured by $L2$ norm falls below a very small value such as 10^{-9}.

We extend the network propagation on a single network to our integrated network. The nodes receive the probabilities from other nodes in the same homo-network, and also can get the probabilities from nodes in other homo-networks through hetero-networks [29]. Therefore, the initial probability would be replaced by adding the additional information from its immediate neighbors through hetero-networks [30]. The new initial probability vector p_i^0 in each homogeneous network i is proposed by a weighted sum which is formulated as follows:

$$p_i^0 = a_i p_i^0 + b_{ij} \sum_{i \neq j} S_{ij} p_j^0$$

Where a_i and b_{ij} denote the weights in homo-network i and those between two homo-networks i and j, respectively. S_{ij} denotes the normalized transition matrix of the hetero-network. Then, we should keep the probability of the node be equal to one and we must further ensure the parameters a_i and b_{ij} to satisfy:

$$a_i + \sum_{i \neq j} b_{ij} = 1$$

If homo-network i connect to other k homo-networks, we adopt the weight of b_{ij} the same as diffusion parameter α_i to the immediate neighbors in the other homo-networks. We calculate $a_i + k\alpha_i = 1$ and obtain the weight $a_i = 1 - k\alpha_i$. Finally, we further elaborate the network propagation method on a homo-network i into:

$$p_i^t = (1 - \alpha_i)\left[(1 - k\alpha_i)\, p_i^0 + \alpha_i \sum_{i \neq j} S_{ij} p_j^0 \right]$$
$$+ \alpha_i S_i p_i^{t-1}$$

Thus, network propagation is calculated with an enriched initialization from the other homo-networks through hetero-subnetworks and the proof of convergence is in [31].

Given a query phenotype, we first set the initial probability distribution over nodes where the probabilities to

the query disease nodes set to one and other nodes in the other homo-networks to zero. Second, we apply our network propagation method iteratively until the probability converges on each homo-network. Finally, we use the coverged probabilities of the nodes in homo-networks as initail probability distribution and then repeated the network propagation processes until all homo-networks converge to a final probability distribution.

Evaluation of association specificity between drug and disease

In our method, we apply chemical similarity, gene expression, and phenotype similarity data and the transition in the network propagation processes may skew the visitation frequencies towards those supplied data values. Since the frequencies of node visitations may also be highly biased by the linkages in the network topology, the probabilities of the nodes may directly reflect the relative centralities in the network based on the local network connectivity [32]. In order to control the topological biases in PIN, we calculate the reference visitation frequencies without taking the gene expression profile of the specific disease phenotype into consideration. Therefore, we set all the edge weights among genes to one in PIN and get the reference probabilities distribution P_i^{ref} over nodes in each homo-network i using our method. Then, we evaluate the specificity of the probability of a node using Z-score as the final normalized score distribution which reflects the relevance nodes related to the query phenotype.

$$Z_i(v) = \frac{P_i(v) - avg(P_i^{ref})}{std\left(P_i^{ref}\right)}$$

Here, $P_i(v)$ denote the probability of node v in homo-network i using genomic data calculated by our method. The functions *avg* and *std* denote the average and standard deviation for the set of reference probabilities P_i^{ref} in homo-network i respectively.

Results

Gene expression profile

We adopt microarray data taken from [33] that consists of 62 primary prostate tumors and 41 normal tissues from Stanford Microarray Database (SMD) [34]. We use genome-wide gene expression profiles from tissue samples of 18 healthy normal controls and 27 patients with colorectal cancer evaluated by HG-U133 Plus 2.0 platform

microarrays (Affymetrix, Santa Clara) through from Gene Expression Omnibus (GEO) database (GSE4183 and GSE4107) [35,36].

Protein interaction network

We successfully obtained 137,037 interactions among 13,388 genes by integrating five protein interaction network databases (HPRD, BIND, IntAct, MINT, and OPHID) and by mapping the UniProt protein ID to the human Entrez gene ID, erase the duplicated interaction pairs.

Phenotype network and Phenotype-genotype hetero-network

The OMIM database constructed the catalogue of genetic diseases in human and provides the phenotype-genotype association for 14,433 genes and 5,080 diseases [18]. The gene-phenotype hetero-network contains 275 disease phenotypes and 649 genes from 877 relations while mapping the genes in microarray data and PIN.

Drug network and drug-target hetero-network

We collect 1,571 FDA-approved drugs and 1,410 of them with available SMILES data in DrugBank database [21]. There are 4,456 relations between 1,215 drugs and 1,141 targets to be the drug-target hetero-network.

Benchmark of drug-disease associations

We extract 53 and 106 known associations between drug, prostate cancer and colorectal cancer extracted from CTD database [6] in May 2013 as our benchmark.

The performance of our method

We compare our method with the previous state-of-the-art Cmap project to evaluate the performance [10]. We first prepare the ranked lists of over-expressed genes and down-expressed genes in both prostate and colorectal cancers which are conducted with affymetrix HG-U133A platform for Cmap project. The number of over-expressed and under-expressed gene signature in prostate cancer and colorectal cancer microarray data with different fold changes are shown in Table 1. Given the disease-specific gene signature as input query, the drug compounds ranked lists obtained from Cmap which are scored based on how well they are correlated with the input query. The score of 1 represents the input query perfect matches the changes among drug expression profiles in the database,

Table 1 The number of gene signature with different fold changes in prostate cancer and colorectal cancer

Cancer	Prostate cancer					Colorectal cancer				
Fold change	1.0	1.1	1.2	1.3	1.4	1.2	1.4	1.6	1.8	2
# over-expressed genes	89	60	45	31	27	539	308	175	106	68
# down-expressed genes	232	156	115	86	58	72	42	27	12	5

-1 represents perfect anti-correlation between them, and 0 represents a null match. Both negative and positive enrichment belong to the drug-disease associations. Therefore, we take the absolute values of scores and re-rank them and higher score represents the stronger drug-disease associations. Due to Cmap project is derived from the expression profile of a drug compound to the isolated cell lines, multiple instances may correspond to same drug and even with same dose in the obtained results. We calculate the maximum score of multiple instances that correspond to the same drug as the associated score of a specific drug for dealing with such cases. In prostate and colorectal cancer data, we use 601 drug compounds which are overlapped between DrugBank and Cmap results and the drug rank list have shown in the previous works [37]. We compute observed the area under the receiver operator characteristic (ROC) curve (AUC) for analyzing the quality of performance in Figures 2 and 3, respectively. The Cmap method obtains much lower AUC of 0.59 ± 0.04 and 0.59 ± 0.03 under the gene signature with different fold changes while our method obtains 0.94 and 0.89 in prostate cancer and colorectal cancer respectively. Previous study showed that only depending on the drug response expression profile is incapable to acquire the drug-disease associations accurately due to the profiles generated under different conditions [38]. There are several problems that limit the performance in Cmap project: First, the set of differentially expressed genes that constitute disease signatures or drug signatures were chosen empirically that cannot guarantee the biological relevance of the selected signatures. A bad selection of signatures may tend to capture similarities in the experimental settings rather than revealing the underlying mechanisms. Second, only the overlap among genes between disease state and drug treatments are not quantified as the overall effect of a drug and the values of differential expression are also not taken into account in Cmap project. Therefore, the feasibility and benefits of using network-based analyses of chemical, genomic and phenotype data provides a good chance to reveal drug-disease associations.

The AUC with varying diffusion parameters

We investigate the systematic effect of the different diffusion parameters among networks in our approach. We apply the diffusion parameters $\alpha_i \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ used in the experiments and 125 combinations of the parameters are tested. The ranking performances of our method with different combinations are measured by AUC values in Figure 4. The results show that the diffusion parameters set in drug, gene, and phenotype homo-network α_d, α_g, and α_p as 0.1, 0.7 and 0.3 can get a higher AUC in both prostate and colorectal cancer. A higher similarity score between the chemical structures of a pair of drugs may sometimes have different targets to affect different functional modules in PIN. The drugs with similar chemical structures without binding to similar enzymes would reduce the predictive accuracy for drug
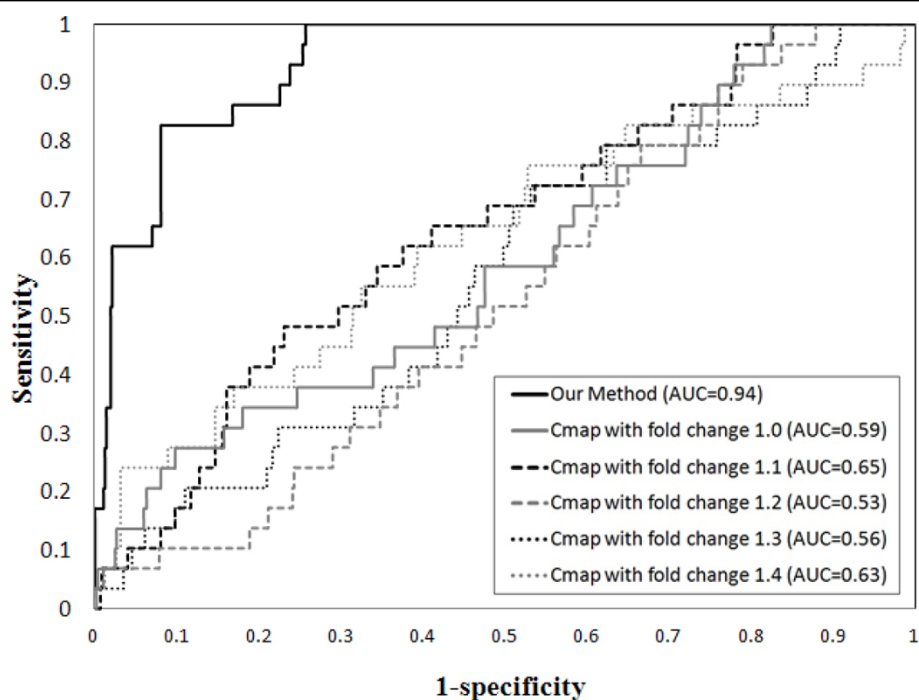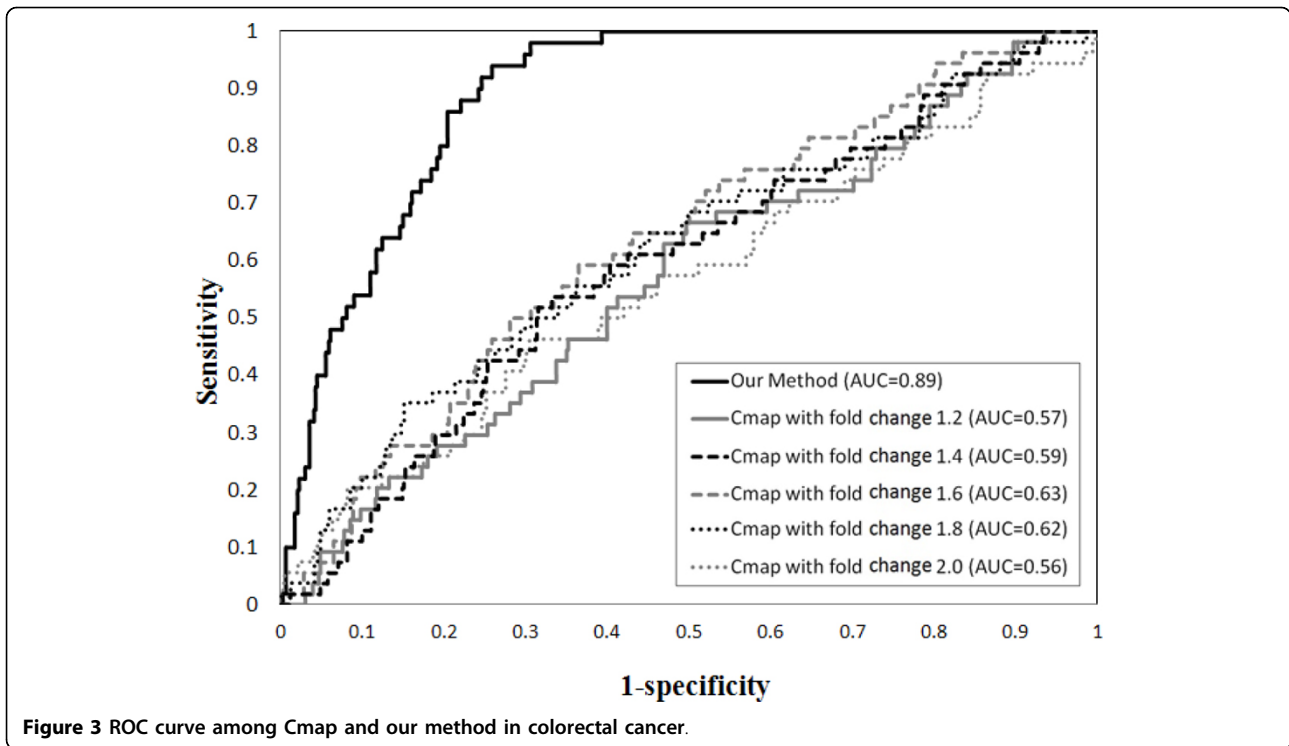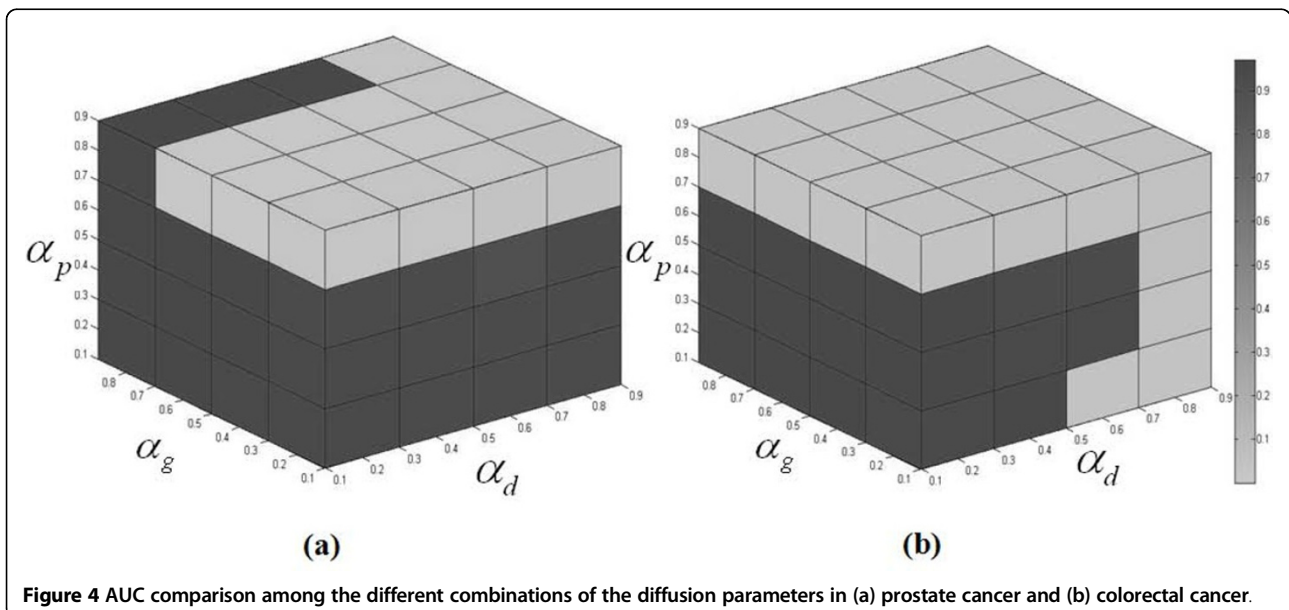

Figure 2 ROC curve among Cmap and our method in prostate cancer.

Figure 3 ROC curve among Cmap and our method in colorectal cancer.

similarity [39]. On the other hand, the bit-comparison method of chemical structures using Tanimoto coefficient also has its limitations [40]: First, this kind of method does not include biochemical information at the atomic level of the representation. Second, sometimes it may yield low similarity values and it has an inherent bias related to the size of the molecules that are being sought. Furthermore, many physiological effects cannot be predicted accurately by chemical structure properties alone without more detailed information of metabolic and pharmacokinetic transformations of drugs [41]. Those may be the reason that why the diffusion parameter value in the drug network should be relatively smaller in our experiment. PIN tends to have much higher diffusion parameters to get higher AUC and it is reasonable to infer drug-disease associations not only by targeting the specific proteins

Figure 4 AUC comparison among the different combinations of the diffusion parameters in (a) prostate cancer and (b) colorectal cancer.

directly but also by modulating the pathways involved in the pathological process [42].

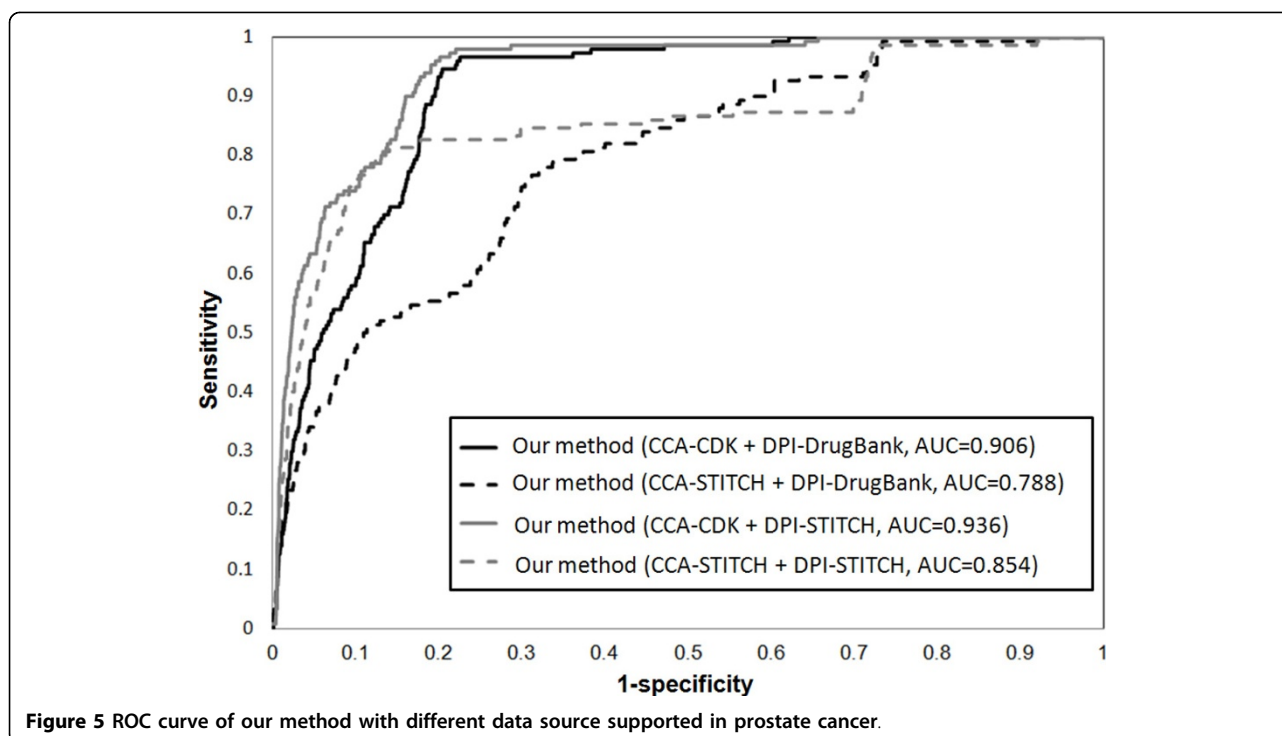**The performance of our method with different data source**
The drug is designed to target on the primary targets, but it also may interact with unexpected proteins (called off-targets) to trigger the associated biological processes and pathways. It may display undesired off-target toxicity or drug reposition for the disease phenotype. Therefore, we aim to add additional drug targets information to better understanding of dynamic processes under drug exposure. The database STITCH contains the relations between chemicals as chemical-chemical associations (CCA), and chemicals and their interacted proteins as drug-protein interactions (DPI) which are all curated by the evidence derived from experiments, publicly databases and the literature extraction [43]. However, the textual co-occurrence from text mining does not necessarily indicate meaningful relationships [44]. Therefore, we take 29,275 chemical-chemical interactions and 45,567 chemical-protein pairs from STITCH excluded the relations extracted from the text mining method. In the rank lists of 1,410 drugs, our method with the chemical structures similarity from CDK in drug homo-network and the chemical-protein interactions from STITCH in drug-protein hetero-network obtain highest AUC of 0.936 and 0.861 in prostate cancer and colorectal cancer dataset in Figure 5 and 6, respectively. It shows that our method with more potential off-targets from STITCH gets higher

AUC than that with only primary drug targets from DrugBank in both prostate cancer and colorectal cancer data. Due to only a limited number of chemical-chemical associations have been identified in STITCH [45], its performance is worse than the associations constructed by chemical structures similarity from CDK in drug homo-network.
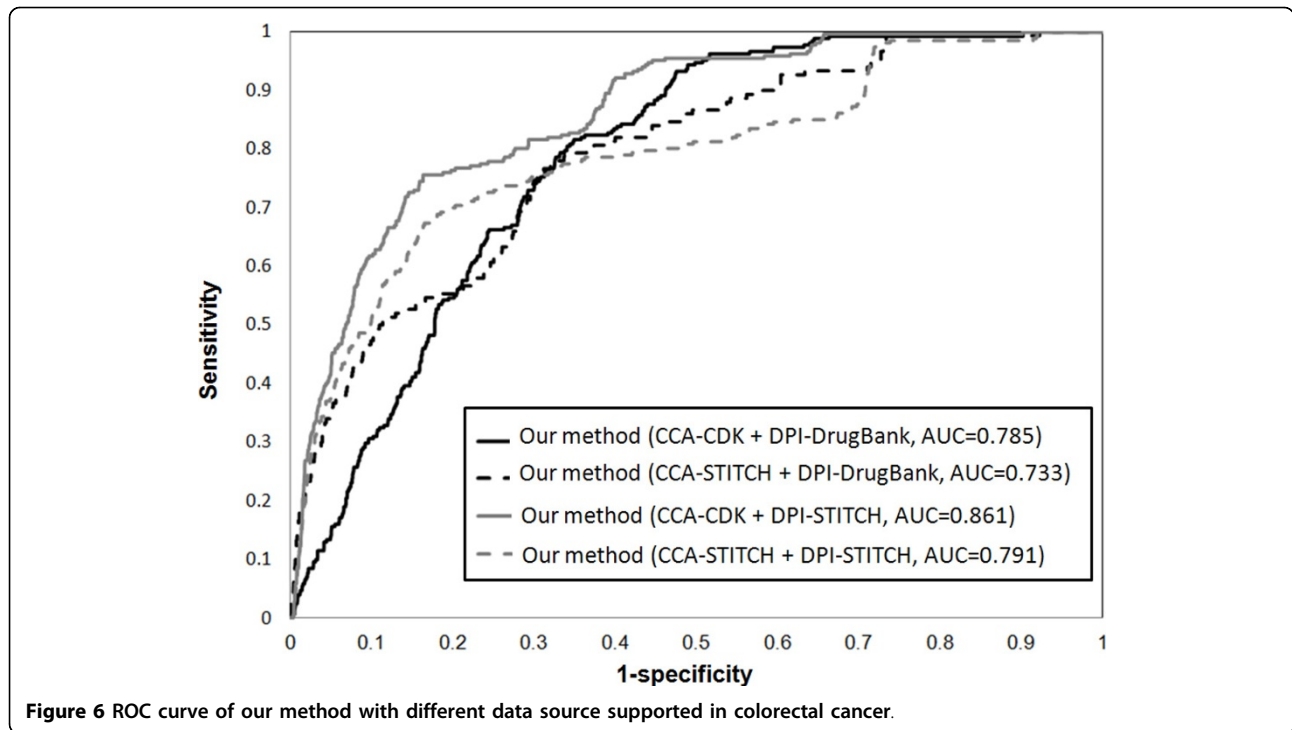
Case study: prostate cancer
Potential drug and prostate cancer relations
We run our method with diffusion parameters α_d, α_g, and α_p as 0.1, 0.7 and 0.3 in drug, gene/protein, and phenotype homo-networks and display 26 drug-prostate cancer associations by evaluating the one-tailed test at 0.01 level of significance with Z-score > 2.33 in Table 2. Since the predictions are not regarded as true and we need to be further validated using external literature support, 17 known associations are approved by the benchmark and other 7 associations are supported by the literature. The man with geriatric cancer may have prostate enlargement or prostate cancer at a higher risk while taking fluoxymesterone. On the other hand, if people has known or suspected prostate cancer, oxandrolone and drostanolone related to androgen receptors are not recommended to be taken in certain medical conditions. Previous clinical studies showed that prostate-specific antigen (PSA) plays an important tumor marker for prostate cancer in male and nandrolone phenpropionate is an androgen receptor agonist and previous study showed that it can decrease

Figure 5 ROC curve of our method with different data source supported in prostate cancer.

Figure 6 ROC curve of our method with different data source supported in colorectal cancer.

cell growth of prostate cancer LNCaP cells [46]. The primary action of cyproterone is to suppress the activity of the androgen hormones via competitive antagonism of the androgen receptor and inhibition of enzymes in the androgen biosynthesis pathway [47]. While the patients with metastatic prostate cancer were given mitomycin, the clinical results show good anti-tumour activity in metastatic prostate cancer and low toxicity in a phase II chemotherapy study [48]. Due to the androgens can stimulate the growth of prostate cancer cells, previous phase II trial study in 2012 demonstrated the hormone

therapy using exemestane with or without bicalutamide may fight prostate cancer [49].

The significant functional modules related to the drug-prostate cancer association

There are 31 genes with Z-score > 2.33 which are strongly related to 18 drugs and prostate cancer associations. We use the functional annotation analysis to investigate the functional enrichment of them via GSEA toolkits [50]. We select 4 annotation categories related to the functional pathways and processes including Biological Process (BP) in Gene Ontology (GO), KEGG, BIOCARTA, and

Table 2 Drug-prostate cancer associations

Drug ID	Drug Name	Score	Drug ID	Drug Name	Score
DB01420[a]	Testosterone Propionate	13.69	DB01216[a]	Finasteride	4.85
DB00621[b]	Oxandrolone	12.76	DB00367[a]	Levonorgestrel	4.46
DB00984[b]	Nandrolone phenpropionate	9.69	DB00262[a]	Carmustine	4.16
DB00858[b]	Drostanolone	8.14	DB06710[a]	Methyltestosterone	4.00
DB04839[b]	Cyproterone	7.78	DB00227[a]	Lovastatin	3.90
DB00687[a]	Fludrocortisone	7.44	DB00783[a]	Estradiol	3.77
DB00665[a]	Nilutamide	6.56	DB01599	Probucol	3.77
DB01185[b]	Fluoxymesterone	6.31	DB00279	Liothyronine	3.22
DB01128[a]	Bicalutamide	6.20	DB00421[a]	Spironolactone	3.01
DB01395[a]	Drospirenone	6.17	DB01406[a]	Danazol	2.99
DB00305[b]	Mitomycin	5.24	DB00624[a]	Testosterone	2.96
DB00499[a]	Flutamide	5.07	DB00396[a]	Progesterone	2.77
DB00990[b]	Exemestane	5.03	DB00928[a]	Azacitidine	2.61

a. Primary therapeutic function approved by our benchmark; b. supported by literature

REACTOME pathways to do the functional enrichment. The enriched functional biological processes and pathways of those genes with p-value < 0.05 are shown in Table 3 and the detailed information is obtained from GSEA and DAVID gene functional classification tool [51] see Additional file 1. The most significantly enriched terms of the regulation of apoptosis, cell cycle, p53 signaling and DNA repair indicate that those genes serve important roles in drug and prostate cancer association. The BRCA2 mutation contributing to the young-onset prostate cancer has shown to be related to the Fanconi anemia pathway and DNA repair processes [52,53]. The prostate apoptosis response (Par) factor-related proapoptotic function is associated with prostate tumor progression and sheds light on the effects of the AR pathway on cell survival and apoptosis [54]. Interestingly, this functional module may be also for programmed cell death in response to induce apoptosis in neurodegenerative disorders such as Alzheimer's and Huntington's disease pathway [55]. With literature support, Par-4 has also been initially characterized in prostate cancer and also linked with the direct induction of apoptosis and even recognized to be a new target in pancreatic cancer [56].

Case study: colorectal cancer
Potential drug and colorectal cancer relations
We display 37 significant drug-colorectal cancer associations with Z-score > 2.33 in Table 4. There are 28 known associations between drugs and colorectal cancer approved by our benchmark and 4 associations with literature support. Pyridoxal phosphate (PLP) is an important cofactor in the reactions of amino acid metabolism and previous studies indicated that increased blood PLP levels are associated with a reduced risk of colorectal cancer [57]. The previous findings showed the association between dietary fats and colorectal cancer, and a significant inverse association was found between colorectal cancer and

alpha-linoleic acid [58]. The results of this study suggest that substituting alpha-linoleic acid in the diet may reduce the risk of the colorectal cancer [58]. Previous works examined the effect of arsenic trioxide (As_2O_3) at various concentrations on the cell growth of the colon cancer cell lines and showed that the growth of all cell lines was gradually suppressed with As_2O_3 in comparison with that obtained without treatment [59,60]. Polyamines, organic compounds having two or more amino groups, have been shown to play an important role in the growth and survival in colorectal cancer [61]. We also find that spermine, a polyamine, may be a candidate target for therapeutic intervention in colorectal cancer [62]. The activity of the polyamine-synthesising enzyme, ornithine decarboxylase (ODC), is very highly expressed in proliferative HT-29 colon cancer cells comparing to those from control samples, as well as in our case study [63]. L-ornithine is apparently efficiently utilized in the ODC pathway and is also considered as growth factors involved in cell proliferation and differentiation regulated by amino acids metabolism [64]. The growing studies indicated the potential effectiveness of bortezomib in treatment of patients with HCT116 colon cancer by significantly increasing survivin expression [65,66]. The increasing evidences determined that the relationships among certain bacteria and cancer exist but the detail mechanism still unclear [67]. In our results, we interestingly find that an antibiotic drug zanamivir for bacterial infection may cause or cure colorectal cancer [68].

The significant functional modules related to the drug-colorectal cancer association
The enriched functional biological processes of 44 significant genes in drug-colorectal cancer associations by investigating the functional enrichment are shown in Table 5 and the detailed information see Additional file 2. The popular enriched functional terms of the gene modules are regulation of Glycogen synthase kinase 3 (GSK-3) pathway, innate immune system, Wnt signalling pathway [69] and apoptosis pathway which all appear to be promising biological processes associated with colorectal cancer. The activity of GSK3b expression in colorectal cancer patients is higher than those in their normal samples and the GSK3b inhibitor may induce apoptosis in human colorectal cancer cells [70,71]. Lipopolysaccharide (LPS) elicits several immediate proinflammatoy responses including CD14, Toll-like receptors, phosphatidylinositol-3'-kinase (PI3K)/Akt signaling, myeloid differentiation factor, nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB) transcription factors, and also promotes downstream b1 integrin function in tumor growth and progression thereby increasing the adhesiveness and metastatic capacity of colorectal cancer cells [72-74]. Here, we present evidence that S100A8/A9 actives the downstream genes associated with Mitogen-activated protein kinases

Table 3 Functional modules related to the drug-prostate cancer associations

Database	Functional modules	p-value
REACTOME	Fanconi anemia pathway	4.65E-5
KEGG	Huntington's disease pathway	1.60E-4
KEGG	p53 signaling pathway	9.72E-4
BIOCARTA	BRCA1, BRCA2 and ATR pathway	1.61E-3
GO, KEGG	Cell cycle	1.81E-3
REACTOME	DNA Repair	3.89E-3
GO	Negative regulation of cell proliferation	9.75E-3
KEGG	Alzheimer's disease pathway	1.21E-2
GO	Apoptosis	1.26E-2
KEGG	Pancreatic cancer pathway	1.69E-2
KEGG	Viral myocarditis	1.83E-2
REACTOME	Electron transport	2.13E-2

Table 4 Drug-colorectal cancer associations

Drug ID	Drug Name	Score	Drug ID	Drug Name	Score
DB0017[a]	Adenosine triphosphate	12.35	DB00313[a]	Valproic Acid	4.20
DB01593[a]	Zinc	8.66	DB00162	Vitamin A	3.77
DB00591	Fluocinolone Acetonide	8.25	DB00131	Adenosine monophosphate	3.74
DB00163[a]	Vitamin E	7.13	DB00129[b]	L-Ornithine	3.72
DB01262[a]	Decitabine	7.02	DB00947[a]	Fulvestrant	3.46
DB00435[a]	Nitric Oxide	6.70	DB00741[a]	Hydrocortisone	3.34
DB00783[a]	Estradiol	6.40	DB01064[a]	Isoproterenol	3.21
DB00328[a]	Indomethacin	6.35	DB01128[a]	Bicalutamide	3.17
DB00755[a]	Tretinoin	6.26	DB00773[a]	Etoposide	3.15
DB00114[b]	Pyridoxal Phosphate	5.91	DB00481[a]	Raloxifene	3.10
DB00544[a]	Fluorouracil	5.62	DB00305[a]	Mitomycin	3.05
DB00132[b]	Alpha-Linolenic Acid	5.61	DB00188[b]	Bortezomib	2.77
DB00396[a]	Progesterone	5.59	DB01005[a]	Hydroxyurea	2.66
DB00997[a]	Doxorubicin	5.52	DB00619[a]	Imatinib	2.61
DB00179[a]	Masoprocol	5.09	DB02546[a]	Vorinostat	2.53
DB00558	Zanamivir	5.03	DB00928[a]	Azacitidine	2.47
DB01169[a]	Arsenic trioxide	4.38	DB00498[a]	Phenindione	2.44
DB00127[a]	Spermine	4.38	DB00548	Azelaic Acid	2.39
DB00834[a]	Mifepristone	4.21			

a. Primary therapeutic function approved by our benchmark; b. supported by literature

(MAPK) and NF-kB signaling pathways to promote tumor growth and metastasis and the expression of S100A8/A9 on myeloid cells is also essential for development of colon tumors in mice model [75]. Untersmayr detected Fc epsilon RI (FcɛRI)-positive epithelial cells in the colon cancer patients [76]. There has been a growing importance of the neurotrophin signalling in a variety of human cancers including colorectal cancer and malignant gliomas in particular [77]. Our study also denotes that the significances of p53, hypoxia inducible factor-1 alpha pathway, and vascular endothelial growth factor (VEGF) expression in colorectal cancer [78].

Table 5 Functional modules related to the drug-colorectal cancer associations

Database	Functional modules	p-value
BIOCARTA	GSK3 pathway	7.29E-14
KEGG	Endometrial cancer	2.30E-11
KEGG	Colorectal cancer	4.32E-9
KEGG, BIOCARTA, REACTOME	Toll like receptor pathway	5.79E-9
KEGG	Prostate cancer	1.36E-6
BIOCARTA	P53hypoxia pathway	1.93E-6
KEGG	Lung cancer	2.30E-6
KEGG	Glioma	5.82E-6
KEGG	Renal cell carcinoma	8.40E-6
KEGG	Melanoma	9.01E-6
KEGG	Chronic myeloid leukemia	1.03E-5
KEGG	VEGF signaling pathway	1.26E-5
KEGG	Innate Immune System	1.29E-5
KEGG	Fc epsilon RI signaling pathway	1.52E-5
REACTOME	Regulation of ornithine decarboxylase (ODC)	4.28E-5
KEGG	Pancreatic cancer	1.74E-4
KEGG	Wnt signaling pathway	3.34E-4
GO	Apoptosis	1.54E-3
KEGG	Neurotrophin signaling pathway	1.62E-3

Conclusions

We integrate the information of drug, genomic and disease phenotype from available experimental data and knowledge as weighted networks and their connected relationships together. We apply disease-oriented network propagation approach for inferring and evaluating the likelihood of the probability between drugs and query disease. In our experiment, we adopt the prostate cancer and colorectal cancer as our case study and the results clearly outperform previous Cmap project. Our results are also found to be significantly enriched in both the biomedical literature and clinical trials. The success of our methods can be attributed as follows: First, we integrate heterogeneous data and knowledge about disease phenotype, chemical structure of drugs, and gene expression into our model. Second, our network propagation method combines the information not only from the single network but also derived the information from other connected homo-networks to infer the drug-disease association. Finally, our method with off-targets information gets higher performance than that with only primary drug targets in both test data. We believe that the combination of network and heterogeneous data source could help us to generate new hypotheses to infer the drug-disease associations and even speed up the drug development processes. Our study provides opportunities for future toxicogenomics and drug discovery applications but the limitation is the difficulty in distinguishing the positive and negative associations between drug and disease. In the future, we can choose different methods to calculate the chemical structural similarity between drugs, which could improve the limitations by using Tanimoto coefficient. On the other hand, our approach heavily relies on the weights for the edges in each network derived from the existing knowledge of drugs, targets, protein, disease or reported databases, or experimental results from the public database and the incompleteness of such information would limit our prediction power. We can also integrate various data sources such as drug response profile, side effect and pharmacological data and therapeutic/toxicological expression profiles to verify the reliability and confidence of the interactions.

Additional material

Additional file 1: The functional enrichment canonical pathways of the genes in prostate cancer. We filter the functional enrichment canonical pathways of the overlap of the significant genes related to the drug-prostate cancer associations with at least 3 members in each functional category and p-value<0.05 using GSEA and DAVID online toolkit.

Additional file 2: The functional enrichment canonical pathways of genes in colorectal cancer. We filter the functional enrichment canonical pathways of the overlap of the significant genes related to the drug-colorectal cancer associations with at least 3 members in each functional category and p-value<0.05 using GSEA and DAVID online toolkit.

Authors' details
[1]Institute of Information Systems and Applications, National Tsing Hua University, HsinChu, Taiwan. [2]Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan.

References
1. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network**. *Proc Natl Acad Sci USA* 2007, **104(21)**:8685-8690.
2. Dimasi JA: **New drug development in the United States from 1963 to 1999**. *Clin Pharmacol Ther* 2001, **69**:286-96.
3. Adams CP, Brantner VV: **Estimating the cost of new drug development: Is it really 802 million dollars?** *Health Aff (Millwood)* 2006, **25**:402-8.
4. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: **How to improve R&D productivity: the pharmaceutical industry's grand challenge**. *Nat Rev Drug Discov.* 2010, **9(3)**:203-14.
5. DiMasi J, Hansen R, Grabowski H: **R&D Costs and Returns by Therapeutic Category**. *Drug Information Journal* 2004, **38**:211-223.
6. Mattingly CJ, Rosenstein MC, Colby GT, Forrest JN, Boyer JL: **The Comparative toxicogenomics database (CTD): a resource for comparative toxicological studies**. *J Exp Zool A Comp Exp Biol* 2006, **305**:689-692.
7. Cheng F, Li W, Zhou Y, Lee PW, Tang Y: **Prediction of Human Genes and Diseases Targeted by Xenobiotics Using Predictive Toxicogenomics-Derived Models (PTDMs)**. *Mol BioSyst* 2013.
8. Gong L, Owen RP, Gor W, Altman RB, Klein TE: **PharmGKB: an integrated resource of pharmacogenomic data and knowledge**. *Curr Protoc Bioinformatics* 2008, Unit 14.7, Chapter 14.
9. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R: **PROMISCUOUS: a database for network-based drugrepositioning**. *Nucleic Acids Res* 2011, **39**:D1060-6.
10. Lamb J: **The Connectivity Map: a new tool for biomedical research**. *Nat Rev Cancer* 2007, **7**:54-60.
11. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science* 2006, **313**:1929-1935.

12. Li J, Zhu X, Chen JY: **Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts.** *PLoS Comput Biol* 2009, **5**:e1000450.

13. Chiang AP, Butte AJ: **Systematic evaluation of drug-disease relationships to identify leads for novel drug uses.** *Clin Pharmacol Ther* 2009, **86**:507-510.

14. Gottlieb A, Stein GY, Ruppin E, Sharana R: **PREDICT: a method for inferring novel drug indications with application to personalized medicine.** *Mol Syst Biol* 2011, **7**:496.

15. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ: **Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets.** *PLoS Comput Biol* 2010, **6**:e1000662.

16. Daminelli S, Haupt VJ, Reimann M, Schroeder M: **Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network.** *Integr Biol (Camb)* 2012, **4(7)**:778-88.

17. Ye H, Yang LL, Cao ZW, Tang KL, Li YX: **A pathway profile-based method for drug repositioning.** *Chinese Science Bulletin* 2012, **57**:2106-2112.

18. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30**:52-55.

19. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA: **A text-mining analysis of the human phenome.** *Eur J Hum Genet* 2006, **14**:535-542.

20. Gottlieb A, Magger O, Berman I, Ruppin E: **PRINCIPLE: A tool for associating genes with diseases via network propagation.** *Bioinformatics* 2011, **27**:3325-6.

21. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets.** *Nucleic Acids Res* 2008, **36**:D901-D906.

22. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28**:31-36.

23. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL: **Recent developments of the chemistry development kit (CDK)–an open-source java library for chemo- and bioinformatics.** *Curr Pharm Des* 2006, **12**:2111-2120.

24. Tanimoto T: *Internal report: IBM Technical Report Series* 1957.

25. Prieto C, De Las Rivas J: **APID: Agile Protein Interaction Data Analyzer.** *Nucleic Acids Research* 2006, **34**:W298-302.

26. Kohler S, Bauer S, Horn D, Robinson PN: **Walking the Interactome for Prioritization of Candidate Disease Genes.** *Am J Hum Genet* 2008, **82(4)**:949-958.

27. Zhou D, Orshanskiy SA, Zha H, Giles CL: **Co-ranking authors and documents in a heterogeneous network.** *In ICDM* 2007, 739-744.

28. Hwang TH, Kuang R: **A Heterogeneous Label Propagation Algorithm for Disease Gene Discovery.** *Proc of SIAM International Conference on Data Mining (SDM)* 2010, 583-594.

29. Xi W, Zhang B, Chen Z, Lu Y, Yan S, Ma WY, Fox EA: **Link Fusion: A Unified Link Analysis Framework for Multi-Type Interrelated Data Objects.** *World Wide Web Conference Series - WWW* 2004, 319-327.

30. Pinski G, Narin N: **Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics.** *Information Process and Management* 1976, **12**:297-312.

31. Zhou D, Weston J, Gretton A, Bousquet O, Scholkopf B: **Learning with local and global consistency.** *In NIPS* 2004, **16**:321-328.

32. Komurov K, White MA, Ram PT: **Use of Data-Biased Random Walks on Graphs for the Retrieval of Context-Specific Networks from Genomic Data.** *PLoS Comput Biol* 2010, **6(8)**:e1000889.

33. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer.** *Proc Natl Acad Sci U S A* 2004, **101(3)**:811-816.

34. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM: **The stanford microarray database.** *Nucleic Acids Research* 2001, **29**:152-155.

35. Gyorffy B, Molnar B, Lage H, Szallasi Z, Eklund AC: **Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples.** *PLoS One* 2009, **4(5)**:e5645.

36. Hong Y, Ho KS, Eu KW, Cheah PY: **A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis.** *Clin Cancer Res* 2007, **13(4)**:1107-14.

37. Huang YF, Yeh HY, Soo VW: **Network-based inferring drug-disease associations from chemical, genomic and phenotype data.** *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on: 4-7 October 2012* 2012, 1-6.

38. Hu G, Agarwal P: **Human Disease-Drug Network Based on Genomic Expression Profiles.** *PLoS ONE* 2009, **4(8)**:e6536.

39. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y: **Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference.** *PLoS Comput Biol.* 2012, **8(5)**:e1002503.

40. Godden JW, Xue L, Bajorath J: **Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients.** *J Chem Inf Comput Sci* 2000, **40**:163-166.

41. Martin YC, Jacson RM: **Do structurally similar molecules have similar biological activity?** *J Med Chem* 2002, **45**:4350-4358.

42. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M: **Drug-target network.** *Nat Biotechnol* 2007, **25(10)**:1119-26.

43. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P: **STITCH 3: zooming in on protein-chemical interactions.** *Nucleic Acids Res.* 2012, **40(Database)**:D876-80.

44. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J: **Extraction of gene-disease relations from Medline using domain dictionaries and machine learning.** *Pac Symp Biocomput* 2006, 4-15.

45. Chen L, Lu J, Zhang J, Feng KR, Zheng MY, Cai YD: **Predicting Chemical Toxicity Effects Based on Chemical-Chemical Interactions.** *PLoS One.* 2013, **8(2)**:e56517.

46. Fujii Y, Kawakami S, Okada Y, Kageyama Y, Kihara K: **Regulation of prostate-specific antigen by activin A in prostate cancer LNCaP cells.** *Am J Physiol Endocrinol Metab* 2004, **286(6)**:E927-31.

47. Giorgi E, Shirley I, Grant J, Stewart J: **Androgen dynamics in vitro in the human prostate gland. Effect of cyproterone and cyproterone acetate.** *Biochem J* 1973, **132(3)**:465-74.

48. Jones WG, Fosså SD, Bono AV, Croles JJ, Stoter G, Pauw MD, Sylvester R: **Mitomycin-C in the treatment of metastatic prostate cancer: report on an EORTC phase II study.** *World Journal of Urology* 1986, **4(3)**:182-185.

49. ClinicalTrials.gov. [http://clinicaltrials.gov/ct2/show/NCT00031889].

50. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *PNAS* 2005, **102(43)**:15545-15550.

51. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome Biology* 2007, **8(9)**:R183.

52. Edwards SM, Kote-Jarai Z, Meitz J, Hamoudi R, Hope Q, Osin P, Jackson R, Southgate C, Singh R, Falconer A, Dearnaley DP, Ardern-Jones A, Murkin A, Dowe A, Kelly J, Williams S, Oram R, Stevens M, Teare DM, Ponder BA, Gayther SA, Easton DF, Eeles RA, Cancer Research UK/Bristish Prostate Group UK Familial Prostate Cancer Study Collaborators, British Association of Urological Surgeons Section of Oncology: **Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene.** *Am J Hum Genet* 2003, **72**:1-12.

53. Hucl T, Gallmeier E: **DNA repair: exploiting the Fanconi anemia pathway as a potential therapeutic target.** *Physiol Res.* 2011, **60(3)**:453-65.

54. Mattson MP, Duan W, Chan SL, Camandola S: **Par-4: an emerging pivotal player in neuronal apoptosis and neurodegenerative disorders.** *J Mol Neurosci* 1999, **3(1-2)**:17-30.

55. Azmi AS, Wang Z, Burikhanov R, Rangnekar VM, Wang G, Chen J, Wang S, Sarkar FH, Mohammad RM: **Critical role of prostate apoptosis response-4 in determining the sensitivity of pancreatic cancer cells to small-molecule inhibitor-induced apoptosis.** *Mol Cancer Ther* 2008, **7**:2884-2893.

56. Azmi AS, Philip PA, Zafar SF, Sarkar FH, Mohammad RM: **PAR-4 as a possible new target for pancreatic cancer therapy.** *Expert Opin Ther Targetsm* 2010, **14(6)**:611-20.

57. Larsson SC, Orsini N, Wolk A: **Vitamin B6 and risk of colorectal cancer: a meta-analysis of prospective studies.** *JAMA* 2010, **303(11)**:1077-83.

58. Nkondjock A, Shatenstein B, Maisonneuve P, Ghadirian P: **Assessment of risk associated with specific fatty acids and colorectal cancer among French-Canadians in Montreal: a case-control study.** *Int J Epidemiol.* 2003, **32(2)**:200-9.

59. Nakagawa Y, Akao Y, Morikawa H, Hirata I, Katsu K, Naoe T, Ohishi N, Yagi K: Arsenic trioxide-induced apoptosis through oxidative stress in cells of colon cancer cell lines. *Life Sci* 2002, **70**(19):2253-69.
60. Lee HR, Cheong HJ, Kim SJ, Lee NS, Park HS, Won JH: Sulindac enhances arsenic trioxide-mediated apoptosis by inhibition of NF-kappaB in HCT116 colon cancer cells. *Oncol Rep* 2008, **20**(1):41-7.
61. Milovic V, Turchanowa L: Polyamines and colon cancer. *Biochem Soc Trans* 2003, **31**(2):381-3.
62. Allen WL, McLean EG, Boyer J, McCulla A, Wilson PM, Coyle V, Longley DB, Casero RA Jr, Johnston PG: The role of spermidine/spermine N1-acetyltransferase in determining response to chemotherapeutic agents in colorectal cancer cells. *Mol Cancer Ther* 2007, **6**(1):128-37.
63. Zell JA, Ziogas A, Ignatenko N, Honda J, Qu N, Bobbs AS, Neuhausen SL, Gerner EW, Anton-Culver H: Associations of a polymorphism in the ornithine decarboxylase gene with colorectal cancer survival. *Clin Cancer Res* 2009, **15**(19):6208-16.
64. Selamnia M, Robert V, Mayeur C, Delpal S, Blachier F: De novo synthesis of arginine and ornithine from citrulline in human colon carcinoma cells: metabolic fate of L-ornithine. *Biochim Biophys Acta* 1998, **1425**(1):93-102.
65. Mackay H, Hedley D, Major P, Townsley C, Mackenzie M, Vincent M, Degendorfer P, Tsao MS, Nicklee T, Birle D, Wright J, Siu L, Moore M, Oza A: A phase II trial with pharmacodynamic endpoints of the proteasome inhibitor bortezomib in patients with metastatic colorectal cancer. *Clin Cancer Res* 2005, **11**:5526-5533.
66. Kozuch PS, Rocha-Lima CM, Dragovich T, Hochster H, O'Neil BH, Atiq OT, Pipas JM, Ryan DP, Lenz HJ: Bortezomib with or without irinotecan in relapsed or refractory colorectal cancer: results from a randomized phase II study. *J Clin Oncol* 2008, **26**:2320-2326.
67. Mager DL: Bacteria and cancer: cause, coincidence or cure? A review. *J Transl Med* 2006, **4**:14.
68. Yoshioka T, Morimoto Y, Iwagaki H, Itoh H, Saito S, Kobayashi N, Yagi T, Tanaka N: Bacterial lipopolysaccharide induces transforming growth factor beta and hepatocyte growth factor through toll-like receptor 2 in cultured human colon cancer cells. *J Int Med Res* 2001, **29**(5):409-20.
69. Leung JY, Kolligs FT, Wu R, Zhai Y, Kuick R, Hanash S, Cho KR, Fearon ER: Activation of AXIN2 expression by beta-catenin-T cell factor. A feedback repressor pathway regulating Wnt signaling. *J Biol Chem* 2002, **277**(24):21657-65.
70. Shakoori A, Ougolkov A, Yu ZW, Zhang B, Modarressi MH, Billadeau DD, Mai M, Takahashi Y, Minamoto T: Deregulated GSK3beta activity in colorectal cancer: its association with tumor cell survival and proliferation. *Biochem Biophys Res Commun.* 2005, **334**(4):1365-73.
71. Tan J, Zhuang L, Leong HS, Iyer NG, Liu ET, Yu Q: Pharmacologic modulation of glycogen synthase kinase-3beta promotes p53-dependent apoptosis through a direct Bax-mediated mitochondrial pathway in colorectal cancer cells. *Cancer Res* 2005, **65**(19):9012-20.
72. Hsu RY, Chan CH, Spicer JD, Rousseau MC, Giannias B, Rousseau S, Ferri LE: LPS-induced TLR4 signaling in human colorectal cancer cells increases beta1 integrin-mediated cell adhesion and liver metastasis. *Cancer Res* 2011, **71**(5):1989-98.
73. Doan HQ, Bowen KA, Jackson LA, Evers BM: Toll-like receptor 4 activation increases Akt phosphorylation in colon cancer cells. *Anticancer Res.* 2009, **29**(7):2473-8.
74. Slattery ML, Herrick JS, Bondurant KL, Wolff RK: Toll-like receptor genes and their association with colon and rectal cancer development and prognosis. *Int J Cancer* 2012, **130**(12):2974-80.
75. Ichikawa M, Williams R, Wang L, Vogl T, Srikrishna G: S100A8/A9 activate key genes and pathways in colon tumor progression. *Mol Cancer Res* 2011, **9**(2):133-48.
76. Untersmayr E, Bises G, Starkl P, Bevins CL, Scheiner O, Boltz-Nitulescu G, Wrba F, Jensen-Jarolim E: The high affinity IgE receptor Fc epsilonRI is expressed by human intestinal epithelial cells. *PLoS One* 2010, **5**:e9023.
77. Johnston AL, Lun X, Rahn JJ, Liacini A, Wang L, Hamilton MG, Parney IF, Hempstead BL, Robbins SM, Forsyth PA, Senger DL: The p75 neurotrophin receptor is a central regulator of glioma invasion. *PLoS Biol* 2007, **5**(8): e212.
78. Kwon HC, Kim SH, Oh SY, Lee S, Kwon KA, Choi HJ, Park KJ, Kim HJ, Roh MS: Clinicopathological significance of p53, hypoxia-inducible factor 1alpha, and vascular endothelial growth factor expression in colorectal cancer. *Anticancer Res.* 2010, **30**(10):4163-8.